SIAM J. CONTROL OPTIM.
Vol. 44, No. 1, pp. 1–28

DISTRIBUTED CONTROL OF SPATIALLY REVERSIBLE INTERCONNECTED SYSTEMS WITH BOUNDARY CONDITIONS*

CÉDRIC LANGBORT† AND RAFFAELLO D'ANDREA‡

Abstract. We present a class of spatially interconnected systems with boundary conditions that have close links with their spatially invariant extensions. In particular, well-posedness, stability, and performance of the extension imply the same characteristics for the actual, finite extent system. In turn, existing synthesis methods for control of spatially invariant systems can be extended to this class. The relation between the two kinds of systems is proved using ideas based on the "method of images" of partial differential equations theory and uses symmetry properties of the interconnection as a key tool.

Key words. spatially distributed control, symmetric systems, boundary conditions

AMS subject classifications. 93A14, 93A15, 93B36

DOI. 10.1137/S0363012902415803

1. Introduction.

Many systems consist of the interconnection of a large number of identical subunits which interact with their nearest neighbors. Examples of such interconnected systems include formations of autonomous vehicles [11], [22], cross-directional control in the pulp and paper and chemical process industry [13], [14], "smart structures" (large arrays of distributed micro electromechanical actuators and sensors) [2], and semidiscretized partial differential equations [3].

Over the years, several frameworks for control of interconnected systems have been proposed that all assumed the existence of a particular mathematical structure induced by the interconnection. Early works [3], [17] showed that some systems, especially semidiscretized partial differential equations, can sometimes be treated as systems over modules. More recently, the papers [1], [5], [15] have considered so-called spatially invariant systems and used Fourier techniques or algebraic transformations to derive implementable and scalable optimal control algorithms, even in the limit of an infinite number of subunits. While some practical systems can be accurately modelled as being spatially invariant (e.g., circular plastic extrusion machines or very large arrays of sensors and actuators), most examples do not fall into this category because they are of finite extent and possess boundary conditions. This is why much of the current research is geared toward spatially varying systems, in an effort to adapt methods from the monodimensional time-varying case [8].

The approach taken in this paper is different, as we show that analysis and synthesis for the actual finite extent system with boundary conditions can sometimes be performed by studying a larger, spatially invariant system. The key assumption for this result is another structural property which we call spatial reversibility. In short,

*Received by the editors October 5, 2002; accepted for publication (in revised form) August 23, 2004; published electronically June 14, 2005. This work was supported in part by a National Science Foundation CAREER Award for "Robust and Optimal Control of Interconnected Systems" and in part by the Air Force Office of Scientific Research under grant F49620-01-1-0119.

http://www.siam.org/journals/sicon/44-1/41580.html

†Center for the Mathematics of Information, California Institute of Technology, 1200 E. California Blvd., MS 136-93, Pasadena, CA 91125 (clangbort@ist.caltech.edu). This work was performed while the author was affiliated with T&AM, Cornell University.

‡Mechanical and Aerospace Engineering, Cornell University, 101 Rhodes Hall, Ithaca, NY 14853 (rd28@cornell.edu).

we prove that a lack of spatial invariance can be made up for by spatial reversibility of the finite extent system with boundary conditions and that, in turn, any technique designed for spatially invariant systems can be used in that case too. A major difference between spatial invariance and spatial reversibility is that the former is a property of the interconnection while the latter is a property of the subsystems. In the language of [6], spatial reversibility is an *internal symmetry* while spatial invariance is a *global symmetry* of the system.

Our method borrows concepts from two different lines of thought. First, the idea of associating a larger spatially invariant system to the actual finite extent system is very similar in nature to the "lifting technique" introduced in [12] to relate linear time-periodic to linear time-invariant systems or to the method of [13] used to prove robustness of cross-directional controllers. Second, the motivation for considering symmetries of the system comes from the so-called "method of images" used in potential theory. The main issue that has to be addressed in order to establish a real link between finite extent and spatially invariant systems is that the boundary conditions are lost in this correspondence. One would like to consider finite extent systems, the solution of which can be recovered from the spatially invariant system in spite of this information loss. The method of images gives an example in which such a situation is at hand, although in a different context. It essentially states that boundary conditions for Laplace's equation

$$\Delta u = 0 \text{ in } U; \qquad u = g \text{ on } \partial U$$

on some simple domains $U \subset \mathbb{R}^n$ (e.g., half-spaces) can be dropped as such since its solution can be determined by solving a similar equation on the whole of $\mathbb{R}^n$, provided "mirror-image singularities" are introduced [20], [9]. The main reason why this technique works is that the Laplacian has some symmetry properties—namely, it commutes with any isometry of $\mathbb{R}^n$. It is thus natural to hope that spatial symmetry is also relevant for our problem.

It should be noted that an "embedding technique" similar to the method of images was already used in [3] to handle boundary conditions for the particular example of the semidiscretized heat equation on a finite interval. However, it was not emphasized that the possibility of using such a technique was due to symmetries of the problem.

The paper is organized as follows. After giving general preliminaries and notions on finite extent and spatially invariant systems in sections 2 and 3, we define spatial reversibility in section 4 and explain how it allows us to relate well-posedness, stability, and performance of these two kinds of systems. Section 5 presents practical examples of spatially reversible systems and section 6 is devoted to synthesis of distributed controllers for an $\mathcal{H}_\infty$ criterion. In particular, section 6.2 is largely independent from the rest, as it illustrates how the specific results of [5] can be adapted to this finite extent problem. Finally section 7 contains some generalizations of the core results, while concluding remarks can be found in section 8.

2. Modelling spatially interconnected systems. In this section, we introduce our notation and define the basic objects of interest. The goal is to provide a framework in which infinite, periodic, and finite extent systems can be handled simultaneously.

2.1. Signal spaces. Unless otherwise stated, $\mathbb{M}$ will stand for any one of the following three sets: $\{1, \dots, L\}$ for some integer $L > 0$, $\mathbb{Z}_{2L}$ (the group of integers

modulo $2L$), and $\mathbb{Z}$. We define $\ell_2^q(\mathbb{M})$ as the space of functions $x : \mathbb{M} \to \mathbb{R}^q$ such that

$$\|x\|^2_{\ell_2^q(\mathbb{M})} := \sum_{s\in\mathbb{M}} x(s)^* x(s) < \infty.$$

Then $\mathcal{L}_2^q(\mathbb{M})$ is defined as the Hilbert space of functions $x : \mathbb{R}^+ \to \ell_2^q(\mathbb{M})$ such that

$$\|x\|^2_{\mathcal{L}_2^q(\mathbb{M})} := \int_0^\infty \|x(t)\|^2_{\ell_2^q(\mathbb{M})} dt < \infty.$$

When the dimension of the target space is clear from the context or is irrelevant, we omit the superscript q and simply write $\ell_2(\mathbb{M})$ and $\mathcal{L}_2(\mathbb{M})$. Finally, if J is a matrix, we will abuse notation and identify it with the operator that associates $y \in \ell_2(\mathbb{M})$ to $x \in \ell_2(\mathbb{M})$ such that $y(s) = Jx(s)$ for all $s \in \mathbb{M}$.

2.2. Systems. Let a linear time-invariant, finite dimensional, dynamical system with input (d, v^+, v^-) and output (z, w^+, w^-) be given in state space by

$$\frac{d}{dt}x(t) = A_{\mathrm{TT}}x(t) + A_{\mathrm{TS}_+}v^+(t) + A_{\mathrm{TS}_-}v^-(t) + B_{\mathrm{T}}d(t)\ ;\ x(0) = x^0, \tag{1a}$$

$$w^+(t) = A_{\mathrm{ST}_+}x(t) + A_{\mathrm{SS}_{+,+}}v^+(t) + A_{\mathrm{SS}_{+,-}}v^-(t) + B_{\mathrm{S}_+}d(t), \tag{1b}$$

$$w^-(t) = A_{\mathrm{ST}_-}x(t) + A_{\mathrm{SS}_{-,+}}v^+(t) + A_{\mathrm{SS}_{-,-}}v^-(t) + B_{\mathrm{S}_-}d(t), \tag{1c}$$

$$z(t) = C_{\mathrm{T}}x(t) + C_{\mathrm{S}_+}v^+(t) + C_{\mathrm{S}_-}v^-(t) + Dd(t), \tag{1d}$$

where $x(t)$, $d(t)$, and $z(t)$ belong to $\mathbb{R}^{n_{\mathrm{T}}}$, $\mathbb{R}^m$, and $\mathbb{R}^p$, respectively, for all $t \geq 0$, and $v^+(t), v^-(t), w^+(t), w^-(t)$ all belong to $\mathbb{R}^n$. We also let $n_{\mathrm{S}} := 2n$. To such a system, which we call the *basic building block*, we can associate three different spatially interconnected systems as follows.

2.2.1. Infinite system. Let the *shift operator* $\mathbf{S}$ be defined on $\ell_2(\mathbb{Z})$ by

$$(\mathbf{S}v)(s) := v(s+1) \text{ for all } s. \tag{2}$$

$\mathbf{S}$ is clearly an isometry. We introduce the operator $\Delta_{\mathrm{S}} := \mathbf{diag}(\mathbf{S}I_n, \mathbf{S}^{-1}I_n)$ on $\ell_2^{n_{\mathrm{S}}}(\mathbb{Z})$. The infinite system associated to building block (1) is described by

$$\frac{d}{dt}x(t) = A_{\mathrm{TT}}x(t) + A_{\mathrm{TS}}v(t) + B_{\mathrm{T}}d(t) \quad \text{for all } t \geq 0\ ;\ x(0) = x^0, \tag{3a}$$

$$(\Delta_{\mathrm{S}} - A_{\mathrm{SS}})\,[v(t)] = A_{\mathrm{ST}}x(t) + B_{\mathrm{S}}d(t) \ \text{ for all } t \geq 0, \tag{3b}$$

$$z(t) = C_{\mathrm{T}}x(t) + C_{\mathrm{S}}v(t) + Dd(t) \quad \text{for all } t \geq 0, \tag{3c}$$

where we have used the shorthand

$$A_{\mathrm{ST}} := \begin{pmatrix} A_{\mathrm{ST}_+} \\ A_{\mathrm{ST}_-} \end{pmatrix},\ A_{\mathrm{SS}} := \begin{pmatrix} A_{\mathrm{SS}_{+,+}} & A_{\mathrm{SS}_{+,-}} \\ A_{\mathrm{SS}_{-,+}} & A_{\mathrm{SS}_{-,-}} \end{pmatrix},\ B_{\mathrm{S}} := \begin{pmatrix} B_{\mathrm{S}_+} \\ B_{\mathrm{S}_-} \end{pmatrix},$$

$$A_{\mathrm{TS}} := \begin{pmatrix} A_{\mathrm{TS}_+} & A_{\mathrm{TS}_-} \end{pmatrix},\ C_{\mathrm{S}} := \begin{pmatrix} C_{\mathrm{S}_+} & C_{\mathrm{S}_-} \end{pmatrix}.$$

In (3), the triple $(x(t), v(t), z(t))$ is sought in $\ell_2^{n_{\mathrm{T}}}(\mathbb{Z}) \times \ell_2^{n_{\mathrm{S}}}(\mathbb{Z}) \times \ell_2^p(\mathbb{Z})$ for all $t \geq 0$, when an initial condition $x^0 \in \ell_2^{n_{\mathrm{T}}}(\mathbb{Z})$ and a disturbance d such that $d(t) \in \ell_2^m(\mathbb{Z})$ for all $t \geq 0$ are given. The question of whether such a triple exists is addressed in section 3. We answer it by rewriting the infinite set of differential-algebraic equations

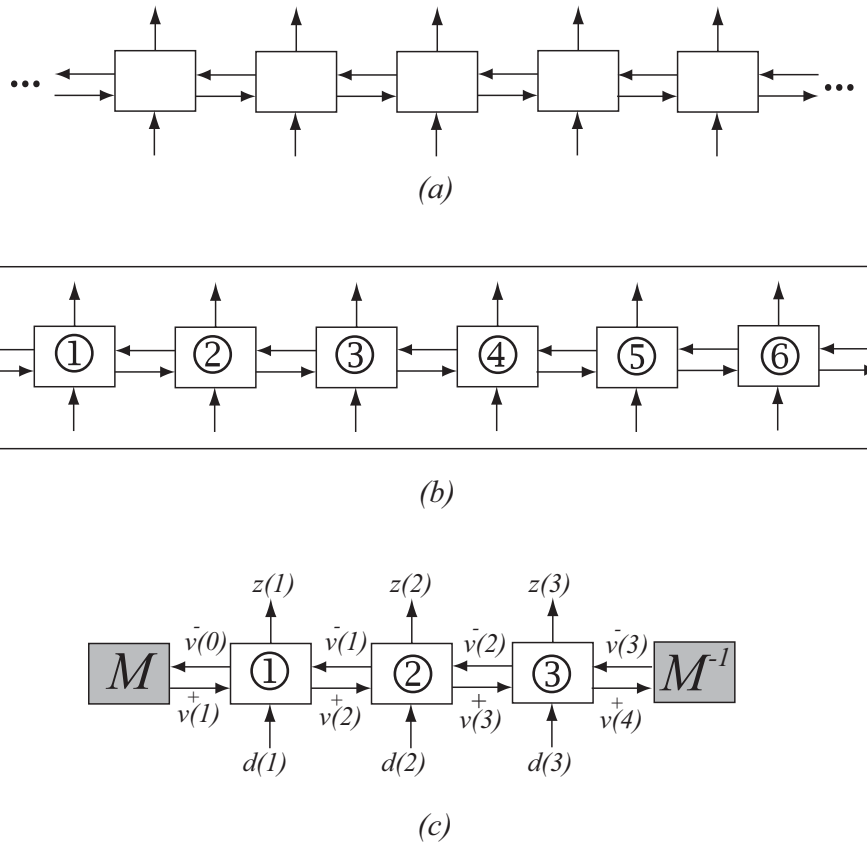


FIG. 1. *The three types of interconnected systems.* (a) *Infinite system.* (b) *Periodic system for* $L = 3$. (c) *Finite extent system for* $L = 3$ *(time dependence of signals is not indicated to simplify notation).*

(3) as an abstract differential equation on the Hilbert space $\ell_2^{n_{\mathrm{T}}}(\mathbb{Z})$. However, it should be noted that infinite interconnected systems appear more naturally in the form of (3) than as abstract differential equations. Indeed, consider the block diagram pictured in Figure 1(a). Each box stands for an instance of the basic building block (1) that exchanges signals v^+, v^-, w^+, and w^- with its neighbors according to the interconnection relation

$$[v^+(t)](s+1) = [w^+(t)](s), \ s \in \mathbb{Z}, \tag{4a}$$

$$[v^-(t)](s-1) = [w^-(t)](s), \ s \in \mathbb{Z}, \tag{4b}$$

where we have indexed the subsystems by $s \in \mathbb{Z}$ and considered all signals mentioned before as vector-valued functions on $\mathbb{Z}$ for all t. Introducing v such that

$$[v(t)](s) := \big([v^+(t)](s), [v^-(t)](s)\big) \ \text{ for all } t \geq 0 \tag{5}$$

and recalling the definition of the operator Δ_{S}, it is easy to see that conditions (4) and the state space description of each subsystem yield differential-algebraic equations (3).

2.2.2. Periodic system. A periodic interconnected system is also captured by (3) but with the operators $\mathbf{S}$ and Δ_{S} now defined, respectively, on $\ell_2(\mathbb{Z}_{2L})$ and

$\ell_2^{n_s}(\mathbb{Z}_{2L})$. In particular, the shift operator $\mathbf{S}$ is still defined by (2) but addition should now be understood modulo $2L$. Accordingly, the triple $(x(t), v(t), z(t))$ is sought in $\ell_2^{n_T}(\mathbb{Z}_{2L}) \times \ell_2^{n_s}(\mathbb{Z}_{2L}) \times \ell_2^{p}(\mathbb{Z}_{2L})$ for all $t \geq 0$, when an initial condition $x^0 \in \ell_2^{n_T}(\mathbb{Z}_{2L})$ and a disturbance d such that $d(t) \in \ell_2^{m}(\mathbb{Z}_{2L})$ for all $t \geq 0$ are given.

The physical interconnection corresponding to a periodic system is illustrated in Figure 1(b). The subsystems are again instances of the basic building block and are interconnected according to the relation

$$[v^+(t)](s+1) = [w^+(t)](s),\ s \in \mathbb{Z}_{2L}, \tag{6a}$$

$$[v^-(t)](s-1) = [w^-(t)](s),\ s \in \mathbb{Z}_{2L}. \tag{6b}$$

For reasons that should become clear in section 3, we will say that periodic and infinite systems are spatially invariant.

2.2.3. Finite extent system. Unlike infinite and periodic systems that can be readily defined in a formal setting, finite extent systems are easier to introduce through the physical interconnection they describe. Consider the block diagram of Figure 1(c). As in the infinite and periodic case, each box represents an instance of the basic building block, except the two end ones, which specify boundary conditions. More precisely, if we index subsystems by $1 \leq s \leq L$, the interconnection relation between neighboring subsystems now is

$$[v^+(t)](s+1) = [w^+(t)](s),\ 1 \leq s \leq L-1, \tag{7a}$$

$$[v^-(t)](s-1) = [w^-(t)](s),\ 2 \leq s \leq L, \tag{7b}$$

$$[v^+(t)](1) = M[w^-(t)](1), \tag{7c}$$

$$[v^-(t)](L) = M^{-1}[w^+(t)](L), \tag{7d}$$

where M is a nonsingular matrix called the *boundary conditions matrix*. We can represent such a finite extent system by a set of differential-algebraic equations formally similar to that describing infinite and periodic systems. To this end, we need to introduce the operator Δ_{BC} as follows. First, if $v = (v^+, v^-)$ belongs to $\ell_2^{n_s}(\{1, \ldots, L\})$, we define the vector $\overrightarrow{v} \in \mathbb{R}^{n_s L}$ by

$$\overrightarrow{v} = \left(v^+(1), \ldots, v^+(L), v^-(1), \ldots, v^-(L)\right).$$

The map $\overrightarrow{\cdot}$ is an isomorphism of $\mathbb{R}$-vector spaces and its inverse will be denoted $\overleftarrow{\cdot}$. As a consequence, we can define another isomorphism, also denoted $\overleftarrow{\cdot}$, between the space of $n_s L \times n_s L$ real matrices and the space of endomorphisms of $\ell_2^{n_s}(\{1, \ldots, L\})$ by

$$\overleftarrow{J}\, v := \overleftarrow{(J\, \overrightarrow{v})} \text{ for all } J \in \mathbb{R}^{n_s L \times n_s L},\ v \in \ell_2^{n_s}(\{1, \ldots, L\}).$$

With this notation, we can rewrite the interconnection relation (7) as

$$w = \overleftarrow{\mathcal{C}}\, v$$

for the invertible *interconnection matrix* $\mathcal{C}$. In the remainder of this paper, we will let $\overleftarrow{\mathcal{C}} =: \Delta_{\text{BC}}$. Then, introducing again signal v as per (5), a finite extent system can be represented by the following set of differential-algebraic equations:

$$\frac{d}{dt}x(t) = A_{\text{TT}}x(t) + A_{\text{TS}}v(t) + B_{\text{T}}d(t) \ \text{ for all } t \geq 0\ ;\ x(0) = x^0, \tag{8a}$$

$$(\Delta_{\text{BC}} - A_{\text{SS}})[v(t)] = A_{\text{ST}}x(t) + B_{\text{S}}d(t) \ \text{ for all } t \geq 0, \tag{8b}$$

$$z(t) = C_{\text{T}}x(t) + C_{\text{S}}v(t) + Dd(t) \ \text{ for all } t \geq 0, \tag{8c}$$

which is formally similar to (3). The triple $(x(t), v(t), z(t))$ is sought in $\ell_2^{n_{\mathrm{T}}}(\{1,\dots,L\}) \times \ell_2^{n_{\mathrm{S}}}(\{1,\dots,L\}) \times \ell_2^{p}(\{1,\dots,L\})$ for all $t \geq 0$, when an initial condition $x^0 \in \ell_2^{n_{\mathrm{T}}}(\{1,\dots,L\})$ and a disturbance d such that $d(t) \in \ell_2^{m}(\{1,\dots,L\})$ for all $t \geq 0$ are given.

As already mentioned in the introduction, analysis is much more tractable for spatially invariant systems than for finite extent systems, especially if L is large. Hence, it would be desirable to know what relationships exist between them. The main goal of the next sections is to show that stability and performance of the spatially invariant systems actually imply similar properties for the corresponding finite extent system, provided some reversibility properties are satisfied. A proof of this statement, which we call the method of images, as well as a precise definition of what we call "reversibility" are given in section 4. Before presenting these results, more should be said about well-posedness, stability, and performance.

3. Well-posedness, stability, and performance. We have just seen that all interconnected systems of interest can be captured by the following equations:

$$\frac{d}{dt}x(t) = A_{\mathrm{TT}}x(t) + A_{\mathrm{TS}}v(t) + B_{\mathrm{T}}d(t) \quad \text{for all } t \geq 0;\ x(0) = x^0, \tag{9a}$$

$$(\Delta - A_{\mathrm{SS}})\,[v(t)] = A_{\mathrm{ST}}x(t) + B_{\mathrm{S}}d(t) \ \text{ for all } t \geq 0, \tag{9b}$$

$$z(t) = C_{\mathrm{T}}x(t) + C_{\mathrm{S}}v(t) + Dd(t) \quad \text{for all } t \geq 0, \tag{9c}$$

where

$$\begin{array}{ll} \Delta = \Delta_{\mathrm{BC}} & \text{for a finite extent system,} \\ \Delta = \Delta_{\mathrm{S}} & \text{for an infinite or periodic system.} \end{array}$$

System (9) is said to be *well-posed* if the bounded linear operator $(\Delta - A_{\mathrm{SS}}) : \ell_2^{n_{\mathrm{S}}}(\mathbb{M}) \to \ell_2^{n_{\mathrm{S}}}(\mathbb{M})$ is invertible. Assume system (9) is well-posed and let an initial state $x^0 \in \ell_2(\mathbb{M})$ and a disturbance $d \in \mathcal{L}_2(\mathbb{M})$ be given. We can write

$$v(t) = (\Delta - A_{\mathrm{SS}})^{-1}\left(A_{\mathrm{ST}}x(t) + B_{\mathrm{S}}d(t)\right) \quad \text{for all } t \geq 0 \tag{10}$$

and, in turn, x will satisfy

$$\frac{d}{dt}x(t) = \mathbf{A}x(t) + \mathbf{B}d(t) \ \text{ for all } t \geq 0, \tag{11a}$$

$$x(0) = x^0, \tag{11b}$$

where

$$\mathbf{A} = \left(A_{\mathrm{TT}} + A_{\mathrm{TS}}(\Delta - A_{\mathrm{SS}})^{-1}A_{\mathrm{ST}}\right);\ \mathbf{B} = \left(B_{\mathrm{T}} + A_{\mathrm{TS}}(\Delta - A_{\mathrm{SS}})^{-1}B_{\mathrm{S}}\right). \tag{12}$$

Note that operators $\mathbf{A}$ and $\mathbf{B}$ are bounded and thus, in particular, $\mathbf{A}$ generates a strongly continuous semigroup of operators $\{\Phi(t)\}_{t\geq 0}$ on $\ell_2(\mathbb{M})$. We can even write

$$\Phi(t) = e^{t\mathbf{A}}, \tag{13}$$

where the exponential is defined by the usual power series. As a result [4], (11a) has a unique weak solution on $[0,T]$ for any $x^0 \in \ell_2(\mathbb{M})$ and $d \in \mathcal{L}_2(\mathbb{M})$, which is the mild solution given by

$$x(t) = e^{t\mathbf{A}}x^0 + \int_0^t e^{(t-\tau)\mathbf{A}}(\mathbf{B}d(\tau))d\tau\ .$$

The *solution* of well-posed system (9) on the interval $[0,T]$ is the unique triple (x,v,z) such that x is the mild solution of (11a) on $[0,T]$ and (9c) and (10) are satisfied for all $0 \le t \le T$. It is not hard to see that if (x,v,z) is the solution of a periodic or infinite system for initial condition x^0 and disturbance $d(t)$, then $(\tilde{x},\tilde{v},\tilde{z})$ is another solution for initial condition $\mathbf{S}x^0$ and disturbance $\mathbf{S}d(t)$, where

$$\tilde{x}(t) = \mathbf{S}[x(t)],\ \tilde{v}(t) = \mathbf{S}[v(t)],\ \tilde{z}(t) = \mathbf{S}[z(t)] \text{ for all t.}$$

This is the reason why we chose to call these systems spatially invariant. The physical explanation of this invariance is that all subsystems in block diagrams 1(a) and 1(b) are identical and interconnected to their neighbors in the same way.

We now show that well-posedness can be characterized algebraically.

PROPOSITION 3.1. (i) *A finite extent system is well-posed if and only if* $(I-N)$ *is invertible, where* N *is defined as*

(14)
$$\left(\begin{array}{c|c} MA_{\mathrm{SS}_{-,+}} \quad 0^{n\times n(L-1)} & MA_{\mathrm{SS}_{-,-}} \quad 0^{n\times n(L-1)} \\ \hline \mathcal{T}_{\mathrm{L}}(A_{\mathrm{SS}_{+,+}}) & \mathcal{T}_{\mathrm{L}}(A_{\mathrm{SS}_{+,-}}) \\ \hline \mathcal{T}_{\mathrm{U}}(A_{\mathrm{SS}_{-,+}}) & \mathcal{T}_{\mathrm{U}}(A_{\mathrm{SS}_{-,-}}) \\ \hline 0^{n\times n(L-1)} \quad M^{-1}A_{\mathrm{SS}_{+,+}} & 0^{n\times n(L-1)} \quad M^{-1}A_{\mathrm{SS}_{+,-}} \end{array}\right),$$

and the rectangular Toeplitz matrices $\mathcal{T}_{\mathrm{U}}(K)$ *and* $\mathcal{T}_{\mathrm{L}}(K)$ *in* $\mathbb{R}^{n(L-1)\times nL}$ *satisfy*

$$\mathcal{T}_{\mathrm{U}}(K) := \begin{pmatrix} 0 & K & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & K \end{pmatrix};\ \mathcal{T}_{\mathrm{L}}(K) := \begin{pmatrix} K & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & K & 0 \end{pmatrix}$$

for any given $K \in \mathbb{R}^{n\times n}$.

(ii) *A periodic (respectively, infinite) system is well-posed if and only if* $(\Delta(\lambda) - A_{\mathrm{SS}})$ *is invertible for all* $\lambda \in \mathbb{U}$ *(respectively, for all* $\lambda \in \partial\mathbb{D}$*), where*

$$\Delta(\lambda) := \mathbf{diag}(\lambda I_n, \lambda^{-1} I_n) \text{ and } \partial\mathbb{D} := \{z \in \mathbb{C}, |z| = 1\}, \mathbb{U} := \{z \in \mathbb{C}, z^{2L} = 1\}.$$

Proof. (i) Since the interconnection matrix $\mathcal{C}$ is invertible, so is Δ_{BC}. It is easy to see that N satisfies $\overleftarrow{N} = \Delta_{\mathrm{BC}}^{-1} A_{\mathrm{SS}}$. Hence matrix $(I-N)$ is nonsingular if and only if $(\Delta_{\mathrm{BC}} - A_{\mathrm{SS}})$ is. Since $\ell_2(\{1,\ldots,L\})$ is finite dimensional, $(\Delta_{\mathrm{BC}} - A_{\mathrm{SS}})^{-1}$ is bounded whenever it exists.

(ii) We first study the periodic case. Let $v \in \ell_2(\mathbb{Z}_{2L})$ be given. It has a discrete Fourier transform $\hat{v} \in \ell_2(\mathbb{U})$ defined by

$$\hat{v}(\lambda) = \sum_{s=1}^{2L} v(s)\lambda^s \text{ for all } \lambda \in \mathbb{U}.$$

Now assume $(\Delta(\lambda) - A_{\mathrm{SS}})$ is invertible for all $\lambda \in \mathbb{U}$. Then n defined by

$$n(s) = \frac{1}{2L} \sum_{\lambda\in\mathbb{U}} \lambda^{-s} \left(\Delta(\lambda) - A_{\mathrm{SS}}\right)^{-1} \hat{v}(\lambda)$$

is a well-defined function on $\mathbb{Z}_{2L}$. Also, noting that

$$\hat{n}(\lambda) = \left(\Delta(\lambda) - A_{\mathrm{SS}}\right)^{-1} \hat{v}(\lambda) \text{ for all } \lambda \tag{15}$$

and using Parseval's identity, we get that

$$\|n\|_{\ell_2(\mathbb{Z}_{2L})} \leq \max_{\lambda\in\mathbb{U}} \overline{\sigma}((\Delta(\lambda) - A_{\rm ss})^{-1})\|v\|_{\ell_2(\mathbb{Z}_{2L})} < \infty.$$

Finally, it is easy to check that it satisfies

$$(\Delta_{\rm s} - A_{\rm ss})\, n = v. \tag{16}$$

This is the unique solution in $\ell_2(\mathbb{Z}_{2L})$ since any such solution must satisfy (15) and elements of $\ell_2(\mathbb{Z}_{2L})$ are fully specified by their Fourier coefficients. Hence, $(\Delta_{\rm s} - A_{\rm ss})^{-1}$ is well defined on $\ell_2(\mathbb{Z}_{2L})$ and has norm less than $\max_{\lambda\in\mathbb{U}} \overline{\sigma}((\Delta(\lambda) - A_{\rm ss})^{-1})$.

Conversely, assume $(\Delta(\lambda) - A_{\rm ss})$ is not invertible for some $\lambda = \lambda_0$ in $\mathbb{U}$ and let $n_0 \neq 0$ be in the corresponding null-space. Then n defined by

$$n(s) = \lambda_0^s n_0 \ \text{ for all } s$$

belongs to $\ell_2(\mathbb{Z}_{2L})$ and satisfies (16) with $v \equiv 0$. Hence operator $(\Delta_{\rm s} - A_{\rm ss})$ is not invertible on $\ell_2(\mathbb{Z}_{2L})$.

For the infinite system case, the proof of sufficiency is identical, replacing the discrete Fourier transform with a two-sided $\mathcal{Z}$-transform. For necessity, assuming that there exists $\lambda_0 = e^{i\omega_0} \in \partial\mathbb{D}$ such that $(\Delta(\lambda_0) - A_{\rm ss})$ is singular, we will show that $(\Delta_{\rm s} - A_{\rm ss})$ is not bounded below on $\ell_2(\mathbb{Z})$ and hence not injective. Let ξ_0 be a unitary vector in the null-space of $(\Delta(\lambda_0) - A_{\rm ss})$. Define the sequence of functions $\{u_k\}$ from $[0, 2\pi)$ to $\mathbb{R}$ by

$$u_k(\omega) = \begin{cases} 2^{\frac{k-1}{2}} \text{ if } |\omega - \omega_0| < \left(\frac{1}{2}\right)^k, \\ 0 \text{ otherwise,} \end{cases}$$

and, for each k, let $\hat{n}_k \in \ell_2(\partial\mathbb{D})$ be defined by $\hat{n}_k(\lambda) = u_k(\omega)\xi_0$ for all $\lambda = e^{i\omega}$ and n_k be the inverse $\mathcal{Z}$-transform of $\hat{n}_k$, which thus belongs to $\ell_2(\mathbb{Z})$. We have $\|n_k\|_{\ell_2(\mathbb{Z})} = 1$ for all k. Now, if we let $v_k = (\Delta_{\rm s} - A_{\rm ss})\, n_k$ for each k, we get

$$\|v_k\|^2_{\ell_2(\mathbb{Z})} = \|\hat{v}_k\|^2_{\ell_2(\partial\mathbb{D})} = 2^{(k-1)} \int_{\omega-(\frac{1}{2})^k}^{\omega+(\frac{1}{2})^k} |\left(\Delta(e^{i\omega}) - A_{\rm ss}\right) n_0|^2 d\omega.$$

Let $\epsilon > 0$. Since $\omega \mapsto \|\left(\Delta(e^{i\omega}) - A_{\rm ss}\right) n_0\|^2$ is continuous and $\left(\Delta(e^{i\omega_0}) - A_{\rm ss}\right) n_0 = 0$, there exists K such that

$$\|\left(\Delta(e^{i\omega}) - A_{\rm ss}\right) n_0\|^2 < \epsilon^2 \ \text{ for all } |\omega - \omega_0| < \left(\frac{1}{2}\right)^k,$$

provided $k > K$. Hence, $\|v_k\|_{\ell_2(\mathbb{Z})} < \epsilon$ for $k > K$. Since this holds for any $\epsilon > 0$, the sequence $\left\{\|v_k\|_{\ell_2(\mathbb{Z})}\right\}_k$ converges to zero, showing that $(\Delta_{\rm s} - A_{\rm ss})$ is not bounded below. □

Since $\mathbb{U} \subset \partial\mathbb{D}$, well-posedness of the infinite system implies well-posedness of the corresponding periodic system. For a well-posed system, one can define stability as follows.

Definition 3.2. *A well-posed system is stable if, in the absence of input* ($d \equiv 0$), *the weak solution* $x(t) \in \ell_2(\mathbb{M})$ *of* (11a) *is defined on* $\mathbb{R}^+$ *and satisfies*

$$\|x(t)\|_{\ell_2(\mathbb{M})} \xrightarrow[t\to\infty]{} 0 \ \textit{exponentially, irrespective of the initial condition } x^0 \ .$$

Equivalently, this means that there exist $M,\ \alpha > 0$ *such that*

$$\|\Phi(t)\|_{\ell_2(\mathbb{M})} \le Me^{-\alpha t} \quad \textit{for all } t \ge 0,$$

where the norm in the latter equation is the $\ell_2(\mathbb{M})$*-induced norm of an operator.*

It follows from the results of [1] that stability of periodic and infinite systems can be checked by looking at the corresponding Fourier-transformed systems. More precisely, if we associate the operator $\widehat{\mathbf{A}} := \mathcal{F}\mathbf{A}\mathcal{F}^{-1}$ on $\ell_2(\widehat{\mathbb{M}})$ to the operator $\mathbf{A}$ on $\ell_2(\mathbb{M})$, where

$$\begin{array}{ll} \widehat{\mathbb{M}} = \partial\mathbb{D}, & \mathcal{F} \text{ is the two-sided } \mathcal{Z}\text{-transform for an infinite system,} \\ \widehat{\mathbb{M}} = \mathbb{U}, & \mathcal{F} \text{ is the discrete Fourier transform for a periodic system,} \end{array}$$

then we have the following.

PROPOSITION 3.3 (see [1]). *The following hold:*

(i) $\widehat{\mathbf{A}}$ *is a multiplication operator; i.e., there exists a matrix-valued function* A *such that* $(\widehat{\mathbf{A}}\hat{f})(\lambda) = A(\lambda)\hat{f}(\lambda)$ *for all* $\hat{f} \in \ell_2(\widehat{\mathbb{M}})$.

(ii) *The periodic and infinite system* (9) *is stable if and only if* $A(\lambda)$ *is Hurwitz for all* $\lambda \in \widehat{\mathbb{M}}$.

Note that (ii) is in fact simpler than the general condition of [1] for stability, owing to the compactness of $\widehat{\mathbb{M}}$. Once again, since $\mathbb{U} \subset \partial\mathbb{D}$, stability of the infinite system implies stability of the corresponding periodic system.

It is easy to see that if a system is well-posed and stable then, for any $d \in \mathcal{L}_2(\mathbb{M})$, x and in turn z also belong to $\mathcal{L}_2(\mathbb{M})$. Such a system thus has a well-defined input/output map, T_{dz}. It is a bounded linear map from $\mathcal{L}_2(\mathbb{M})$ to $\mathcal{L}_2(\mathbb{M})$ and its induced norm, $\|T_{dz}\|_{\mathcal{L}_2(\mathbb{M})}$, characterizes the performance of the system. If it is strictly less than 1, we will say that the system is contractive.

4. Spatial reversibility and the method of images. We now turn our attention to a particular class of finite extent systems, as defined below.

DEFINITION 4.1. *Given a basic building block as per* (1) *and a nonsingular matrix* M*, we say that the block is (spatially)* M*-reversible if there exist matrices* $R \in \mathbb{R}^{m\times m}$*,* $P \in \mathbb{R}^{n_{\mathrm{T}}\times n_{\mathrm{T}}}$*, and* $U \in \mathbb{R}^{p\times p}$ *such that*

(i) $R^2 = I_m$, $P^2 = I_{n_{\mathrm{T}}}$, $U^2 = I_p$*; i.e.,* R*,* U*, and* P *are involutions,*

(ii)
$$\begin{pmatrix} P & 0 & 0 \\ 0 & Q & 0 \\ 0 & 0 & U \end{pmatrix} \begin{pmatrix} A_{\mathrm{TT}} & A_{\mathrm{TS}} & B_{\mathrm{T}} \\ A_{\mathrm{ST}} & A_{\mathrm{SS}} & B_{\mathrm{S}} \\ C_{\mathrm{T}} & C_{\mathrm{S}} & D \end{pmatrix} = \begin{pmatrix} A_{\mathrm{TT}} & A_{\mathrm{TS}} & B_{\mathrm{T}} \\ A_{\mathrm{ST}} & A_{\mathrm{SS}} & B_{\mathrm{S}} \\ C_{\mathrm{T}} & C_{\mathrm{S}} & D \end{pmatrix} \begin{pmatrix} P & 0 & 0 \\ 0 & Q & 0 \\ 0 & 0 & R \end{pmatrix},$$

where $Q := \left(\begin{smallmatrix} 0 & M \\ M^{-1} & 0 \end{smallmatrix}\right)$.

We will say that a finite extent, periodic, or infinite system is M-reversible if the basic building block is. When the finite extent system at hand has boundary conditions matrix M and is M-reversible, we will simply say that it is reversible without referring to the matrix. Our goal in this section is to relate the properties of reversible finite extent and periodic systems. This will require several properties that we explain in turn. We start with a result that motivates our use of the notion of spatial reversibility.

First, we introduce the reflection $\Upsilon : \ell_2(\mathbb{Z}_{2L}) \to \ell_2(\mathbb{Z}_{2L})$ such that $(\Upsilon x)(s) =$

$x(2L+1-s)$ for all s. Then we consider the following spaces of *reversible* signals:

$$\begin{aligned}\Re_{\mathrm{T}} &:= \{x \in \ell_2^{n_{\mathrm{T}}}(\mathbb{Z}_{2L}) : x = P\Upsilon x\}\,,\\ \Re_{\mathrm{S}} &:= \{v \in \ell_2^{n_{\mathrm{S}}}(\mathbb{Z}_{2L}) : v = Q\Upsilon v\ \}\,,\\ \Re_d &:= \{d \in \ell_2^{m}(\mathbb{Z}_{2L}) : d = R\Upsilon d\}\,,\\ \Re_z &:= \{z \in \ell_2^{p}(\mathbb{Z}_{2L}) : z = U\Upsilon z\}\,.\end{aligned}$$

Then the fact that Q and Υ, as seen as operators on $\ell_2^{n_{\mathrm{S}}}(\mathbb{Z}_{2L})$, satisfy

$$\Delta_{\mathrm{S}} Q = Q\Delta_{\mathrm{S}}^{-1}\ ;\ \Delta_{\mathrm{S}}\Upsilon = \Upsilon\Delta_{\mathrm{S}}^{-1} \tag{17}$$

yields the following property.

PROPOSITION 4.2. *Assume periodic system* (3) *is M-reversible and well-posed. Let the initial state x^0 belong to $\Re_{\mathrm{T}}$ and disturbance $d \in \mathcal{L}_2(\mathbb{Z}_{2L})$ satisfy $d(t) \in \Re_d$ for all $t \geq 0$. Then the corresponding solution (x, v, z) of* (8) *on $\mathbb{R}^+$ is spatially reversible, i.e., $x(t) \in \Re_{\mathrm{T}}$, $v(t) \in \Re_{\mathrm{S}}$, and $z(t) \in \Re_z$ for all $t \geq 0$.*

The proof relies on manipulations very similar to those used later for Theorem 4.5 and we thus omit it. Physically, Proposition 4.2 means that, for the right type of inputs and initial conditions, the signals flowing to the right from the Lth subsystem are related to those flowing to the left from the $(L+1)$th subsystem. Hence switching left and right is equivalent to operating Q on v, P on x, and R on z. This property allows us to draw a parallel between spatially reversible and *time-reversible* dynamical systems. Recall that a nonlinear autonomous dynamical system

$$\dot{x} = f(x) \tag{18}$$

is time-reversible if there exists an involution R that anticommutes with f. Then for every solution x of the differential equation (18) we have another, i.e., $\tilde{x} : t \mapsto \tilde{x}(t) = Rx(-t)$. If there exists $t^\star$ such that $x(t^\star) \in Fix(R) = \{\xi : \xi = R\xi\}$, then the solution x is reversible.

In both spatial and temporal cases, the key property is some kind of anticommutation of an involution with an evolution operator ((17) in the spatial case) and the result is that solutions either "come in pairs" or are reversible. It is because of this analogy that the denomination "spatially reversible" was used in our definition, although "symmetric" has sometimes been used in the literature with a somewhat similar meaning [16], [21]. This latter denomination is acceptable because (17) and Definition 4.1 essentially mean that the set of equations describing the periodic system is equivariant under the action of $\mathbb{Z}_2$. However, we feel that it is desirable to keep the adjective "symmetric" for systems that are invariant under the action of more general groups, as is done in [10].

The second useful result is given by the following proposition.

PROPOSITION 4.3. *A spatially reversible finite extent system is well-posed if the corresponding periodic system is well-posed.*

Proof. We use a contrapositive. Assume the finite extent system is not well-posed and let M be its boundary conditions matrix. We want to show that $(\Delta(\theta) - A_{\mathrm{ss}})$ is singular for some $\theta \in \mathbb{U}$. Let $\omega = e^{i\pi/L}$ so that

$$\mathbb{U} = \{1, \omega, \omega^2, \ldots, \omega^{(2L-1)}\}.$$

According to Proposition 3.1, there exists $x \neq 0$ such that $Nx = x$, where N is defined as in (14). Also, because of spatial reversibility, we have that $QA_{\mathrm{ss}} = A_{\mathrm{ss}}Q$,

namely,

$$\begin{pmatrix} MA_{\mathrm{SS}_{-,-}}M^{-1} & MA_{\mathrm{SS}_{-,+}}M \\ M^{-1}A_{\mathrm{SS}_{+,-}}M^{-1} & M^{-1}A_{\mathrm{SS}_{+,+}}M \end{pmatrix} = \begin{pmatrix} A_{\mathrm{SS}_{+,+}} & A_{\mathrm{SS}_{+,-}} \\ A_{\mathrm{SS}_{-,+}} & A_{\mathrm{SS}_{-,-}} \end{pmatrix}. \tag{19}$$

As a result, $\mathcal{Q}$ defined by

$$\mathcal{Q} := \left(\begin{array}{ccc|ccc} & & & 0 & & M \\ & 0 & & & \diagup & \\ & & & M & & 0 \\ \hline 0 & & M^{-1} & & & \\ & \diagup & & & 0 & \\ M^{-1} & & 0 & & & \end{array}\right)$$

commutes with N and $\mathcal{Q}x$ is also an eigenvector of N with eigenvalue 1. There are two cases as follows.

Case 1. $x = -\mathcal{Q}x$.

Since the Vandermonde matrix $V(\omega, \omega^3, \ldots, \omega^{(2L-1)})$ defined by

$$V(\omega, \omega^3, \ldots, \omega^{(2L-1)}) := \begin{pmatrix} I_n & \frac{1}{\omega}I_n & \cdots & \left(\frac{1}{\omega}\right)^{(L-1)} I_n \\ I_n & \frac{1}{\omega^3}I_n & \cdots & \left(\frac{1}{\omega^3}\right)^{(L-1)} I_n \\ \vdots & \vdots & & \vdots \\ I_n & \frac{1}{\omega^{(2L-1)}}I_n & \cdots & \left(\frac{1}{\omega^{(2L-1)}}\right)^{(L-1)} I_n \end{pmatrix}$$

is invertible, there exists $1 \le k \le L$ such that $\sum_{l=1}^{L} \left(\frac{1}{\omega^{(2k-1)}}\right)^{(l-1)} x_l \neq 0$, for otherwise we would have $x_1 = \cdots = x_L = 0$ and, in turn, since $x = -\mathcal{Q}x$, $x = 0$, a contradiction.

For this k, let

$$x^+ = \sum_{l=1}^{L} \left(\frac{1}{\omega^{(2k-1)}}\right)^{(l-1)} x_l \; ; \; x^- = \sum_{l=1}^{L} \left(\frac{1}{\omega^{(2k-1)}}\right)^{(l-1)} x_{(L+l)} \; ; \; X = \begin{pmatrix} x^+ \\ x^- \end{pmatrix}.$$

Note that $X = -QX \neq 0$ and

$$\begin{aligned} \omega^{(2k-1)}x^+ &= A_{\mathrm{SS}_{+,+}} \left[\omega^{(2k-1)} M x_{(L+1)} + x_1 + \cdots + \left(\frac{1}{\omega^{(2k-1)}}\right)^{(L-2)} x_{(L-1)}\right] \\ &\quad + A_{\mathrm{SS}_{+,-}} \left[\omega^{(2k-1)} M^{-1} x_1 + x_{(L+1)} + \cdots + \left(\frac{1}{\omega^{(2k-1)}}\right)^{(L-2)} x_{(2L-1)}\right] \\ &= A_{\mathrm{SS}_{+,+}} x^+ + A_{\mathrm{SS}_{+,-}} x^- \end{aligned}$$

since $x_L = -Mx_{(L+1)}, x_1 = -Mx_{2L}$, and $-\omega^{(2k-1)} = \left(\frac{1}{\omega^{(2k-1)}}\right)^{(L-1)}$.

Likewise, we get

$$\left(\frac{1}{\omega^{(2k-1)}}\right) x^- = A_{\mathrm{SS}_{-,+}} x^+ + A_{\mathrm{SS}_{-,-}} x^- \quad .$$

Hence, $(\Delta(\omega^{(2k-1)}) - A_{\mathrm{SS}})X = 0$ with $X \neq 0$.

Case 2. $x \neq -\mathcal{Q}x$.
Define $z := x + \mathcal{Q}x$. Since it is nonzero, it is an eigenvector of N with eigenvalue 1. Also, $z = \mathcal{Q}z$. Considering $V(1, \omega^2, \ldots, \omega^{(2L-2)})$ as for Case 1, one deduces that there exists $0 \leq k \leq (L-1)$ such that $\sum_{l=1}^{L} \left(\frac{1}{\omega^{(2k)}}\right)^{(l-1)} z_l \neq 0$.
Then, if we let

$$z^+ = \sum_{l=1}^{L} \left(\frac{1}{\omega^{(2k)}}\right)^{(l-1)} z_l;\ z^- = \sum_{l=1}^{L} \left(\frac{1}{\omega^{(2k)}}\right)^{(l-1)} z_{(L+l)};\ Z = \begin{pmatrix} z^+ \\ z^- \end{pmatrix},$$

we get $(\Delta(\omega^{(2k)}) - A_{\mathrm{SS}})Z = 0$ with $Z \neq 0$, after calculations similar to those of Case 1.

Hence, in any case, the periodic system is not well-posed. □

Remark 1. Note that the condition in Proposition 4.3 is sufficient but not necessary, as can be seen by considering the following case where $n = 1$, $M = 1$, and $L = 2$:

$$A_{\mathrm{SS}} = \begin{pmatrix} 3 & 2 \\ 2 & 3 \end{pmatrix}.$$

A_{SS} has an eigenvalue at 1, which means that the periodic system is not well-posed. However, the corresponding matrix $I - N$ is

$$I_4 - \begin{pmatrix} 2 & 0 & 3 & 0 \\ 3 & 0 & 2 & 0 \\ 0 & 2 & 0 & 3 \\ 0 & 3 & 0 & 2 \end{pmatrix} = \begin{pmatrix} -1 & 0 & -3 & 0 \\ -3 & 1 & -2 & 0 \\ 0 & -2 & 1 & -3 \\ 0 & -3 & 0 & -1 \end{pmatrix},$$

which is invertible.

In particular, this means that all analysis results pertaining to periodic systems only yield sufficient conditions for finite extent systems.

Finally, let $\mathbf{H} : \ell_2(\mathbb{Z}_{2L}) \to \ell_2(\{1, \ldots, L\})$ be defined by

$$(\mathbf{H}v)(s) = v(L + s) \text{ for all } s = 1, \ldots, L.$$

Note that the restriction of $\mathbf{H}$ to the reversible subspaces $\mathfrak{R}_{\mathrm{T}}$, $\mathfrak{R}_{\mathrm{S}}$, $\mathfrak{R}_d$, and $\mathfrak{R}_z$ is invertible with, for example,

$$(\mathbf{H}^{-1}_{|\mathfrak{R}_{\mathrm{s}}} v)(s) = \begin{cases} Qv(L + 1 - s) \text{ if } s = 1, \ldots, L, \\ v(s - L) \text{ if } s = (L + 1), \ldots, 2L, \end{cases} \tag{20}$$

and similar relations in the other cases.

PROPOSITION 4.4. *Assume the basic building block* (1) *is reversible. Then*

$$\mathbf{H}\, K = K\, \mathbf{H} \text{ for all matrix } K, \tag{21a}$$

$$\Delta_{\mathrm{S}} \mathbf{H}^{-1}_{|\mathfrak{R}_{\mathrm{s}}} = \mathbf{H}^{-1}_{|\mathfrak{R}_{\mathrm{s}}} \Delta_{\mathrm{BC}}, \tag{21b}$$

$$A_{\mathrm{SS}} \mathbf{H}^{-1}_{|\mathfrak{R}_{\mathrm{s}}} = \mathbf{H}^{-1}_{|\mathfrak{R}_{\mathrm{s}}} A_{\mathrm{SS}}, \tag{21c}$$

Proof. Equation (21a) is clear. Equation (21c) simply follows from the fact that A_{SS} and Q commute. For (21b), we start by showing that $\Delta_{\mathrm{BC}} = \mathbf{H}\Delta_{\mathrm{S}}\mathbf{H}^{-1}_{|\mathfrak{R}_{\mathrm{s}}}$. Let

$w = \mathbf{H}_{|\Re_S}^{-1} v$, $y = \Delta_S w$ and $z = \mathbf{H} y$. Then, for all $1 \le s \le L$,

$$z^+(s) = y^+(L+s) = w^+(L+s+1), \tag{22a}$$

$$z^-(s) = y^-(L+s) = w^-(L+s-1). \tag{22b}$$

If $1 \le s \le L-1$, (22a) means that $z^+(s) = v^+(s+1)$. For $s = L$, $L+s+1 = 2L+1 = 1 \bmod 2L$, and hence $z^+(L) = w^+(1) = Mv^-(L)$, using (20). Likewise, for $2 \le s \le L$, $z^-(s) = v^-(s-1)$ while $z^-(1) = w^-(L) = M^{-1}v^+(1)$. All in all, recalling interconnection relation (7), we see that $z = \Delta_{BC} v$. This shows that

$$\Delta_{BC} = \mathbf{H} \Delta_S \mathbf{H}_{|\Re_S}^{-1}. \tag{23}$$

Finally, (17) implies that $\Re_S$ is a stable subspace for operator Δ_S (i.e., $\Delta_S v \in \Re_S$ if $v \in \Re_S$). We can thus left-multiply (23) by $\mathbf{H}_{|\Re_S}^{-1}$ to get (ii). ∎

We are now in a position to state and prove the main theorem of this section.

THEOREM 4.5 (method of images). *Let a spatially M-reversible finite extent system be such that the corresponding periodic system is well-posed. For an input $d \in \mathcal{L}_2(\{1,\dots,L\})$ and initial state $x^0 \in \ell_2(\{1,\dots,L\})$, let $d^P(t) := \mathbf{H}_{|\Re_d}^{-1} d(t)$ and $(x^0)^P := \mathbf{H}_{|\Re_T}^{-1} x^0$. Let (x^P, v^P, z^P) be the spatially reversible solution of the periodic system with input d^P, initial state $(x^0)^P$. Then (x, v, z) defined by $x(t) := \mathbf{H} x^P(t)$, $v(t) := \mathbf{H} v^P(t)$, and $z(t) := \mathbf{H} z(t)^P$ for all $t \ge 0$ is the unique solution of the finite extent system with input d and initial condition x^0.*

Proof. First, according to Proposition 4.3, the finite extent system is well-posed since the periodic one is. The finite system thus has a unique solution (x, v, z), where x is the mild solution of (11a) for $\Delta = \Delta_{BC}$. Now, note that $x^P(t)$ satisfies

$$x^P(t) = e^{t\mathbf{A}} (x^0)^P + \int_0^t e^{(t-\tau)\mathbf{A}} (\mathbf{B} d^P)(\tau)\, d\tau \tag{24}$$

for $\mathbf{A} = A_{TT} + A_{TS}(\Delta_S - A_{SS})^{-1} A_{ST}$ and $\mathbf{B} = B_T + A_{TS}(\Delta_S - A_{SS})^{-1} B_S$. Hence

$$\left(\mathbf{H} x^P\right)(t) = \mathbf{H} e^{t\mathbf{A}} (x^0)^P + \mathbf{H} \int_0^t e^{(t-\tau)\mathbf{A}} (\mathbf{B} d^P)(\tau)\, d\tau. \tag{25}$$

Using relation (17) and the fact that A_{SS} commutes with Q, it is easy to see that the subspace $\Re_S$ is invariant for the mapping $(\Delta_S - A_{SS})$. Hence, we can write

$$\begin{aligned} \mathbf{H}(\Delta_S - A_{SS})_{|\Re_S}^{-1} &= \left((\Delta_S - A_{SS})_{|\Re_S} \mathbf{H}_{|\Re_S}^{-1}\right)^{-1} \\ &= \left(\mathbf{H}_{|\Re_S}^{-1} (\Delta_{BC} - A_{SS})_{|\Re_S}\right)^{-1}, \end{aligned}$$

where we have used Proposition 4.4. This, coupled with the fact that the basic building block is reversible, yields $\mathbf{H}\mathbf{A}x = \mathbf{A}_{BC}\mathbf{H}x$ for all $x \in \Re_T$ and $\mathbf{H}\mathbf{B}d = \mathbf{B}_{BC}\mathbf{H}d$ for all $d \in \Re_d$, where $\mathbf{A}_{BC} = A_{TT} + A_{TS}(\Delta_{BC} - A_{SS})^{-1} A_{ST}$ and $\mathbf{B}_{BC} = B_T + A_{TS}(\Delta_{BC} - A_{SS})^{-1} B_S$. In particular, this also implies that $\mathbf{H} e^{t\mathbf{A}} (x^0)^P = e^{t\mathbf{A}_{BC}} \mathbf{H} (x^0)^P$, since $(x^0)^P \in \Re_T$. All in all, plugging this back into (24) gives that

$$\mathbf{H} x^P(t) = e^{t\mathbf{A}_{BC}}\, \mathbf{H} (x^0)^P + \int_0^t e^{(t-\tau)\mathbf{A}_{BC}} \mathbf{B}_{BC} (\mathbf{H} d^P)(\tau) d\tau,$$

i.e., that $x(t) = \mathbf{H}x^P(t)$ for all $t \geq 0$ since $(x^0)^P$ and d^P are reversible by construction. Once this is known, it is clear that $v = \mathbf{H}v^P$ and $z = \mathbf{H}z^P$. ∎

An easy but fundamental corollary is that stability and performance of a spatially reversible finite extent system are related to similar properties of the corresponding periodic and infinite system.

COROLLARY 4.6. *If a spatially M-reversible, well-posed, periodic system is stable, then the corresponding finite extent system, with boundary conditions matrix M, is stable. Moreover, the input/output gains of the two systems satisfy*

$$\|T_{dz}\|_{\mathcal{L}_2(\{1,\dots,L\})} \leq \sqrt{\frac{1+\bar{\sigma}(R)^2}{1+\underline{\sigma}(U)^2}}\|T^P_{dz}\|_{\mathcal{L}_2(\mathbb{Z}_{2L})},$$

where T_{dz} and T^P_{dz} are the input/output map of the finite extent and periodic system, respectively. In particular, if R and U are unitary,

$$\|T^P_{dz}\|_{\mathcal{L}_2(\mathbb{Z}_{2L})} < 1 \Rightarrow \|T_{dz}\|_{\mathcal{L}_2(\{1,\dots,L\})} < 1\ .$$

Proof. We first prove stability. First assume that the periodic system is stable and pick an initial condition $x^0 \in \ell_2(\{1,\dots,L\})$ for the finite extent system. Let (x,v,z) be the corresponding solution in the absence of an input, which is uniquely determined since we assumed well-posedness. Let $(x^0)^P$ and x^P be defined as in Theorem 4.5. Then, Theorem 4.5 implies that, for all $t \geq 0$,

$$\|x(t)\|_{\ell_2(\{1,\dots,L\})} \leq \|x^P(t)\|_{\ell_2(\mathbb{Z}_{2L})} \xrightarrow[t\to\infty]{} 0 \text{ exponentially}$$

since the periodic system is stable. Since this holds for any x^0, the finite extent is stable.

For performance, using the notation of Theorem 4.5, we note that

$$\|d^P\|^2_{\mathcal{L}_2(\mathbb{Z}_{2L})} \leq \left(1+\bar{\sigma}(R)^2\right)\|d\|^2_{\mathcal{L}_2(\{1,\dots,L\})}\ .$$

Also $\|z^P\|^2_{\mathcal{L}_2(\mathbb{Z}_{2L})} \geq \left(1+\underline{\sigma}(U)^2\right)\|z\|^2_{\mathcal{L}_2(\{1,\dots,L\})}$. ∎

5. Examples. We now give some practical examples of spatially reversible systems and their corresponding boundary conditions.

Example 1 (two-sided platoon). The following is adapted from [19]. Consider the problem of controlling a platoon of L vehicles such that each has a constant velocity V and is halfway between its predecessor and successor in the line, in spite of external noise. This design requirement captures the notion of "safety" since it ensures that each vehicle is as far away as possible from its two closest neighbors. This system can be described by

$$\dot{e}(t,s) = -\mathrm{v}(t,s) + \frac{1}{2}\left(\mathrm{v}(t,s+1) + \mathrm{v}(t,s-1)\right), \tag{26a}$$

$$\dot{\mathrm{v}}(t,s) = a(t,s), \tag{26b}$$

$$\dot{a}(t,s) = -a(t,s) + u(t,s) + m(t,s), \tag{26c}$$

$$z(t,s) = e(t,s) \ \text{ for all } s = 1,\dots,L, \qquad t \geq 0, \tag{26d}$$

where, in a frame moving with constant velocity V, $\mathrm{v}(.,s)$, $a(.,s)$, $u(.,s)$, $m(.,s)$ are the velocity, acceleration, control, and external noise of the sth vehicle, respectively.

$e(.,s)$ is the difference between the position of the sth vehicle and the middle of its closest neighbors. This system can be put into the standard form (8) by choosing

$$x = \begin{pmatrix} e \\ \mathrm{v} \\ a \end{pmatrix},\ d = \begin{pmatrix} u \\ m \end{pmatrix}, \qquad n_{\mathrm{T}} = 3,\ n = 1,$$

$$A_{\mathrm{TT}} = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & -1 \end{pmatrix},\ A_{\mathrm{TS}} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 0 \\ 0 & 0 \end{pmatrix},\ B_{\mathrm{T}} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 1 \end{pmatrix},$$

$$A_{\mathrm{ST}} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix},\ A_{\mathrm{SS}} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},\ B_{\mathrm{S}} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

$$C_{\mathrm{T}} = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix},\ C_{\mathrm{S}} = \begin{pmatrix} 0 & 0 \end{pmatrix},\ D = \begin{pmatrix} 0 & 0 \end{pmatrix}.$$

Two different sets of relevant boundary conditions can be thought of that will yield a spatially reversible system. Both involve a virtual leader, located in front of the first vehicle of the platoon, and a virtual follower located behind the Lth vehicle of the platoon as follows:

- The virtual leader and follower have the same velocity as the first and last vehicle of the platoon, respectively. This case can be captured by the boundary conditions matrix $M = 1$ and can be used to specify that the platoon should follow the virtual leader. The corresponding finite extent system is then reversible with $P = U = R = I$.
- The virtual leader's (respectively, follower's) velocity is the opposite of the first (respectively, last) vehicle's. This boundary condition can be captured by taking $M = -1$ and corresponds to a case where the leader is reversing in front of the platoon. The corresponding finite extent system is reversible with $P = U = R = -I$.

Example 2 (heat equation). The following is the partial differential equation describing the diffusion of heat in a bar of unit length:

$$\frac{\partial x}{\partial t} = \frac{\partial^2 x}{\partial l^2} + d \ \text{ for all } l \in (0,1), \qquad t \geq 0, \tag{27a}$$

$$x(0,l) = x^0(l) \ \text{ for all } l \in (0,1), \tag{27b}$$

$$x(t,0) = x(t,1) = 0 \ \text{ for all } t \geq 0. \tag{27c}$$

In (27), $x(t,l) \in \mathbb{R}$ is the temperature at time t and position $l \in [0,1]$, and $d(t,l)$ is a distributed heat source. The Dirichlet boundary conditions (27c) mean that the temperature is held constant at both ends, while initial conditions (27b) specify that the initial temperature profile is x^0.

We discretize this equation in the spatial direction using a centered finite-difference method with step δ_l such that $\frac{1}{\delta_l} = L \in \mathbb{N}$. If we write $\bar{x}(t,s)$ for the approximation of $x\left(t, \left(s - \frac{1}{2}\right)\delta_l\right)$ and approximate $\frac{\partial x}{\partial l}(t,l)$ to second order in δ_l by

$$\frac{x(t, l + \frac{\delta_l}{2}) - x(t, l - \frac{\delta_l}{2})}{\delta_l}$$

for all $l \in [0, L]$, we get the following semidiscretized system, [3], [5]:

$$\frac{d\bar{x}}{dt}(t,s) = \frac{\bar{x}(t,s+1) - 2\bar{x}(t,s) + \bar{x}(t,s-1)}{\delta_l^2} + d\left(t, \left(s - \frac{1}{2}\right)\delta_l\right) \text{ for all } s = 1, \ldots, L, \qquad t \geq 0, \tag{28a}$$

$$\bar{x}(0,s) = x^0\left(\left(s - \frac{1}{2}\right)\delta_l\right) \tag{28b}$$

with boundary conditions

$$\bar{x}(t,0) = -\bar{x}(t,1);\ \bar{x}(t,L) = -\bar{x}(t,L+1). \tag{29}$$

The latter approximate the original boundary conditions (27c) up to order δ_l^2, which is also the order of accuracy of (28a). This system can be written as a finite extent system in standard form (8) with

$$x = z = \bar{x},\ n_{\mathrm{T}} = 1,\ n = 1,$$

$$A_{\mathrm{TT}} = -\frac{2}{\delta_l^2},\ A_{\mathrm{TS}} = \frac{1}{\delta_l^2}\begin{pmatrix} 1 & 1 \end{pmatrix},\ B_{\mathrm{T}} = 1,$$

$$A_{\mathrm{ST}} = \begin{pmatrix} 1 \\ 1 \end{pmatrix},\ A_{\mathrm{SS}} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},\ B_{\mathrm{S}} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

$$C_{\mathrm{T}} = 1,\ C_{\mathrm{S}} = \begin{pmatrix} 0 & 0 \end{pmatrix},\ D = 0.$$

Then if we let $v^+(t,1) := \bar{x}(t,0)$ and $v^-(L) := \bar{x}(t,L+1)$, boundary conditions (29) can be rewritten as

$$v^+(t,1) = -w^-(t,1);\ v^-(t,L) = -w^+(t,L),$$

which corresponds to the boundary conditions matrix $M = -1$. It is then easy to see that the system is spatially reversible with $P = U = R = -I$.

6. Control synthesis for finite extent systems. We are now interested in solving the following $\mathcal{H}_\infty$ synthesis problem.

PROBLEM 1. *Given a finite extent system with boundary conditions (the plant), find another such system (the controller) such that the closed-loop is well-posed, stable, and contractive.*

It is also desirable that the algorithm for determining a satisfactory controller be computationally tractable, irrespective of the number of subsystems in the plant. Also note that we explicitly require the controller to have the same spatial structure as the plant, as shown in Figure 2. Hence we are aiming for a *distributed control strategy*, as opposed to a centralized strategy (in which all the subsystems of the plant are connected to the *same* controller) or decentralized strategy (in which subsystems of the controller are not interconnected with each other).

As is the case for analysis, Problem 1 has counterparts for the periodic and infinite systems corresponding to the plant. Their statements are obvious and they will be referred to as the periodic synthesis problem and infinite synthesis problem,

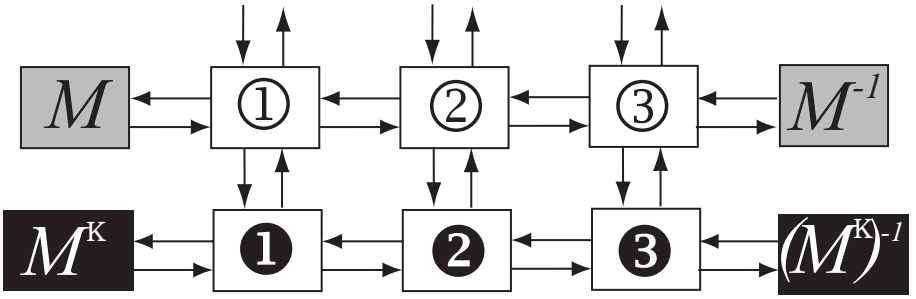


FIG. 2. *Interconnection of the finite extent plant and controller for $L = 3$. Note that feedback is distributed.*

respectively. Both the periodic synthesis problem and the infinite synthesis problem can be efficiently solved using the methods developed in [1] and [5]. The first involves solving a family of synthesis problems, parameterized by spatial frequency while, for the second, synthesis conditions take the form of a *single* linear matrix inequality (LMI). However, these are only sufficient conditions. In the recent past [18], [14], Problem 1 has been tackled in the following way:

1. Solve the periodic (respectively, infinite) synthesis problem for the periodic (respectively, infinite) system corresponding to the given, finite extent, plant. This results in a periodic (respectively, infinite) controller.
2. Supplement the periodic or infinite controller's realization with some "well-chosen" boundary conditions to obtain a finite extent one solving Problem 1.

This procedure implicitly assumes that the stability and performance of the finite extent closed-loop system can be derived from the properties of the spatially invariant one and that the influence of the controller's boundary conditions can be evaluated. As already noted in [18], the last point is delicate. Bluntly stated, it is not clear what "well-chosen" boundary conditions should be in the general case.

However, we have just established that such a link between finite extent and periodic systems exists in the case of spatial reversibility. In this framework, "well-chosen" also gains a clear meaning: given a realization of the closed-loop periodic system, a well-chosen boundary conditions matrix M^{K} for the controller should be such that the closed-loop system is spatially reversible with some boundary conditions matrix M^{C}. Indeed, if this is the case, Corollary 4.6 will guarantee stability and performance when the controller, which solves the periodic synthesis problem, is implemented on the finite plant, with boundary conditions M^{K}.

In the remaining sections, we develop tools to show that $M^{\mathrm{K}} = (M^*)^{-1}$ is a well-chosen boundary condition for the controller if the plant, with boundary conditions matrix M, is spatially reversible. More precisely, we show the following.

THEOREM 6.1. *Given a finite extent, spatially reversible plant, there exists a spatially reversible, finite extent controller, with $n_{\mathrm{T}}^{\mathrm{K}} = n_{\mathrm{T}}$, $n^{\mathrm{K}} = n$, and boundary conditions matrix $M^{\mathrm{K}} = (M^*)^{-1}$, that solves Problem* 1 *if the LMI conditions of* [5] (*equations* (34)) *are satisfied and if the plant's involutions R and U satisfy $R = \mathbf{diag}(R^u, R^d)$ and $U = \mathbf{diag}(U^y, U^z)$ with $R^d = \left(R^d\right)^* = \left(R^d\right)^{-1}$, $U^z = (U^z)^* = (U^z)^{-1}$.*

This means that $\mathcal{H}_\infty$ synthesis for a spatially reversible finite extent system (with boundary conditions) can be achieved by solving a convex problem, to determine

the controller's basic building block, and by "reading off" the controller's boundary conditions from the plant. The first step toward this synthesis result, which is also needed to make the statement of Problem 1 more rigorous, is the interconnection of reversible systems.

6.1. Interconnection of systems and spatial reversibility. The interconnection of two finite extent, periodic, or infinite systems is obtained by performing a linear fractional transformation of every pair of subsystems with the same index. This is depicted in Figure 2 in the case of finite extent systems.

More precisely, suppose we are given a plant with two sets of inputs (the exogenous disturbance $d \in \mathbb{R}^{m_d}$ and the control input $u \in \mathbb{R}^{m_u}$) and outputs (the performance output $z \in \mathbb{R}^{p_z}$ and the measured output $y \in \mathbb{R}^{p_y}$), as described by

$$\frac{d}{dt}x(t) = A_{\mathrm{TT}}x(t) + A_{\mathrm{TS}}v(t) + \begin{pmatrix} B_{\mathrm{T}}^u & B_{\mathrm{T}}^d \end{pmatrix}\begin{pmatrix} u(t) \\ d(t) \end{pmatrix},\ x(0) = x^0, \tag{30a}$$

$$(\Delta - A_{\mathrm{SS}})\,[v(t)] = A_{\mathrm{ST}}x(t) + \begin{pmatrix} B_{\mathrm{S}}^u & B_{\mathrm{S}}^d \end{pmatrix}\begin{pmatrix} u(t) \\ d(t) \end{pmatrix}, \tag{30b}$$

$$\begin{pmatrix} y(t) \\ z(t) \end{pmatrix} = \begin{pmatrix} C_{\mathrm{T}}^y \\ C_{\mathrm{T}}^z \end{pmatrix}x(t) + \begin{pmatrix} C_{\mathrm{S}}^y \\ C_{\mathrm{S}}^z \end{pmatrix}v(t) + \begin{pmatrix} D^{yu} & D^{yd} \\ D^{zu} & D^{zd} \end{pmatrix}\begin{pmatrix} u(t) \\ d(t) \end{pmatrix}. \tag{30c}$$

Then its interconnection with the controller given by

$$\frac{d}{dt}x^{\mathrm{K}}(t) = A_{\mathrm{TT}}^{\mathrm{K}}x^{\mathrm{K}}(t) + A_{\mathrm{TS}}^{\mathrm{K}}v^{\mathrm{K}}(t) + B_{\mathrm{T}}^{\mathrm{K}}y(t),\ x^{\mathrm{K}}(0) = (x^{\mathrm{K}})^0, \tag{31a}$$

$$(\Delta - A_{\mathrm{SS}})\,[v^{\mathrm{K}}(t)] = A_{\mathrm{ST}}^{\mathrm{K}}x^{\mathrm{K}}(t) + B_{\mathrm{S}}^{\mathrm{K}}y(t), \tag{31b}$$

$$u(t) = C_{\mathrm{T}}^{\mathrm{K}}x^{\mathrm{K}}(t) + C_{\mathrm{S}}^{\mathrm{K}}v^{\mathrm{K}}(t) + D^{\mathrm{K}}y(t) \ \text{ for all } t \geq 0 \tag{31c}$$

is the system obtained by eliminating u and y in (30) and (31). We emphasize that $[x^{\mathrm{K}}(t)](s) \in \mathbb{R}^{n_{\mathrm{T}}^{\mathrm{K}}}$ and $[(v^+)^K(t)](s),\ [(v^-)^K(t)](s) \in \mathbb{R}^{n^{\mathrm{K}}}$ for all $t \geq 0$, and s with $n_{\mathrm{T}}^{\mathrm{K}} \neq n_{\mathrm{T}}$ and $n^{\mathrm{K}} \neq n$ a priori. The dimensions of the matrices defining operator Δ in (31b) are thus chosen accordingly.

The corresponding closed-loop system equations can be put in standard form (8) using the permutation matrix

$$\Pi = \begin{pmatrix} I_n & 0 & 0 & 0 \\ 0 & 0 & I_{n^{\mathrm{K}}} & 0 \\ 0 & I_n & 0 & 0 \\ 0 & 0 & 0 & I_{n^{\mathrm{K}}} \end{pmatrix}$$

in order to group the spatial variables properly. This means that the closed-loop system is also a finite extent, periodic, or infinite system, depending on the case. A realization of the corresponding basic building block, when $D^{yu} = 0$, is given below. Note that this is not a restrictive assumption since one can always use loop-shifting

if this situation is not at hand (see [5] and references therein):

$$A^{\mathrm{C}}_{\mathrm{TT}} = \begin{pmatrix} A_{\mathrm{TT}} + B^u_{\mathrm{T}} D^{\mathrm{K}} C^y_{\mathrm{T}} & B^u_{\mathrm{T}} C^{\mathrm{K}}_{\mathrm{T}} \\ B^{\mathrm{K}}_{\mathrm{T}} C^y_{\mathrm{T}} & A^{\mathrm{K}}_{\mathrm{TT}} \end{pmatrix}, \quad A^{\mathrm{C}}_{\mathrm{TS}} = \begin{pmatrix} A_{\mathrm{TS}} + B^u_{\mathrm{T}} D^{\mathrm{K}} C^y_{\mathrm{S}} & B^u_{\mathrm{T}} C^{\mathrm{K}}_{\mathrm{S}} \\ B^{\mathrm{K}}_{\mathrm{T}} C^y_{\mathrm{S}} & A^{\mathrm{K}}_{\mathrm{TS}} \end{pmatrix} \Pi,$$

$$A^{\mathrm{C}}_{\mathrm{ST}} = \Pi \begin{pmatrix} A_{\mathrm{ST}} + B^u_{\mathrm{S}} D^{\mathrm{K}} C^y_{\mathrm{T}} & B^u_{\mathrm{S}} C^{\mathrm{K}}_{\mathrm{T}} \\ B^{\mathrm{K}}_{\mathrm{S}} C^y_{\mathrm{T}} & A^{\mathrm{K}}_{\mathrm{ST}} \end{pmatrix}, \quad A^{\mathrm{C}}_{\mathrm{SS}} = \Pi \begin{pmatrix} A_{\mathrm{SS}} + B^u_{\mathrm{S}} D^{\mathrm{K}} C^y_{\mathrm{S}} & B^u_{\mathrm{S}} C^{\mathrm{K}}_{\mathrm{S}} \\ B^{\mathrm{K}}_{\mathrm{S}} C^y_{\mathrm{S}} & A^{\mathrm{K}}_{\mathrm{SS}} \end{pmatrix} \Pi,$$

$$B^{\mathrm{C}}_{\mathrm{T}} = \begin{pmatrix} B^d_{\mathrm{T}} + B^u_{\mathrm{T}} D^{\mathrm{K}} D^{yd} \\ B^{\mathrm{K}}_{\mathrm{T}} D^{yd} \end{pmatrix}, \qquad B^{\mathrm{C}}_{\mathrm{S}} = \Pi \begin{pmatrix} B^d_{\mathrm{S}} + B^u_{\mathrm{S}} D^{\mathrm{K}} D^{yd} \\ B^{\mathrm{K}}_{\mathrm{S}} D^{yd} \end{pmatrix},$$

$$C^{\mathrm{C}}_{\mathrm{T}} = \begin{pmatrix} C^z_{\mathrm{T}} + D^{zu} D^{\mathrm{K}} C^y_{\mathrm{T}} & D^{zu} C^{\mathrm{K}}_{\mathrm{T}} \end{pmatrix}, \quad C^{\mathrm{C}}_{\mathrm{S}} = \begin{pmatrix} C^z_{\mathrm{S}} + D^{zu} D^{\mathrm{K}} C^y_{\mathrm{S}} & D^{zu} C^{\mathrm{K}}_{\mathrm{S}} \end{pmatrix} \Pi,$$

$$D^{\mathrm{C}} = D^{zd} + D^{zu} D^{\mathrm{K}} D^{yd}. \tag{32}$$

Using (32), it is easy to show the following.

PROPOSITION 6.2. *Let the plant, with boundary conditions matrix M, and the controller, with boundary conditions matrix M^{K}, be spatially reversible. Assume further that the involutions R and U for the plant and R^{K} and U^{K} for the controller satisfy $R^{\mathrm{K}} = U^y$ and $U^{\mathrm{K}} = R^u$, where U and R are partitioned conformably to the inputs and outputs as $R = \mathbf{diag}(R^u, R^d)$ and $U = \mathbf{diag}(U^y, U^z)$. Then their interconnection, which has $M^{\mathrm{C}} = \mathbf{diag}(M, M^{\mathrm{K}})$ as boundary conditions matrix, is spatially reversible with $R^{\mathrm{C}} = R^d$, $P^{\mathrm{C}} = \mathbf{diag}(P, P^{\mathrm{K}})$, and $U^{\mathrm{C}} = U^z$.*

Proposition 6.2, combined with Corollary 4.6, already gives a way to solve Problem 1: if the periodic synthesis problem can be (tractably) solved *by any means*, and if the resulting controller can be shown to be M^{K}-reversible for some matrix M^{K}, with $R^{\mathrm{K}} = U$ and $U^{\mathrm{K}} = R$, then this is the boundary conditions matrix that should be used for the finite extent controller.

6.2. Reversible infinite controllers for reversible infinite plants. In this section we give the second element needed to establish Theorem 6.1, namely, the following.

PROPOSITION 6.3. *Consider an M-reversible infinite plant. Assume the involutions R and U for the plant are of the type indicated in Theorem* 6.1. *Then it is always possible to solve the infinite synthesis problem with an $(M^*)^{-1}$-*reversible *controller, provided the LMI condition of* [5] (*equation* (34)) *is satisfied. Moreover, $R^{\mathrm{K}} = U^y$, $U^{\mathrm{K}} = R^u$, and $P^{\mathrm{K}} = P^*$ for this controller.*

It should be noted that this reversible controller is *not* necessarily the solution that one would obtain by directly solving (34)–(37) with a numerical solver such as those included in the LMI toolbox for MATLAB. However, if one has a solution, one can construct such an $(M^*)^{-1}$-reversible controller by following the steps of the proof.

Before proving Proposition 6.3, we should clarify why this implies Theorem 6.1. First, thanks to a theorem of [1] stating that the input/output gain of well-posed, stable systems over a group can be determined by a frequency-grid search, one can prove that contractiveness of the infinite system implies contractiveness of the corresponding periodic system, using arguments very similar to those of Proposition 3.3. As a result, the periodic system corresponding to the $(M^*)^{-1}$-reversible controller of Proposition 6.3 solves the periodic synthesis problem. The periodic closed-loop system is thus well-posed, stable, and contractive. It is also M^{C}-reversible with $M^{\mathrm{C}} = \mathbf{diag}(M, (M^*)^{-1})$ by virtue of Proposition 6.2. One can then apply the method of images to show that the corresponding finite extent closed-loop system is also well-posed and stable. Finally, since R^d and U^z are assumed to be unitary, Corollary 4.6 yields contractiveness of the finite extent closed-loop.

It might seem artificial to introduce the infinite system in order to solve Problem 1, while the method of images refers only to the periodic system. The main practical reason for using a finite controller corresponding to a reversible solution of the *infinite* synthesis problem is the following. Imagine the number L of subsystems in the finite extent system is changed to $L' \neq L$. Then the size of the corresponding periodic system also changes (from $2L$ to $2L'$) and so does the corresponding group $\mathbb{U}$, which now becomes $\mathbb{U}'$, the group of $(2L')$th root of unity. A solution of the periodic synthesis problem for $2L$ subsystems does not necessarily solve the same problem for $2L'$ subsystems since well-posedness, stability, and performance of the closed-loop all depend on the group $\mathbb{U}$, and $\mathbb{U}'$ may or may not be a subgroup of $\mathbb{U}$. Hence, if one uses a finite extent controller corresponding to a reversible solution of the periodic synthesis problem, one has to redo a synthesis if the number of subsystems in the finite extent plant changes. This is not desirable since this number is in fact irrelevant for a spatially reversible plant (only the boundary conditions matter). A reversible solution of the infinite synthesis problem, on the other hand, solves it irrespective of L.

Proof. The proof is by construction. For the reader's convenience and because they are used extensively in this proof, we first recall the notation and main results of [5].

Given a well-posed infinite plant with basic building block (1), let

$$H = \begin{pmatrix} I_n & 0 \\ 0 & -I_n \end{pmatrix}$$

and define the bilinear algebraic transformed system by

$$\overline{A_{\mathrm{SS}}} := H(A_{\mathrm{SS}} - I)(A_{\mathrm{SS}} + I)^{-1}, \tag{33a}$$

$$\begin{pmatrix} \overline{A_{\mathrm{ST}}} & \overline{B_{\mathrm{S}}} \end{pmatrix} := \sqrt{2}H(A_{\mathrm{SS}} + I)^{-1}\begin{pmatrix} A_{\mathrm{ST}} & B_{\mathrm{S}} \end{pmatrix}, \tag{33b}$$

$$\begin{pmatrix} \overline{A_{\mathrm{TS}}} \\ \overline{C_{\mathrm{S}}} \end{pmatrix} := \sqrt{2}\begin{pmatrix} A_{\mathrm{TS}} \\ C_{\mathrm{S}} \end{pmatrix}(A_{\mathrm{SS}} + I)^{-1}, \tag{33c}$$

$$\begin{pmatrix} \overline{A_{\mathrm{TT}}} & \overline{B_{\mathrm{T}}} \\ \overline{C_{\mathrm{T}}} & \overline{D} \end{pmatrix} := \begin{pmatrix} A_{\mathrm{TT}} & B_{\mathrm{T}} \\ C_{\mathrm{T}} & D \end{pmatrix} - \begin{pmatrix} A_{\mathrm{TS}} \\ C_{\mathrm{S}} \end{pmatrix}(A_{\mathrm{SS}} + I)^{-1}\begin{pmatrix} A_{\mathrm{ST}} & B_{\mathrm{S}} \end{pmatrix}, \tag{33d}$$

$$\overline{A}^{\mathrm{G}} = \begin{pmatrix} \overline{A_{\mathrm{TT}}} & \overline{A_{\mathrm{TS}}} \\ \overline{A_{\mathrm{ST}}} & \overline{A_{\mathrm{SS}}} \end{pmatrix},\ \overline{B}^{\mathrm{G}} = \begin{pmatrix} \overline{B_{\mathrm{T}}} \\ \overline{B_{\mathrm{S}}} \end{pmatrix},\ \overline{C}^{\mathrm{G}} = \begin{pmatrix} \overline{C_{\mathrm{T}}} & \overline{C_{\mathrm{S}}} \end{pmatrix},\ \overline{D}^{\mathrm{G}} = \overline{D}.$$

We also define several sets of scaling matrices:

$$\begin{aligned}
\mathcal{X}^{\mathrm{G}} &= \{X^{\mathrm{G}} = \mathbf{diag}(X_{\mathrm{T}}^{\mathrm{G}}, X_{\mathrm{S}}^{\mathrm{G}}),\ X_{\mathrm{T}}^{\mathrm{G}} \in \mathbb{R}^{n_{\mathrm{T}} \times n_{\mathrm{T}}},\ X_{\mathrm{T}}^{\mathrm{G}} > 0,\ X_{\mathrm{S}}^{\mathrm{G}} \in \mathbb{R}^{n_{\mathrm{S}} \times n_{\mathrm{S}}},\ X_{\mathrm{S}}^{\mathrm{G}} \text{ is symmetric}\}, \\
\mathcal{X}^{\mathrm{K}} &= \{X^{\mathrm{K}} = \mathbf{diag}(X_{\mathrm{T}}^{\mathrm{K}}, X_{\mathrm{S}}^{\mathrm{K}}),\ X_{\mathrm{T}}^{\mathrm{K}} \in \mathbb{R}^{n_{\mathrm{T}}^{\mathrm{K}} \times n_{\mathrm{T}}^{\mathrm{K}}},\ X_{\mathrm{T}}^{\mathrm{K}} > 0,\ X_{\mathrm{S}}^{\mathrm{K}} \in \mathbb{R}^{n_{\mathrm{S}}^{\mathrm{K}} \times n_{\mathrm{S}}^{\mathrm{K}}},\ X_{\mathrm{S}}^{\mathrm{K}} \text{ is symmetric}\}, \\
\mathcal{X}^{\mathrm{GK}} &= \{X = \mathbf{diag}(X_{\mathrm{T}}^{\mathrm{GK}}, X_{\mathrm{S}}^{\mathrm{GK}}),\ X_{\mathrm{T}}^{\mathrm{GK}} \in \mathbb{R}^{n_{\mathrm{T}} \times n_{\mathrm{T}}^{\mathrm{K}}},\ X_{\mathrm{S}}^{\mathrm{GK}} \in \mathbb{R}^{n_{\mathrm{S}} \times n_{\mathrm{S}}^{\mathrm{K}}}\}.
\end{aligned}$$

Then, the main synthesis result of [5] is the following.

THEOREM 6.4. *Let the columns of $\mathcal{N}_{\mathrm{Y}}$ span the null space of $((\overline{B^u}^{\mathrm{G}})^* \quad (\overline{D^{zu}}_g)^*)$ and let those of $\mathcal{N}_{\mathrm{X}}$ span the null space of $((\overline{C^y}^{\mathrm{G}})^* \quad (\overline{D^{yd}}^{\mathrm{G}})^*)$, respectively. Then there exists an infinite controller such that the closed-loop is well-posed, stable, and*

contractive if there exist X^{G} *and* Y^{G} *in* $\mathcal{X}^{\mathrm{G}}$ *satisfying the following LMI:*

$$\text{(34a)}\quad \begin{pmatrix} \mathcal{N}_{\mathrm{Y}} & 0 \\ 0 & I \end{pmatrix}^{*} \left(\begin{array}{cc} \begin{pmatrix} \overline{A}^{\mathrm{G}} Y^{\mathrm{G}} + Y^{\mathrm{G}} (\overline{A}^{\mathrm{G}})^{*} & Y^{\mathrm{G}} (\overline{C^{z}}^{\mathrm{G}})^{*} \\ \overline{C^{z}}^{\mathrm{G}} Y^{\mathrm{G}} & -I_{p_z} \end{pmatrix} & \begin{pmatrix} \overline{B^{d}}^{\mathrm{G}} \\ \overline{D^{zd}}^{\mathrm{G}} \end{pmatrix} \\ \begin{pmatrix} (\overline{B^{d}}^{\mathrm{G}})^{*} & (\overline{D^{zd}}^{\mathrm{G}})^{*} \end{pmatrix} & -I_{m_d} \end{array} \right) \begin{pmatrix} \mathcal{N}_{\mathrm{Y}} & 0 \\ 0 & I \end{pmatrix} < 0,$$

$$\text{(34b)}\quad \begin{pmatrix} \mathcal{N}_{\mathrm{X}} & 0 \\ 0 & I \end{pmatrix}^{*} \left(\begin{array}{cc} \begin{pmatrix} (\overline{A}^{\mathrm{G}})^{*} X^{\mathrm{G}} + X^{\mathrm{G}} \overline{A}^{\mathrm{G}} & X^{\mathrm{G}} \overline{B^{d}}^{\mathrm{G}} \\ (\overline{B^{d}}^{\mathrm{G}})^{*} X^{\mathrm{G}} & -I_{m_d} \end{pmatrix} & \begin{pmatrix} (\overline{C^{z}}^{\mathrm{G}})^{*} \\ (\overline{D^{zd}}^{\mathrm{G}})^{*} \end{pmatrix} \\ \begin{pmatrix} \overline{C^{z}}^{\mathrm{G}} & \overline{D^{zd}}^{\mathrm{G}} \end{pmatrix} & -I_{p_z} \end{array} \right) \begin{pmatrix} \mathcal{N}_{\mathrm{X}} & 0 \\ 0 & I \end{pmatrix} < 0,$$

$$\text{(34c)}\quad \begin{pmatrix} X_{\mathrm{T}}^{\mathrm{G}} & I \\ I & Y_{\mathrm{T}}^{\mathrm{G}} \end{pmatrix} \geq 0.$$

Because (34c) *is satisfied, there exist* X^{K} *and* Y^{K} *in* $\mathcal{X}^{\mathrm{K}}$ *and* X^{GK} *and* Y^{GK} *in* $\mathcal{X}^{\mathrm{GK}}$ *such that*

$$\text{(35)}\quad \begin{pmatrix} X^{\mathrm{G}} & X^{\mathrm{GK}} \\ (X^{\mathrm{GK}})^{*} & X^{\mathrm{K}} \end{pmatrix} = \begin{pmatrix} Y^{\mathrm{G}} & Y^{\mathrm{GK}} \\ (Y^{\mathrm{GK}})^{*} & Y^{\mathrm{K}} \end{pmatrix}^{-1}$$

and $n_{\mathrm{T}}^{\mathrm{K}} = n_{\mathrm{T}}$, $n_{\mathrm{S}}^{\mathrm{K}} = n_{\mathrm{S}}$.

Then defining

$$\text{(36)}\quad \overline{X} = \begin{pmatrix} X^{\mathrm{G}} & X^{\mathrm{GK}} \\ (X^{\mathrm{GK}})^{*} & X^{\mathrm{K}} \end{pmatrix},$$

we can construct a controller that solves the infinite synthesis problem in two steps as follows:

1. *Solve the LMI*

$$\text{(37)}\quad \begin{pmatrix} (\overline{A}^{\mathrm{C}})^{*} \overline{X} + \overline{X A}^{\mathrm{C}} & \overline{X B}^{\mathrm{C}} & (\overline{C}^{\mathrm{C}})^{*} \\ (\overline{B}^{\mathrm{C}})^{*} \overline{X} & -I & (\overline{D}^{\mathrm{C}})^{*} \\ \overline{C}^{\mathrm{C}} & \overline{D}^{\mathrm{C}} & -I \end{pmatrix} < 0,$$

which is affine in the unknown

$$\Theta := \begin{pmatrix} \overline{A}^{\mathrm{K}} & \overline{B}^{\mathrm{K}} \\ \overline{C}^{\mathrm{K}} & \overline{D}^{\mathrm{K}} \end{pmatrix},$$

where

$$\begin{pmatrix} \overline{A}^{\mathrm{C}} & \overline{B}^{\mathrm{C}} \\ \overline{C}^{\mathrm{C}} & \overline{D}^{\mathrm{C}} \end{pmatrix} = \begin{pmatrix} \overline{A}^{\mathrm{G}} & 0 & \overline{B^{d}}^{\mathrm{G}} \\ 0 & 0 & 0 \\ \overline{C^{z}}^{\mathrm{G}} & 0 & \overline{D^{zd}}^{\mathrm{G}} \end{pmatrix} + \begin{pmatrix} 0 & \overline{B^{u}}^{\mathrm{G}} \\ I & 0 \\ 0 & \overline{D^{zu}}^{\mathrm{G}} \end{pmatrix} \Theta \begin{pmatrix} 0 & I & 0 \\ \overline{C^{y}}^{\mathrm{G}} & 0 & \overline{D^{yd}}^{\mathrm{G}} \end{pmatrix}.$$

2. *Once* Θ *is known, make a change of coordinates that puts* $\overline{A_{\mathrm{SS}}}^{\mathrm{K}}$ *into the form*

$$\overline{A_{\mathrm{SS}}}^{\mathrm{K}} = \begin{pmatrix} A^{+} & 0 \\ 0 & A^{-} \end{pmatrix},$$

where both A^- and $-(A^+)$ are Hurwitz. This can always be achieved if $\overline{A_{\rm SS}}^{\rm K}$ has no eigenvalue on the imaginary axis. If this situation is not at hand, one can perturb $\overline{A_{\rm SS}}^{\rm K}$ so that it holds and LMI (34) will still be satisfied. Then let

$$H^{\rm K} = \begin{pmatrix} I_{n^+} & 0 \\ 0 & -I_{n^-} \end{pmatrix},$$

where $n^{\pm} = \dim(A^{\pm})$, and invert the bilinear algebraic transformation by

$$A^{\rm K}_{\rm SS} := (H^{\rm K} - \overline{A_{\rm SS}}^{\rm K})^{-1}(H^{\rm K} + \overline{A_{\rm SS}}^{\rm K}), \tag{38a}$$

$$(A^{\rm K}_{\rm ST} \quad B^{\rm K}_{\rm S}) := \sqrt{2}(H^{\rm K} - \overline{A_{\rm SS}}^{\rm K})^{-1}(\overline{A_{\rm ST}}^{\rm K} \quad \overline{B_{\rm S}}^{\rm K}), \tag{38b}$$

$$\begin{pmatrix} A^{\rm K}_{\rm TS} \\ C^{\rm K}_{\rm S} \end{pmatrix} := \sqrt{2} \begin{pmatrix} \overline{A_{\rm TS}}^{\rm K} \\ \overline{C_{\rm S}}^{\rm K} \end{pmatrix} (H^{\rm K} - \overline{A_{\rm SS}}^{\rm K})^{-1} H^{\rm K}, \tag{38c}$$

(38d)
$$\begin{pmatrix} A^{\rm K}_{\rm TT} & B^{\rm K}_{\rm T} \\ C^{\rm K}_{\rm T} & D^{\rm K} \end{pmatrix} := \begin{pmatrix} \overline{A_{\rm TT}}^{\rm K} & \overline{B_{\rm T}}^{\rm K} \\ \overline{C_{\rm T}}^{\rm K} & \overline{D}^{\rm K} \end{pmatrix} + \begin{pmatrix} \overline{A_{\rm TS}}^{\rm K} \\ \overline{C_{\rm S}}^{\rm K} \end{pmatrix} (H^{\rm K} - \overline{A_{\rm SS}}^{\rm K})^{-1} (\overline{A_{\rm ST}}^{\rm K} \quad \overline{B_{\rm S}}^{\rm K}),$$

to find the building block of an infinite controller solving the infinite synthesis problem.

Now assume that the plant at hand is M-reversible for some boundary conditions matrix M with $R^d = \left(R^d\right)^* = \left(R^d\right)^{-1}$ and $U^z = (U^z)^* = (U^z)^{-1}$. Because of spatial reversibility and since $QH = -HQ$, it is easy to see that

$$V\overline{A}^{\rm G}W = \overline{A}^{\rm G},\ \overline{C}^{\rm G}W = U\overline{C}^{\rm G},\ V\overline{B}^{\rm G} = \overline{B}^{\rm G}R,\ U\overline{D}^{\rm G} = \overline{D}^{\rm G}R, \tag{39}$$

where $V := \left(\begin{smallmatrix} P & 0 \\ 0 & -Q \end{smallmatrix}\right)$ and $W := \left(\begin{smallmatrix} P & 0 \\ 0 & Q \end{smallmatrix}\right)$.

Let $X^{\rm G}$, $Y^{\rm G}$ solve (34a)–(34b) for $\mathcal{N}_{\rm X}$ and $\mathcal{N}_{\rm Y}$. Then, pre- and postmultiplying (34a) by

$$\begin{pmatrix} I & 0 \\ 0 & \left(R^d\right)^* \end{pmatrix} \text{ and } \begin{pmatrix} I & 0 \\ 0 & R^d \end{pmatrix},$$

(34b) by

$$\begin{pmatrix} I & 0 \\ 0 & U^z \end{pmatrix} \text{ and } \begin{pmatrix} I & 0 \\ 0 & (U^z)^* \end{pmatrix},$$

and using (39), we get that $V^*X^{\rm G}W$, $WY^{\rm G}V^* \in \mathcal{X}^{\rm G}$ also satisfy (34a)–(34b) but with $\tilde{\mathcal{N}}_{\rm X} = \left(\begin{smallmatrix} W & 0 \\ 0 & R^d \end{smallmatrix}\right)\mathcal{N}_{\rm X}$ and $\tilde{\mathcal{N}}_{\rm Y} = \left(\begin{smallmatrix} V^* & 0 \\ 0 & (U^z)^* \end{smallmatrix}\right)\mathcal{N}_{\rm Y}$, the columns of which also satisfy the assumptions of Theorem 6.4.

Now an important point is that $X^{\rm G}$ and $Y^{\rm G}$ *also* satisfy (34a)–(34b) for $\tilde{\mathcal{N}}_{\rm X}$ and $\tilde{\mathcal{N}}_{\rm Y}$. In fact, the matrices do not matter as long as their columns span the appropriate null-spaces. (The reason why it is so can be easily understood if one follows the usual procedure for formulating $\mathcal{H}_\infty$ synthesis as a convex problem, as presented, e.g., in Chapter 7 of [7]. See Lemma 7.2 in particular.)

Hence averaging the two sets of LMIs, we see that

$$\tilde{X}^{\rm G} = \frac{1}{2}(X^{\rm G} + V^*X^{\rm G}W) \text{ and } \tilde{Y}^{\rm G} = \frac{1}{2}(Y^{\rm G} + WY^{\rm G}V^*)$$

solve (34) for $\tilde{\mathcal{N}}_{\mathrm{X}}$ and $\tilde{\mathcal{N}}_{\mathrm{Y}}$. Note that

$$Q^*\tilde{X}^{\mathrm{G}}Q = -\tilde{X}^{\mathrm{G}} \text{ and } Q\tilde{Y}^{\mathrm{G}}Q^* = -\tilde{Y}^{\mathrm{G}}.$$

In order to solve (35), one can choose a *full-rank* controller by picking

$$X_{\mathrm{S}}^{\mathrm{GK}} = \left(I - \tilde{X}_{\mathrm{S}}^{\ \mathrm{G}}\tilde{Y}_{\mathrm{S}}^{\ \mathrm{G}}\right),\ X_{\mathrm{S}}^{\mathrm{K}} = -(X_{\mathrm{S}}^{\mathrm{GK}})^*\tilde{Y}_{\mathrm{S}}^{\ \mathrm{G}},\ Y_{\mathrm{S}}^{\mathrm{GK}} = I.$$

Then easy algebra yields

$$Q^*X_{\mathrm{S}}^{\mathrm{GK}}Q^* = X_{\mathrm{S}}^{\mathrm{GK}},\ QX_{\mathrm{S}}^{\mathrm{K}}Q^* = -X_{\mathrm{S}}^{\mathrm{K}},$$

and, in turn, that $\overline{X}$ as per (36) satisfies

$$\begin{pmatrix} W^* & 0 \\ 0 & V \end{pmatrix}\overline{X}\begin{pmatrix} V & 0 \\ 0 & W^* \end{pmatrix} = \overline{X} = \overline{X}^* = \begin{pmatrix} V^* & 0 \\ 0 & W \end{pmatrix}\overline{X}\begin{pmatrix} W & 0 \\ 0 & V^* \end{pmatrix}.$$

Using this scaling and pre- and postmultiplying (37) by

$$\begin{pmatrix} W^* & 0 & 0 & 0 \\ 0 & V & 0 & 0 \\ 0 & 0 & \left(R^d\right)^* & 0 \\ 0 & 0 & 0 & U^z \end{pmatrix} \text{ and } \begin{pmatrix} W & 0 & 0 & 0 \\ 0 & V^* & 0 & 0 \\ 0 & 0 & R^d & 0 \\ 0 & 0 & 0 & (U^z)^* \end{pmatrix},$$

we see that if Θ is a solution, so is $\hat{\Theta} := \left(\begin{smallmatrix} W^* & 0 \\ 0 & R^u \end{smallmatrix}\right)\Theta\left(\begin{smallmatrix} V^* & 0 \\ 0 & U^y \end{smallmatrix}\right)$ because of (39). Hence, $\tilde{\Theta} := \frac{1}{2}(\Theta + \hat{\Theta})$ also satisfies LMI (37). Note that $\left(\begin{smallmatrix} W^* & 0 \\ 0 & R^u \end{smallmatrix}\right)\tilde{\Theta}\left(\begin{smallmatrix} V^* & 0 \\ 0 & U^y \end{smallmatrix}\right) = \tilde{\Theta}$, which means that the corresponding controller is such that

$$\begin{pmatrix} P^*\overline{A_{\mathrm{TS}}}^{\mathrm{K}} \\ R^u\overline{C_{\mathrm{S}}}^{\mathrm{K}} \end{pmatrix}Q^* = -\begin{pmatrix} \overline{A_{\mathrm{TS}}}^{\mathrm{K}} \\ \overline{C_{\mathrm{S}}}^{\mathrm{K}} \end{pmatrix}, \tag{40a}$$

$$Q^*\begin{pmatrix} \overline{A_{\mathrm{ST}}}^{\mathrm{K}}P^* & \overline{B_{\mathrm{S}}}^{\mathrm{K}}U^y \end{pmatrix} = \begin{pmatrix} \overline{A_{\mathrm{ST}}}^{\mathrm{K}} & \overline{B_{\mathrm{S}}}^{\mathrm{K}} \end{pmatrix},$$

$$Q^*\overline{A_{\mathrm{SS}}}^{\mathrm{K}}Q^* = -\overline{A_{\mathrm{SS}}}^{\mathrm{K}},$$

$$P^*\overline{A_{\mathrm{TT}}}^{\mathrm{K}}P^* = \overline{A_{\mathrm{TT}}}^{\mathrm{K}},\ P^*\overline{B_{\mathrm{T}}}^{\mathrm{K}} = \overline{B_{\mathrm{T}}}^{\mathrm{K}}U^y, \tag{40b}$$

$$\overline{C_{\mathrm{T}}}^{\mathrm{K}}P^* = R^u\overline{C_{\mathrm{T}}}^{\mathrm{K}},\ R^u\overline{D}^{\mathrm{K}}U^y = \overline{D}^{\mathrm{K}}.$$

This last relation implies that the spectrum of $\overline{A_{\mathrm{SS}}}^{\mathrm{K}}$ is symmetric with respect to the origin. Thus, if it does not contain any point on the imaginary axis, we will have $n^+ = \dim(A^+) = \dim(A^-) = n^-$. Also, because we picked a full-rank controller, we have $n_{\mathrm{S}}^{\mathrm{K}} = n_{\mathrm{S}}$ and thus $n^{\pm} = n$ and $H^{\mathrm{K}} = H$.

Now, plugging (40) into (38), we get

$$\begin{pmatrix} P^*A_{\mathrm{TS}}^{\mathrm{K}} \\ R^uC_{\mathrm{S}}^{\mathrm{K}} \end{pmatrix}Q^* = -\begin{pmatrix} A_{\mathrm{TS}}^{\mathrm{K}} \\ C_{\mathrm{S}}^{\mathrm{K}} \end{pmatrix},$$

$$Q^*\begin{pmatrix} A_{\mathrm{ST}}^{\mathrm{K}}P^* & B_{\mathrm{S}}^{\mathrm{K}}U^y \end{pmatrix} = -\begin{pmatrix} A_{\mathrm{ST}}^{\mathrm{K}} & B_{\mathrm{S}}^{\mathrm{K}} \end{pmatrix},$$

$$Q^*A_{\mathrm{SS}}^{\mathrm{K}}Q^* = A_{\mathrm{SS}}^{\mathrm{K}}, \tag{41a}$$

$$P^*A_{\mathrm{TT}}^{\mathrm{K}}P^* = A_{\mathrm{TT}}^{\mathrm{K}},\ P^*B_{\mathrm{T}}^{\mathrm{K}} = B_{\mathrm{T}}^{\mathrm{K}}U^y,$$

$$C_{\mathrm{T}}^{\mathrm{K}}P^* = C_{\mathrm{T}}^{\mathrm{K}},\ R^uD^{\mathrm{K}}U^y = D^{\mathrm{K}}. \tag{41b}$$

Finally, perform a state transformation on v^{K},

$$v^{\mathrm{K}} \to Hv^{\mathrm{K}},$$

to yield an $(M^*)^{-1}$-reversible controller with $R^{\mathrm{K}} = U^y$, $U^{\mathrm{K}} = R^u$, and $P^{\mathrm{K}} = P^*$ ∎

7. Some generalizations. In this section, we explain how some assumptions can be relaxed and our results extended to more general, spatially multidimensional, reversible systems.

7.1. The case where R^d and U^z are not unitary. Although Theorem 6.1 treats the case where both R^d and U^z are unitary, it is possible to handle cases where $\left(R^d\right)^* \neq (R^d)^{-1}$ or $(U^z)^* \neq (U^z)^{-1}$ as well. There are two cases as follows:

- $\overline{\sigma}(R^d) \leq \underline{\sigma}(U^z)$:
  If one replaces the $-I_{m_d}$ and $-I_{p_z}$ blocks in (34a)–(34b) by $-\left(R^d\right)^* R^d$ and $-U^z \left(U^z\right)^*$, respectively, i.e., if one starts with an infinite controller such that

$$\|T_{dz}\|_{\mathcal{L}_2(\mathbb{Z})} < \frac{\overline{\sigma}(R^d)}{\underline{\sigma}(U^z)},$$

  then following all the steps of the proof will yield an $(M^*)^{-1}$-*reversible* infinite controller such that the closed-loop system is well-posed, stable, and contractive. In turn, the corresponding periodic closed-loop system will also be contractive and, since

$$\sqrt{\frac{1+\underline{\sigma}(U^z)^2}{1+\overline{\sigma}(R^d)^2}} \geq 1,$$

  Corollary 4.6 implies that the finite extent closed-loop system is also contractive.
- $\overline{\sigma}(R^d) > \underline{\sigma}(U^z)$:
  In this case, it is possible to construct an $(M^*)^{-1}$-*reversible* infinite controller that guarantees contractiveness of the closed-loop system if the LMIs (34a)–(34b) have a solution. The corresponding periodic closed-loop will also be contractive but the finite extent one need not be. We have only the upper-bound

$$\|T_{dz}\|_{\mathcal{L}(\{1,\dots,L\})} < \sqrt{\frac{1+\underline{\sigma}(U^z)^2}{1+\overline{\sigma}(R^d)^2}}.$$

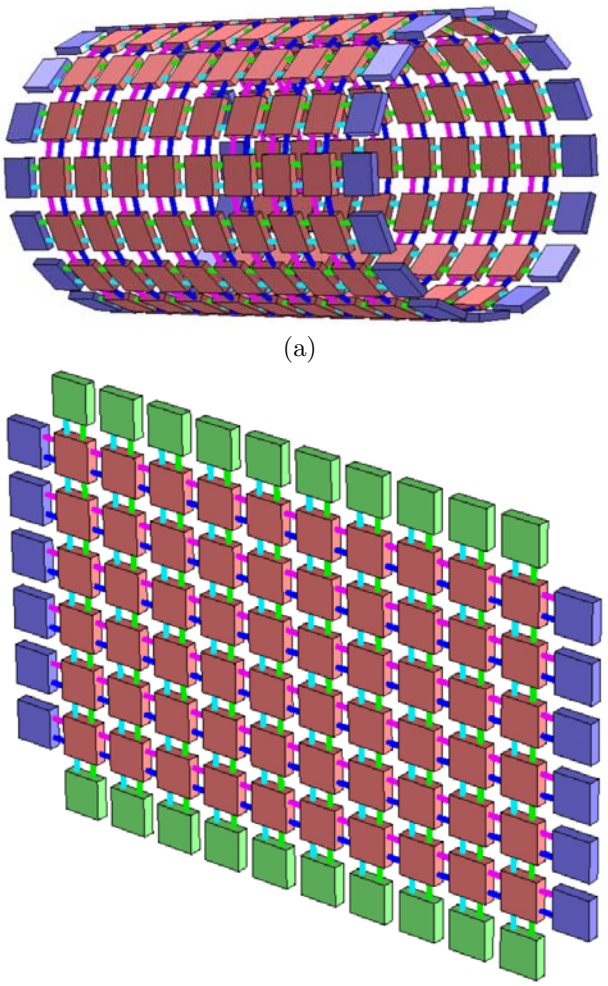
7.2. Multiple spatial dimensions. It is straightforward to extend our present results to cases where the subsystems are distributed on a multidimensional grid instead of a line. The basic building block then has two interconnection inputs (v_i^+, $v_i^- \in \mathbb{R}^{n_i}$) and outputs ($w_i^+$, $w_i^- \in \mathbb{R}^{n_i}$) per spatial dimension. The index s used to describe the finite extent interconnection now belongs to a cartesian product set of the form

$$\mathbb{M} = \mathbb{Z}_{l_1} \times \cdots \times \mathbb{Z}_{l_k} \times \{1,\dots,L_1\} \times \cdots \times \{1,\dots,L_d\}$$

where $k \geq 0,\ d > 0$ and there is a boundary conditions matrix $M_i,\ 1 \leq i \leq d$ associated with every nonperiodic spatial dimension. Likewise, the corresponding periodic and infinite systems are indexed over the set

$$\mathbb{M} = \mathbb{Z}_{l_1} \times \cdots \times \mathbb{Z}_{l_k} \times \mathbb{Z}_{2L_1} \times \cdots \times \mathbb{Z}_{2L_d} \text{ and } \mathbb{M} = \mathbb{Z}_{l_1} \times \cdots \times \mathbb{Z}_{l_k} \times \mathbb{Z}^d,$$

respectively. Examples of such spatially multidimensional finite extent systems are given in Figure 3 for $k = 1,\ d = 1$ and $k = 0,\ d = 2$.



(a)

(b)

FIG. 3. *Examples of spatially multidimensional interconnections. The inputs and outputs have been omitted for clarity.* (a) *All boundary conditions matrices are equal to* M_1. (b) *All boundary conditions matrices for lines (respectively, columns) are equal to* M_1 *(respectively,* M_2*).*

These spatially multidimensional systems can be represented by (9) if we let

$$v(t) := \left(v_1^+(t), v_1^-(t), \ldots, v_{k+d}^+(t), v_{k+d}^-(t)\right) \in \ell_2^{n_\mathrm{s}}(\mathbb{M})$$

for a suitable n_s and replace Δ_{BC} and Δ_s by multidimensional spatial operators that capture all $k+d$ dimensions.

For example, if $k = 1$, $d = 1$, and $\mathbf{S}_j$ is the shift operator in the jth spatial dimension ($1 \leq j \leq k+d$), Δ_s should then be taken to be the structured operator

$$\Delta_\mathrm{s} = \mathbf{diag}\left(\mathbf{S}_1 I_{n_1}, \mathbf{S}_1^{-1} I_{n_1}, \mathbf{S}_2 I_{n_2}, \mathbf{S}_2^{-1} I_{n_2}\right).$$

Spatial reversibility can then be defined as in the spatially monodimensional case, the only difference being that the basic building block's realization must now commute with several different matrices, one for each of the d spatial dimensions.

DEFINITION 7.1. *For $1 \leq i \leq d$, let $\mathcal{M}_i := \left(\begin{smallmatrix} 0 & M_i \\ M_i^{-1} & 0 \end{smallmatrix}\right)$ and*

$$Q_i := \mathbf{diag}(I_{n_1}, \ldots, I_{n_k}, I_{n_{k+1}}, \ldots, \mathcal{M}_i, \ldots, I_{n_{k+d}}).$$

We say that the basic building block and, in turn, the interconnections are spatially reversible if, for all i, there exist matrices $R_i \in \mathbb{R}^{m\times m}$, $P_i \in \mathbb{R}^{n_{\mathrm{T}}\times n_{\mathrm{T}}}$, and $U_i \in \mathbb{R}^{p\times p}$ such that

(i) *R_i, U_i, and P_i are involutions;*
(ii) *$R_iR_j = R_jR_i$, $P_iP_j = P_jP_i$, and $U_iU_j = U_jU_i$ for all $1 \leq i, j \leq d$;*
(iii)

$$\begin{pmatrix} P_i & 0 & 0 \\ 0 & Q_i & 0 \\ 0 & 0 & U_i \end{pmatrix} \begin{pmatrix} A_{\mathrm{TT}} & A_{\mathrm{TS}} & B_{\mathrm{T}} \\ A_{\mathrm{ST}} & A_{\mathrm{SS}} & B_{\mathrm{S}} \\ C_{\mathrm{T}} & C_{\mathrm{S}} & D \end{pmatrix} = \begin{pmatrix} A_{\mathrm{TT}} & A_{\mathrm{TS}} & B_{\mathrm{T}} \\ A_{\mathrm{ST}} & A_{\mathrm{SS}} & B_{\mathrm{S}} \\ C_{\mathrm{T}} & C_{\mathrm{S}} & D \end{pmatrix} \begin{pmatrix} P_i & 0 & 0 \\ 0 & Q_i & 0 \\ 0 & 0 & R_i \end{pmatrix}.$$

Condition (ii) is essential for the application of the method of images: it ensures that one can extend the finite extent system by reflection in the d spatial dimensions to yield a periodic one.

One can then proceed to analyze and perform distributed control synthesis for multidimensional spatially reversible systems. All proofs are similar to the spatially monodimensional case but require more intensive notational bookkeeping. The two most important results, corresponding to Corollary 4.6 and Theorem 6.1, are given below for the case where R_i and U_i are unitary for all $1 \leq i \leq d$.

THEOREM 7.2. *If a spatially multidimensional, reversible, well-posed, periodic system is stable, then the corresponding finite extent system is stable. Moreover, if R_i and U_i are unitary for all $1 \leq i \leq d$, the input/output gains of the two systems satisfy*

$$\|T^P_{dz}\|_{\mathcal{L}_2(\mathbb{Z}_{l_1}\times\cdots\times\mathbb{Z}_{l_k}\times\mathbb{Z}_{2L_1}\times\cdots\times\mathbb{Z}_{2L_d})} < 1 \Rightarrow \|T_{dz}\|_{\mathcal{L}_2(\mathbb{Z}_{l_1}\times\cdots\times\mathbb{Z}_{l_k}\times\{1,\ldots,L_1\}\times\cdots\times\{1,\ldots,L_d\})} < 1.$$

THEOREM 7.3. *Given a finite extent, spatially multidimensional reversible plant, there exists a spatially reversible, finite extent controller, with $n^{\mathrm{K}}_{\mathrm{T}} = n_{\mathrm{T}}$, $n^{\mathrm{K}} = n$, and boundary conditions matrix $M^{\mathrm{K}}_i = (M^*_i)^{-1}$, for each $1 \leq i \leq d$, that solves Problem 1 if*

(i) *the LMI conditions of* [5] *for the spatially multidimensional case are satisfied;*
(ii) *the plant's involutions R_i and U_i satisfy $R_i = \mathbf{diag}(R^u_i, R^d_i)$, $U_i = \mathbf{diag}(U^y_i, U^z_i)$ with R^d_i and U^z_i unitary for all $1 \leq i \leq d$.*

The conditions of item (i) are (34a), (34b), and (34c), supplemented by a fourth LMI (equation (92) in [5]) needed to guarantee that the matrix $\overline{A_{\mathrm{SS}}}^{\mathrm{K}}$ yields an *implementable* controller. This LMI is always trivially satisfied in the spatially monodimensional case or when $\overline{A_{\mathrm{SS}}}^{\mathrm{K}}$ is block diagonal. We refer to [5] for more details.

The spatially multidimensional reversible controller is constructed iteratively as follows: First, starting with any satisfactory controller and following steps that are identical to those of the proof of Proposition 6.3, we get controller number 1 such that

$$\begin{pmatrix} P_1 & 0 & 0 \\ 0 & Q^*_1 & 0 \\ 0 & 0 & U_1 \end{pmatrix} \begin{pmatrix} {A_{\mathrm{TT}}}^{\mathrm{K}}_1 & {A_{\mathrm{TS}}}^{\mathrm{K}}_1 & {B_{\mathrm{T}}}^{\mathrm{K}}_1 \\ {A_{\mathrm{ST}}}^{\mathrm{K}}_1 & {A_{\mathrm{SS}}}^{\mathrm{K}}_1 & {B_{\mathrm{S}}}^{\mathrm{K}}_1 \\ {C_{\mathrm{T}}}^{\mathrm{K}}_1 & {C_{\mathrm{S}}}^{\mathrm{K}}_1 & D^{\mathrm{K}}_1 \end{pmatrix} = \begin{pmatrix} {A_{\mathrm{TT}}}^{\mathrm{K}}_1 & {A_{\mathrm{TS}}}^{\mathrm{K}}_1 & {B_{\mathrm{T}}}^{\mathrm{K}}_1 \\ {A_{\mathrm{ST}}}^{\mathrm{K}}_1 & {A_{\mathrm{SS}}}^{\mathrm{K}}_1 & {B_{\mathrm{S}}}^{\mathrm{K}}_1 \\ {C_{\mathrm{T}}}^{\mathrm{K}}_1 & {C_{\mathrm{S}}}^{\mathrm{K}}_1 & D^{\mathrm{K}}_1 \end{pmatrix} \begin{pmatrix} P_1 & 0 & 0 \\ 0 & Q^*_1 & 0 \\ 0 & 0 & R_1 \end{pmatrix}.$$

Then *starting with controller number* 1 and proceeding similarly, we get controller number 2 such that

$$\begin{pmatrix} P_2 & 0 & 0 \\ 0 & Q_2^* & 0 \\ 0 & 0 & U_2 \end{pmatrix} \begin{pmatrix} A_{\mathrm{TT}2}^{\mathrm{K}} & A_{\mathrm{TS}2}^{\mathrm{K}} & B_{\mathrm{T}2}^{\mathrm{K}} \\ A_{\mathrm{ST}2}^{\mathrm{K}} & A_{\mathrm{SS}2}^{\mathrm{K}} & B_{\mathrm{S}2}^{\mathrm{K}} \\ C_{\mathrm{T}2}^{\mathrm{K}} & C_{\mathrm{S}2}^{\mathrm{K}} & D_2^{\mathrm{K}} \end{pmatrix} = \begin{pmatrix} A_{\mathrm{TT}2}^{\mathrm{K}} & A_{\mathrm{TS}2}^{\mathrm{K}} & B_{\mathrm{T}2}^{\mathrm{K}} \\ A_{\mathrm{ST}2}^{\mathrm{K}} & A_{\mathrm{SS}2}^{\mathrm{K}} & B_{\mathrm{S}2}^{\mathrm{K}} \\ C_{\mathrm{T}2}^{\mathrm{K}} & C_{\mathrm{S}2}^{\mathrm{K}} & D_2^{\mathrm{K}} \end{pmatrix} \begin{pmatrix} P_2 & 0 & 0 \\ 0 & Q_2^* & 0 \\ 0 & 0 & R_2 \end{pmatrix}.$$

The final controller, obtained after d such iterations, is spatially reversible because for $1 \le i, j \le d$, P_i and P_j, Q_i and Q_j, R_i and R_j, and U_i and U_j commute.

8. Conclusion. We have shown that a finite extent spatially reversible system is closely related to its periodic and infinite extensions and have demonstrated that synthesis for such systems can be performed with existing tools developed in the context of spatial invariance.

The synthesis LMI conditions that we use—which were already sufficient only for well-posedness, stability, and contractiveness of an infinite system [5]— are even more conservative for finite extent systems since well-posedness of the spatially invariant systems is not necessary for well-posedness of the finite extent system. However, when these LMI are feasible, the obtained controller guarantees stability and performance irrespective of the number of subsystems in the finite extent interconnection, with obvious consequences for system reconfiguration and fault tolerance. The boundary conditions matrix of the plant is the only relevant parameter.

REFERENCES

[1] B. Bamieh, F. Paganini, and M. Dahleh, *Distributed control of spatially invariant systems*, IEEE Trans. Automat. Control, 47 (2002), pp. 1091–1118.

[2] H. T. Banks, R. C. Smith, and Y. Wang, *Smart Material Structures Modeling, Estimation and Control*, John Wiley and Sons, New York, 1996.

[3] R. W. Brockett and J. L. Willems, *Discretized partial differential equations: Examples of control systems defined on modules*, Automatica, 10 (1974), pp. 507–515.

[4] R. F. Curtain and H. J. Zwart, *An Introduction to Infinite-Dimensional Linear System Theory*, Texts in Applied Mathematics 21, Springer-Verlag, Berlin, New York, 1995.

[5] R. D'Andrea and G. E. Dullerud, *Distributed control of spatially interconnected systems*, IEEE Trans. Automat. Control, 48 (2003), pp. 1478–1495.

[6] B. Dionne, M. Golubitsky, and I. Stewart, *Coupled cells with internal symmetry* I *and* II, Nonlinearity, 9 (1996), pp. 559–599.

[7] G. E. Dullerud and F. Paganini, *A Course in Robust Control Theory*, Texts in Applied Mathematics 36, Springer-Verlag, Berlin, New York, 1999.

[8] G. E. Dullerud, R. D'Andrea, and S. G. Lall, *Control of spatially varying distributed systems*, in Proceedings of the 37th IEEE Conference on Decision and Control, Tampa, FL, 1998, pp. 1889–1893.

[9] L. C. Evans, *Partial Differential Equations*, Graduate Studies in Mathematics 19, AMS, Providence, RI, 1998.

[10] F. Fagnani and J. C. Willems, *Representations of symmetric linear dynamical systems*, SIAM J. Control Optim., 31 (1993), pp. 1267–1293.

[11] V. Kapila, A. G. Sparks, J. Buffington, and Q. Yan, *Spacecraft formation flying: Dynamics and control*, in Proceedings of the American Control Conference, San Diego, CA, 1999, pp. 4137–4141.

[12] P. P. Khargonekar, K. Poolla, and A. Tannenbaum, *Robust control of linear time-invariant plants using periodic compensation*, IEEE Trans. Automat. Control, 27 (1982), pp. 627–638.

[13] D. Laughlin, M. Morari, and R. D. Braatz, *Robust performance of cross-directional basis-weight control in paper machines*, Automatica, 29 (1993), pp. 1395–1410.

[14] S. Mijanovic, G. E. Stewart, G. A. Dumont, and M. S. Davies, $\mathcal{H}_\infty$ *robustification of a paper machine cross-directional control system*, in Proceedings of the American Control Conference, 2001, Arlington, VA, pp. 2203–2209.

[15] F. Paganini and B. Bamieh, *Decentralization properties of optimal distributed controllers*, in Proceedings of the 37th IEEE Conference on Decision and Control, Tampa, FL, 1998, pp. 1877–1882.
[16] L. Qiu, *On the robustness of symmetric systems*, Systems Control Lett., 27 (1996), pp. 187–190.
[17] E. D. Sontag, *Linear systems over commutative rings: A survey*, Ricerche Automat., 7 (1976), pp. 1–34.
[18] G. E. Stewart, *Analysis and Design of Boundary Conditions for a Spatially Distributed Control System*, Tech. report, Honeywell Industrial Control, 2001.
[19] D. Swaroop and J. K. Hedrick, *Constant spacing strategies for platooning in automated highway systems*, J. Dynam. Systems, Measurement Control, 121 (1999), pp. 462–470.
[20] M. E. Taylor, *Partial Differential Equations* I: *Basic Theory*, Springer-Verlag, Berlin, New York, 1996.
[21] G. Yang, J. L. Wang, and Y. C. Soh, *Decentralized fixed modes of symmetric systems*, in Proceedings of the American Control Conference, Arlington, VA, 2001, pp. 3134–3135.
[22] J. D. Wolfe, D. F. Chichka, and J. L. Speyer, *Decentralized Controllers for Unmanned Aerial Vehicle Formation Flight*, Tech. report 96-3833, American Institute of Aeronautics and Astronautics, Reston, VA, 1996.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 1, pp. 29–48

OPTIMAL CONTROL OF ERGODIC CONTINUOUS-TIME MARKOV CHAINS WITH AVERAGE SAMPLE-PATH REWARDS*

XIANPING GUO† AND XI-REN CAO‡

Abstract. In this paper we study continuous-time Markov decision processes with the *average sample-path reward* (ASPR) criterion and possibly unbounded transition and reward rates. We propose conditions on the system's *primitive data* for the existence of ϵ-ASPR-optimal (deterministic) stationary policies in a class of randomized Markov policies satisfying some additional continuity assumptions. The proof of this fact is based on the *time discretization* technique, the martingale stability theory, and the concept of potential. We also provide both policy and value iteration algorithms for computing, or at least approximating, the ϵ-ASPR-optimal stationary policies. We illustrate with examples our main results as well as the difference between the ASPR and the average expected reward criteria.

Key words. average sample-path reward, continuous-time Markov chain, optimal stationary policy, policy and value iteration algorithms

AMS subject classifications. 90C40, 93E20

DOI. 10.1137/S0363012903420875

1. Introduction.

Markov decision processes (MDPs) with the long-run *average expected reward* (AER) criterion have been widely studied in literature; see, for instance, the books [1, 6, 12, 22, 23, 25, 31, 32, 33, 35], the survey paper [3], and their extensive references. However, the *sample-path* reward corresponding to an optimal policy that maximizes the average expected rewards may have fluctuations from its expected reward value. To take these fluctuations into account, the *average sample-path reward* (ASPR) criterion has been proposed and studied; see, for instance, [3, 10, 15, 23, 24] and their extensive bibliographies. To the best of our knowledge, all the existing works with the ASPR criterion are on *discrete-time* MDPs. On the other hand, many real-world problems, for instance, in communication engineering, queueing systems, and other control problems, require continuous-time models. Therefore, there is a large amount of works in literature on continuous-time MDPs; see, for instance, [4, 5, 16, 18, 19, 20, 21, 26, 27, 29, 31, 35, 37, 39] and their references. All of these works, however, consider only the AER criterion. Our paper is a first attempt to fill the gap between the works on discrete-time MDPs with the ASPR criterion and those on continuous-time MDPs with the AER criterion.

Denumerable continuous-time MDPs are specified by the system's four *primitive data*: a countable state space S; action sets $A(i)$, which may depend on the current state $i \in S$; transition rates $q(j|i,a)$ with $a \in A(i)$ and $j \in S$; and reward rates $r(i,a)$ with $a \in A(i)$. In this paper, we consider these MDPs with the ASPR criterion in the class of randomized Markov policies satisfying some *additional* continuity assumptions.

*Received by the editors January 7, 2003; accepted for publication (in revised form) August 18, 2004; published electronically June 14, 2005. This work was supported by a grant from Hong Kong UGC.

http://www.siam.org/journals/sicon/44-1/42087.html

†The School of Mathematics and Computational Science, Zhongshan University, Guangzhou 510275, People's Republic of China (mcsgxp@zsu.edu.cn). The research of this author has been supported by the Natural Science Foundations of China and Guangdong Province, by EYTP, NCET, and by Zhongshan University Advanced Research Center, China.

‡Corresponding author. Department of Electrical and Electronic Engineering, The Hong Kong University of Science and Technology, Hong Kong (eecao@ust.hk).

The state processes here are possibly *nonhomogeneous* continuous-time Markov chains with possibly *unbounded* transition rates, and the reward rates may have *neither upper nor lower bounds*. Under suitable conditions on the primitive data, we first prove the existence of a solution to the optimality equation. The proof is constructive, using *policy iteration*, which is based on the concept of potentials [8, 7] and is rather different from both the "vanishing discount approach" in [16, 18, 19, 21, 26] and the "uniformization technique" in [29, 31, 36]. We then establish the existence of $\epsilon(\geq 0)$-ASPR-optimal stationary policies by introducing a *time-discretization approach* to continuous-time martingales and by using the *extended generator* technique. This approach is different from those used for the discrete-time case; see, for instance, [3, 6, 10, 15, 23, 24]. Also, we provide both policy and value iteration algorithms for computing, or at least approximating (when the algorithms take infinitely many steps to converge), $\epsilon(\geq 0)$-ASPR-optimal stationary policies. Furthermore, we use several examples to explain our conditions and to show the difference between the AER and ASPR criteria.

The policy iteration approach developed in this paper to establish a solution to the optimality equation does not require any result about discounted continuous-time MDPs. Thus, this approach is simple and direct. Also, our method to prove the existence of an ASPR-optimal stationary policy is straightforward and different from those in [29, 31, 36, 37], which require the *equivalence* between continuous- and discrete-time MDPs as well as results about discrete-time MDPs. Finally, it should be mentioned that the ergodicity results about continuous-time Markov chains and the convergence results for continuous martingales available in the literature *cannot* be applied to our problems because in this paper the Markov chains may be *nonhomogeneous* and the associated reward and transition rates may be *time-dependent* and *unbounded*. In addition, a key feature of our results is that the conditions are imposed on the *primitive data* (see (2.1)) and can be easily verified.

The rest of this paper is organized as follows. In section 2, we introduce the control model and the optimal control problem considered in this paper. After some technical preliminaries developed in section 3, we study the existence of the $\epsilon(\geq 0)$-ASPR optimal stationary policies in section 4. The policy and value iteration algorithms are described in section 5. Our hypotheses and the difference between the AER and ASPR criteria are illustrated with examples in section 6. We conclude in section 7 with some general remarks.

2. The optimal control problem. The control model that we are concerned with can be described by

$$\{S, A(i), q(j|i,a), r(i,a),\ i,j \in S\}, \tag{2.1}$$

where S is the *state space*; $A(i)$ is a set of *admissible actions* at state $i \in S$; $q(j|i,a)$ with $i,j \in S$ and $a \in A(i)$ are the system's *transition rates*; and $r(i,a)$ with $i \in S$ and $a \in A(i)$ are the *reward rates*. Let $K := \{(i,a) : i \in S,\ a \in A(i)\}$ be the set of all state-action pairs.

In this paper we assume that S is denumerable and in fact we write it as the set of nonnegative integers, i.e., $S = \{0, 1, 2, \ldots\}$. Furthermore, we assume that for each $i \in S$ the set $A(i)$ is a Borel space endowed with the Borel σ-algebra $\mathcal{B}(A(i))$. The transition rates $q(j|i,a)$ in (2.1) satisfy $q(j|i,a) \geq 0$ for all $(i,a) \in K$ and $j \neq i$. Moreover, we assume that the matrix $[q(j|i,a)]$ with (i,j)-element $q(j|i,a)$

is *conservative*, i.e.,

$$\sum_{j\in S} q(j|i,a) = 0 \quad \forall (i,a) \in K,$$

and *stable*, which means that

$$q(i) := \sup_{a\in A(i)} q_i(a) < \infty \quad \forall i \in S,$$

where $q_i(a) := -q(i|i,a) \geq 0$, for all $(i,a) \in K$. In addition, $q(j|i,a)$ is measurable in $a \in A(i)$ for each fixed $i, j \in S$.

Finally, the function $r(i,a)$ on K is a real-valued reward rate, and $r(i,a)$ is assumed to be measurable in $a \in A(i)$ for each fixed $i \in S$. (As $r(i,a)$ is allowed to take positive and negative values, it can be interpreted as a *cost rate* rather than a "reward" rate.)

We first introduce randomized Markov policies.

DEFINITION 2.1 (randomized Markov policies). *A randomized Markov policy is a function $\pi_t(B|i)$ that satisfies the following conditions:*

(1) *for each $i \in S$ and $B \in \mathcal{B}(A(i))$, the mapping $t \mapsto \pi_t(B|i)$ is Borel measurable on $[0,\infty)$, and*

(2) *for each $i \in S$ and $t \geq 0$, $B \mapsto \pi_t(B|i)$ is a probability measure on $\mathcal{B}(A(i))$.*

Let $A := \bigcup_{i\in S} A(i)$. A (deterministic) stationary policy is a function $f : S \to A$ such that $f(i)$ is in $A(i)$ for all $i \in S$.

Let Φ be the set of all randomized Markov policies and let F be the set of all stationary policies. Note that a function $f \in F$ can be viewed as a function $\pi_t(B|i) \in \Phi$ for which, for all $t \geq 0$ and $i \in S$, $\pi_t(\cdot|i)$ is the Dirac measure at $f(i)$. Thus, $F \subset \Phi$. We will write a randomized Markov policy $\pi_t(B|i)$ in Φ simply as (π_t). The subscript "t" in π_t indicates the possible dependence on time; it will be dropped for simplicity when there is no confusion.

For each $(\pi_t) \in \Phi$, the associated transition and reward rates are defined, respectively, as follows:

$$q(j|i,\pi_t) := \int_{A(i)} q(j|i,a)\pi_t(da|i) \quad \text{for } i,j \in S \text{ and } t \geq 0, \tag{2.2}$$

$$r(i,\pi_t) := \int_{A(i)} r(i,a)\pi_t(da|i) \quad \text{for } i \in S \text{ and } t \geq 0. \tag{2.3}$$

Obviously, the transition rate $q(j|i,\pi_t)$ and reward rate $r(i,\pi_t)$ can depend on time t if π is not stationary. When $\pi = f \in F$, we write $q(j|i,\pi_t)$ and $r(i,\pi_t)$ as $q(j|i,f(i))$ and $r(i,f(i))$, respectively.

For each $\pi := (\pi_t) \in \Phi$, let $Q(\pi_t) := [q(j|i,\pi_t)]$ with $t \geq 0$ be the transition rate matrices. Any (possibly substochastic and nonhomogeneous) transition function $\tilde{p}(s,i,t,j,\pi)$ such that

$$\lim_{\gamma\to 0^+} \frac{\tilde{p}(t,i,t+\gamma,j,\pi) - \delta_{ij}}{\gamma} = q(j|i,\pi_t) \quad \forall i,j \in S \text{ and } t \geq 0$$

is called a Q-*process* with the transition rate matrices $Q(\pi_t)$, where δ_{ij} is the Kronecker delta. To guarantee the existence of such a Q-process, we now define the class of admissible policies.

DEFINITION 2.2 (admissible policies). *A randomized Markov policy* (π_t) *in* Φ *is said to be admissible if* $q(j|i,\pi_t)$ *is continuous in* $t \geq 0$ *for each fixed* $i, j \in S$. *We denote by* Π *the class of all admissible policies.*

Π is *nonempty* because it contains F. Moreover, as shown in Example 6.3 below, Π contains a randomized Markov policy which is *not* in F.

On the other hand, $Q(\pi_t)$ is also conservative and stable, i.e.,

$$q_i(\pi_t) := -q(i|i,\pi_t) = \sum_{j\neq i} q(j|i,\pi_t) < \infty \quad \forall i \in S \text{ and } t \geq 0.$$

Hence, for each $\pi \in \Pi$, the existence of a Q-process such as the *minimum* Q-process denoted by $p^{\min}(s,i,t,j,\pi)$ (i.e., $p^{\min}(s,i,t,j,\pi) \leq \tilde{p}(s,i,t,j,\pi)$ for any Q-process $\tilde{p}(s,i,t,j,\pi)$) is guaranteed but is not necessarily regular; that is, we might have $\sum_{j\in S} p^{\min}(s,i,t,j,\pi) < 1$ for some $i \in S$ and $t \geq s \geq 0$ (see [13] or Theorem 4.2.6 in [2]).

To ensure the regularity of a Q-process, we use the following ergodicity conditions.

Assumption A. There exist a sequence $\{S_n,\ n \geq 1\}$ of subsets of S, a nondecreasing function $w \geq 1$ on S, and two constants $c > 0$ and $b \geq 0$, such that

(1) $\sup_{i\in S_n} q(i) < \infty$ for each $n \geq 1$, and $S_n \uparrow S$ in the sense of convergence of a set sequence;

(2) $\lim_{n\to\infty}[\inf_{j\notin S_n} w(j)] = +\infty$;

(3) $\sum_{j\in S} q(j|i,a)w(j) \leq -cw(i) + b\delta_{0i} \forall\ (i,a) \in K$; and

(4) for each $f \in F$, the minimum Q-process $p^{\min}(s,i,t,j,f)$ is *monotone*, i.e.,

$$\sum_{j\geq k} q(j|i,f(i)) \leq \sum_{j\geq k} q(j|i+1,f(i+1)) \quad \forall i,k \in S \text{ with } k \neq i+1,$$

and *irreducible*, i.e., for each pair of states i and j, either $q(j|i,f(i)) > 0$, or there are an integer l (which may depend on i, j, and f) and l states $i_1, i_2, \ldots, i_l$ with $i \neq i_1$, $j \neq i_l$, $i_{k-1} \neq i_k$, $k = 2, \ldots, l$, such that

$$q(i_1|i,f(i))q(i_2|i_1,f(i_1))\cdots q(i_l|i_{l-1},f(i_{l-1}))q(j|i_l,f(i_l)) > 0.$$

LEMMA 2.3. (a) *If Assumptions* A(1), A(2), *and* A(3) *hold, then for each* $\pi = (\pi_t) \in \Pi$ *the corresponding Q-process with transition rate matrices* $Q(\pi_t)$ *is regular; that is,*

$$\sum_{j\in S} p^{\min}(s,i,t,j,\pi) = 1 \quad \forall i \in S \text{ and } t \geq s \geq 0.$$

(b) *If Assumption* A *holds, then for each* $f \in F$ *the corresponding Q-process with transition rate matrices* $[Q(j|i,f(i))]$ *is ergodic, and its unique invariant probability measure* μ_f *(with* $\mu_f(i) > 0$ *for all* $i \in S$*) can be determined by the equation*

$$\sum_{i\in S} \mu_f(i)q(j|i,f(i)) = 0 \quad \forall j \in S. \tag{2.4}$$

Moreover, for each $i \in S$ *and* $t \geq 0$

$$\left|\sum_{j\in S} p^{\min}(0,i,t,j,f)h(j) - \mu_f(h)\right| \leq 2e^{-ct}\left[w(i) + \frac{b}{c}\right] \leq 2e^{-ct}\left(1 + \frac{b}{c}\right)w(i) \tag{2.5}$$

for any function h on S such that $|h| \le w$, where $\mu_f(h) := \sum_{j\in S} h(j)\mu_f(j)$.

Proof. (a) Under Assumptions A(1)–A(3), by Theorem 3.1 in [17] we see that (a) is true.

(b) By (a) and Proposition 5.4.1 in [2], we see that (2.4) is true. Moreover, from the proof of (3.9) in [30] we see that the condition (2.1) in [30] is not required for Theorem 2.2(ii) in [30]. Thus, by (3.9) in [30] and Assumption A we see that (2.5) is also true. □

Under Assumptions A(1)–A(3), Lemma 2.3 shows that for each $\pi = (\pi_t) \in \Pi$ a Q-process with transition rate matrices $Q(\pi_t)$ is regular. Thus, under Assumption A, we will denote by $\{x(t,\pi)\}$ the associated right-continuous Markov chain with values in S, and write the regular Q-process $p^{\min}(s,i,t,j,\pi)$ simply as $p(s,i,t,j,\pi)$. Furthermore, for each initial state $i \in S$ at time $s = 0$, we denote by $(\Omega, \mathcal{F}, P_i^\pi)$ the probability measure space determined by $p(s,i,t,j,\pi)$, by E_i^π the corresponding expectation operator, and by $x(t,\pi)(e)$ the value of $x(t,\pi)$ at $e \in \mathcal{F}$.

Remark 2.4. (a) For the case where $\sup_{i\in S} q(i) < \infty$ (see, for instance, [7, 26, 31, 35, 39]), Assumptions A(1) and A(2) are not required because they are used only to guarantee the regularity of a Q-process. For the case of *unbounded* transition rates (e.g., [18, 19]), the conditions for a Q-process to be regular are usually imposed on both the possibly *nonhomogeneous* minimum Q-processes and the transition rates. Hence, our Assumptions A(1)–A(3) are quite different from those in [18, 19].

(b) Assumptions A(1)–A(3) are an extension of both the "drift condition" in [30] and the hypotheses of Corollary 2.2.16 in [2] for a *homogeneous* Q-process to be regular. Assumption A(4) is a variant of the monotonicity conditions in Theorem 7.3.4 and the irreducibility conditions in Proposition 5.3.1 in [2].

(c) It should be mentioned that if there is a set $\bar{S}$ of transient states which is independent of stationary policies, then $\mu_f(i) = 0$ for each $i \in \bar{S}$ and $f \in F$. In this case, Lemma 3.4 below may *not* hold because its proof uses the result $\mu_f(i) > 0$ for all $i \in S$.

Now we define the ASPR criterion $V_{sp}(\cdot,\cdot)$ as follows: for each $\pi = (\pi_t) \in \Pi$ and $i \in S$

$$V_{sp}(\pi,i) := \limsup_{T\to\infty} \frac{1}{T}\left[\int_0^T r(x(t,\pi),\pi_t)dt\right], \tag{2.6}$$

where the subscript "sp" stands for "sample-path." Note that $V_{sp}(\pi,i)$ has been defined by the so-called *sample-path rewards* $r(x(t,\pi),\pi_t)$; therefore, it is a *random variable* rather than a number as in the AER-criterion defined as

$$\bar{V}(\pi,i) := \limsup_{T\to\infty} \frac{1}{T}\left[\int_0^T E_i^\pi r(x(t,\pi),\pi_t)dt\right] \tag{2.7}$$

(see [4, 7, 16, 19, 20, 21, 26, 27, 31, 35, 39], for instance). Thus, the following definition of optimal policies for the ASPR criterion is different from that for the AER criterion.

Definition 2.5. *For a given $\epsilon \ge 0$, a policy $\pi^* \in \Pi$ is said to be ϵ-ASPR-optimal if there exists a constant g^* such that*

$$P_i^{\pi^*}(V_{sp}(\pi^*,i) \ge g^* - \epsilon) = 1 \quad \textit{and} \quad P_i^\pi(V_{sp}(\pi,i) \le g^*) = 1 \quad \forall i \in S \textit{ and } \pi \in \Pi.$$

A 0-ASPR-optimal policy is simply called an ASPR-optimal policy.

The main goal of this paper is to give conditions on the primitive data in (2.1) that ensure the existence of an ASPR-optimal stationary policy.

3. Preliminaries. In this section we present some preliminary facts that are needed to prove our main results.

Let $w \geq 1$ be the function in Assumption A. Following the concept of a weighted supremum norm introduced by Lippman [28] and widely used by many authors (e.g., [23, p. 2]), we define the weighted supremum norm $\|v\|_w$ for a real-valued functions v on S by

$$\|v\|_w := \sup_{i \in S}[w(i)^{-1}|v(i)|]$$

and the Banach space by $B_w(S) := \{v : \|v\|_w < \infty\}$.

LEMMA 3.1. *Let $\bar{w}$ be any nonnegative function on S, and $\bar{c}$, $\bar{b}$ two constants such that $\bar{b} \geq 0$ and $\bar{c} \neq 0$. Then, for each $\pi = (\pi_t) \in \Pi$, the following statements are equivalent:*

(a) $\sum_{j \in S} p^{\min}(s, i, t, j, \pi)\bar{w}(j) \leq e^{-\bar{c}(t-s)}\bar{w}(i) + \frac{\bar{b}}{\bar{c}}[1 - e^{-\bar{c}(t-s)}]$ *for all $i \in S$ and $t \geq s \geq 0$;*

(b) $\sum_{j \in S} q(j|i, \pi_t)\bar{w}(j) \leq -\bar{c}\bar{w}(i) + \bar{b}$ *for all $i \in S$ and $t \geq 0$.*

Proof. See Lemma 3.2 in [16]. □

It should be noted that in Lemma 3.1, Assumption A is *not* required.

To establish the so-called optimality equation, we will use a *policy iteration algorithm* instead of the *vanishing discount approach* in [16, 19, 21, 26]. To state the policy iteration algorithm, in addition to Assumption A we also need the following standard continuity-compactness conditions (Assumption B); see, for instance, [3, 19, 23, 31, 35] and their references.

Assumption B. (1) For each $i \in S$, $A(i)$ is compact.

(2) $r(i, a)$ and $q(j|i, a)$ are continuous in $a \in A(i)$ for each fixed $i, j \in S$.

(3) The function $\sum_{j \in S} q(j|i, a)w(j)$ is continuous in $a \in A(i)$ for each fixed $i \in S$.

(4) There exists a positive constant M such that $|r(i, a)| \leq Mw(i)$ for all $i \in S$ and $a \in A(i)$.

In the spirit of the potential concept in [8, 7], for a given $f \in F$ and the corresponding unique invariant probability measure μ_f, we define the *potential*

$$u(f, i) := \int_0^\infty [E_i^f r(x(t, f), f(x(t, f))) - g(f)]dt \quad \forall\, i \in S, \tag{3.1}$$

where the constant $g(f)$ is defined as

$$g(f) := \sum_{j \in S} r(j, f(j))\mu_f(j). \tag{3.2}$$

LEMMA 3.2. *Let Assumptions* A *and* B(4) *hold. Then*

(a) *$g(f)$ and $\|u(f, \cdot)\|_w$ are both bounded in $f \in F$,*

(b) *the Poisson equation $g(f) = r(i, f(i)) + \sum_{j \in S} q(j|i, f(i))u(f, j)$ holds for all $i \in S$ and $f \in F$.*

Proof. By (3.1) and (2.5) we see that $\|u(f, \cdot)\|_w$ is bounded in $f \in F$. With the constants M, c, and b as in Assumptions A and B(4), by Lemma 3.1, (3.1), (2.5), and (2.7), we have $|\overline{V}(f, \cdot)| = |g(f)| \leq \frac{Mb}{c}$ for all $f \in F$, and so (a) follows. Obviously, (b) follows from Lemma 5.1 in [16]. □

Under Assumptions A and B, we now state the policy iteration algorithm.

POLICY ITERATION ALGORITHM 3.1.

Step I. Take $n = 0$ and $f_n \in F$.

Step II. Solve (2.4) for μ_{f_n} and then calculate $u(f_n, \cdot)$ and $g(f_n)$ as in (3.1) and (3.2).

Step III. Define a new stationary policy f_{n+1} in the following way:

Set $f_{n+1}(i) := f_n(i)$ for all $i \in S$ for which

$$r(i, f_n(i)) + \sum_{j \in S} q(j|i, f_n(i))u(f_n, j) = \max_{a \in A(i)} \left\{ r(i,a) + \sum_{j \in S} q(j|i,a)u(f_n, j) \right\}; \tag{3.3}$$

otherwise (i.e., when (3.3) does not hold), choose $f_{n+1}(i) \in A(i)$ such that

$$\begin{aligned} r(i, f_{n+1}(i)) &+ \sum_{j \in S} q(j|i, f_{n+1}(i))u(f_n, j) \\ &= \max_{a \in A(i)} \left\{ r(i,a) + \sum_{j \in S} q(j|i,a)u(f_n, j) \right\}. \end{aligned} \tag{3.4}$$

Step IV. If $f_{n+1}(i)$ satisfies (3.3) for all $i \in S$, then stop because (by Theorem 4.1 below) f_{n+1} is ASPR-optimal; otherwise, replace f_n with f_{n+1} and go back to *Step* II.

Finally, to prove the existence of an ASPR-optimal stationary policy, in addition to Assumptions A and B we propose the following conditions.

Assumption C. There exist nonnegative functions $w_k^* \geq 1$ on S as well as constants $c_k^* > 0$, $b_k^* \geq 0$, and $M_k^* > 0$ $(k = 1, 2)$ such that, for each $i \in S$ and $a \in A(i)$,

(1) $w^2(i) \leq M_1^* w_1^*(i)$ and $\sum_{j \in S} q(j|i,a)w_1^*(j) \leq -c_1^* w_1^*(i) + b_1^*$, and

(2) $[q(i)w(i)]^2 \leq M_2^* w_2^*(i)$ and $\sum_{j \in S} q(j|i,a)w_2^*(j) \leq -c_2^* w_k^*(i) + b_2^*$.

Remark 3.3. (a) Assumption C allows us to use the martingale stability theorem; see Lemma 3.11 in [22], for instance. However, it is not required when a solution u^* in (4.1) below and the transition rates are both uniformly bounded.

(b) Assumption C(2) is slightly different from Assumption B(4) in [16], but all conclusions in [16] still hold after Assumption B(4) in [16] is replaced by Assumption C(2) here.

For each $n \geq 1$, take f_n as the policy obtained in the policy iteration algorithm 3.1, and for each $i \in S$ let

$$\varepsilon(f_n, i) := r(i, f_n(i)) + \sum_{j \in S} q(j|i, f_n(i))u(f_{n-1}, j) - g(f_{n-1}). \tag{3.5}$$

LEMMA 3.4. *Let Assumptions* A, B, *and* C(2) *hold. Then* $g(f_{n+1}) > g(f_n)$ *when* $f_{n+1} \neq f_n$, *and for each* $i \in S$, $\varepsilon(f_n, i) \to 0$ *as* $n \to \infty$.

Proof. As in the proof of Theorem 5.2 and Lemma 5.3 in [16], by Lemma 3.2 above we obtain Lemma 3.4. □

Lemma 3.4 will be used to establish the optimality equation (4.1) below.

4. The existence of ASPR-optimal stationary policies. In this section, we state and prove our main result, Theorem 4.1.

THEOREM 4.1. *Under Assumptions* A, B, *and* C, *the following statements hold.*

(a) *There exist a unique constant* g^*, *a function* $u^* \in B_w(S)$, *and a stationary policy* $f^* \in F$ *satisfying the optimality equation*

$$g^* = r(i, f^*(i)) + \sum_{j \in S} q(j|i, f^*(i))u^*(j)$$

$$= \max_{a\in A(i)} \left\{ r(i,a) + \sum_{j\in S} q(j|i,a)u^*(j) \right\} \quad \forall i \in S. \tag{4.1}$$

(b) *The policy f^* in* (a) *is ASPR-optimal, and $P_i^{f^*}(V_{sp}(f^*,i) = g^*) = 1$ for all $i \in S$.*

(c) *A policy f in F is ASPR-optimal if and only if it realizes the maximum of* (4.1).

(d) *For given $\epsilon \geq 0$ and $f \in F$, if there is a function $\bar{u} \in B_w(S)$ such that*

$$g^* \leq r(i,f(i)) + \sum_{j\in S} q(j|i,f(i))\bar{u}(j) + \epsilon \quad \forall i \in S,$$

then f is ϵ-ASPR-optimal.

Proof. (a) Let $\{f_n\}$ be the sequence of the stationary policies obtained by the policy iteration algorithm 3.1. By Assumption B(1) and the Tichonoff theorem, the policy class F is compact. Thus, by Lemma 3.2(a), there exist a subsequence $\{f_{n_k}\}$ of $\{f_n\}$ and $u^* \in B_w(S)$ such that for each $i \in S$

$$\lim_{k\to\infty} u(f_{n_k},i) = u^*(i), \quad \lim_{k\to\infty} f_{n_k}(i) =: f^*(i), \quad \text{and} \quad \lim_{k\to\infty} g(f_{n_k}) =: g^*. \tag{4.2}$$

On the other hand, by Lemmas 3.2(b), (3.4), and (3.5), we have

$$\begin{aligned} g(f_{n_k}) &= r(i,f_{n_k}(i)) + \sum_{j\in S} q(j|i,f_{n_k}(i))u(f_{n_k},j) \\ &= \max_{a\in A(i)} \left\{ r(i,a) + \sum_{j\in S} q(j|i,a)u(f_{n_k},j) \right\} - \varepsilon(f_{n_k+1},i) \\ &\geq r(i,a) + \sum_{j\in S} q(j|i,a)u(f_{n_k},j) - \varepsilon(f_{n_k+1},i) \quad \forall i \in S \text{ and } a \in A(i). \end{aligned} \tag{4.3}$$

Letting $k \to \infty$ in (4.3), by the "extension of Fatou's Lemma" 8.3.7 in [23] and our Lemma 3.4 and (4.2), we have

$$\begin{aligned} g^* &= r(i,f^*(i)) + \sum_{j\in S} q(j|i,f^*(i))u^*(j) \\ &\geq r(i,a) + \sum_{j\in S} q(j|i,a)u^*(j) \quad \forall i \in S \text{ and } a \in A(i), \end{aligned}$$

and so

$$\begin{aligned} g^* &= r(i,f^*(i)) + \sum_{j\in S} q(j|i,f^*(i))u^*(j) \\ &= \max_{a\in A(i)} \left\{ r(i,a) + \sum_{j\in S} q(j|i,a)u^*(j) \right\} \quad \forall i \in S, \end{aligned}$$

which gives (4.1). Moreover, the proof of the uniqueness of the constant g^* satisfying (4.1) follows from Theorem 4.1(b) in [16] and Remark 3.3(b).

(b) To prove (b), for each $i \in S$, $\pi = (\pi_t) \in \Pi$, and $t \geq 0$, let

$$\Delta(i,\pi_t) := r(i,\pi_t) + \sum_{j\in S} q(j|i,\pi_t)u^*(j) - g^*, \tag{4.4}$$

$$\mathcal{F}_t(\pi) := \sigma\{x(s,\pi),\ 0 \le s \le t\},$$
$$g(i,\pi_t) := \sum_{j\in S} q(j|i,\pi_t)u^*(j). \tag{4.5}$$

In particular, let $\Delta(i,f(i)) =: r(i,f(i)) + \sum_{j\in S} q(j|i,f(i))u^*(j) - g^*$ for all $f \in F$.

We now define a (continuous-time) stochastic process,

$$M(t,\pi) := \int_0^t g(x(y,\pi),\pi_y)dy - u^*(x(t,\pi)) \quad \text{for } t \ge 0. \tag{4.6}$$

Then $\{M(t,\pi), \mathcal{F}_t(\pi),\ t \ge 0\}$ is a P_i^π-martingale in continuous-time; that is,

$$E_i^\pi[M(t,\pi)|\mathcal{F}_s(\pi)] = M(s,\pi) \quad \forall t \ge s \ge 0. \tag{4.7}$$

Indeed, for each $t \ge s \ge 0$, by (4.6) and the Markov property we have

$$\begin{aligned} E_i^\pi[M(t,\pi)|\mathcal{F}_s(\pi)] &= M(s,\pi) + E_i^\pi\left[\int_s^t g(x(y,\pi),\pi_y)dy|\mathcal{F}_s(\pi)\right] \\ &\quad + u^*(x(s,\pi)) - E^\pi_{x(s,\pi)}u^*(x(t,\pi)). \end{aligned} \tag{4.8}$$

Since $u^* \in B_w(S)$ (by (4.2) and Lemma 3.2(a)), it follows from Assumption A(3) that

$$\begin{aligned} \left|\sum_{j\in S} q(j|i,a)u^*(j)\right| &\le \|u^*\|_w\left[\sum_{j\in S} q(j|i,a)w(j) - 2q(i|i,a)w(i)\right] \\ &\le \|u^*\|_w[-cw(i) + b + 2q(i)w(i)] \\ &\le \|u^*\|_w[b + 2q(i)w(i)] \end{aligned} \tag{4.9}$$

for all $a \in A(i)$ and $i \in S$. Therefore, by (4.5) and (2.2) we obtain

$$|g(i,\pi_y)| \le \|u^*\|_w[b + 2q(i)w(i)] \quad \forall y \ge 0 \text{ and } i \in S. \tag{4.10}$$

On the other hand, by the Markov property we have

$$E_i^\pi\left[\int_s^t g(x(y,\pi),\pi_y)dy|\mathcal{F}_s(\pi)\right] = E^\pi_{x(s,\pi)}\left[\int_s^t g(x(y,\pi),\pi_y)dy\right],$$

which together with (4.10), Assumption C(2), Lemma 3.1, and Fubini's theorem gives

$$E_i^\pi\left[\int_s^t g(x(y,\pi),\pi_y)dy|\mathcal{F}_s(\pi)\right] = \int_s^t\left[E^\pi_{x(s,\pi)}g(x(y,\pi),\pi_y)\right]dy. \tag{4.11}$$

From Lemma 2.1(b) in [21] and (4.10) in [16] about the *extended generator* of a possibly nonhomogeneous continuous-time Markov process, by (4.11) and (4.5) we obtain

$$E_i^\pi\left[\int_s^t g(x(y,\pi),\pi_y)dy|\mathcal{F}_s(\pi)\right] = E^\pi_{x(s,\pi)}u^*(x(t,\pi)) - u^*(x(s,\pi)),$$

which together with (4.8) gives (4.7).

It follows from (4.7) that $\{M(n,\pi), \mathcal{F}_n(\pi),\ n \ge 1\}$ is also a P_i^π-martingale in discrete-time. Moreover, By Assumption C and Lemma 3.1 we have

$$E_i^\pi w_k^*(x(t,\pi)) \le w_k^*(i) + \frac{b_k^*}{c_k^*} \quad \forall t \ge 0 \text{ and } k = 1,2, \tag{4.12}$$

which together with (4.6), (4.10), the Hölder inequality, and Assumption C gives

$$
\begin{aligned}
&E_i^\pi[M(n+1,\pi)-M(n,\pi)]^2\\
&\quad= E_i^\pi\left[\int_n^{n+1} g(x(y,\pi),\pi_y)dy + u^*(x(n,\pi)) - u^*(x(n+1,\pi))\right]^2\\
&\quad\le 2E_i^\pi\left[\int_n^{n+1} g(x(y,\pi),\pi_y)dy\right]^2 + 2E_i^\pi[u^*(x(n+1,\pi)) - u^*(x(n,\pi))]^2\\
&\quad\le 2E_i^\pi\left[\int_n^{n+1} g^2(x(y,\pi),\pi_y)dy\right] \qquad \text{(by the Hölder inequality)}\\
&\qquad + 4\|u^*\|_w^2 E_i^\pi[w^2(x(n+1,\pi)) + w^2(x(n,\pi))]\\
&\quad\le 2E_i^\pi\left[\int_n^{n+1} \|u^*\|_w^2[b+2q(x(y,\pi))w(x(y,\pi))]^2dy\right] \quad \text{(by (4.10))}\\
&\qquad + 4M_1^*\|u^*\|_w^2 E_i^\pi[w_1^*(x(n+1,\pi)) + w_1^*(x(n,\pi))] \qquad \text{(by Assumption C(1))}\\
&\quad\le 4\|u^*\|_w^2 E_i^\pi\left[\int_n^{n+1} (b^2 + 4[q(x(y,\pi))w(x(y,\pi))]^2)dy\right]\\
&\qquad + 4M_1^*\|u^*\|_w^2 E_i^\pi[w_1^*(x(n+1,\pi)) + w_1^*(x(n,\pi))]\\
&\quad\le 4\|u^*\|_w^2 E_i^\pi\left[\int_n^{n+1} (b^2 + 4M_2^* w_2^*(x(y,\pi)))dy\right] \qquad \text{(by Assumption C(2))}\\
&\qquad + 4M_1^*\|u^*\|_w^2 E_i^\pi[w_1^*(x(n+1,\pi)) + w_1^*(x(n,\pi))],
\end{aligned}
$$

which gives

$$
\begin{aligned}
(4.13)\quad &E_i^\pi[M(n+1,\pi)-M(n,\pi)]^2\\
&\le 16\|u^*\|_w^2\left[b^2 + M_2^*\left(w_2^*(i) + \frac{b_2^*}{c_2^*}\right) + M_1^*\left(w_1^*(i) + \frac{b_1^*}{c_1^*}\right)\right] \quad \text{(by (4.12))}.
\end{aligned}
$$

This means that $E_i^\pi[M(n+1,\pi)-M(n,\pi)]^2$ is bounded in $n \ge 1$. Thus, by the martingale stability theorem (e.g., [22, p. 105], or Remark 11.2.6 in [23], for instance), we have

$$
\lim_{n\to\infty} \frac{M(n,\pi)}{n} = 0 \quad \text{a.s.} - P_i^\pi. \tag{4.14}
$$

On the other hand, for any $T \ge 1$, let $[T]$ be the unique integer such that $[T] \le T < [T]+1$. By (4.6) we have

$$
\frac{M(T,\pi)}{T} = \frac{[T]}{T}\left(\frac{M([T],\pi)}{[T]} + \frac{\int_{[T]}^T g(x(y,\pi),\pi_y)dy}{[T]} - \frac{u^*(x(T,\pi))}{[T]} + \frac{u^*(x([T],\pi))}{[T]}\right). \tag{4.15}
$$

Moreover, for any arbitrary $\epsilon > 0$, as in the proof of (4.13), by the Chebyshev's inequality we have

$$
\begin{aligned}
P_i^\pi\left(\frac{\left|\int_{[T]}^T g(x(y,\pi),\pi_y)dy\right|}{[T]} > \epsilon\right) &\le \frac{E_i^\pi\left[\int_{[T]}^T |g(x(y,\pi),\pi_y)|dy\right]^2}{\epsilon^2[T]^2}\\
&\le \frac{16\|u^*\|_w^2\left[b^2 + M_2^*\left(w_2^*(i) + \frac{b_2^*}{c_2^*}\right)\right]}{\epsilon^2[T]^2}. \qquad (4.16)
\end{aligned}
$$

Since $\sum_{[T]=1}^{\infty} \frac{1}{[T]^2} < \infty$, by (4.16) and the Borel–Cantelli lemma, we have

$$P_i^\pi \left(\limsup_{[T]} \left\{ \frac{\left| \int_{[T]}^{T} g(x(y,\pi),\pi_y)dy \right|}{[T]} > \epsilon \right\} \right) = 0.$$

Now let

$$E_{[T]} := \left\{ \frac{|\int_{[T]}^{T} g(x(y,\pi),\pi_y)dy|}{[T]} > \epsilon \right\} \in \mathcal{F},$$

$E := \limsup_{[T]} E_{[T]} \in \mathcal{F}$, and $E^c := \Omega - E$ being the complement of set E. Then $P_i^\pi(E^c) = 1$. Let $e \in E^c$, which means that e is in finitely many sets $E_{[T]}$. So there exists an integer $N_0(e)$ (depending on e) such that $e \notin E_{[T]}$ for all $[T] \geq N_0(e)$, i.e.,

$$\frac{\left| \int_{[T]}^{T} g(x(y,\pi)(e),\pi_y)dy \right|}{[T]} \leq \epsilon \quad \forall\ [T] \geq N_0(e) \text{ and } e \in E^c,$$

which together with $P_i^\pi(E^c) = 1$ yields

$$\lim_{[T]\to\infty} \frac{\int_{[T]}^{T} g(x(y,\pi),\pi_y)dy}{[T]} = 0 \quad \text{a.s.} - P_i^\pi. \tag{4.17}$$

Similarly, we have

$$\lim_{[T]\to\infty} \frac{u^*(x(T,\pi))}{[T]} = \lim_{[T]\to\infty} \frac{u^*(x([T],\pi))}{[T]} = 0 \quad \text{a.s.} - P_i^\pi. \tag{4.18}$$

Since $\lim_{T\to\infty} \frac{[T]}{T} = 1$, by (4.14), (4.15), (4.17), and (4.18), we have

$$\lim_{T\to\infty} \frac{M(T,\pi)}{T} = 0 \quad \text{a.s.} - P_i^\pi. \tag{4.19}$$

By (4.4)–(4.6) it follows that

$$M(t,\pi) = -\int_0^t r(x(y,\pi),\pi_y)dy + \int_0^t \Delta(x(y,\pi),\pi_y)dy - u^*(x(t,\pi)) + tg^*. \tag{4.20}$$

By (4.1), (4.4), (2.2), and (2.3), we have $\Delta(i,\pi_t) \leq 0$ and $\Delta(i, f^*(i)) = 0$ for all $t \geq 0$ and $i \in S$. Thus, by (4.18), (4.19), and (4.20) we obtain

$$P_i^\pi(V_{sp}(\pi,i) \leq g^*) = 1 \ \text{ and} \tag{4.21}$$

$$P_i^{f^*}(V_{sp}(f^*,i) = g^*) = 1, \tag{4.22}$$

which, together with the arbitrariness of π and i, give (b).

(c) By (b), it suffices to prove that $f(\in F)$ realizes the maximum of (4.1) if f is SPAR-optimal. Now suppose that f is SPAR-optimal but does not realize the maximum of (4.1). Then there exist some $i_0 \in S$ and a constant $\alpha(i_0, f) > 0$ (depending on i_0 and f) such that

$$g^* \geq [r(i,f(i)) + \alpha(i_0,f)\delta_{i_0 i}] + \sum_{j\in S} q(j|i,f(i))u^*(j) \quad \forall i \in S. \tag{4.23}$$

On the other hand, since f is SPAR-optimal, by (b) and (4.21) we have $V_{sp}(f,i) = g^*$ a.s. for all $i \in S$. Moreover, as in the proof of (4.22), from Lemma 3.2(b) we also have $V_{sp}(f,i) = g(f)$ a.s., and so

$$g^* = g(f) = \sum_{j\in S} \mu_f(j) r(j, f(j)). \tag{4.24}$$

Also, as in the proof of (4.12) in [16], by (4.23) and (4.24) as well as (2.7) we obtain

$$g^* \geq \sum_{j\in S} \mu_f(j)[r(j,f(j)) + \alpha(i_0,f)\delta_{i_0 j}] = g^* + \mu_f(i_0)\alpha(i_0,f),$$

which gives a contradiction because $\mu_f(i_0)$ and $\alpha(i_0,f)$ are both positive.

(d) Let $\Delta_{\bar{u}}(i,f(i)) := r(i,f(i)) + \sum_{j\in S} q(j|i,f(i))\bar{u}(j) - g^*$. Then, $\Delta_{\bar{u}}(i,f(i)) \geq -\epsilon$ for all $i \in S$. Thus, as in the proof of (4.21), we have

$$P_i^f(V_{sp}(f,i) \geq g^* - \epsilon) = 1,$$

which together with (b) gives (d). □

Theorem 4.1 is an important result: part (a) establishes the optimality equation (4.1) and the existence of a so-called *canonical* policy f^*, whereas part (b) further shows that the canonical policy f^* is ASPR-optimal.

Remark 4.2. (a) Under Assumptions A, B, and C(2) only, from the proof of Theorem 4.1 here and Theorem 4.1 in [16] we see that the canonical policy f^* in Theorem 4.1(a) is also optimal for the AER criterion. However, it is shown that an optimal stationary policy for the AER criterion may *not* be canonical [18]. Therefore, it is natural to *guess* that an ASPR-optimal stationary policy may *not* be canonical either. An attempt to answer this problem faces significant technical difficulties, and the problem remains unsolved to this date.

(b) From the proof of Theorem 4.1(b) and (c) we see that both Assumptions C(1) and C(2) are indeed required for the ASPR criterion. That is because (i) the proof of Theorem 4.1(b) and (c) uses the estimates in (4.13) and (4.16), and (ii) the proof of (4.13) and (4.16) is based on both Assumptions C(1) and C(2); see the proof of (4.12) and (4.24) (In the proof of the "if" part of Theorem 4.1(c), we cannot obtain (4.24) by the dominated convergence theorem because $V_{sp}(i,f)$ is defined via "limsup" instead of "lim.")

(c) To establish the optimality equation (4.1), we have used the *policy iteration algorithm* 3.1, instead of the "vanishing discount factor method" used in [16, 18, 19, 21, 26], for instance. It should be noted that our approach is *direct* because it does not require any result about discounted continuous-time MDPs. This is by way of the same logic introduced in [25] for *discrete-time* unichain MDPs. (A similar approach is adopted for discrete-time general MDPs with finite state and action sets in [9, 38] by using simple algebra and properties of the Cesaro-limit of a transition probability matrix and in [12, 31] by using vanishing discount factors.)

(d) We can also prove Theorem 4.1 by using the "vanishing discount factor method." More precisely, under Assumptions A and B, we can (i) establish the average optimality inequalities by using the α-discounted optimality equation in [17], (ii) obtain the optimality equation, and (iii) prove the existence of ASPR-optimal stationary policies under the additional Assumption C. However, this vanishing factor method needs *additional* results about discounted continuous-time MDPs in [17].

When the transition and reward rates are both *uniformly bounded*, we need to impose conditions only on the *embedded* Markov chains to guarantee the existence of SPAR-optimal stationary policies. This is stated in the following corollary.

COROLLARY 4.3. *Suppose the following conditions* (1)–(3) *are satisfied.*

(1) $\|q\| := \sup_{i\in S} q(i) < \infty$, $\|r\| := \sup_{i\in S, a\in A(i)} |r(i,a)| < \infty$.

(2) *For each* $i \in S$, $A(i)$ *is compact; and* $r(i,a)$ *and* $q(j|i,a)$ *are continuous in* $a \in A(i)$ *for each fixed* $i, j \in S$.

(3) *Either* $\inf_{i\neq j_0, a\in A(i)} q(j_0|i,a) > 0$ *for some* $j_0 \in S$; *or* $\sum_{j\in S} \sup_{i\in S, a\in A(i)} (\frac{q(j|i,a)}{\|q\|} + \delta_{ij}) < 2$.

Then, the following results hold.

(a) *There exists an ASPR-optimal stationary policy.*

(b) *For each* $\epsilon > 0$, *an* ϵ*-ASPR-optimal stationary policy can be obtained in a finite number of steps of the policy iteration algorithm* 3.1.

Proof. Define maps T_k on the set $M(S)$ of *bounded functions* on S as

$$T_k u(i) := \sup_{a\in A(i)} \left\{ \frac{r(i,a)}{\|q\|+1} + \sum_{j\in S} \left[\left(\frac{q(j|i,a)}{\|q\|+1} + \delta_{ij} \right) - \mu_k(j) \right] u(j) \right\} \tag{4.25}$$

for all $i \in S$, $u \in M(S)$, and $k = 1, 2$, where the measures μ_k on S are given by

$$\mu_1(j) := \inf_{i\in S, a\in A(i)} \left[\frac{q(j|i,a)}{\|q\|+1} + \delta_{ij} \right] \text{ and}$$

$$\mu_2(j) := \sup_{i\in S, a\in A(i)} \left[\frac{q(j|i,a)}{\|q\|+1} + \delta_{ij} \right] \text{ for } j \in S,$$

which correspond to the first and second hypotheses in the condition (3), respectively. Thus, the maps T_1 and T_2 are both contractive with contraction factors β_1 and β_2, respectively, where

$$\beta_1 := 1 - \mu_1(S) \in (0,1) \text{ and } \beta_2 := \mu_2(S) - 1 \in (0,1). \tag{4.26}$$

Hence, the Banach's fixed point theorem gives the existence of $u^* \in M(S)$, $f^* \in F$ and a unique constant g^* satisfying (4.1). Then, as in the proof of Theorem 4.1(a) and (d), we see that Corollary 4.3 is true. □

Remark 4.4. (a) The two sets in the condition (3) in Corollary 4.3 are variants of the ergodicity condition in [22] for discrete-time MDPs, and each set implies that the embedded chain with the transition probability $(\frac{q(j|i,f(i))}{1+\|q\|} + \delta_{ij})$ has a unique invariant probability measure; see p. 56 in [22], for instance. The difference between the "monotonicity" condition in Assumption A(4) and the condition (3) in Corollary 4.3 can be shown by examples.

(b) Corollary 4.3 can also be obtained by using the uniformization method in [29, 31, 36] and the equivalence between continuous- and discrete-time MDPs in [31, 36, 37], as well as the results for discrete-time MDPs in [3, 10, 14, 15, 22, 24, 32, 34], for instance.

5. Algorithms. Following the procedure in the proof of Theorem 4.1, we now provide a policy iteration algorithm to obtain ASPR-optimal stationary policies.

PROPOSITION 5.1. *Suppose that Assumptions* A, B, *and* C *hold. Then any limit point* f^* *of the sequence* $\{f_n\}$ *obtained by the policy iteration Algorithm* 3.1 *is ASPR-optimal.*

Proof. The proposition follows directly from the proof of Theorem 4.1. □

Under the conditions in Corollary 4.3, we provide a *value iteration algorithm* to compute $\epsilon(>0)$-ASPR-optimal stationary policies. It should be mentioned that, as in the proof of Corollary 4.3, the choice of $k = 1$ (or 2) corresponds to the first (or second) hypothesis in the condition (3) in Corollary 4.3. Thus, we will understand that k in this algorithm is *fixed*.

VALUE ITERATION ALGORITHM 5.1.

Step I. For a fixed $\epsilon > 0$, take arbitrarily $u_0 \in M(S)$.

Step II. If $T_k u_0 = u_0$, then obtain a policy f (in F) satisfying

$$r(i, f(i)) + \sum_{j\in S} q(j|i, f(i))u_0(j) = \sup_{a\in A(i)} \left\{ r(i,a) + \sum_{j\in S} q(j|i,a)u_0(j) \right\} \quad \forall i \in S,$$

and f is ASPR-optimal (by Theorem 4.1), stop; otherwise, calculate a positive integer $N \geq \frac{1}{\beta_k} \ln \frac{\epsilon(1-\beta_k)}{4(1+\|q\|)\|u_1-u_0\|} + 1$ with β_k as in (4.26), and $u_N := T_k^N u_0 = T_k(T_k^{N-1}u_0)$ (by (4.25)).

Step III. Choose $f_\epsilon(i) \in A(i)$ such that for each $i \in S$

$$r(i, f_\epsilon(i)) + \sum_{j\in S} q(j|i, f_\epsilon(i))u_N(j) \geq \sup_{a\in A(i)} \left\{ r(i,a) + \sum_{j\in S} q(j|i,a)u_N(j) \right\} - \frac{\epsilon}{2}.$$

Then we have the following facts.

PROPOSITION 5.2. *Under the conditions in Corollary* 4.3*, the policy f_ϵ obtained by the value iteration algorithm* 5.1 *is ϵ-ASPR-optimal.*

For the policy iteration algorithm 3.1, if we luckily choose an initial policy such that the algorithm 3.1 stops after a *finite* number of iterations, then Proposition 5.1 shows that an ASPR-optimal stationary policy can be computed. Otherwise, since the policy space F may be infinite, the algorithm 3.1 may not stop in any finite number of iterations. In this case, Proposition 5.1 shows that an ASPR-optimal stationary policy can be approximated. On the other hand, Proposition 5.2 implies that under the conditions in Corollary 4.3 an ϵ-ASPR-optimal stationary policy can indeed be computed in a finite number of iterations, where $\epsilon > 0$.

6. Examples. In this section, we illustrate our conditions and show the difference between the ASPR and AER criteria with examples.

Example 6.1 (a controlled birth-death system). Consider a controlled birth-death system in which the state variable denotes the population size at any time $t \geq 0$. There are "natural" birth and death rates denoted by *positive* constants λ and μ, respectively, as well as *nonnegative* emigration and immigration parameters. The two parameters are assumed to be controlled by a decision-maker and denoted by $h_1(i, a_1)$ and $h_2(i, a_2)$, respectively, which may depend on system's state i and decision variables a_1 and a_2 taken by the decision-maker. When the system is at state $i \in S := \{0, 1, \ldots\}$, the decision-maker takes an action $a := (a_1, a_2)$ from a *compact* set $A(i) =: A_1(i) \times A_2(i)$ of available actions, which increases/decreases the emigration parameter $h_1(i, a_1)$ and may incur a cost with rate $c(i, a_1)$, and also increases/decreases the immigration parameter $h_2(i, a_2)$ and gives a reward with rate $\bar{r}(i, a_2)$. Moreover, suppose that the benefit rate caused by a population is represented by $p > 0$. Then the *net* income rate in this system is $r(i, a) := pi + \bar{r}(i, a_2) - c(i, a_1)$ for each $i \in S$ and $a = (a_1, a_2) \in A(i)$. On the other hand, when there is no population in the system (i.e., $i = 0$), it is impossible to decrease/increase the emigration rate,

and so we have $h_1(0, a_1) \equiv 0$ for all $a_1 \in A_1(0)$. Also, in this case (i.e., $i = 0$) we may assume that the decision-maker hopes to increase the immigration rate, and then $h_2(0, a_2) > 0$ for all $a_2 \in A_2(0)$. (This assumption guarantees the irreducibility condition in Assumption A(4).)

We now formulate this system as a continuous-time Markov decision process. The corresponding transition rates $q(j|i, a)$ and reward rates $r(i, a)$ are given as follows.

For $i = 0$ and each $a = (a_1, a_2) \in A(0)$

$$q(1|0, a) = -q(0|0, a) := h_2(0, a_2) > 0,$$

and for $i \geq 1$ and all $a = (a_1, a_2) \in A(i)$

$$q(j|i, a) := \begin{cases} \mu i + h_1(i, a_1) & \text{if } j = i - 1, \\ -(\mu + \lambda)i - h_1(i, a_1) - h_2(i, a_2) & \text{if } j = i, \\ \lambda i + h_2(i, a_2) & \text{if } j = i + 1, \\ 0 & \text{otherwise;} \end{cases} \tag{6.1}$$

$$r(i, a) := pi + \bar{r}(i, a_2) - c(i, a_1) \quad \text{for } i \in S \text{ and } a = (a_1, a_2) \in A(i). \tag{6.2}$$

We aim to find conditions that ensure the existence of an ASPR-optimal stationary policy. To do this, in the spirit of Assumptions A, B, and C we consider the following conditions:

(E$_1$) (a) $\mu - \lambda > 0$. (b) Either $\kappa := \mu - \lambda + h_2^* - h_{1*} \leq 0$, or $\mu - \lambda > |h_2^* - h_{1*}|$ when $\kappa > 0$, where $h_2^* := \sup_{a_2 \in A_2(i), i \geq 1} h_2(i, a_2)$, $h_{1*} := \inf_{a_1 \in A_1(i), i \geq 1} h_1(i, a_1)$.

(E$_2$) For each fixed $i \in S$, the functions $h_1(i, \cdot)$, $h_2(i, \cdot)$, $c(i, \cdot)$, and $\bar{r}(i, \cdot)$ are all continuous.

(E$_3$) (a) There exist positive constants $L_k (k = 1, 2)$ such that $|c(i, a_1)| \leq L_1(i+1)$ and $|\bar{r}(i, a_2)| \leq L_2(i + 1)$ for all $i \in S$ and $(a_1, a_2) \in A_1(i) \times A_2(i)$. (b) $\|h_k\| := \sup_{i \in S, a_k \in A_k(i)} |h_k(i, a_k)| < \infty$, for $k = 1, 2$.

To further explain Example 6.1, we consider the *special* case of *birth-death processes with controlled immigration*. Consider a pest population in a region which may be isolated to prevent immigration. Let c denote the cost rate when immigration is always prevented, b denote the immigration rate without any control, and action $a \in [0, 1]$ denote the *level* of immigration prevented, where c and b are *fixed positive* constants. When the population size is $i \in S := \{0, 1, \ldots\}$, an action a from a set $A(i)$ consisting of available actions is taken. Then a cost rate ca is incurred, the immigration rate $(1 - a)b$ is permitted, and the evolution of the population depends on birth, death, and immigration with parameters λ, μ, and $(1 - a)b$, respectively, where λ and μ are given *positive constants*. Suppose that the damage rate caused by the pest is represented by $p > 0$. Then the reward rate is of the form $r(i, a) := -pi - ca$ for each $i \in S$ and $a \in A(i)$. Obviously, we have $A(i) := [0, 1]$ for each $i \geq 1$. However, when there is no pest in the region (i.e., $i = 0$), to guarantee the irreducibility condition in Assumption A(4) we need that $A(0) := [0, \beta]$ with a given $\beta \in (0, 1)$. (This, however, can be explained as follows: For the ecological balance of the region, the pest is not permitted to become extinct, and so the immigration rate $(1 - \beta)b > 0$ is left.) Using the notation in Example 6.1, for this model we have $h_1 \equiv 0$ and $h_2(i, a_2) = (1 - a)b$ with $a_2 := a$ here. Hence, when $\mu - \lambda > b$, the conditions E$_1$, E$_2$, and E$_3$ above are all satisfied.

Under E$_1$, E$_2$, and E$_3$, we obtain the following.

PROPOSITION 6.2. *Under conditions* E$_1$, E$_2$, *and* E$_3$, *the above controlled birth-death system satisfies the Assumptions* A, B, *and* C. *Therefore (by Theorem* 4.1*),*

there exists an ASPR-optimal stationary policy, which can be computed or at least approximated by the policy iteration algorithm 3.1.

Proof. We shall first verify Assumption A. Let $S_n := \{0, 1, \dots, n\}$ for each $n \geq 1$, $w(i) := i + 1$ for all $i \in S$, and

$$\rho := \frac{\mu - \lambda - h_2^* + h_{1*}}{2} = \mu - \lambda - \frac{\kappa}{2} > 0 \quad \text{when } \mu - \lambda > |h_2^* - h_{1*}|.$$

Then Assumptions A(1) and A(2) are obviously true. Moreover, for each $a = (a_1, a_2) \in A(i)$ with $i \geq 1$, by condition E_1 and (6.1), we have

$$\begin{aligned} \sum_{j \in S} q(j|i,a)w(j) &= (\lambda - \mu)(i+1) + \mu - \lambda - h_1(i,a_1) + h_2(i,a_2) \\ &\leq -(\mu - \lambda)w(i) + \kappa \\ &\leq \begin{cases} -(\mu - \lambda)w(i) & \text{when } \kappa \leq 0, \\ -\rho w(i) & \text{when } \kappa > 0 \quad (\text{and so } \rho > 0). \end{cases} \end{aligned} \tag{6.3}$$

In particular, for $i = 0$ and each $a = (a_1, a_2) \in A(0)$, we have

$$\sum_{j \in S} q(j|i,a)w(j) = h_2(0,a_2) \leq -(\mu - \lambda)w(0) + b' = -\rho w(0) + b' - \frac{\kappa}{2}, \tag{6.4}$$

where $b' := \mu - \lambda + \|h_2\| > 0$.

By the inequalities (6.3) and (6.4) we see that Assumption A(3) holds with $c := \mu - \lambda$ and $b := b'$ when $\kappa \leq 0$, or $c := \rho$ and $b := b'$ when $\kappa > 0$. Since $h_2(0, a_2) > 0$ for all $a_2 \in A_2(0)$, by (6.1) we see that Assumption A(4) is true. Hence Assumption A follows.

By E_3 and (6.2), we have $|r(i,a)| \leq pi + L_1(i+1) + L_2(i+1) \leq (p + L_1 + L_2)w(i)$ for all $i \in S$ and $a \in A(i)$, which verifies Assumption B(4). Hence, Assumption B is satisfied because Assumptions B(1), B(2), and B(3) follow from E_2 and the model's description.

Finally, to verify Assumption C we let

$$w_1^*(i) := i^2 + 1, \quad w_2^*(i) := i^4 + 1 \quad \forall i \in S. \tag{6.5}$$

Then

$$w^2(i) \leq M_1^* w_1^*(i), \quad [q(i)w(i)]^2 \leq M_2^* w_2^*(i) \quad \forall i \in S, \tag{6.6}$$

with $M_1^* := 3$ and $M_2^* := 8(\lambda + \mu + \|h_1\| + \|h_2\|)$.

Moreover, for each $i \geq 1$ and $a = (a_1, a_2) \in A(i)$, by (6.1), (6.5), and E_3, we have

$$\begin{aligned} \sum_{j \in S} q(j|i,a)w_1^*(j) &= -2i[\mu i + h_1(i,a_1)] + \mu i + h_1(i,a_1) \\ &\quad + 2i[\lambda i + h_2(i,a_2)] + \lambda i + h_2(i,a_2) \\ &\leq -2(\mu - \lambda)(i^2 + 1) + 3(\mu + \lambda + \|h_1\| + \|h_2\|)i. \end{aligned}$$

Hence, for each $i \geq \frac{3(\mu+\lambda+\|h_1\|+\|h_2\|)}{\mu-\lambda} + 1 =: i_*$, we have

$$\sum_{j \in S} q(j|i,a)w_1^*(j) \leq -(\mu - \lambda)w_1^*(i). \tag{6.7}$$

On the other hand, since $A(i)$ is assumed to be compact for each $i \in S$, by (6.1) and (6.5) we see that $\sum_{j\in S} q(j|i,a)w_1^*(j)$ and $(\mu-\lambda)w_1^*(i)$ are both bounded in $a \in A(i)$ and $i \le i_*$. Thus, from (6.7) there exists a positive constant b_1^* such that

$$\sum_{j\in S} q(j|i,a)w_1^*(j) \le -(\mu-\lambda)w_1^*(i) + b_1^* \quad \forall i \in S \text{ and } a \in A(i). \tag{6.8}$$

Also, for each $i \ge 1$ and $a \in A(i)$, by (6.1) and (6.5) we have

$$\sum_{j\in S} q(j|i,a)w_2^*(j) \le -2(\mu-\lambda)(i^4+1) - (\mu-\lambda)i^4 + c_3 i^3 + c_2 i^2 + c_1 i + c_0, \tag{6.9}$$

where the constants $c_k (k = 0, 1, 2, 3)$ are determined completely by λ, μ, $\|h_1\|$, and $\|h_2\|$. Similarly, by (6.9) and (6.1), there exists a positive constant b_2^* such that

$$\sum_{j\in S} q(j|i,a)w_2^*(j) \le -(\mu-\lambda)w_2^*(i) + b_2^* \quad \forall i \in S \text{ and } a \in A(i), \tag{6.10}$$

which, together with (6.8) and (6.6), verifies Assumption C. □

It should be noted that in Example 6.1 both the reward and transition rates are *unbounded*; see (6.1) and (6.2). Next, we will show that our admissible policy class Π can indeed be chosen to be larger than the usual stationary policy class F.

Example 6.3. In Example 6.1, for each $i \in S$ we take arbitrarily two actions $a^k(i)$ $(k = 1, 2)$ from $A(i)$ which may depend on i, and then define an admissible policy $\tilde\pi = (\tilde\pi_t)$ as

$$\tilde\pi_t(B|i) = \begin{cases} \dfrac{1}{2}e^{-\rho_0 it} & \text{if } B = \{a^1(i)\}, \\ 1 - \dfrac{1}{2}e^{-\rho_0 it} & \text{if } B = \{a^2(i)\}, \\ 0 & \text{otherwise} \end{cases} \tag{6.11}$$

for some fixed constant $\rho_0 > 0$.

Then, by (6.1), (6.11), and (2.2), we see that $\tilde\pi$ is in Π but *not* in F. Therefore, we have $\Pi \supset F$, but $\Pi \ne F$. It is also noted that the associated Q-process $p(s, i, t, j, \tilde\pi)$ is *nonhomogeneous*, and so is the associated continuous-time Markov chain $x(t, \tilde\pi)$. Moreover, the corresponding reward rates $r(i, \tilde\pi_t)$ are *time-dependent* and *unbounded*; see (6.2) and (2.3).

Finally, in the following example we show that *in general* the AER and ASPR criteria are different.

Example 6.4. Let $S := \{1, 2\}$. For some $\hat\pi = (\hat\pi_t)$, $f \in \Pi$, suppose that for $0 \le t \le 1$,

$$Q(\hat\pi_t) := \begin{pmatrix} -1+t & 1-t \\ 2-2t & -2+2t \end{pmatrix} \quad \text{and} \quad Q(f) := \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}. \tag{6.12}$$

Let $t_0 := 1$, and define

$$Q(\tilde\pi_t) := \begin{cases} Q(\hat\pi_t) & \text{when } 0 \le t \le t_0, \\ Q(f) & \text{when } t \ge t_0. \end{cases} \tag{6.13}$$

By (6.12), (6.13), and Definition 2.2, we see that the associated policy $\tilde{\pi}$ belongs to Π. For reference, we recall that for any $\pi \in \Pi$ the associated regular Q-process $p(s,i,t,j,\pi)$ can be constructed as follows [13, 16, 17]: for $i,j \in S$ and $n \geq 0$, let

$$p_0(s,i,t,j,\pi) := \delta_{ij} e^{-\int_s^t q_i(\pi_y)dy}, \tag{6.14}$$

$$p_{n+1}(s,i,t,j,\pi) := \int_s^t e^{-\int_s^y q_i(\pi_v)dv} \sum_{k\neq i} q(k|i,\pi_y) p_n(y,k,t,j,\pi)dy. \tag{6.15}$$

Then

$$p(s,i,t,j,\pi) = \sum_{n=0}^{\infty} p_n(s,i,t,j,\pi). \tag{6.16}$$

For each $i,j \in S$, by (6.12)–(6.16), $p(0,i,t_0,j,\hat{\pi}) > 0$. Hence, $0 < p(0,i,t_0,2,\hat{\pi}) < 1$. Moreover,

$$p(s,i,t,j,\tilde{\pi}) = \begin{cases} p(s,i,t,j,\hat{\pi}) & \text{when } 0 \leq t \leq t_0, \\ p(s,i,t,j,f) & \text{when } t \geq s \geq t_0. \end{cases} \tag{6.17}$$

Let $r(1,a) = 0$, $r(2,a) = 1$ for all $a \in A(i)$ with $i = 1,2$. Then, by (6.12) and (6.13) we see that states 1 and 2 are absorbing after time t_0. By (6.12), (6.14)–(6.17), we get

$$p(t_0,i,t,i,\tilde{\pi}) = 1 \quad \forall i \in S \text{ and } t \geq t_0. \tag{6.18}$$

Noting that $r(1,\tilde{\pi}_t) = 0$ and $r(2,\tilde{\pi}_t) = 1$ for each $t \geq 0$, by (6.18) and (2.6) we have that for each $i \in S$

$$V_{sp}(\tilde{\pi},i) = 1 \quad \text{for any sample path in } \{x(t,\tilde{\pi}) = 2,\ t \geq t_0\}. \tag{6.19}$$

On the other hand, by the Chapman–Kolmogorov equation and (6.18), we have

$$\begin{aligned} p(0,i,t,2,\tilde{\pi}) &= p(0,i,t_0,1,\tilde{\pi})p(t_0,1,t,2,\tilde{\pi}) + p(0,i,t_0,2,\tilde{\pi})p(t_0,2,t,2,\tilde{\pi}) \\ &= p(0,i,t_0,2,\tilde{\pi}) < 1 \quad \forall t_0 \leq t. \end{aligned} \tag{6.20}$$

Using again $r(1,\tilde{\pi}_t) = 0$ and $r(2,\tilde{\pi}_t) = 1$ for each $t \geq 0$, by (6.20) and (2.7) we get

$$\begin{aligned} \bar{V}(\tilde{\pi},i) &= \limsup_{T\to\infty} \frac{\int_0^T p(0,i,t,2,\tilde{\pi})dt}{T} \\ &= \limsup_{T\to\infty} \frac{\int_{t_0}^T p(0,i,t,2,\tilde{\pi})dt}{T} \\ &= p(0,i,t_0,2,\hat{\pi}) < 1 \quad \forall i \in S, \end{aligned} \tag{6.21}$$

which together with (6.19) and $P_i^{\tilde{\pi}}(\{x(t,\tilde{\pi}) = 2,\ t \geq t_0\}) = p(0,i,t_0,2,\hat{\pi}) > 0$ shows the difference between the ASPR and AER criteria.

7. Concluding remarks. In the previous sections we have studied ASPR optimality for denumerable continuous-time Markov chains determined by possibly unbounded transition rates. Under suitable assumptions we have shown the existence of a solution to the optimality equation and the existence of an ASPR-optimal stationary policy. In addition, we have presented two algorithms to compute, or at least

approximate, the ASPR-optimal stationary policies. Our formulation and approach are sufficiently general and can be used to analyze other important problems, such as the relation among potentials, perturbation analysis, and Markov decision processes in general spaces, as well as minimax control problems. These problems, as far as we can tell, have not been previously studied for continuous-time Markov chains with unbounded transition or reward rates. It should be mentioned that Example 6.4 shows that *in general* the ASPR and AER criteria are different, and it is an interesting and challenging problem to further show the difference between the two criteria under some ergodicity condition. Also, it remains open to show that an ASPR-optimal stationary policy is *not* necessarily canonical. Research on these topics is in progress.

Acknowledgments. The authors are indebted to the anonymous referees for many valuable comments and suggestions that have helped us in improving the presentation.

REFERENCES

[1] E. Altman, *Constrained Markov Decision Processes*, Chapman & Hall/CRC, Boca Raton, FL, 1999.

[2] W.J. Anderson, *Continuous-Time Markov Chains*, Springer-Verlag, New York, 1991.

[3] A. Arapostathis, V.S. Borkar, E. Fernández-Gaucherand, M.K. Ghosh, and S.I. Marcus, *Discrete-time controlled Markov processes with average cost criterion: A survey*, SIAM J. Control Optim., 31 (1993), pp. 282–344.

[4] J. Bather, *Optimal stationary policies for denumerable Markov chains in continuous time*, Adv. in Appl. Probab., 8 (1976), pp. 144–158.

[5] J. Bather, *Optimal decision procedures for finite Markov chains.* II. *Communicating systems*, Adv. in Appl. Probab., 5 (1973), pp. 521–540.

[6] V.S. Borkar, *Topics in Controlled Markov Chains*, Pitman Research Notes in Math. 240, Longman Scientific and Technical, Harlow, UK, 1991.

[7] X.-R. Cao, *The relations among potentials, perturbation analysis, and Markov decision processes*, Discrete Event Dyn. Syst., 8 (1998), pp. 71–87.

[8] X.-R. Cao and H.F. Chen, *Potentials, perturbation realization and sensitivity analysis of Markov processes*, IEEE Trans. Automat. Control, 42 (1997), pp. 1382–1397.

[9] X.-R. Cao and X.P. Guo, *A unified approach to Markov decision problems and performance sensitivity analysis with discounted and average criteria: Multichain cases*, Automatica, 40 (2004), pp. 1749–1759.

[10] R. Cavazos-Cadena and E. Fernández-Gaucherand, *Denumerable controlled Markov chains with average reward criterion: Sample path optimality*, ZOR—Math. Methods Oper. Res., 41 (1995), pp. 89–108.

[11] S.P. Coraluppi and S.I. Marcus, *Risk-sensitive, minimax, and mixed risk-neutral/minimax control of Markov decision processes*, in Stochastic Analysis, Control, Optimization and Applications, Systems Control Found. Appl., Birkhäuser Boston, Boston, 1999, pp. 21–40.

[12] E.B. Dynkin and A.A. Yushkevich, *Controlled Markov Processes*, Springer-Verlag, New York, 1979.

[13] W. Feller, *On the integro-differential equations of purely discontinuous Markoff processes*, Trans. Amer. Math. Soc., 48 (1940), pp. 488–515.

[14] L.G. Gubenko and E.S. Statland, *On discrete time Markov decision processes*, Teor. Verojatnost. i Mat. Statist., 7 (1972), pp. 51–64 (in Russian).

[15] X.P. Guo and P. Shi, *Limiting average criteria for nonstationary Markov decision processes*, SIAM J. Optim., 11 (2001), pp. 1037–1053.

[16] X.P. Guo and O. Hernández-Lerma, *Drift and monotonicity conditions for continuous-time controlled Markov chains with an average criterion*, IEEE Trans. Automat. Control, 48 (2003), pp. 236–245.

[17] X.P. Guo and O. Hernández-Lerma, *Continuous-time controlled Markov chains with discounted rewards*, Acta Appl. Math., 79 (2003), pp. 195–216.

[18] X.P. Guo and K. Liu, *A note on optimality conditions for continuous-time Markov decision processes with average cost criterion*, IEEE Trans. Automat. Control, 46 (2001), pp. 1984–1989.

[19] X.P. Guo and W.P. Zhu, *Denumerable state continuous-time Markov decision processes with unbounded cost and transition rates under average criterion*, ANZIAM J., 43 (2002), pp. 541–557.
[20] M. Haviv and M.L. Puterman, *Bias optimality in controlled queueing systems*, J. Appl. Probab., 35 (1998), pp. 136–150.
[21] O. Hernández-Lerma, *Lectures on Continuous-Time Markov Control Processes*, Aportaciones Matemáticas 3, Sociedad Matematica Mexicana, México City, 1994.
[22] O. Hernández-Lerma, *Adaptive Markov Control Processes*, Springer-Verlag, New York, 1989.
[23] O. Hernández-Lerma and J.B. Lasserre, *Further Topics on Discrete-Time Markov Control Processes*, Springer-Verlag, New York, 1999.
[24] O. Hernández-Lerma, O. Vega-Amaya, and G. Carrasco, *Sample-path optimality and variance-minimization of average cost Markov control processes*, SIAM J. Control Optim., 38 (1999), pp. 79–93.
[25] R.A. Howard, *Dynamic Programming and Markov Processes*, MIT Press, Cambridge, MA, 1960.
[26] P. Kakumanu, *Nondiscounted continuous-time Markov decision processes with countable state space*, SIAM J. Control, 10 (1972), pp. 210–220.
[27] M.E. Lewis and M.L. Puterman, *A note on bias optimality in controlled queueing systems*, J. Appl. Probab., 37 (2000), pp. 300–305.
[28] S.A. Lippman, *On dynamic programming with unbounded rewards*, Management Sci., 21 (1974/75), pp. 1225–1233.
[29] S.A. Lippman, *Applying a new device in the optimization of exponential queuing systems*, Operations Res., 23 (1975), pp. 687–710.
[30] R.B. Lund, S.P. Meyn, and R.L. Tweedie, *Computable exponential convergence rates for stochastically ordered Markov processes*, Ann. Appl. Probab., 6 (1996), pp. 218–237.
[31] M.L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Wiley, New York, 1994.
[32] S.M. Ross, *Applied Probability Models with Optimization Applications*, Holden-Day, San Francisco, 1970.
[33] S.M. Ross, *Introduction to Stochastic Dynamic Programming*, Academic Press, New York, 1983.
[34] S.M. Ross, *Non-discounted denumerable Markovian decision models*, Ann. Math. Statist., 39 (1968), pp. 412–423.
[35] L.I. Sennott, *Stochastic Dynamic Programming and the Control of Queueing Systems*, Wiley, New York, 1999.
[36] R. Serfozo, *Optimal control of random walks, birth and death processes, and queues*, Adv. in Appl. Probab., 13 (1981), pp. 61–83.
[37] R. Serfozo, *An equivalence between continuous and discrete time Markov decision processes*, Oper. Res., 27 (1979), pp. 616–620.
[38] A.F. Veinott, *On finding optimal policies in discrete dynamic programming with no discounting*, Ann. Math. Statist., 37 (1966), pp. 1284–1294.
[39] A.A. Yushkevich and E.A. Feinberg, *On homogeneous Markov model with continuous-time and finite or countable state space*, Theory Probab. Appl., 24 (1979), pp. 156–161.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 1, pp. 49–74

L^P-OPTIMAL BOUNDARY CONTROL FOR THE WAVE EQUATION*

M. GUGAT†, G. LEUGERING†, AND G. SKLYAR‡

Abstract. We study problems of boundary controllability with minimal L^p-norm ($p \in [2, \infty]$) for the one-dimensional wave equation, where the state is controlled at both boundaries through Dirichlet or Neumann conditions. The problem is to reach a given terminal state and velocity in a given finite time, while minimizing the L^p-norm of the controls. We give necessary and sufficient conditions for the solvability of this problem. We show as follows how this infinite-dimensional optimization problem can be transformed into a problem which is much simpler: The feasible set of the transformed problem is described by a finite number of simple pointwise equality constraints for the control function in the Dirichlet case while, in the Neumann case, an additional integral equality constraint appears. We provide explicit complete solutions of the problems for all $p \in [2, \infty]$ in the Dirichlet case and solutions for some typical examples in the Neumann case.

Key words. optimal control, boundary control, wave equation, analytic solution, distributed parameter systems, robust optimization, controllability, state constraints, sensitivity, test examples

AMS subject classifications. 49K20, 90C25, 90C31

DOI. 10.1137/S0363012903419212

1. Introduction.

In this paper, we discuss two-sided Dirchlet or Neumann controls for the one-dimensional wave equation for p between 2 and ∞. We consider the problem of exact control; that is, starting from the zero position we want to reach a given terminal state in a given finite time. Our aim is to find control functions with minimal L^p-norm that steer the system to the target. For certain typical cases, we present explicit representations of such optimal control functions in terms of the given target state.

It is well known that, in the L^2-case, the optimal control functions can be characterized as the L^2-norm minimal solutions of a trigonometric moment problem, which has been analyzed in depth (see [4], [19], [15]). For the L^p-case ($p > 2$) there are only a few publications on the subject (see [2], [16], [12], [11], [14], [10]), and even the question of existence of solutions, which is equivalent to the question of L^p-controllability, has not been solved completely.

In the present paper, we give a complete analysis of this problem for the boundary control of the one-dimensional wave equation. The problem can be reduced to the case of the minimal time interval, where controllability is possible. This allows an answer to be given to the question of solvability of the problem of L^p-controllability in terms of the properties of the target states. We use the control function for the minimal time interval to transform the infinite-dimensional problem into a problem, which is much simpler because it has only a finite number of simple pointwise equality constraints with an additional integral equality constraint (see (3.29) below) in the Neumann case. The transformation is based upon the representation of the state as a trigonometric series and on the corresponding description of the feasible set by

*Received by the editors February 26, 2003; accepted for publication (in revised form) September 8, 2004; published electronically June 27, 2005. This work has been supported by DAAD and Polish KBN grant 5 PO3A 030 21.

http://www.siam.org/journals/sicon/44-1/41921.html

†Lehrstuhl für angewandte Mathematik, Martensstr. 3, 91058 Erlangen, Germany (gugat@am.uni-erlangen.de, leugering@am.uni-erlangen.de).

‡University Szczecin, Institute of Mathematics, Wielkopolska 15, 70451 Szczecin, Poland (sklar@sus.univ.szczecin.pl).

a sequence of moment equations on the control time interval that is transformed into a sequence of moment equations on a shorter time interval, namely, the interval corresponding to the time that a characteristic curve needs to travel from one end of the string to the other.

In the Dirichlet case, solutions of our problem of optimal control exist for all time intervals which are at least as long as the time that waves need to travel from one boundary of the system to the other, provided that the functions that describe the target are sufficiently regular. The required regularity is that the initial state and the primitive of the initial velocity are both in the space $L^p(0, L)$. The optimal controls are given explicitly in terms of these functions for all $p \in [2, \infty)$. In general, the L^∞-norm minimal control is not determined uniquely. Hence, in general there is a set of L^∞-norm minimal controls that contains more than one element. This convex set contains a unique element with minimal L^2-norm. In Theorem 2.2, we give this element explicitly.

For Neumann boundary controls, the situation is more complicated. On the time interval that allows a characteristic curve to travel from one boundary of the system interval to the other, in general, controllability is possible up to a constant only, in the sense that instead of the desired target a state can be reached that differs from the desired target by an appropriate constant. If this time interval is enlarged by an arbitrarily small time, L^p-controllability is possible on the elongated time interval if and only if the target functions are sufficiently regular. In this case, the required regularity is that the initial velocity and the derivative of the initial state are both in the space $L^p(0, L)$. We transform the optimal control problem to a simpler problem, where the feasible set is defined by a finite number of constraints. For some cases, we give explicit expressions for the solution of the optimal control problem in terms of the given target state.

The relation between the L^p regularity of the controls and the data in the Neumann and Dirichlet cases is consistent with what is known in the classical L^2 theory. The form of the optimal solutions depends on the relation between the length of the time interval and the time that the waves need to travel from one boundary of the system to the other. The structure of the optimal control functions is, in general, quite complicated. The optimal controls usually do not show bang-bang behavior; this is also true in the L^∞-case. This may be surprising, since the optimal solutions of discretized problems often are of bang-bang type. The difference between the structure of the controls that solve the PDE-constrained optimization problem and controls that are solutions of discretized problems has been the subject of recent research; see [9], [21], and the references therein.

This paper is also a contribution to robust optimal control: We have found controls whose optimality is robust with respect to perturbations of the objective function. For symmetric targets in the case of Dirichlet boundary controls and for antisymmetric target states in the case of Neumann boundary controls, the controls that are optimal for L^p $(p > 2)$ are exactly the controls that are optimal with respect to the L^2-norm and sufficiently regular to be contained in the space L^p.

Our results are also interesting from the point of view of sensitivity analysis (see [3]), since they allow us to analyze how the solutions of the optimization problems depend on p, the target state, and the time interval.

The explicit solutions of problems of optimal control that we present provide valuable test examples for numerical methods.

In the linear case, L^p-boundary controls have been considered in Krabs and

Leugering [16], where $W^{1,p}$ and also L^p-Dirichlet controls for $p \in [2, \infty]$ are applied at one of the boundary points. Optimal control problems for first-order-in-time equations with distributed L^p-controls have been considered in Fabre, Puel, and Zuazua [5]. See also Glowinski and Lions [7], [8].

Our work is related to existing controllability results for nonlinear problems; see [6] for approximate controllability results for semilinear parabolic equations with L^p-interior or boundary controls and [20] for controllability results for a semilinear wave equation for a class of nonlinearities that grow superlinearly at infinity.

We hope that our results will prove helpful for future analysis of optimal control problems with nonlinear systems.

This paper is organized as follows. Section 2 considers the problem of Dirichlet boundary control. We define the exact optimal boundary control problem, give an exact controllability result (Theorem 2.1), and state Theorem 2.2, where boundary controls that solve the optimal control problem are given in terms of the target state. This result allows a sensitivity analysis for the optimal control problem that is the subject of the next section. Then examples for the solutions presented in Theorem 2.2 are given. For the proof of the results, a series representation of weak solutions of the initial value problem is given and the moment problem describing the successful controls is defined. Then the minimal time interval, where controllability holds, is studied (see Lemma 2.6). Longer time intervals are studied section 2.10. The corresponding moment equations are transformed to moment equations on the minimal time interval. This transformation allows us to prove Theorem 2.1. Also by transformation to a problem on the minimal time interval, Theorem 2.2 is proved. At the end of section 2, the cases of symmetric and antisymmetric targets are discussed.

In section 3, we consider the problem of Neumann boundary control. The optimal control problem is defined and an exact controllability result (Theorem 3.1) is given. A series representation of the weak solution of the initial value problem and the moment problem that describes the successful controls is stated. Exact controllability up to a constant is proved for the minimal time interval (see Lemma 3.2). For larger time intervals, exact controllability is proved. Theorem 3.4 on the solutions of the optimization problem for a certain range of control times is stated. Finally, the cases of symmetric and antisymmetric targets are considered.

2. Dirichlet boundary control. In this section we present a complete solution of the problem of L^p-norm minimal Dirichlet boundary control of our system.

2.1. The initial-value problem. Let an interval $[0, L]$, a time interval $[0, T]$, and a wave speed $c > 0$ be given. We consider the initial-value problem for the wave equation

$$y_{tt}(x,t) = c^2 y_{xx}(x,t), \quad (x,t) \in [0,L] \times [0,T], \tag{2.1}$$

subject to the initial conditions

$$y(x,0) = 0, \quad y_t(x,0) = 0, \quad x \in [0,L], \tag{2.2}$$

and the Dirichlet boundary conditions

$$y(0,t) = f_1(t), \quad y(L,t) = f_2(t), \quad t \in [0,T]. \tag{2.3}$$

For the description of the desired target state, we add the following end conditions:

$$y(x,T) = y_0(x), \quad y_t(x,T) = y_1(x), \quad x \in [0,L]. \tag{2.4}$$

The function y_0 is in the Hilbert space $L^2(0,L)$ of square integrable functions on the interval $(0,L)$, and the function y_1 is integrable such that $Y_1(x) = \int_0^x y_1(z)\,dz$ is in $L^2(0,L)$, so y_1 is contained in the corresponding Sobolev space $W^{-1,2}(0,L)$.

2.2. The optimization problem. For a fixed time $T > 0$ and a given value of $p \in [2,\infty)$, we consider the following optimization problem:

$$C(p):\ \inf\ \|f_1\|^p_{p,(0,T)} + \|f_2\|^p_{p,(0,T)} \text{ s.t. } f_1, f_2 \in L^p(0,T)$$

and the solution y of the initial boundary-value problem (2.1)–(2.3) satisfies the end conditions (2.4).

In the case $p = \infty$, our optimization problem is the following:

$$C(\infty):\ \inf\ \max\{\|f_1\|_{\infty,(0,T)},\ \|f_2\|_{\infty,(0,T)}\} \text{ s.t. } f_1, f_2 \in L^\infty(0,T)$$

and the solution y of (2.1)–(2.3) satisfies the end conditions (2.4).

Here, for $p \in [2,\infty)$, we use the norm

$$\|f\|_{p,(0,T)} = \left(\int_0^T |f(t)|^p\,dt\right)^{1/p},$$

and for $p = \infty$ we use the norm $\|f\|_{\infty,(0,T)} = \text{ess sup}\{|f(t)| :\ t \in (0,T)\}$.

2.3. Exact controllability.

THEOREM 2.1. *Let $p \in [2,\infty]$ and $T \geq L/c$ be given. The initial boundary-value problem* (2.1)–(2.3) *has a weak solution that satisfies the end conditions* (2.4) *with f_1, $f_2 \in L^p(0,T)$ if and only if the target states y_0, y_1 satisfy the following conditions: $y_0 \in L^p(0,L)$ and $Y_1 \in L^p(0,L)$, where $Y_1(x) = \int_0^x y_1(z)\,dz$, that is, $y_1 \in W^{-1,p}(0,L)$. This implies that problem $C(p)$ is solvable if and only if y_0 and Y_1 are in $L^p(0,L)$.*

A standard method for proving an exact controllability result of this type is to reduce the exact controllability problem to a moment problem and to prove the solvability of the moment problem using Ingham's classical inequalities (see [13]) or its generalizations (see [18], [16]). For the cases $p = 2$ and $p = \infty$, Ingham's results provide an alternative proof of Theorem 2.1. In this paper, we use a different approach: We give a solution of the moment problem explicitly (see section 2.9).

We expect that Theorem 2.1 holds for all $p \geq 1$, but for the case $1 \leq p < 2$ a different method of proof should be used.

In the next section we will state our main result, which gives the solution of problem $C(p)$ in terms of two functions that depend on the target states.

2.4. Solution of the optimal control problem. In this section we present optimal control functions that solve problem $C(p)$. For $p < \infty$, the solution is unique, and for $p = \infty$, we present one element on the set of L^∞-norm minimal controls, namely, the element of this convex set with minimal L^2-norm.

THEOREM 2.2. *Let $p \in [2,\infty]$ and a time $T \geq L/c$ be given. Choose a natural number k such that $kL/c \leq T < (k+1)L/c$. Define the function*

$$Y_1(x) = \int_0^x y_1(t)\,dt.$$

Assume that y_0, $Y_1 \in L^p(0,L)$ and define the functions g_1, g_2 in $L^p(0,L/c)$ by

$$\begin{aligned} g_1(t) &= \ \ y_0(ct)/2 \quad - (1/(2c))\ Y_1(ct), \\ g_2(t) &= y_0(L-ct)/2 + (1/(2c))\ Y_1(L-ct). \end{aligned}$$

If $p < \infty$, let $\hat{r}$ denote the real number that minimizes the function

$$h_p(r) = \left[\int_0^{T-kL/c} \frac{1}{(k+1)^{p-1}} \left[\left| g_1(t) + r \right|^p + \left| g_2(t) - r \right|^p \right] dt \right.$$

$$\left. + \int_{T-kL/c}^{L/c} \frac{1}{k^{p-1}} \left[\left| g_1(t) + r \right|^p + \left| g_2(t) - r \right|^p \right] dt \right],$$

while, if $p = \infty$, let $\hat{r}$ be the real number that minimizes

$$h_\infty(r) = \max \left[\|(g_1(t) + r)/(k+1)\|_{\infty,(0,T-kL/c)}, \|(g_2(t) - r)/(k+1)\|_{\infty,(0,T-kL/c)}, \right.$$

$$\left. \|(g_1(t) + r)/k\|_{\infty,(T-kL/c,L/c)}, \|(g_2(t) - r)/k\|_{\infty,(T-kL/c,L/c)} \right].$$

For $j \in \{0, \ldots, k\}$, define the intervals

$$I_j^1 = [jL/c,\, T - k(L/c) + jL/c],$$

and for $j \in \{0, \ldots, k-1\}$, define the intervals

$$I_j^2 = [T - k(L/c) + jL/c,\, (j+1)L/c].$$

For natural numbers j and n let $g_{b(n+j)} = g_1$ if $n + j$ is odd and $g_{b(n+j)} = g_2$ if $n + j$ is even. Then a solution of problem $C(p)$ is given by the pair of control functions $(f_1,\, f_2)$ defined as follows:

$$f_n(T-t) = \frac{(-1)^j\, g_{b(n+j)}(t - jL/c) - (-1)^n\, \hat{r}}{k+1} \tag{2.5}$$

for $n \in \{1, 2\}$, $j \in \{0, \ldots, k\}$, $t \in I_j^1$, and

$$f_n(T-t) = \frac{(-1)^j\, g_{b(n+j)}(t - jL/c) - (-1)^n \hat{r}}{k} \tag{2.6}$$

for $n \in \{1, 2\}$, $j \in \{0, \ldots, k-1\}$, $t \in I_j^2$.

If $p < \infty$, this is the unique solution of problem $C(p)$. If $p = \infty$, this is a solution of $C(\infty)$, namely, the element of the set of solutions that has the smallest L^2-norm.

For certain interesting target states, the value of $\hat{r}$ can be computed explicitly. If y_0 and y_1 are symmetric, $\hat{r} = Y_1(L)/(4c)$. In particular, in this case the value of $\hat{r}$ is independent of p. This implies that the solution is also independent of p. Hence for a symmetric target state with $y_0,\, Y_1 \in L^p(0, L)$, our optimal control that solves problem $C(p)$ also solves problem $C(q)$ for all $q \in [2, p]$.

Later we will characterize the feasible controls $(f_1,\, f_2)$ that steer the system to the desired target state as the solutions of a trigonometric moment problem. To do this, we need a series representation of the solution of the initial boundary-value problem that we obtain from the weak form of the problem.

Then we show that the set of successful controls can be described by a set of two equations for each $t \in [0, L/c]$. This leads to a family of optimization problems, with parameter $t \in [0, L/c]$, whose solutions are coupled by a constant r. The solutions of these optimization problems yield the values of the optimal controls up to the constant r, and thus we obtain a parametric family of successful controls with parameter r. Inserting the elements of this family into the objective function yields the values $h_p(r)$. The optimal control is the element of this family of controls for which the value $h_p(r)$ is minimal.

2.5. Sensitivity analysis for the optimal control problem. The explicit solutions that we have obtained allow a detailed study of their sensitivity with respect to data perturbations, which is useful for obtaining some idea about what might hold in the general case of optimal control problems with hyperbolic PDEs. Here we study only the continuity of the solutions of $C(p)$ as functions of the parameter p.

LEMMA 2.3. *The number $\hat{r}(p)$ that minimizes the function h_p depends continuously on p. In fact, if y_0 and Y_1 are in $L^q(0,L)$ for some $q \in [2,\infty]$, we have $\lim_{p\to q-} \hat{r}(p) = \hat{r}(q)$ and for $p_1 < q$ we have $\lim_{p_2\to p_1} \hat{r}(p_2) = \hat{r}(p_1)$.*

Proof. Case 1: If $h_q(\hat{r}(q)) = 0$, we have $\hat{r}(p) = \hat{r}(q)$ for all $p < q$. Case 2: Assume that $q < \infty$ and $h_q(\hat{r}(q)) > 0$. Then for all $p \in [2,q]$, $h_p(\hat{r}(p)) > 0$ and $h_p''(\hat{r}(p)) > 0$. Consider the function $F : [2,q] \times R \to R$, $F(p,r) = h_p'(r)$. Then for all $p \in [2,q]$, $F(p,\hat{r}(p)) = 0$ and $\partial_r F(p,\hat{r}(p)) = h_p''(\hat{r}(p)) > 0$. Hence the implicit function theorem implies that the function $\hat{r}$ is continuously differentiable on $[2,q]$. Case 3: $q = \infty$. Let $f_1(r)$, $f_2(r)$ denote the control functions defined in Theorem 2.2 that correspond to $r \in R$. Then for all $p \in [2,\infty)$ we have $h_p(r) = \|f_1(r)\|_{p,(0,T)}^p + \|f_2(r)\|_{p,(0,T)}^p$ and $h_\infty(r) = \max\{\|f_1(r)\|_{\infty,(0,T)} + \|f_2(r)\|_{\infty,(0,T)}\}$. Thus for all r, $\lim_{p\to\infty} h_p(r)^{1/p} = h_\infty(r)$. Moreover, the triangle inequality for the p-norm implies that for all $p \in [2,\infty)$, r_1, $r_2 \in R$, we have

(2.7)
$$|h_p(r_1)^{1/p} - h_p(r_2)^{1/p}| \le (\|f_1(r_1) - f_1(r_2)\|_{p,(0,T)}^p + \|f_2(r_1) - f_2(r_2)\|_{p,(0,T)}^p)^{1/p}.$$

For all $p \in [2,\infty]$ we have $h_p(\hat{r}(p))^{1/p} \le h_p(\hat{r}(\infty))^{1/p}$ and $\lim_{p\to\infty} h_p(\hat{r}(\infty))^{1/p} = h_\infty(\hat{r}(\infty))$, and hence the set $\{h_p(\hat{r}(p))^{1/p},\, p \in [2,\infty]\}$ is bounded. With the definition of h_p, this implies that the set $\{\hat{r}(p),\, p \in [2,\infty]\}$ is also bounded. Suppose that a sequence (p_k) converging to ∞ with $p_k \in [2,\infty)$ for all k is given and $\lim_k \hat{r}(p_k) = r_*$. Using (2.7) it can be shown that $h_\infty(r_*) = \lim_{p\to\infty} h_p(r_*)^{1/p} \le \limsup_{p\to\infty} h_p(\hat{r}(p))^{1/p} \le h_\infty(\hat{r}(\infty))$. Thus $h_\infty(r_*) \le h_\infty(\hat{r}(\infty))$. Since $\hat{r}(\infty)$ is the minimizer of h_∞ this implies that $h_\infty(r_*) = h_\infty(\hat{r}(\infty))$, and since the minimizer of h_∞ is determined uniquely, this implies that $r_* = \hat{r}(\infty)$, and the assertion follows. ∎

Lemma 2.3 and Theorem 2.2 imply the following proposition.

PROPOSITION 2.4. *Let $p \in [2,\infty]$ be given. Assume that y_0 and Y_1 are in $L^p(0,L)$. Consider a sequence $(q_k)_k$ ($q_k \le p$) that converges to $q_0 \le p$. Then for the solutions $(f_{1,k}, f_{2,k})^T$ of the optimization problems $C(q_k)$ presented in Theorem 2.2, we have*

$$\lim_{k\to\infty} \|f_{1,k} - f_{1,0}\|_{p,(0,T)} + \|f_{2,k} - f_{2,0}\|_{p,(0,T)} = 0,$$

where $(f_{1,0}, f_{2,0})$ is the solution of $C(q_0)$ presented in Theorem 2.2.

2.6. Examples. For our examples, let $L = 1$, $c = 1$, and $T = 3.25$, and hence $k = 3$.

2.6.1. Example 1. Let $y_0(x) = x - L/2$, $y_1(x) = 1$. For $p = \infty$, the optimal $\hat{r}$ is $5/28$ and we have $h_\infty(5/28) = 1/7$. Figure 2.1(a) shows the optimal controls. The thick lines show f_1 and the dotted line shows f_2. A plot of the corresponding optimal state y in the interior of the rectangle $[0,L] \times [0,T]$ is shown in Figure 2.1(b). Here the optimal state is piecewise linear and the optimal velocity is piecewise constant on areas that are bounded by characteristic curves in the interior of the rectangle $[0,L] \times [0,T]$.

2.6.2. Example 2. The desired state is $y_0(x) = x - L/2$, $y_1(x) = 0$. For $p = \infty$, the optimal $\hat{r}$ equals $-1/28$. For $p = 2$, the optimal $\hat{r}$ is $-1/80$. Figure 2.2(a) shows the optimal controls for $p = \infty$. The thick lines show f_1, and the dotted line shows f_2. Figure 2.2(b) is a plot of the corresponding state y in the interior of the rectangle $[0, L] \times [0, T]$. Due to the form of the target, the optimal state is piecewise linear and the optimal velocity is piecewise constant.

2.6.3. Example 3. In [21], [6] it is pointed out that the optimal boundary controls can also be determined as boundary traces of solutions of adjoint optimal control problems. This example illustrates that for $p \in [2, \infty)$, this approach yields the same controls as Theorem 2.1.

Consider the following optimal control problem:

$$\text{(A)} \qquad \inf \|v\|_{2,(0,2)} \text{ s.t. } y_{tt}(x,t) = y_{xx}(x,t),\ y(0,t) = 0,\ y(1,t) = v(t),$$

$y(x,0) = y^0(x),\ y_t(x,0) = 0,\ y(x,T) = y_t(x,T) = 0,\ (x,t) \in (0,1) \times (0,2)$.

Let $y^0 \in L^2(0,1)$ be continued to the interval $(-1,1)$ as the antisymmetric function y^0_a, that is, $y^0_a(x) = y^0(x)$ for $x \in (0,1)$, $y^0_a(x) = -y^0(-x)$ for $x \in (-1,0)$. Then problem (A) has the same solutions as problem $C(2)$ with $c = 1$, $L = T = 2$, $y_0(x) = y^0_a(x+1)$, $y_1(x) = 0$ in the sense that the optimal controls satisfy $f_2(t) = v(T-t) = -f_1(t)$. The adjoint optimization problem presented in [21] is

$$\inf \frac{1}{2} \int_0^2 |u_x(1,t)|^2\, dt + \int_0^1 y^0(x)u^1(x) - y^1(x)u^0(x)\, dx \text{ s.t. } u_{tt}(x,t) = u_{xx}(x,t),$$

$u(0,t) = u(1,t) = 0,\ u(x,0) = u^0(x),\ u_t(x,0) = u^1(x), (x,t) \in (0,1) \times (0,2)$.

The solution of this adjoint optimization problem is $u^0(x) = \text{const}$, $u^1(x) = -y^0(x)/2$, which yields the optimal control $v(t) = u_x(1,t) = y^0(1-t)/2$ for $t \in (0,1)$, $v(t) = u_x(1,t) = -y^0(t-1)/2$ for $t \in (1,2)$.

Theorem 2.2 yields exactly the same solutions. With $k = 1$, $g_1(t) = y^0_a(t-1)/2$, $g_2(t) = y^0_a(1-t)/2$, we have $\hat{r} = 0$, and hence for $t \in (0,1) = I^2_0$ we have $f_2(T-t) = g_2(t)$ and for $t \in (1,2) = I^2_1$ we have $f_2(T-t) = -g_1(t)$.

2.6.4. Example 4. For $p = \infty$, consider the following optimal control problem:

$$\text{(B)} \qquad \inf \|v\|_{\infty,(0,2)} \text{ s.t. } y_{tt}(x,t) = y_{xx}(x,t),\ y(0,t) = 0,\ y(1,t) = v(t),$$

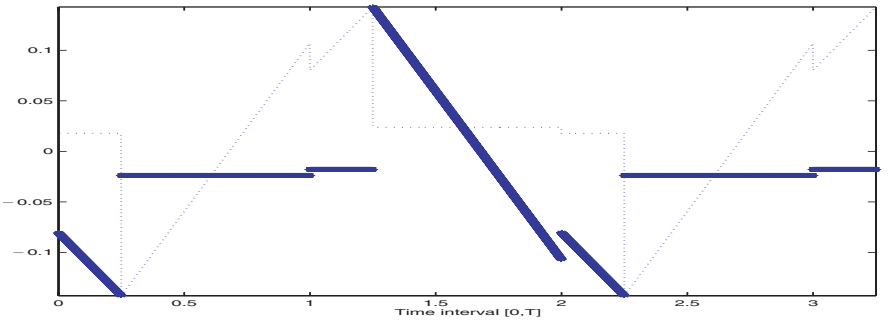
$y(x,0) = y^0(x),\ y_t(x,0) = 0,\ y(x,T) = y_t(x,T) = 0,\ (x,t) \in (0,1) \times (0,2)$.

Let $y^0 \in L^2(0,1)$ be continued to the interval $(-1,1)$ as the antisymmetric function y^0_a. Then problem (B) has the same solutions as problem $C(\infty)$ with $c = 1$, $L = T = 2$, $y_0(x) = y^0_a(x+1)$, $y_1(x) = 0$ in the sense that the optimal controls satisfy $f_2(t) = v(T-t) = -f_1(t)$.

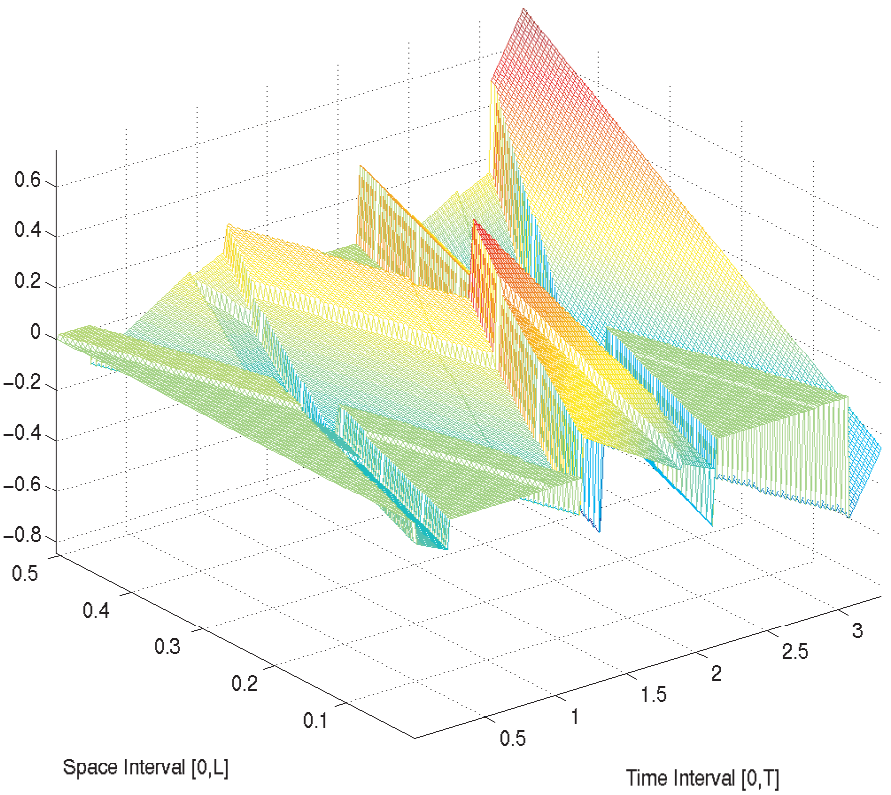
Following [6], for a given solution u of the adjoint optimization problem

$$\inf \frac{1}{2} \left(\int_0^2 |u_x(1,t)|\, dt \right)^2 + \int_0^1 y^0(x)u^1(x) - y^1(x)u^0(x)\, dx \text{ s.t. } u_{tt}(x,t) = u_{xx}(x,t),$$

$u(0,t) = u(1,t) = 0,\ u(x,0) = u^0(x),\ u_t(x,0) = u^1(x), (x,t) \in (0,1) \times (0,2)$ a solution v of problem (B) is quasi bang-bang in the sense that $v(t) \in \ \text{sign}(u_x(1,t))$.

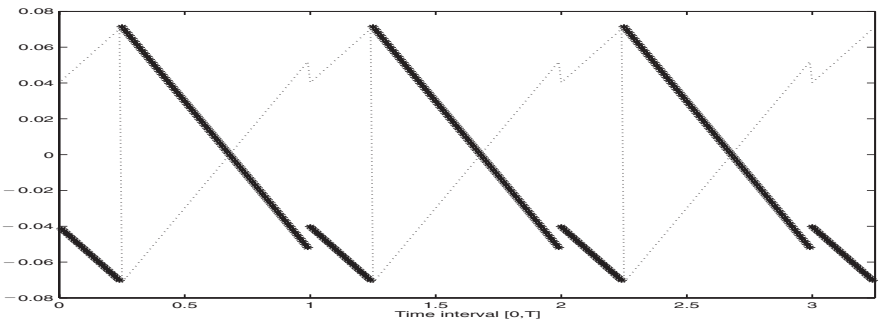


(a) The optimal controls

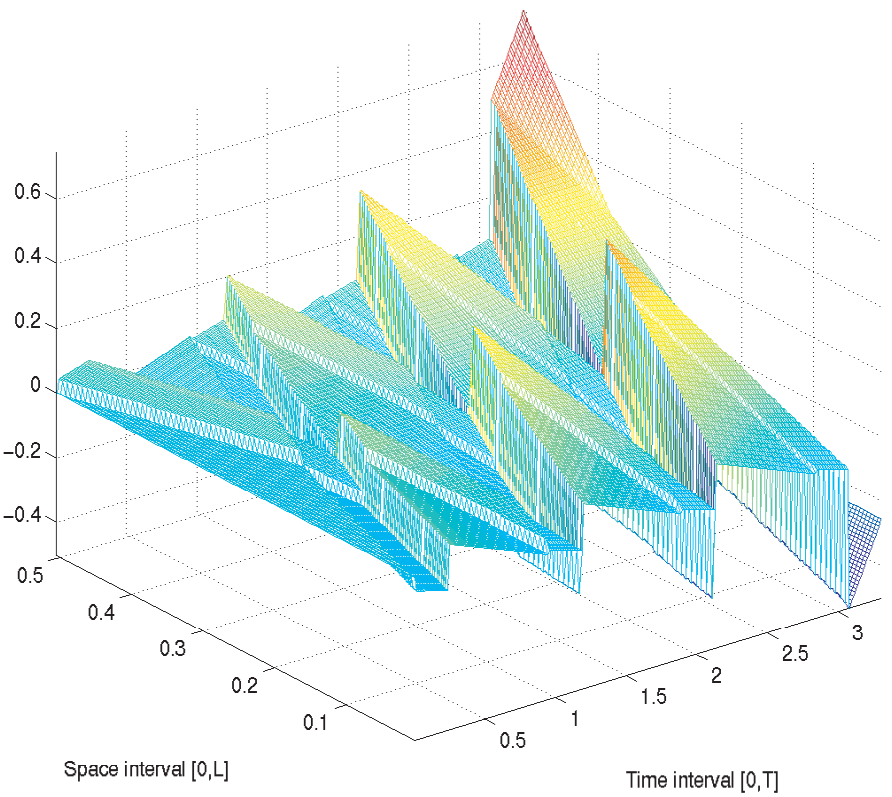


(b) The optimal state

FIG. 2.1.



(a) The optimal controls



(b) The optimal state

FIG. 2.2.

Assume that y^0 attains its infinity norm on a set of measure greater than zero. Then the necessary optimality conditions for the adjoint optimization problem imply that there is a constant $c_0 > 0$ such that for $t \in (0,1)$, $u_x(1,t) = y^0(1-t)c_0$ if $y^0(1-t) \in \{\|y^0\|_{\infty,(0,1)}, -\|y^0\|_{\infty,(0,1)}\}$, $u_x(1,t) = 0$ otherwise, and for $t \in (1,2)$, $u_x(1,t) = -y^0(t-1)c_0$ if $y^0(t-1) \in \{\|y^0\|_{\infty,(0,1)}, -\|y^0\|_{\infty,(0,1)}\}$, $u_x(1,t) = 0$ otherwise.

Theorem 2.2 yields exactly the same solutions (f_1, f_2) as in Example 3. The solution $v(t) = f_2(2-t)$ satisfies the quasi-bang-bang characterization given above, which is a restriction only on the set where y^0 attains its infinity norm. Note that $v(t)$ satisfies at the same time the quasi-bang-bang characterization and is given as the boundary trace of the optimal solution of the adjoint optimization problem for the L^2-case considered in Example 3.

2.7. Weak solutions of the initial-value problem. A general description of the weak form of the initial boundary-value problems with Dirichlet or Neumann boundary conditions can be found in [17]. The solution y of our initial boundary-value problem (2.1)–(2.3) has the series representation

$$y(x,t) = \sum_{j=1}^{\infty} (2c/L) \int_0^t [f_1(s) - (-1)^j f_2(s)] \sin\left((c\pi j/L)(t-s)\right)\, ds \sin\left((j\pi/L)x\right),$$

and for the time-derivative y_t we have

$$y_t(x,t) = \sum_{j=1}^{\infty} \frac{2c^2 j\pi}{L^2} \int_0^t [f_1(s) - (-1)^j f_2(s)] \cos\left((c\pi j/L)(t-s)\right)\, ds \sin\left((j\pi/L)x\right).$$

2.8. End conditions and a trigonometric moment problem. For $j \in \mathbb{N}$, define the function $\varphi_j(x) = (\sqrt{2}/\sqrt{L}) \sin(j\pi x/L)$ and the numbers

$$y_0^j = \int_0^L y_0(x)\varphi_j(x)\, dx,\ y_1^j = \int_0^L y_1(x)\varphi_j(x)\, dx.$$

Inserting the series representations of the solution y and its time derivative y_t into the end conditions (2.4) yields the trigonometric moment equations

$$\int_0^T (\sqrt{2}c/\sqrt{L})[f_1(s) - (-1)^j f_2(s)] \sin\left((c\pi j/L)(T-s)\right)\, ds = y_0^j, \tag{2.8}$$

$$\int_0^T (\sqrt{2}c^2\pi j/L^{3/2})[f_1(s) - (-1)^j f_2(s)] \cos\left((c\pi j/L)(T-s)\right)\, ds = y_1^j \tag{2.9}$$

for $j \in \mathbb{N}$. Hence, we have described the set of feasible controls as the solution set of a trigonometric moment problem. This approach to controllability via moment problems is well established (see, for example, [19], [1]).

2.9. The minimal time interval with controllability. In this section we study controllability on the time interval with $T = L/c$. Since this is the time that a characteristic curve starting at one end of the system needs to reach the other end, it is clear that this is the minimal time interval, where controllability for general target states $y_0 \in L^2(0,L)$, $y_1 \in W_2^{-1}$ can possibly hold.

DEFINITION 2.5. *A function $f \in L^2(0,L)$ is* symmetric with respect to the midpoint $L/2$ *if $f(L/2 - x) = f(L/2 + x)$ for all $x \in (0, L/2)$. The function f is*

antisymmetric *on the interval* $[0, L]$ *with respect to the midpoint* $L/2$ *if* $f(L/2 - x) = -f(L/2 + x)$ *for all* $x \in (0, L/2)$.

Remark 2.1. Each function $f \in L^2(0, L)$ can be written as a sum $f = f^{even} + f^{odd}$ with a symmetric function $f^{even} \in L^2(0, L)$ and an antisymmetric function $f^{odd} \in L^2(0, L)$. The functions f^{even} and f^{odd} are determined uniquely. Moreover, $f \in L^p(0, L)$ if and only if f^{even} and f^{odd} are in $L^p(0, L)$. In fact, we have

$$f^{even}(x) = (f(x) + f(L - x))/2, \qquad f^{odd}(x) = (f(x) - f(L - x))/2.$$

Note that $f^{even}(x) - f^{odd}(x) = f(L - x)$.

For given control functions f_1 and f_2 we introduce the sum

$$S(t) = (f_1(T - t) + f_2(T - t))/2 \tag{2.10}$$

and the difference

$$D(t) = (f_1(T - t) - f_2(T - t))/2. \tag{2.11}$$

The trigonometric moment equations (2.8), (2.9) are equivalent to two moment problems for the functions S and D. We start with the moment problem for D:

$$\int_0^{Tc} D(t/c)(\sqrt{2}/\sqrt{L}) \sin(2\pi j t/L)\, dt = y_0^{2j}/2, \tag{2.12}$$

$$\int_0^{Tc} D(t/c)(\sqrt{2}/\sqrt{L}) \cos(2\pi j t/L)\, dt = L y_1^{2j}/(4c\pi j). \tag{2.13}$$

This means that we know all the Fourier coefficients of the function $D(\cdot/c)$ except the coefficient that corresponds to the constant function. Hence there exists a real number r such that for all $x \in [0, L]$ we have

$$D\left(\frac{x}{c}\right) = r + \sum_{j=1}^{\infty} \frac{y_0^{2j}}{2} \sqrt{\frac{2}{L}} \sin\left(\frac{2\pi j x}{L}\right) - \frac{1}{2c} \sum_{j=1}^{\infty} \frac{-y_1^{2j} L}{2\pi j} \sqrt{\frac{2}{L}} \cos\left(\frac{2\pi j x}{L}\right).$$

We define the symmetric function

$$Y_1^{even}(x) = \sum_{j=1}^{\infty} -y_1^{2j} (L/(2\pi j)) \sqrt{(2/L)} \cos((2\pi j/L)x)$$

and the antisymmetric functions

$$y_1^{odd} = \sum_{j=1}^{\infty} (y_1^{2j}) \sqrt{(2/L)} \sin(2\pi j x/L), \; y_0^{odd} = \sum_{j=1}^{\infty} (y_0^{2j}) \sqrt{(2/L)} \sin(2\pi j x/L).$$

We have $(Y_1^{even})'(x) = y_1^{odd}$, and for the function D for all $x \in [0, L/c]$, we have

$$D(x) = r + y_0^{odd}(cx)/2 - (1/(2c))\, Y_1^{even}(cx). \tag{2.14}$$

Now we consider the moment problem for the function S:

$$\int_0^{Tc} S(t/c)(\sqrt{2}/\sqrt{L}) \sin((2j - 1)\pi t/L)\, dt = y_0^{2j-1}/2, \tag{2.15}$$

$$\int_0^{Tc} S(t/c)(\sqrt{2}/\sqrt{L}) \cos((2j - 1)\pi t/L)\, dt = L y_1^{2j-1}/(2(2j - 1)\pi c). \tag{2.16}$$

We define the symmetric functions

$$y_0^{even}(x) = \sum_{j=1}^{\infty} (y_0^{2j-1}) \sqrt{(2/L)} \sin((2j-1)\pi x/L),$$
$$y_1^{even}(x) = \sum_{j=1}^{\infty} (y_1^{2j-1}) \sqrt{(2/L)} \sin((2j-1)\pi x/L)$$

and the antisymmetric function

$$Y_1^{odd}(x) = \sum_{j=1}^{\infty} -y_1^{2j-1} (L/((2j-1)\pi)) \sqrt{(2/L)} \cos(((2j-1)\pi/L)x).$$

Then we have $(Y_1^{odd})'(x) = y_1^{even}(x)$. For the function S for all $x \in [0, L/c]$, we have

$$S(x) = y_0^{even}(cx)/2 - (1/(2c))\, Y_1^{odd}(cx). \tag{2.17}$$

Thus, for the control functions f_1, f_1 that steer the system to the target state in the time $T = L/c$, we have

$$f_1(T-t) = S(t) + D(t) = \quad r + y_0(ct)/2 - (1/(2c))\, Y_1(ct), \tag{2.18}$$
$$f_2(T-t) = S(t) - D(t) = -r + y_0(L-ct)/2 + (1/(2c))Y_1(L-ct), \tag{2.19}$$

where r is a real number.

This representation of the functions f_1 and f_2 implies that if y_0 and Y_1 are in the space $L^p(0, L)$, then the functions f_1 and f_2 are in the space $L^p(0, L/c)$. On the other hand, if f_1 and f_2 are in the space $L^p(0, L/c)$, then D and S also are in the space $L^p(0, L/c)$, which implies that y_0^{odd}, Y_1^{even}, y_0^{even}, Y_1^{odd} are in $L^p(0, L)$. This, in turn, is equivalent to the statement that y_0 and Y_1 are in $L^p(0, L)$. Thus, we have the following.

LEMMA 2.6. *Let $p \in [2, \infty]$ and $T = L/c$. If the control functions f_1 and f_2 are in $L^p(0, T)$, then the state $y(\cdot, T)$, $y_t(\cdot, T)$ that the system has reached at time T has the following regularity: $y(\cdot, T)$ is in $L^p(0, L)$ and $h(x) = \int_0^x y_t(z, T)\, dz$ is also in $L^p(0, L)$.*

For a given target state (y_0, y_1) with y_0, $Y_1 \in L^p(0, L)$, there exist control functions f_1 and f_2 in $L^p(0, T)$ that steer the system to this target; moreover, these controls are uniquely determined up to the constant r in (2.18), (2.19).

The uniqueness follows from the fact that the moment problem for S has a unique solution in $L^2(0, T)$ and the moment problem for D determines D up to a constant.

2.10. Controllability on larger time intervals. In this section we show how the question of controllability for a time interval $[0, T]$ with $T > L/c$ can be reduced to the question for the minimal time interval $[0, L/c]$ that was considered in the last section. This reduction depends upon the fact that all the trigonometric functions that appear in the moment equations have similar periodicity properties.

2.10.1. Transformation of the moment equations. Assume that $T \geq L/c$. Choose the natural number k such that $kL/c \leq T < (k+1)L/c$. Let the function $\varphi(s)$ be an element of the set $\{\sin((c\pi j/L)s), \cos((c\pi j/L)s)$ with $j \in \mathbb{N}$, j odd $\}$. Then we have $\varphi(s + L/c) = -\varphi(s)$, and for all functions $v \in L^2(0, T)$, the following equation

is valid:

$$\int_0^T v(s)\varphi(s)\,ds = \int_0^{T-kL/c} \left[\sum_{j=0}^{k}(-1)^j v(s+jL/c)\right]\varphi(s)\,ds + \int_{T-kL/c}^{L/c} \left[\sum_{j=0}^{k-1}(-1)^j v(s+jL/c)\right]\varphi(s)\,ds.$$

Define the function

$$\hat{v}(t) = \sum_{j=0}^{k}(-1)^j v(t+jL/c) \text{ for } t \in (0, T-kL/c), \tag{2.20}$$

$$\hat{v}(t) = \sum_{j=0}^{k-1}(-1)^j v(t+jL/c) \text{ for } t \in (T-kL/c, L/c). \tag{2.21}$$

Then

$$\int_0^{L/c} \hat{v}(s)\varphi(s)\,ds = \int_0^T v(s)\varphi(s)\,ds.$$

As in the last section, let the functions S and D be defined by (2.10) and (2.11). Then for a function S that satisfies the moment equations (2.15), (2.16) for all $j \in \mathbb{N}$, the corresponding function $\hat{v}$ must satisfy these moment equations with integrals on the interval $(0, L/c)$. In Lemma 2.6, we have stated that the moment equations (2.15), (2.16) with $T = L/c$ determine a unique solution $\hat{S}$, which is given by (2.17).

For a function D that satisfies the moment equations (2.12), (2.13), the corresponding function $\hat{v}$ is defined as in (2.20), (2.21) but the numbers $(-1)^j$ are replaced by 1 and must satisfy (2.12), (2.13) on the interval $(0, L/c)$. These moment equations determine $\hat{D}$, which is given by (2.14) up to a constant.

In what follows, let $\hat{D}$ be defined by the equation

$$\hat{D}(x) = y_0^{odd}(cx)/2 - (1/(2c))\, Y_1^{even}(cx) \tag{2.22}$$

and $\hat{S}$ by

$$\hat{S}(x) = y_0^{even}(cx)/2 - (1/(2c))\, Y_1^{odd}(cx). \tag{2.23}$$

So we see that we can describe the feasible controls, that is, the controls that steer the system to the target, by the following equations (with $\Delta = T - kL/c$):

$$\hat{S}(t) = \sum_{j=0}^{k}(-1)^j S(t+jL/c), \qquad \hat{D}(t) + \hat{r} = \sum_{j=0}^{k} D(t+jL/c), t \in (0, \Delta), \tag{2.24}$$

$$\hat{S}(t) = \sum_{j=0}^{k-1}(-1)^j S(t+jL/c), \qquad \hat{D}(t) + \hat{r} = \sum_{j=0}^{k-1} D(t+jL/c), t \in (\Delta, L/c), \tag{2.25}$$

where $\hat{r}$ can be any real number.

This means that we have reduced our problem of optimal control to an optimization problem with four affine linear pointwise equality constraints.

2.10.2. Proof of Theorem 2.1. Now we proceed to the proof of Theorem 2.1. It is clear that f_1 and f_2 are in $L^p(0,T)$ if and only if S and D are in $L^p(0,T)$.

If y_0 and Y_1 are in $L^p(0,L)$, Lemma 2.6 implies that we can find $\hat{S}$ and $\hat{D}$ in $L^p(0,L/c)$ that satisfy the moment equations (2.15), (2.16) and (2.12), (2.13), respectively. Then we can find functions S and D in $L^p(0,T)$ such that (2.24)–(2.25) hold, for example, with $\hat{r}=0$ and the definitions $D(t)=\hat{D}(t)$ and $S(t)=\hat{S}(t)$ for $t\in(0,L/c)$ and $0=D(t)=S(t)$ for $t\geq L/c$. In this way, we obtain feasible controls f_1 and f_2 in $L^p(0,T)$.

Now we prove the converse. Let f_1 and f_2 in the space $L^p(0,T)$ be given. Then the corresponding functions $\hat{D}$ and $\hat{S}$ defined by (2.24)–(2.25) with $\hat{r}=0$ are in the space $L^p(0,L/c)$. The corresponding controls on the time interval $(0,L/c)$ reach the same target as f_1 and f_2 on the time interval $(0,T)$ since they solve the corresponding moment problem on the shorter time interval $(0,L/c)$. Hence Lemma 2.6 implies that the corresponding target state has the desired regularity, namely, y_0 and Y_1 are in $L^p(0,L)$. So the proof of Theorem 2.1 is complete.

2.11. Transformation of the optimization problem. In section 2.10.1 we showed that the set of controls f_1, f_2 for which the corresponding state y satisfies the end conditions (2.4) can be described by (2.24)–(2.25), with S defined by (2.10), D defined by (2.11), $\hat{S}$ given by (2.23), and $\hat{D}$ as in (2.22).

Thus for $p<\infty$, problem $C(p)$ can be transformed into the following form:

$$\inf \|f_1\|^p_{p,(0,T)}+\|f_2\|^p_{p,(0,T)} \text{ s.t. } f_1,\ f_2\in L^p[0,T] \text{ and } S \text{ defined by (2.10)}$$

and D defined by (2.11) satisfy the constraints (2.24)–(2.25) for some $\hat{r}\in R$ with $\hat{S}$ given by (2.23) and $\hat{D}$ given by (2.22). For $p=\infty$, $C(p)$ is equivalent to the corresponding problem with objective function

$$\max\{\|f_1\|_{\infty,(0,T)},\ \|f_2\|_{\infty,(0,T)}\}.$$

2.11.1. Proof of Theorem 2.2. In this section, we use the transformed form of problem $C(p)$ that was given in the last section to prove Theorem 2.2.

First we consider the case $p<\infty$. Define the function

$$J(f_1,f_2)=\|f_1\|^p_{p,(0,T)}+\|f_2\|^p_{p,(0,T)},$$

which is the objective function of problem $C(p)$ for $p<\infty$. We have the representation

$$J(f_1,f_2)=\int_0^{T-kL/c}\sum_{j=0}^{k}\big|f_1(T-t-jL/c)\big|^p+\big|f_2(T-t-jL/c)\big|^p\,dt \tag{2.26}$$

$$+\int_{T-kL/c}^{L/c}\sum_{j=0}^{k-1}\big|f_1(T-t-jL/c)\big|^p+\big|f_2(T-t-jL/c)\big|^p\,dt.$$

Since $f_1(T-t)=S(t)+D(t)$ and $f_2(T-t)=S(t)-D(t)$, the constraints (2.24) imply (for a natural number n, let $b(n)=1$ if j is odd and $b(n)=2$ if j is even)

$$\sum_{j=0}^{k}(-1)^j f_{b(j+1)}(T-t-jL/c)=\hat{S}(t)+\hat{D}(t)+\hat{r}, \tag{2.27}$$

$$\sum_{j=0}^{k}(-1)^j f_{b(j)}(T-t-jL/c)=\hat{S}(t)-\hat{D}(t)-\hat{r} \tag{2.28}$$

for all $t \in (0, T - kL/c)$, and the constraints (2.25) imply

$$\sum_{j=0}^{k-1} (-1)^j f_{b(j+1)}(T - t - jL/c) = \hat{S}(t) + \hat{D}(t) + \hat{r}, \tag{2.29}$$

$$\sum_{j=0}^{k-1} (-1)^j f_{b(j)}(T - t - jL/c) = \hat{S}(t) - \hat{D}(t) - \hat{r} \tag{2.30}$$

for all $t \in (T - kL/c, L/c)$.

In our optimization problem, each point of the time interval $[0, T]$ corresponds to two equality constraints. The objective function J is given by an integral over the time interval $[0, T]$, where for each $t \in [0, T]$ the integrand is the sum of two terms, each of which depends only on function values that appear in exactly one of the constraints. The idea of our proof is that we can minimize the objective function J subject to the two pointwise constraints by minimizing for each point in time both parts of the integrand separately subject to the corresponding equality constraint.

The solutions of the resulting parametric family of optimization problems are given in the following lemma.

LEMMA 2.7. *Let $p \geq 2$, a natural number d, and a real number g be given. Consider the optimization problem*

$$H(p, d, g): \quad \min_{(f_0, \ldots, f_d) \in R^{d+1}} \sum_{j=0}^{d} |f_j|^p \ s.t. \quad \sum_{j=0}^{d} (-1)^j f_j = g.$$

The unique solution of $H(p, d, g)$ has the components $f_j = (-1)^j g/(d+1)$ and the optimal value is $|g|^p/(d+1)^{p-1}$.

Proof. $H(p, d, g)$ is a convex optimization problem with a strictly convex objective function, and hence it has at most one solution. The point with the components $(-1)^j g/(d+1)$ is feasible and satisfies the necessary optimality conditions, and hence it is the unique solution of $H(p, d, g)$. ∎

Let the number $\hat{r}$ be given. Representation (2.26) of the objective function J shows that in order to minimize J subject to our pointwise constraints, it suffices to choose the values of our control functions f_1, f_2 as follows: For $t \in [0, T - kL/c]$, let $f_j = f_{b(j+1)}(T - t - jL/c)$ $(j \in \{0, \ldots, k\})$ be such that they solve problem $H(p, k, \hat{S}(t) + \hat{D}(t) + \hat{r})$, that is,

$$f_{b(j+1)}(T - t - jL/c) = (-1)^j (\hat{S}(t) + \hat{D}(t) + \hat{r})/(k+1),$$

and let $f_j = f_{b(j)}(T - t - jL/c)$ be the solution of problem $H(p, k, \hat{S}(t) - \hat{D}(t) - \hat{r})$, that is,

$$f_{b(j)}(T - t - jL/c) = (-1)^j (\hat{S}(t) - \hat{D}(t) - \hat{r})/(k+1).$$

Similarly, for $t \in [T - kL/c, L/c]$, let $f_j = f_{b(j+1)}(T - t - jL/c)$ $(j \in \{0, \ldots, k-1\})$ be such that they solve problem $H(p, k-1, \hat{S}(t) + \hat{D}(t) + \hat{r})$, that is,

$$f_{b(j+1)}(T - t - jL/c) = (-1)^j (\hat{S}(t) + \hat{D}(t) + \hat{r})/k,$$

and let $f_j = f_{b(j)}(T - t - jL/c)$ be such that they solve problem $H(p, k-1, \hat{S}(t) - \hat{D}(t) - \hat{r})$, that is,

$$f_{b(j)}(T - t - jL/c) = (-1)^j (\hat{S}(t) - \hat{D}(t) - \hat{r})/k.$$

By Lemma 2.7, this yields the following value of the objective function:

$$J(f_1,f_2)=\int_0^{T-kL/c}(1/(k+1)^{p-1})\left[|\hat{S}(t)+\hat{D}(t)+\hat{r}|^p+|\hat{S}(t)-\hat{D}(t)-\hat{r}|^p\right]\,dt$$
$$+\int_{T-kL/c}^{L/c}(1/k^{p-1})\left[|\hat{S}(t)+\hat{D}(t)+\hat{r}|^p+|\hat{S}(t)-\hat{D}(t)-\hat{r}|^p\right]\,dt.$$

Since this value still depends on our choice of the real number $\hat{r}$, we define this value as $h_p(\hat{r})$. Now the problem remains to find the value of $\hat{r}$ for which the corresponding value of the objective function is minimal. Since the function h_p is strictly convex and differentiable, the equation $h_p'(\hat{r})=0$ uniquely determines the optimal value of $\hat{r}$.

Remember that due to (2.18) and (2.19), we can compute $\hat{S}+\hat{D}$ and $\hat{S}-\hat{D}$ from the given functions y_0 and Y_1, so the optimal value of $\hat{r}$ can be determined.

Now we come to the case $p=\infty$. Also in this case, we can transform our problem $C(\infty)$ into a problem with the four simple pointwise equality constraints (2.27)–(2.30):

$$\inf\ \ \max\{\|f_1\|_{\infty,(0,T)},\,\|f_2\|_{\infty,(0,T)}\}\ \text{s.t.}\ \ f_1,\,f_2\in L^\infty(0,T)$$

and there is a real number $\hat{r}$ such that f_1, f_2 satisfy (2.27)–(2.30).

In order to solve this problem, we look for solutions at each $t\in(0,T-kL/c)$ of the problems to minimize

$$\max\{|f_{b(j+1)}(T-t-jL/c)|,\ j\in\{0,\dots,k\}\}\ \text{s.t. (2.27) is satisfied,}$$
$$\max\{|f_{b(j)}(T-t-jL/c)|,\ j\in\{0,\dots,k\}\}\ \text{s.t. (2.28) is satisfied,}$$

and for each $t\in(T-kL/c,L/c)$ for solutions of the analogous problems with (2.29), (2.30), respectively. We present the solutions of the resulting parametric family of optimization problems in the following.

Lemma 2.8. *Let a natural number d and a real number g be given. Consider the optimization problem*

$$H(d,g):\quad \min_{(f_0,\dots,f_d)\in R^{d+1}}\max\{|f_j|,\ j\in\{0,\dots,d\}\}\ \ s.t.\ \ \sum_{j=0}^{d}(-1)^j f_j=g.$$

The unique solution of $H(d,g)$ has the components $f_j=(-1)^j g/(d+1)$ and the optimal value is $|g|/(d+1)$.

Proof. The point with the components $(-1)^j g/(d+1)$ is feasible. Hence the optimal value of $H(d,g)$ is $\le|g|/(d+1)$. Suppose that there exists a point $(h_0,\dots,h_d)$ with $\sum_{j=0}^d(-1)^j h_j=g$ and $\max|h_j|<|g|/(d+1)$. Then $|\sum_{j=0}^d(-1)^j h_j|<\sum_{j=0}^d|g|/(d+1)=|g|$, a contradiction. So the optimal value of $H(d,g)$ is $|g|/(d+1)$. Using a similar contradiction argument we see that for every solution $(h_0,\dots,h_d)$ of $H(d,g)$ we have $|h_j|=|g|/(d+1)$ for all j. Inserting this condition into the equation $\sum_{j=0}^d(-1)^j h_j=g$ yields $\sum_{j=0}^d(-1)^j\mathrm{sign}h_j=(d+1)\mathrm{sign}g$, and hence for all j we have $\mathrm{sign}h_j=(-1)^j\mathrm{sign}g$, and the assertion follows. □

In analogy to the case $p<\infty$, Lemma 2.8 yields the desired solutions. In order to obtain an optimal control in this case, we choose $\hat{r}$ such that it minimizes the function h_∞ defined as

$$h_\infty(r)=\max\Big\{\|(\hat{S}(t)+\hat{D}(t)+r)/(k+1)\|_{\infty,(0,T-kL/c)},$$
$$\|(\hat{S}(t)-\hat{D}(t)-r)/(k+1)\|_{\infty,(0,T-kL/c)},$$
$$\|(\hat{S}(t)+\hat{D}(t)+r)/k\|_{\infty,(T-kL/c,L/c)},\|(\hat{S}(t)-\hat{D}(t)-r)/k\|_{\infty,(T-kL/c,L/c)}\Big\}.$$

This determines the value of $\hat{r}$ uniquely. However, in the L^∞-case, the optimal control is in general not uniquely determined. For the given value of $\hat{r}$, our construction above yields the solution of (2.27)–(2.30) with minimal L^2-norm, which is in fact the same as our solution for the case $p = \infty$, so we have constructed the solution of problem $C(\infty)$ with minimal L^2-norm.

2.11.2. Symmetric targets. In this subsection, we assume that for all even $j \in \mathbb{N}$ we have

$$\int_0^L y_0(x)\varphi_j(x)\,dx = 0 = \int_0^L y_1(x)\varphi_j(x)\,dx. \tag{2.31}$$

This means that the functions y_0 and y_1 are even on the interval $[0, L]$ with respect to the midpoint $L/2$. This implies that Y_1 is antisymmetric. Thus, (2.22) implies that $\hat{D} = 0$. We have $h'(0) = 0$, and hence, in this case, the number $\hat{r} = 0$ is the optimal choice. On the time interval $(0, L/c)$ this yields the control functions $f_1(T-t) = y_0(ct)/2 - (1/(2c))Y_1(ct) = f_2(T-t)$. Also on larger time intervals $(0, T)$, for the optimal controls we have $f_1 = f_2$, since $g_1 = g_2$.

2.11.3. Antisymmetric targets. In this subsection, we assume that for all odd $j \in \mathbb{N}$ (2.31) holds. This means that the functions y_0 and y_1 are antisymmetric on the interval $[0, L]$ with respect to the midpoint $L/2$. Then Y_1 is symmetric, and in the statement of Theorem 2.2 we have $g_1(t) = -g_2(t)$. Therefore, for the optimal controls we have $f_1 = -f_2$.

3. Neumann boundary control. In this section we study the problem in which the system is controlled by Neumann boundary conditions.

3.1. The initial-value problem. Let a wave speed $c > 0$ be given. We consider the initial-value problem with the wave equation

$$y_{tt}(x,t) = c^2 y_{xx}(x,t), \quad (x,t) \in [0,L] \times [0,T], \tag{3.1}$$

subject to the initial conditions

$$y(x,0) = 0, \ y_t(x,0) = 0, \quad x \in [0,L], \tag{3.2}$$

and the Neumann boundary conditions

$$y_x(0,t) = -f_1(t), \ y_x(L,t) = f_2(t), \quad t \in [0,T]. \tag{3.3}$$

The desired target state is given in the following end conditions:

$$y(x,T) = y_0(x), \ y_t(x,T) = y_1(x), \quad x \in [0,L]. \tag{3.4}$$

The functions y_0, y_1 are in the space $L^2(0, L)$.

3.2. The optimization problem. For a fixed time $T > 0$ and a given value of $p \in [2, \infty)$, we consider the following optimization problem:

$$C(p): \ \inf \ \|f_1\|^p_{p,(0,T)} + \|f_2\|^p_{p,(0,T)} \text{ s.t. } f_1, f_2 \in L^p[0,T]$$

and the solution y of the initial boundary-value problem (3.1)–(3.3) satisfies the end conditions (3.4).

In the case $p = \infty$, the objective function is $\max\{\|f_1\|_{\infty,(0,T)}, \ \|f_2\|_{\infty,(0,T)}\}$.

3.3. Exact controllability.

THEOREM 3.1. *Let $p \in [2,\infty]$ and $T > L/c$ be given. The initial boundary-value problem* (3.1)–(3.3) *has a weak solution satisfying the end conditions* (3.4) *with $f_1,\ f_2 \in L^p(0,T)$ if and only if the target states $y_0,\ y_1$ satisfy the following conditions: $y_1 \in L^p(0,L)$ and $y_0 \in L^2(0,L)$ is such that the derivative y_0' in the sense of distributions is in the space $L^p(0,L)$, that is, $y_0 \in W_p^1(0,L)$. This implies that the optimization problem $C(p)$ has a solution if and only if y_0' and y_1 are in $L^p(0,L)$.*

3.4. Weak solution of the initial-value problem. The solution y of the initial boundary-value problem (3.1)–(3.3) has the series representation

$$y(x,t) = (c^2/L)\int_0^t [f_1(s)+f_2(s)](t-s)\,ds$$

$$+\sum_{j=1}^{\infty}(2/(cj\pi))\int_0^t [f_1(s)+(-1)^j f_2(s)]\,\sin\left((c\pi j/L)(t-s)\right)\,ds\cos\left((j\pi/L)x\right)$$

and for the time derivative y_t we obtain the series

$$y_t(x,t) = (c^2/L)\int_0^t [f_1(s)+f_2(s)]\,ds$$

$$+\sum_{j=1}^{\infty}(2/L)\int_0^t [f_1(s)+(-1)^j f_2(s)]\,\cos\left((c\pi j/L)(t-s)\right)\,ds\cos\left((j\pi/L)x\right).$$

3.5. End conditions and a trigonometric moment problem. For $j \in \mathbb{N}$, define the functions

$$\varphi_0(x) = 1/\sqrt{L},\ \varphi_j(x) = (\sqrt{2}/\sqrt{L})\,\cos(j\pi x/L),$$

and for $j \in \mathbb{N}\cup\{0\}$, define the numbers

$$y_0^j = \int_0^L y_0(x)\varphi_j(x)\,dx,\ y_1^j = \int_0^L y_1(x)\varphi_j(x)\,dx.$$

Inserting the series representation of the solution y and its time derivative y_t into the end conditions (3.4) yields the moment equations

$$\int_0^T (c^2/\sqrt{L})(f_1(T-s)+f_2(T-s))\,s\,ds = y_0^0, \tag{3.5}$$

$$\int_0^T (\sqrt{2}\sqrt{L}/(c\pi j))(f_1(T-s)+(-1)^j f_2(T-s))\sin\left((c\pi j/L)s\right)\,ds = y_0^j, \qquad j\in\mathbb{N}. \tag{3.6}$$

$$(c^2/\sqrt{L})\int_0^T f_1(T-s)+f_2(T-s)\,ds = y_1^0, \tag{3.7}$$

$$\int_0^T (\sqrt{2}/\sqrt{L})(f_1(T-s)+(-1)^j f_2(T-s))\cos\left((c\pi j/L)s\right)\,ds = y_1^j, \qquad j\in\mathbb{N}. \tag{3.8}$$

3.6. The minimal time interval with controllability up to a constant. In this section we study controllability on the time interval with $T = L/c$, which is the minimal time interval, where controllability for all target states y_0, y_1 in $L^2(0, L)$ can be possible.

For given control functions f_1 and f_2 we introduce the sum S as in (2.10) and the difference D as in (2.11). The trigonometric moment equations (3.5)–(3.8) are equivalent to two moment problems for the functions S and D.

The moment problem for the difference function D is

$$\int_0^{Tc} (L/(c^2(2j-1)\pi))D(x/c)\sqrt{(2/L)}\sin((2j-1)\pi x/L)\,dx = y_0^{2j-1}/2, \tag{3.9}$$

$$\int_0^{Tc} (1/c)D(x/c)\sqrt{(2/L)}\cos((2j-1)\pi x/L)\,dx = y_1^{2j-1}/2. \tag{3.10}$$

We define the antisymmetric functions

$$y_0^{odd}(x) = \sum_{j=1}^{\infty} (y_0^{2j-1})\sqrt{(2/L)}\cos((2j-1)\pi x/L),$$

$$y_1^{odd}(x) = \sum_{j=1}^{\infty} (y_1^{2j-1})\sqrt{(2/L)}\cos((2j-1)\pi x/L)$$

and the symmetric function

$$Y_0^{even}(x) = -\sum_{j=1}^{\infty} y_1^{2j-1}((2j-1)\pi/L)\sqrt{(2/L)}\sin(((2j-1)\pi/L)x).$$

Then $Y_0^{even}(x) = (y_0^{odd})'(x)$ and for all $t \in [0, L/c]$ we have

$$D(t) = (c/2)y_1^{odd}(ct) - (c^2/2)(y_0^{odd})'(ct). \tag{3.11}$$

Consider the moment problem for the function S:

$$\int_0^{Tc} (1/\sqrt{L})S(x/c)\,x\,dx = y_0^0/2, \tag{3.12}$$

$$\int_0^{Tc} S(x/c)(\sqrt{2/L})\sin\left((2\pi j/L)x\right)\,dx = c^2(2\pi j/L)y_0^{2j}/2, \tag{3.13}$$

$$\int_0^{Tc} (1/\sqrt{L})S(x/c)\,dx = y_1^0/(2c), \tag{3.14}$$

$$\int_0^{Tc} S(x/c)(\sqrt{2/L})\cos\left((2\pi j/L)x\right)\,dx = cy_1^{2j}/2, \qquad j \in \mathbb{N}. \tag{3.15}$$

This system of moment equations is in fact overdetermined. If we omit (3.12), the remaining system determines a unique solution.

We define the symmetric functions

$$y_0^{even}(x) = \sum_{j=1}^{\infty} (y_0^{2j})\sqrt{(2/L)}\cos(2j\pi x/L) + y_0^0/\sqrt{L},$$

$$y_1^{even}(x) = \sum_{j=1}^{\infty} (y_1^{2j})\sqrt{(2/L)}\cos(2j\pi x/L) + y_1^0/(c^2\sqrt{L})$$

and the antisymmetric function

$$Y_0^{odd}(x) = \sum_{j=1}^{\infty} -y_0^{2j}(2j\pi/L)\sqrt{(2/L)}\sin((2j\pi/L)x).$$

Note that in the definition of y_1^{even} in the constant term the wave speed c^2 appears. We have $(y_0^{even})'(x) = Y_0^{odd}(x)$. For the function S, we have for all $t \in [0, L/c]$

$$S(t) = (c/2)y_1^{even}(ct) - (c^2/2)\,(y_0^{even})'(ct) \tag{3.16}$$

since all the Fourier coefficients of S are determined by (3.13)–(3.15).

Thus the unique solution of the moment problem (3.6)–(3.8) with $T = L/c$ is

$$f_1(T-t) = S(t) + D(t) = (c/2)\big[y_1(ct) + ((1/c^2) - 1)(y_1^0/\sqrt{L})\big] - (c^2/2)y_0'(ct), \tag{3.17}$$

$$f_2(T-t) = S(t) - D(t) = (c/2)y_1(L-ct) - (c^2/2)y_0'(L-ct). \tag{3.18}$$

With these control functions at time $T = L/c$ the system reaches a target state of the form $y_0 + c_0$, y_1 since in the representation of the state $y(\cdot, T)$ as a series of the functions φ_j ($j \in \mathbb{N} \cup \{0\}$), all the coefficients for $j \neq 0$ are determined by (3.6). To find control functions that satisfy the first moment equation (3.5) with $c_0 = 0$, in general we need a longer time interval. (However, for antisymmetric targets it is possible; see section 3.8.2.) Thus we see that controllability to all target states in (y_0, y_1) with y_0' and y_1 in $L^2(0, L)$ is *not* possible; we have only the following result.

LEMMA 3.2. *Let $p \in [2, \infty]$ and $T = L/c$. If the control functions f_1 and f_2 are in $L^p(0, T)$, then the state $y(\cdot, T)$, $y_t(\cdot, T)$ that the system has reached at time T has the following regularity: $\partial_x y(\cdot, T)$ and $y_t(\cdot, T)$ are in $L^p(0, L)$.*

For a given target state (y_0, y_1) with y_0', $y_1 \in L^p(0, L)$, there exist control functions f_1 and f_2 in $L^p(0, T)$ that steer the system to a state of the form $(y_0 + c_0, y_1)$ with a real constant c_0; moreover, these controls are uniquely determined.

The states that can be reached at the time $T = L/c$ are exactly the states y_0, y_1 with y_0', $y_1 \in L^p(0, L)$ for which the controls f_1, f_2 given in (3.17), (3.18) satisfy (3.5). In this case, f_1, f_2 given in (3.17), (3.18) are the unique solution of $C(p)$.

3.7. Controllability on larger time intervals. In this section we show how the question of controllability for a time interval $(0, T)$ with $T > L/c$ can be solved by transformation of the moment equations to moment equations on the interval $(0, L/c)$. This reduction depends on the fact that all the trigonometric functions that appear in the moment equations have the same periodicity properties.

3.7.1. Transformation of the moment equations. Assume that $T > L/c$. Choose the natural number k such that $kL/c \leq T < (k+1)L/c$.

Let the function $\varphi(s)$ be an element of the set $\{1, \sin((c\pi 2j/L)s), \cos((c\pi 2j/L)s)$ with $j \in \mathbb{N}\}$. Then we have $\varphi(s + L/c) = \varphi(s)$, and for all functions $v \in L^2(0, T)$, the following equation is valid:

$$\int_0^T v(s)\varphi(s)\,ds \tag{3.19}$$

$$= \int_0^{T-kL/c} \left[\sum_{j=0}^{k} v(s + jL/c)\right] \varphi(s)\,ds + \int_{T-kL/c}^{L/c} \left[\sum_{j=0}^{k-1} v(s + jL/c)\right] \varphi(s)\,ds.$$

Define the function

$$\hat{v}(t) = \sum_{j=0}^{k} v(t + jL/c) \text{ for } t \in (0, T - kL/c), \tag{3.20}$$

$$\hat{v}(t) = \sum_{j=0}^{k-1} v(t + jL/c) \text{ for } t \in (T - kL/c, L/c). \tag{3.21}$$

Then

$$\int_0^{L/c} \hat{v}(s)\varphi(s)\,ds = \int_0^T v(s)\varphi(s)\,ds.$$

Again, let the functions S and D be defined by (2.10) and (2.11). Then for a function S that satisfies the moment equations (3.13)–(3.15) for all $j \in \mathbb{N}$, the corresponding function $\hat{v}$ must satisfy these moment equations with integrals on the interval $(0, L/c)$. In section 3.6 we have stated that these moment equations with $T = L/c$ determine a unique solution $\hat{S}$, which is given by (3.16).

For a function D that satisfies (3.9), (3.10) for all $j \in \mathbb{N}$, the corresponding function $\hat{v}$ is defined as in (2.20)–(2.21). Since $\hat{v}$ satisfies (2.20)–(2.21) on the interval $(0, L/c)$, as stated in section 3.6, it is determined uniquely and is given by (3.11). In what follows we call it $\hat{D}$. Thus we have

$$\hat{S}(t) = (c/2)y_1^{even}(ct) - (c^2/2)\,(y_0^{even})'(ct),$$
$$\hat{D}(t) = (c/2)y_1^{odd}(ct) - (c^2/2)\,(y_0^{odd})'(ct).$$

Let $\Delta = T - kL/c$. The set of feasible controls, that is, the controls that steer the system to the target, can be described by the equations

$$\hat{S}(t) = \sum_{j=0}^{k} S(t + jL/c),\ \hat{D}(t) = \sum_{j=0}^{k} (-1)^j D(t + jL/c), \qquad t \in (0, \Delta), \tag{3.22}$$

$$\hat{S}(t) = \sum_{j=0}^{k-1} S(t + jL/c),\ \hat{D}(t) = \sum_{j=0}^{k-1} (-1)^j D(t + jL/c), \qquad t \in (\Delta, L/c), \tag{3.23}$$

and the moment equation (3.12).

This means that we have reduced our problem of optimal control to an optimization problem with a finite number of simple pointwise equality constraints and one integral constraint.

In terms of f_1 and f_2, the constraints (3.22)–(3.23) can be written as

$$\hat{S}(t) + \hat{D}(t) = \sum_{j=0}^{k} f_{b(j+1)}(t + jL/c), \qquad t \in (0, \Delta), \tag{3.24}$$

$$\hat{S}(t) - \hat{D}(t) = \sum_{j=0}^{k} f_{b(j)}(t + jL/c), \qquad t \in (0, \Delta), \tag{3.25}$$

$$\hat{S}(t) + \hat{D}(t) = \sum_{j=0}^{k-1} f_{b(j+1)}(t + jL/c), \qquad t \in (\Delta, L/c), \tag{3.26}$$

$$\hat{S}(t) - \hat{D}(t) = \sum_{j=0}^{k-1} f_{b(j)}(t + jL/c), \qquad t \in (\Delta, L/c), \tag{3.27}$$

and $f_1 + f_2$ must satisfy (3.5). To transform (3.5), we use the equation

$$\int_0^T tS(t)\,dt$$

$$= \int_0^{T-kL/c} \sum_{j=0}^{k} (t + jL/c)S(t + jL/c)\,dt + \int_{T-kL/c}^{L/c} \sum_{j=0}^{k-1} (t + jL/c)S(t + jL/c)\,dt$$

$$= \int_0^{L/c} t\hat{S}(t)\,dt + \int_0^{L/c} \overline{S}(s)\,ds,$$

with

$$\overline{S}(t) = \sum_{j=0}^{k} j(L/c)S(t + jL/c) \text{ for } t \in [0, T - kL/c], \tag{3.28}$$

$$\overline{S}(t) = \sum_{j=0}^{k-1} j(L/c)v(t + jL/c) \text{ for } t \in [T - kL/c, L/c].$$

Since the function $\hat{S}$ is known, we can replace the moment equation (3.5) by

$$\int_0^{L/c} \overline{S}(t)\,dt = \sqrt{L}y_0^0/(2c^2) - \int_0^{L/c} t\hat{S}(t)\,dt =: R_0. \tag{3.29}$$

The description of the feasible controls by the equality constraints (3.24)–(3.29) allows us to prove Theorem 3.1.

3.7.2. Proof of Theorem 3.1. Now we come to the proof of Theorem 3.1. We use the fact that f_1 and f_2 are in $L^p(0,T)$ if and only if S and D are in $L^p(0,T)$.

Let f_1 and f_2 in the space $L^p(0,T)$ be given such that at time T, the system has reached the state $y(\cdot,T) = y_0$, $y_t(\cdot,T) = y_1$. The corresponding functions $\hat{D}$ and $\hat{S}$ defined by (3.22)–(3.23) are in the space $L^p(0,L/c)$. The corresponding controls on the time interval $(0,L/c)$ reach at time L/c a state of the form $y_0 + c_0$, y_1 since they solve the moment problem (3.13)–(3.15). Lemma 3.2 implies that y_1 and y_0' are in $L^p(0,L)$.

Now we show the converse. If y_0' and y_1 are in $L^p(0,L)$, Lemma 3.2 implies that we can find $\hat{S} \in L^p(0,L/c)$ that satisfies the moment equations (3.13)–(3.15) and $\hat{D} \in L^p(0,L/c)$ that satisfies (3.9), (3.10). Then we can find functions f_1 and f_2 in $L^p(0,T)$ such that (3.24)–(3.29) hold, for example, if $T - kL/c > 0$ with the definition

$$f_1(t + L/c) = f_2(t + L/c) = R_0\ c/(L(T - kL/c)),$$

$$f_1(t) = \hat{S}(t) + \hat{D}(t) - f_2(t + L/c),\ f_2(t) = \hat{S}(t) - \hat{D}(t) - f_1(t + L/c)$$

for $t \in (0, T-kL/c)$ and $f_1(t) = \hat{S}(t)+\hat{D}(t)$, $f_2(t) = \hat{S}(t)-\hat{D}(t)$ for $t \in (T-kL/c, L/c)$ and $f_1(t) = f_2(t) = 0$ otherwise. Then the constraints (3.24)–(3.29) hold, and thus we have found a successful control in the space $L^p(0,L)$. So we have proved Theorem 3.1.

3.8. Solution of the optimization problem $C(p)$. In this section we consider the case $L/c < T < 2L/c$, that is, $k = 1$. For $p = \infty$, this case has also been considered in [10], but here we provide a solution for $p < \infty$. We work with the transformed

form of problem $C(p)$, where the feasible set $F(p)$ is described by pointwise equality constraints and one integral constraint:

$$F(p) = \{f_1, f_2 \in L^p(0,T) : (3.24)\text{–}(3.27) \text{ hold}, \overline{S} \text{ defined by } (3.28) \text{ satisfies } (3.29)\}.$$

In our case $k = 1$, (3.26), (3.27) reduce to the equations

$$f_1(t) = \hat{S}(t) + \hat{D}(t),\ f_2(t) = \hat{S}(t) - \hat{D}(t),\ t \in (T - L/c,\ L/c). \tag{3.30}$$

This means that the values of all feasible controls, and thus also of the optimal control, are prescribed on the interval $(T - L/c,\ L/c)$. This fact has important consequences for the structure of the optimal controls: We see that the functions f_1 and f_2 can have any form, so in general there is no reason why they should have a bang-bang, bang-off, or a similar structure. The optimization only takes place on the two intervals $(0, T - L/c)$ and $(L/c, T)$. The constraints (3.24), (3.25) can be written as

$$f_2(t + L/c) = \hat{S}(t) + \hat{D}(t) - f_1(t),\ f_1(t + L/c) = \hat{S}(t) - \hat{D}(t) - f_2(t) \tag{3.31}$$

for all $t \in (0, T - L/c)$. On the middle interval $(T - L/c, L/c)$ we have $\overline{S}(t) = 0$ and on $(0, T - L/c)$ we have $\overline{S}(t) = (L/c)S(t + L/c)$, so (3.29) becomes

$$\int_0^{T-L/c} (L/c)(f_1(t + L/c) + f_2(t + L/c))/2\, dt = R_0.$$

We insert (3.31) and obtain the constraint

$$\int_0^{T-L/c} (L/c) \left[\hat{S}(t) - (f_1(t) + f_2(t))/2\right]\, dt = R_0,$$

and hence for $p < \infty$, problem $C(p)$ reduces to the problem of minimizing

$$\|f_1\|^p_{p,(0,\Delta)} + \|f_2\|^p_{p,(0,\Delta)} + \|\hat{S} + \hat{D} - f_1\|^p_{p,(0,\Delta)} + \|\hat{S} - \hat{D} - f_2\|^p_{p,(0,\Delta)}$$

$$\text{s.t. } f_1, f_2 \in L^p(0,\Delta), \int_0^{\Delta} (f_1(t) + f_2(t))/2\, dt$$

$$= \int_0^{\Delta} \hat{S}(t)\, dt - (c/L)R_0 (\text{with } \Delta = T - L/c).$$

We set $G = \hat{S} + \hat{D}$, $H = \hat{S} - \hat{D}$, and $C_0 = \int_0^{\Delta} \hat{S}(t)\, dt - (c/L)R_0$. Then we can write the above optimization problem in the form

$$\min_{f_1, f_2} \|f_1\|^p_{p,(0,\Delta)} + \|f_2\|^p_{p,(0,\Delta)} + \|G - f_1\|^p_{p,(0,\Delta)} + \|H - f_2\|^p_{p,(0,\Delta)}$$

s.t. $f_1,\ f_2 \in L^p(0,\Delta),\ \int_0^{\Delta} (f_1(t) + f_2(t))/2\, dt = C_0$. The corresponding necessary optimality condition states that there exists a Lagrange multiplier $\lambda \in R$ such that for all $t \in (0,\Delta)$ the following equations hold:

$$|f_1(t)|^{p-1}\text{sign}(f_1(t)) + |f_1(t) - G(t)|^{p-1}\text{sign}(f_1(t) - G(t)) = \lambda,$$
$$|f_2(t)|^{p-1}\text{sign}(f_2(t)) + |f_2(t) - H(t)|^{p-1}\text{sign}(f_2(t) - H(t)) = \lambda.$$

For the solution of the optimality system, we use the following lemma.

LEMMA 3.3. *Let $p \in [2, \infty)$. For a real number a, define the function*

$$h_a(x) = |x|^{p-1}\text{sign}(x) + |x-a|^{p-1}\text{sign}(x-a).$$

Then h_a is strictly increasing and $\lim_{x\to\infty} h_a(x) = \infty$, $\lim_{x\to-\infty} h_a(x) = -\infty$. So the inverse function $\psi_a = h_a^{-1}$ exists and is strictly increasing with $\lim_{\lambda\to\infty} \psi_a(\lambda) = \infty$, $\lim_{\lambda\to-\infty} \psi_a(\lambda) = -\infty$, and for all $\lambda \in R$ the equation $h_a(x) = \lambda$ has the unique solution $x = \psi_a(\lambda)$. For fixed λ, the function $a \mapsto \psi_a(\lambda)$ is continuous. If $p = 2$, we have $\psi_a(\lambda) = (\lambda + a)/2$.

Proof. Consider the function $g(x) = |x|^{p-1}\text{sign}(x)$. Then g is strictly increasing and $\lim_{x\to\infty} g(x) = \infty$, $\lim_{x\to-\infty} g(x) = -\infty$. Since $h_a(x) = g(x) + g(x-a)$, the assertions for h_a follow, except the continuity of the map $a \mapsto \psi_a(\lambda)$.

Define $F : R^2 \to R$, $F(a, x) = h_a(x)$. Then F is continuously differentiable and $F_x(a, x) = g'(x) + g'(x-a) > 0$ for $(a, x) \neq (0, 0)$. We have $F(a, \psi_a(\lambda)) = \lambda$, and hence the implicit function theorem implies the continuity of the map $a \mapsto \psi_a(\lambda)$ for $(a, \lambda) \neq (0, 0)$. Since $h_a(a) = g(a)$ and $h_a(0) = -g(a)$ we have $\psi_a(0) \in (-|a|, |a|)$, so the continuity for $a = \lambda = 0$ also follows. ∎

Hence for $p < \infty$, the solution of problem $C(p)$ can be characterized in the following form.

THEOREM 3.4. *Assume that $L/c < T < 2L/c$, $p \in [2, \infty)$, and that y_1 and y_0' are in $L^p(0, L)$. Let $G = \hat{S} + \hat{D}$, $H = \hat{S} - \hat{D}$, and $C_0 = \int_0^{T-L/c} \hat{S}(t)\,dt - (c/L)R_0$. Let $\lambda \in R$ be the uniquely determined solution of the equation*

$$\int_0^{T-L/c} \psi_{G(t)}(\lambda) + \psi_{H(t)}(\lambda)\,dt = 2C_0. \tag{3.32}$$

Then the unique solution of problem $C(p)$ is given by

$$f_1(t) = \psi_{G(t)}(\lambda),\ f_2(t) = \psi_{H(t)}(\lambda)$$

for $t \in (0, T - L/c)$ and, on the interval $(L/c, T)$, the control functions f_1, f_2 are defined by (3.31) *and on $(T - L/c, L/c)$ by* (3.30).

For $p = \infty$, $C(p)$ can be reduced to the following problem: Minimize

$$\max\{\|f_1\|_{\infty,(0,T-L/c)},\ \|f_2\|_{\infty,(0,T-L/c)},\ \|G - f_1\|_{\infty,(0,T-L/c)},\ \|H - f_2\|_{\infty,(0,T-L/c)}\}$$

s.t. $f_1,\ f_2 \in L^\infty(0, T - L/c)$, $\int_0^{T-L/c} (f_1(t) + f_2(t))/2\,dt = C_0$.

Again let $\Delta = T - L/c$. It is easy to see that the functions $f_1 = f_2 = C_0/\Delta$ on $(0, \Delta)$ satisfy the integral constraint. In fact, the number C_0/Δ is a lower bound for the optimal value of $C(\infty)$. It can happen that the L^∞-norm of the control functions is attainted in the middle interval $(\Delta, L/c)$, where their values are prescribed by (3.30). These observations yield the following lemma.

LEMMA 3.5. *Assume that $L/c < T < 2L/c$ and y_1 and y_0' are in $L^\infty(0, L)$. Set $C_1 = C_0/\Delta$. Assume that $\max\{\|G - C_1\|_{\infty,(0,\Delta)}, \|H - C_1\|_{\infty,(0,\Delta)}\} \le C_1$ or that $\max\{\|G\|_{\infty,(\Delta,L/c)}, \|H\|_{\infty,(\Delta,L/c)}\} \ge \max\{\|G - C_1\|_{\infty,(0,\Delta)}, \|H - C_1\|_{\infty,(0,\Delta)}\}$.*

Then a solution of $C(\infty)$ is $f_1 = f_2 = C_1$ on $(0, \Delta)$. On the interval $(L/c, T)$, the control functions f_1, f_2 are defined by (3.31) *and on $(\Delta, L/c)$ by* (3.30).

3.8.1. Symmetric targets. In this section we assume that y_0 and y_1 are symmetric with respect to $L/2$. Then we have $\hat{D} = 0$, which implies that $G = H = \hat{S}$. Theorem 3.4 yields the equation $f_1 = f_2$, which is also true for $p = \infty$ (see [10]).

If y_0 and y_1 are constant functions, we have $\hat{S}(t) = (c/2)y_1^{even}(ct) = K$, which is also a constant function. Equation (3.32) in Theorem 3.4 yields $\psi_K(\lambda) = C_0/(T - L/c)$; hence for $t \in (0, T - L/c)$ we have $f_1(t) = C_0/(T - L/c)$, $f_1(t + L/c) = K - C_0/(T - L/c)$ and $f_1(t) = K$ for $t \in (T - L/c, L/c)$. Note that this solution is independent of $p < \infty$. Since f_1 and f_2 are in $L^\infty(0, L)$, this implies that it is also the solution of $C(\infty)$ with minimal L^2-norm. If $K = 0$ (that is, $y_1 = 0$), this yields the bang-off-bang control presented in [2] for the case $p = \infty$.

3.8.2. Antisymmetric targets. In this section we assume that y_0 and y_1 are antisymmetric with respect to $L/2$. Then we have $\hat{S} = 0$, which implies that $\hat{D} = G = -H$ and $y_0^0 = 0$, and hence $R_0 = C_0 = 0$. Since $\psi_{-a}(0) = -\psi_a(0)$, Theorem 3.4 yields with $\lambda = 0$ the equation $f_1 = -f_2$, and for $p < \infty$ the optimal control satisfies $f_1(t) = \psi_{\hat{D}(t)}(0)$ on the interval $(0, T - L/c)$. Since $h_a(a/2) = 0$, this yields $f_1 = \hat{D}/2$ on $(0, T - L/c) \cup (L/c, T)$ and $f_1 = \hat{D}$ on $(T - L/c, L/c)$. Note that this solution is again independent of $p < \infty$. If y_1 and y_0' are in $L^\infty(0, L)$, this is also the solution with minimal L^2-norm of $C(\infty)$; this follows from the next lemma.

LEMMA 3.6. *If f_1, f_2 in $L^\infty(0, T)$ solve $C(p)$ for all $p \in [2, \infty)$, then f_1, f_2 also solve $C(\infty)$ and are the solution of $C(\infty)$ with minimal L^2-norm.*

For antisymmetric targets, we have $y_0^0 = 0$. Thus in this case, for the control functions with $f_1 = -f_2$, (3.5) is valid. This implies that for this class of target states, controllability is also possible on the time interval $[0, L/c]$.

Acknowledgments. The authors thank the referees for their comments.

REFERENCES

[1] S. A. AVDONIN AND S. S. IVANOV, *Families of Exponentials*, Cambridge University Press, Cambridge, UK, 1995.

[2] J. K. BENNIGHOF AND R. L. BOUCHER, *Exact minimum-time control of a distributed system using a traveling wave formulation*, J. Optim. Theory Appl., 73 (1992), pp. 149–167.

[3] J. F. BONNANS AND A. SHAPIRO, *Perturbation Analysis of Optimization Problems*, Springer, New York, 2000.

[4] A. G. BUTKOVSKI, *The method of moments in optimal control theory with distributed parameter systems*, Avtomat. i Telemeh., 24 (1963), pp. 1217–1225.

[5] C. FABRE, J.-P. PUEL, AND E. ZUAZUA, *Contrôlabilité approchée de l'équation de la chaleur linéaire avec des contrôles de norme L^∞ minimale*, C.R. Acad. Sci. Paris, Sér. I Math., 316 (1993), pp. 679–684.

[6] C. FABRE, J.-P. PUEL, AND E. ZUAZUA, *Approximate controllability of the semilinear heat equation*. Proc. Roy. Soc. Edinburgh Sect. A, 125 (1995), pp. 31–61.

[7] R. GLOWINSKI AND J.-L. LIONS, *Exact and approximate controllability for distributed parameter systems*, Acta Numer., (1994), pp. 269–378.

[8] R. GLOWINSKI AND J.-L. LIONS, *Exact and approximate controllability for distributed parameter systems*, Acta Numer., (1995), pp. 159–333.

[9] R. GLOWINSKI, C. H. LI, AND J.-L. LIONS, *A numerical approach to the exact boundary controllability of the wave equation*. I. *Dirichlet controls: Description of the numerical methods*, Japan J. Appl. Math., 7 (1990), pp. 1–76.

[10] M. GUGAT, *Analytic solutions of L^∞-optimal control problems for the wave equation*, J. Optim. Theory Appl., 114 (2002), pp. 397–421.

[11] M. GUGAT AND G. LEUGERING, *Regularization of L^∞-optimal control problems for distributed parameter systems*, Comput. Optim. Appl., 22 (2002), pp. 151–192.

[12] M. GUGAT AND G. LEUGERING, *Solutions of L^p-norm-minimal control problems for the wave equation*, Comput. Appl. Math., 21 (2002), pp. 227–244.

[13] A. E. INGHAM, *Some trigonometrical inequalities with applications to the theory of series*, Math. Z., 41 (1936), pp. 367–379.

[14] V. I. KOROBOV, W. KRABS, AND G. M. SKLYAR, *Construction of the control realizing the rotation of a Timoshenko beam*, J. Optim. Theory Appl., 107 (2000), pp. 51–68.

[15] W. KRABS, *Optimal Control of Undamped Linear Vibrations*, Heldermann Verlag, Lemgo, 1995.

[16] W. Krabs and G. Leugering, *On boundary controllability of one-dimensional vibrating systems by $W_0^{1,p}$-controls for $p \in [2, \infty]$*, Math. Methods Appl. Sci., 17 (1994), pp. 77–93.

[17] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer, Berlin, 1971.

[18] S. Micu and E. Zuazua, *Boundary controllability of a linear hybrid systems arising in the control of noise*, SIAM J. Control Optim., 35 (1997), pp. 1614–1637.

[19] D. L. Russell, *Nonharmonic Fourier series in the control theory of distributed parameter systems*, J. Math. Anal. Appl., 18 (1967), pp. 542–560.

[20] E. Zuazua, *Exact controllability for semilinear wave equations in one space dimension*, Ann. Inst. H. Poincaré Anal. NonLinéaire, 10 (1993), pp. 109–129.

[21] E. Zuazua, *Optimal and approximate control of finite-difference approximation schemes for the $1D$ wave equation*, Rend. Mat. Appl. (7), 24 (2004), pp. 201–237.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 1, pp. 75–98

ALMOST SURE STABILIZABILITY OF CONTROLLED DEGENERATE DIFFUSIONS*

MARTINO BARDI† AND ANNALISA CESARONI†

Abstract. We develop a direct Lyapunov method for the almost sure open-loop stabilizability and asymptotic stabilizability of controlled degenerate diffusion processes. The infinitesimal decrease condition for a Lyapunov function is a new form of Hamilton–Jacobi–Bellman partial differential inequality of second order. We give local and global versions of the first and second Lyapunov theorems, assuming the existence of a lower semicontinuous Lyapunov function satisfying such an inequality in the viscosity sense. An explicit formula for a stabilizing feedback is provided for affine systems with smooth Lyapunov function. Several examples illustrate the theory.

Key words. degenerate diffusion, almost sure stability, asymptotic stability, asymptotic controllability, stabilizability, stochastic control, viability, viscosity solutions, Hamilton–Jacobi–Bellman inequalities, nonsmooth analysis

AMS subject classifications. 93E15, 49L25, 93D05, 93D20

DOI. 10.1137/S0363012903438672

1. Introduction.

For controlled diffusion processes in $\mathbb{R}^N$,

$$(CSDE)\begin{cases} dX_t = f(X_t,\alpha_t)dt + \sigma(X_t,\alpha_t)dB_t, \quad \alpha_t \in A, \quad t>0, \\ X_0 = x, \end{cases}$$

there are various possible notions of Lyapunov stability of an equilibrium, say, the origin. The stability in probability has been studied for a long time; we recall here the contributions of Kushner [31, 32], Has'minskii [26], and the recent book of Mao [36] for uncontrolled systems, and the work of Florchinger [21, 22, 23] and Deng, Krstić, and Williams [18] on feedback stabilization for $(CSDE)$; see also the references therein. The almost sure exponential stability was introduced and studied by Kozin [29] (see also [26]), and it implies that, for each fixed sample in a set of probability 1, the (uncontrolled) system is exponentially stable in the usual sense. In this paper we consider a property that we call almost sure stability, or uniform stability with probability 1. For an uncontrolled system it says that for any $\eta>0$ there exists $\delta>0$ such that, for any x with $|x|\le\delta$, the process satisfies $|X_t|\le\eta$ for all $t\ge 0$ almost surely (a.s.). Equivalently, for some increasing, continuous function γ null at 0, and for small $|x|$,

$$|X_t| \le \gamma(|x|) \quad \forall t \ge 0 \text{ a.s.} \tag{1.1}$$

This property describes a behavior very similar to a stable deterministic system. It is stronger than stability in probability and pathwise stability and, in fact, it is never verified by a nondegenerate process. More precisely, we study the *almost sure*

*Received by the editors December 11, 2003; accepted for publication (in revised form) August 18, 2004; published electronically June 27, 2005. This research was partially supported by M.I.U.R., project "Viscosity, metric, and control theoretic methods for nonlinear partial differential equations" and by GNAMPA-INDAM, project "Partial differential equations and control theory."
http://www.siam.org/journals/sicon/44-1/43867.html

†Dipartimento di Matematica P. e A., Università di Padova, via Belzoni 7, 35131 Padova, Italy (bardi@math.unipd.it, acesar@math.unipd.it).

(stochastic open-loop) stabilizability of (*CSDE*), namely, that for each x as above there exists an admissible control function whose trajectory $\overline{X}_\cdot$ verifies a.s. $|\overline{X}_t| \le \eta$ (and $|\overline{X}_t| \le \gamma(|x|)$) for all t. If, in addition, $\lim_{t\to+\infty} \overline{X}_t = 0$ a.s., we say the system is *a.s. (stochastic open-loop) asymptotically stabilizable*. For deterministic systems ($\sigma \equiv 0$) the last property reduces to the well-known *asymptotic controllability*.

We follow the Lyapunov direct method and find that the *infinitesimal decrease condition* to be satisfied by a Lyapunov function V for our problem is

$$\max_{\alpha\in A,\, \sigma(x,\alpha)^T DV(x)=0} \left\{-DV(x)\cdot f(x,\alpha) - trace\left[a(x,\alpha)D^2V(x)\right]\right\} \ge l(x), \tag{1.2}$$

with $l \ge 0$ for mere Lyapunov stability and $l > 0$ for $x \ne 0$ for asymptotic stability, where $a := \sigma\sigma^T/2$. This is not a standard Hamilton–Jacobi–Bellman inequality, because the constraint on the control α depends on V. In fact it should be viewed rather as a system of PDEs and inequalities which, in the special case of uncontrolled diffusion, i.e., $\sigma = \sigma(x)$, reads

$$\begin{cases} \max_{\alpha\in A}\{-DV(x)\cdot f(x,\alpha)\} - trace\left[a(x)D^2V(x)\right] \ge l(x), \\ \sigma_i(x)\cdot DV(x) = 0 \quad \forall\, i, \end{cases} \tag{1.3}$$

where σ_i denotes the ith column of the matrix σ. To motivate the infinitesimal decrease condition (1.3), let us give a formal argument in the case V is of class $\mathcal{C}^2$. By applying Ito's formula to the inequality $dV(X_t)/dt \le l(X_t)$, we get

$$\left[DV(X_t)\cdot f(X_t,\alpha_t) + trace\left(a(X_t)D^2V(X_t)\right)\right]dt + \sigma^T(X_t)DV(X_t)dB_t \le l(X_t).$$

Now the properties of the Brownian motion lead to the conditions

$$\begin{aligned} DV(X_t)\cdot f(X_t,\alpha_t) + trace\left(a(X_t)D^2V(X_t)\right) &\le l(X_t), \\ \sigma^T(X_t)DV(X_t) &= 0, \end{aligned}$$

and the existence of a control α_t verifying this is clearly related to (1.3). A more detailed, yet still formal, derivation of (1.3) is the following. The Dynkin formula gives, for any control,

$$\mathbf{E}V(X_t) - V(x) = \mathbf{E}\int_0^t \left[DV(X_s)\cdot f(X_s,\alpha_s) + trace(a(X_s)D^2V(X_s))\right]ds,$$

and from the inequality in (1.3) one argues the existence of a control function such that

$$\mathbf{E}V(X_t) - V(x) \le -\mathbf{E}\int_0^t l(X_s)ds \le 0. \tag{1.4}$$

Therefore, the process $V(X_t)$ is a positive supermartingale. Following this argument, it can be proved that a function satisfying merely the Hamilton–Jacobi–Bellman inequality in (1.3) is a Lyapunov function for the stability in probability. The additional equalities $\sigma_i(x)\cdot DV(x) = 0$ in (1.3) say that there is diffusion only in the directions tangential to the level sets of V, and they are necessary conditions for the invariance of the sublevel sets of V for the process (*CSDE*). It turns out that the whole set (1.3) of equalities and inequalities implies the weak invariance, or viability, of the sublevel sets of V, i.e., the existence of a control that maintains forever a.s. the system in such

a set if the initial position is in the set. From this property it is possible to infer that, for some control,

$$V(X_t) - V(x) \leq -\int_0^t l(X_s)ds \leq 0 \quad \text{almost sure},$$

a stronger monotonicity-type property than (1.4), which allows us to prove the almost sure stability.

We define a Lyapunov function for the almost sure stability as a *lower semicontinuous* proper function V, continuous at 0 and satisfying (1.2) in the *viscosity sense*, and we call it a strict Lyapunov function if $l > 0$ off 0; see Definitions 2.3 and 2.4 below. Our main results are the natural extensions of the first and second Lyapunov theorems to the controlled diffusions:

> *The existence of a local Lyapunov function implies the almost sure (open-loop) stabilizability of (CSDE); a strict Lyapunov function implies the almost sure (open-loop) asymptotic stabilizability.*

The same proof provides their global versions as well: if V satisfies (1.2) in $\mathbb{R}^N \setminus \{0\}$, then $(CSDE)$ is also a.s. (open-loop) *Lagrange stabilizable*, i.e., for all initial points x there is a control such that (1.1) holds; moreover, if V is strict, then the system is *globally* a.s. (open-loop) asymptotically stabilizable. We also give sufficient conditions for the stability of viable (controlled invariant) sets more general than an equilibrium point, and for the a.s. exponential stability.

These facts are much easier to prove when the Lyapunov function is smooth, but this assumption is not necessary and would limit considerably their applicability. The nonexistence of smooth Lyapunov functions is well known in the deterministic case; see [30, 6] for stable uncontrolled systems, and see the surveys [43, 6] for asymptotically stable controlled systems. Here we give an example of an uncontrolled degenerate diffusion process that is a.s. stable but cannot have a continuous Lyapunov function (Example 1 in section 6). Moreover, in a companion paper [12] the second author proves a *converse Lyapunov theorem*, stating that any a.s. stabilizable system $(CSDE)$ has a lower semicontinuous (l.s.c.) local viscosity Lyapunov function.

All the results listed above refer to open-loop almost sure stabilizability. They raise the question of the existence of a stabilizing feedback. Here we give an answer only for affine systems with a smooth strict Lyapunov function. We adapt Sontag's method [41] to the stochastic setting and find an explicit formula for a feedback that renders the system a.s. asymptotically stable. The feedback stabilizability of controlled diffusions in the case of nonsmooth Lyapunov functions seems considerably harder and we are not aware of any paper on the subject.

In the last section we study some simple applications and examples. For instance, we consider a deterministic, asymptotically controllable system $\dot{X}_t = f(X_t, \alpha_t)$ with Lyapunov pair (V, L) and look for conditions on a stochastic perturbation that keep the system a.s. stabilizable with the same Lyapunov function V for some $l \leq L$.

Our proof of the first Lyapunov-type theorem is based on the observation that the infinitesimal decrease condition (1.2) has the rescaling property of the geometric PDEs arising in the level set approach to front propagation (see, e.g., [9, 40] and the references therein), and on a recent result of the first author and Jensen [11] on the viability, or controlled invariance, of general closed sets for controlled diffusions (see [3, 4] and the references therein for earlier work on viability for stochastic processes). For the second Lyapunov-type theorem we use also martingale inequalities and other properties of diffusions.

The first Lyapunov-type theorem on local almost sure stabilizability was announced in [8], where we presented the simpler proof for uncontrolled processes. In the forthcoming paper [13], the second author shows that the existence of an l.s.c. viscosity solution of the Hamilton–Jacobi–Bellman inequality,

$$\max_{\alpha\in A}\left\{-DV(x)\cdot f(x,\alpha)-trace\left[a(x,\alpha)D^2V(x)\right]\right\}\geq l(x),$$

implies the open-loop stabilizability in probability of $(CSDE)$. Converse theorems in this setting appears in the Ph.D. thesis [14] of the second author.

We conclude with some additional references. Nonsmooth Lyapunov functions for uncontrolled diffusion processes were studied by Ladde and Lakshmikantham [33] with Dini-type derivatives along sample paths, and by Aubin and Da Prato [5] by means of a stochastic contingent epiderivative. Recently, Arnold and Schmalfuss [1] gave an extension of Lyapunov's second method to random dynamical systems. Turning to deterministic controlled systems, we recall that Soravia [45] gave direct and inverse Lyapunov theorems for the open-loop stabilizability by means of viscosity solutions (in the more general context of differential games); Sontag and Sussmann [41, 44] did it for the asymptotic controllability (i.e., asymptotic open-loop stabilizability) by using Dini directional derivatives. Viscosity methods for stability problems were also used in [28, 46, 24]. There is a large literature on feedback stabilization: see [2, 42, 16], the surveys [43, 15, 6], and the references therein. We refer to [17, 7] for the basic theory of viscosity solutions, and to [34, 35, 9, 20, 48] for its applications to deterministic and stochastic optimal control.

The paper is organized as follows. In section 2 we give the main definitions and state the first and second Lyapunov-type theorems. Section 3 recalls some viability theory and then gives the proofs of the two main theorems. Section 4 covers feedback stabilization of affine systems with smooth Lyapunov functions. Section 5 contains some extensions to exponential stability, general equilibrium sets, and target problems. Section 6 is devoted to the examples.

2. Lyapunov functions for almost sure stabilizability and asymptotic stabilizability. We consider a controlled Ito stochastic differential equation,

$$(CSDE)\begin{cases} dX_t=f(X_t,\alpha_t)dt+\sigma(X_t,\alpha_t)dB_t, & t>0,\\ X_0=x, \end{cases}$$

where B_t is an M-dimensional Brownian motion. Throughout the paper we assume that f,σ are continuous functions defined in $\mathbb{R}^N\times A$, where A is a compact metric space, which take values, respectively, in $\mathbb{R}^N$ and in the space of $N\times M$ matrices, and satisfy

$$|f(x,\alpha)-f(y,\alpha)|+\|\sigma(x,\alpha)-\sigma(y,\alpha)\|\leq C|x-y| \quad \forall x,y\in\mathbb{R}^N, \quad \forall\alpha\in A. \tag{2.1}$$

We adopt the definition of admissible control function, or admissible system, of Haussmann and Lepeltier [27, Def. 2.2, p. 853]. For a given $x\in\mathbb{R}^N$ we denote by $\mathcal{A}_x$ the set of admissible control functions, by $\alpha_\cdot$ its generic element (although it is not a standard function $\mathbb{R}\to A$), and by $X_\cdot$ the corresponding solution of $(CSDE)$.

We define

$$a(x,\alpha):=\frac{1}{2}\sigma(x,\alpha)\sigma(x,\alpha)^T$$

and assume

$$\{(a(x,\alpha), f(x,\alpha)) \,:\, \alpha \in A\} \quad \text{is convex } \forall x \in \mathbb{R}^N. \tag{2.2}$$

DEFINITION 2.1 (almost sure stabilizability). *The system* $(CSDE)$ *is* a.s. (stochastic open-loop Lyapunov) stabilizable *at the origin if for every* $\eta > 0$ *there exists* $\delta > 0$ *such that, for any initial point* x *with* $|x| \le \delta$, *there exists an admissible control function* $\overline{\alpha}_\cdot \in \mathcal{A}_x$ *whose corresponding trajectory* $\overline{X}_\cdot$ *verifies* $|\overline{X}_t| \le \eta$ *for all* $t \ge 0$ *a.s.*

The system is a.s. (stochastic open-loop) Lagrange stabilizable, *or it has the property of* uniform boundedness of trajectories, *if for each* $R > 0$ *there is* $S > 0$ *such that for any initial point* x *with* $|x| \le R$ *there exists an admissible control function* $\overline{\alpha}_\cdot \in \mathcal{A}_x$ *whose corresponding trajectory* $\overline{X}_\cdot$ *verifies* $|\overline{X}_t| \le S$ *for all* $t \ge 0$ *a.s.*

Remark 1. The almost sure stabilizability implies that the origin is a *controlled equilibrium* of $(CSDE)$, i.e.,

$$\exists\, \overline{\alpha} \in A \,:\, f(0,\overline{\alpha}) = 0,\ \sigma(0,\overline{\alpha}) = 0.$$

In fact, the definition gives for any $\varepsilon > 0$ an admissible control such that the corresponding trajectory starting at the origin satisfies a.s. $|X_t| \le \varepsilon$ for all t, so $\mathbf{E}_x \int_0^{+\infty} |X_t| e^{-\lambda t} dt \le \frac{\varepsilon}{\lambda}$ for any $\lambda > 0$. Then $\inf_{\alpha_\cdot \in \mathcal{A}_x} \mathbf{E}_x \int_0^{+\infty} |X_t| e^{-\lambda t} dt = 0$. The convexity assumption (2.2) and an existence theorem for optimal controls [27] imply that the inf is attained, and the minimizing control produces a trajectory satisfying a.s. $|X_t| = 0$ for all $t \ge 0$. The conclusion follows from standard properties of stochastic differential equations.

Remark 2. As is common in the modern deterministic stability theory, the previous definitions can be reformulated in terms of the *comparison functions* introduced by Hahn [25]. We will use the class $\mathcal{K}$ of continuous functions $\gamma : [0,+\infty) \to [0,+\infty)$ strictly increasing and such that $\gamma(0) = 0$ and the class $\mathcal{K}_\infty$ of functions $\gamma \in \mathcal{K}$ such that $\lim_{r\to+\infty} \gamma(r) = +\infty$.

The system $(CSDE)$ is a.s. (open-loop) stabilizable at 0 if there exists $\gamma \in \mathcal{K}$ and $\delta_o > 0$ such that for any starting point x with $|x| \le \delta_o$

$$\exists\, \overline{\alpha}_\cdot \in \mathcal{A}_x \,:\quad |\overline{X}_t| \le \gamma(|x|) \quad \forall t \ge 0 \ \text{ a.s.}, \tag{2.3}$$

where $\overline{X}_t$ is the trajectory corresponding to $\overline{\alpha}_\cdot$. If (2.3) holds for some $\gamma \in \mathcal{K}_\infty$ and for all $x \in \mathbb{R}^N$, then the system is also a.s. (open-loop) Lagrange stabilizable.

DEFINITION 2.2 (almost sure asymptotic stabilizability). *The system* $(CSDE)$ *is* a.s. (stochastic open-loop) locally asymptotically stabilizable *(or* a.s. locally asymptotically controllable*) at the origin if for every* $\eta > 0$ *there exists* $\delta > 0$ *such that, for all* $|x| \le \delta$, *there exists an admissible control function* $\overline{\alpha}_\cdot \in \mathcal{A}_x$ *whose corresponding trajectory* $\overline{X}_\cdot$ *verifies a.s.*

$$|\overline{X}_t| \le \eta \quad \forall t \ge 0, \qquad \lim_{t\to+\infty} |\overline{X}_t| = 0.$$

The system is a.s. (stochastic open-loop) globally asymptotically stabilizable *(or* a.s. asymptotically controllable*) at the origin if there is* $\gamma \in \mathcal{K}_\infty$ *and for all* $x \in \mathbb{R}^N$ *there exists* $\overline{\alpha}_\cdot \in \mathcal{A}_x$ *whose trajectory* $\overline{X}_\cdot$ *satisfies a.s.*

$$|\overline{X}_t| \le \gamma(|x|) \quad \forall t \ge 0, \qquad \lim_{t\to+\infty} |\overline{X}_t| = 0.$$

Next we give the appropriate definition of a Lyapunov function for the study of almost sure stabilizability. We recall the definition of the second order semijet of an l.s.c. function V at a point x:

$$\mathcal{J}^{2,-}V(x) := \Big\{ (p,Y) \in \mathbb{R}^N \times S(N) : \text{ for } y \to x$$
$$V(y) \geq V(x) + p \cdot (y-x) + \frac{1}{2}(y-x) \cdot Y(y-x) + o(|y-x|^2) \Big\}.$$

DEFINITION 2.3 (control Lyapunov function). *Let $\mathcal{O} \subseteq \mathbb{R}^N$ be an open set containing the origin. A function $V : \mathcal{O} \to [0, +\infty)$ is a* control Lyapunov function *for the almost sure stability of $(CSDE)$ if*

(i) *V is lower semicontinuous;*

(ii) *V is continuous at 0 and* positive definite*, i.e., $V(0) = 0$ and $V(x) > 0$ for all $x \neq 0$;*

(iii) *V is* proper*, i.e., $\lim_{|x| \to +\infty} V(x) = +\infty$ or, equivalently, the level sets $\{x | V(x) \leq \mu\}$ are bounded for every $\mu \in [0, \infty)$;*

(iv) *for all $x \in \mathcal{O} \setminus \{0\}$ and $(p,Y) \in \mathcal{J}^{2,-}V(x)$ there exists $\overline{\alpha} \in A$ such that*

$$\sigma(x,\overline{\alpha})^T p = 0 \qquad \text{and} \qquad -p \cdot f(x,\overline{\alpha}) - \mathit{trace}\,[a(x,\overline{\alpha})Y] \geq 0. \tag{2.4}$$

Remark 3. The conditions (ii) and (iii) in the previous definition can be stated as

$$\exists\, \gamma_1, \gamma_2 \in \mathcal{K}_\infty \ : \ \gamma_1(|x|) \leq V(x) \leq \gamma_2(|x|) \ \forall\, x \in \mathbb{R}^N. \tag{2.5}$$

Therefore the level sets $\{V(x) \leq \mu\}$ of the Lyapunov function form a basis of neighborhoods of 0.

Remark 4. If the dispersion matrix σ does not depend on the control, then condition (iv) can be reformulated as follows:

V is a solution in viscosity sense in $\mathcal{O} \setminus \{0\}$ of the system

$$\begin{cases} \sigma(x)^T DV(x) = 0, \\ \max_{\alpha \in A} \left\{ -DV(x) \cdot f(x,\alpha) - \mathit{trace}\left[a(x,\alpha) D^2 V(x)\right] \right\} \geq 0. \end{cases}$$

In the general case, we can observe that if condition (iv) holds, then V in particular is a viscosity supersolution of

$$\max_{\alpha \in A} \left\{ -DV(x) \cdot f(x,\alpha) - \mathit{trace}\left[a(x,\alpha) D^2 V(x)\right] \right\} = 0. \tag{2.6}$$

Moreover, if the function V is at least differentiable, then condition (iv) can be stated more concisely as follows:

V is a supersolution in viscosity sense in $\mathcal{O} \setminus \{0\}$ of the equation

$$\max_{\{\alpha \in A \ | \ \sigma(x,\alpha)^T DV(x) = 0\}} \left\{ -DV(x) \cdot f(x,\alpha) - \mathit{trace}\left[a(x,\alpha) D^2 V(x)\right] \right\} = 0.$$

DEFINITION 2.4 (strict control Lyapunov function). *A function $V : \mathcal{O} \to [0, +\infty)$ is a* strict control Lyapunov function *for the almost sure stability of $(CSDE)$ if it satisfies conditions* (i), (ii), (iii) *in Definition* 2.3 *and* (iv)$'$ *for all $x \in \mathcal{O} \setminus \{0\}$ and $(p,Y) \in \mathcal{J}^{2,-}V(x)$ there exists $\overline{\alpha} \in A$ such that*

$$\sigma^T(x,\overline{\alpha}) p = 0 \quad \text{and} \quad -p \cdot f(x,\overline{\alpha}) - \mathit{trace}\,[a(x,\overline{\alpha})Y] - l(x) \geq 0 \tag{2.7}$$

for some positive definite and Lipschitz continuous $l : \mathcal{O} \to \mathbb{R}$.

Remark 5. In the inequality in (iv)$'$ we could take

$$p \cdot f(x, \overline{\alpha}) - trace\,[a(x, \overline{\alpha})Y] - l(x, \overline{\alpha}) \geq 0$$

for some continuous $l : \mathcal{O} \times A \to \mathbb{R}$, Lipschitz continuous in x uniformly in α, with $l(x, A)$ convex for all $x \in \mathcal{O}$, and such that $\tilde{l}(x) := \min_{\alpha \in A} l(x, \alpha)$ is positive definite. However, this would not increase the generality of the definition because V would also satisfy condition (2.7) with l replaced by $\tilde{l}$.

Our main results are the following versions for stochastic controlled systems of the first and the second Lyapunov theorems.

THEOREM 2.5 (almost sure stabilizability). *Assume* (2.1), (2.2), *and the existence of a control Lyapunov function* V. *Then*

(i) *the system* $(CSDE)$ *is a.s. stabilizable at the origin;*

(ii) *if, in addition, the domain* $\mathcal{O}$ *of* V *is all* $\mathbb{R}^N$, *the system is also a.s. Lagrange stabilizable, and for all* $x \in \mathbb{R}^N$ *there exists* $\overline{\alpha}_\cdot \in \mathcal{A}_x$ *such that the corresponding trajectory* $\overline{X}_\cdot$ *satisfies*

$$|\overline{X}_t| \leq \gamma_1^{-1}(\gamma_2(|x|)) \quad \forall\, t \geq 0 \quad a.s. \tag{2.8}$$

with $\gamma_1, \gamma_2 \in \mathcal{K}_\infty$ *verifying* (2.5).

THEOREM 2.6 (almost sure asymptotic stabilizability). *Assume* (2.1), (2.2), *and the existence of a strict control Lyapunov function* V. *Then*

(i) *the system* $(CSDE)$ *is a.s. locally asymptotically stabilizable at the origin;*

(ii) *if, in addition, the domain* $\mathcal{O}$ *of* V *is all* $\mathbb{R}^N$, *the system is a.s. globally asymptotically stabilizable.*

3. A viability theorem and the proofs of stabilizability. In this section we prove Theorems 2.5 and 2.6. Our main tool is a recent result in [11] about the almost sure viability (called also *controlled invariance* and *weak invariance*) of an arbitrary closed set for a controlled diffusion process. (See [3, 4] and the references therein for earlier related results.)

DEFINITION 3.1 (viable set). *A closed set* $K \subset \mathbb{R}^N$ *is* viable *or* controlled invariant *or* weakly invariant *for the stochastic system* $(CSDE)$ *if for all initial points* $x \in K$ *there exists an admissible control* $\alpha_\cdot \in \mathcal{A}_x$ *such that the corresponding trajectory* $X_\cdot$ *satisfies* $X_t \in K$ *for all* $t > 0$ *a.s.*

It is easy to see from its definition that the almost sure stabilizability follows from the viability of all the sublevel sets of any function satisfying conditions (i)–(iii) of Definition 2.3. The next result gives a geometric characterization of viable sets. It will allow us to check that the sublevel sets of a control Lyapunov function are viable by means of condition (iv) in Definition 2.3. The Nagumo-type geometric condition in the viability theorem is given in terms of the following *second order normal cone* to a closed set $K \subset \mathbb{R}^N$, first introduced in [10]:

$$\mathcal{N}_K^2(x) := \Big\{ (p, Y) \in \mathbb{R}^N \times S(N) : \text{ for } y \to x,\ y \in K,$$
$$p \cdot (y - x) + \frac{1}{2}(y - x) \cdot Y(y - x) \geq o(|y - x|^2) \Big\},$$

where $S(N)$ is the set of symmetric $N \times N$ matrices. Note that, if $(p, Y) \in \mathcal{N}_K^2(x)$ and $x \in \partial K$, the vector p is a generalized (proximal or Bony) interior normal to the set K at x. In particular, if ∂K is a smooth surface in a neighborhood of x, $p/|p|$ is

the interior normal and Y is related to the second fundamental form of ∂K at x; see [10].

THEOREM 3.2 (viability theorem [11]). *Assume* (2.1) *and* (2.2). *Then a closed set $K \subseteq \mathbb{R}^N$ is viable for* $(CSDE)$ *if and only if*

$$\forall x \in \partial K,\ \forall (p,Y) \in \mathcal{N}_K^2(x),\ \exists \alpha \in A\ :\ \ f(x,\alpha)\cdot p + trace\,[a(x,\alpha)Y] \geq 0. \tag{3.1}$$

The second tool for the proof of the Lyapunov-type theorem, Theorem 2.5, is the following lemma on the change of unknown for second order PDEs. It says that the Hamilton–Jacobi–Bellman inequality in condition (2.4) in the definition of a control Lyapunov function behaves as a *geometric equation* if the unknown satisfies also the condition in (2.4) of orthogonality between its gradient and the columns of the dispersion matrix σ. We refer the interested reader to the chapters by Evans and Souganidis in the book [9] for an introduction to the geometric PDEs of the theory of front propagation.

LEMMA 3.3. *Let v satisfy condition* (2.4) *for all $(p,Y) \in \mathcal{J}^{2,-}V(x)$, $x \in \mathbb{R}^N \setminus \{0\}$. Let ϕ be a twice continuously differentiable strictly increasing real map. Then $w = \phi \circ v$ is a viscosity supersolution of*

$$\max_{\alpha \in A} \left\{ -DV(x)\cdot f(x,\alpha) - trace\left[a(x,\alpha)D^2V(x)\right] \right\} = 0. \tag{3.2}$$

Proof. It is easy to check that, if $(p,Y) \in \mathcal{J}^{2,-}w(x)$, then

$$(\psi'(w(x))p, \psi'(w(x))Y + \psi''(w(x))p\otimes p) \in \mathcal{J}^{2,-}v(x),$$

where ψ is the inverse of ϕ and $p \otimes p$ is the $N \times N$ matrix whose (i,j) entry is $p_i p_j$. Then, for $(p,Y) \in \mathcal{J}^{2,-}w(x)$ and $x \neq 0$ there exists $\overline{\alpha}$ such that

$$\{-\psi'(w(x))p\cdot f(x,\overline{\alpha}) - trace\,[a(x,\overline{\alpha})\cdot(\psi'(w(x))Y + \psi''(w(x))p\otimes p)]\} \geq 0$$

and

$$trace\,[a(x,\overline{\alpha})\cdot\psi''(w(x))p\otimes p] = \frac{\psi''(w(x))}{(\psi'(w(x)))^2}|\sigma(x,\overline{\alpha})^T\psi'(w(x))p|^2 = 0.$$

Therefore

$$-\psi'(w(x))p\cdot f(x,\overline{\alpha}) - trace\,[a(x,\overline{\alpha})\cdot\psi'(w(x))Y] \geq 0$$

and we can conclude that

$$\sup_{\alpha\in A}\{-p\cdot f(x,\alpha) - trace\,[a(x,\alpha)\cdot Y]\} \geq 0. \qquad \square$$

Proof of Theorem 2.5. We begin with the proof of (ii). We fix an arbitrary $\mu > 0$ and consider the sublevel set of the function V,

$$K := \{x \,|\, V(x) \leq \mu\}.$$

We claim that K is viable. Then for all initial points $x \in \mathbb{R}^N$ there exists $\overline{\alpha}_\cdot \in \mathcal{A}_x$ such that the associated trajectory $\overline{X}_\cdot$ satisfies

$$\gamma_1(|\overline{X}_t|) \leq V(\overline{X}_t) \leq V(x) \leq \gamma_1(|x|) \quad \forall t \geq 0 \quad \text{a.s.},$$

which gives estimate (2.8). Then the system is a.s. stabilizable and Lagrange stabilizable because $\gamma_1^{-1} \circ \gamma_2 \in \mathcal{K}_\infty$.

To prove that K is viable we will check condition (3.1) of the viability theorem, Theorem 3.2. For a given $\lambda > 0$ we define the nondecreasing continuous real function

$$\psi_\lambda(t) = \begin{cases} 0, & t \le \mu, \\ \lambda(t-\mu), & \mu \le t, \le \mu + \frac{1}{\lambda}, \\ 1, & t \ge \mu + \frac{1}{\lambda}. \end{cases}$$

We claim that the function $\psi_\lambda \circ V$ is a viscosity supersolution of (3.2) for every λ. To prove the claim we choose a sequence ψ_n of strictly increasing, smooth real maps that converge uniformly on compact sets to ψ_λ. Then, for every n, the map $\psi_n \circ V$ is a viscosity supersolution of (3.2) by Lemma 3.3. By the stability of viscosity supersolutions with respect to uniform convergence, we get the claim.

Next we observe that the net $\psi_\lambda \circ V$ is increasing and converges as $\lambda \to +\infty$ to the indicator function

$$C(x) = \begin{cases} 0, & x \in K, \\ 1, & x \notin K. \end{cases}$$

Viscosity supersolutions are stable with respect to the pointwise increasing convergence (see, e.g., Prop. V.2.16, p. 306 of [7]). Therefore the indicator function C of K is a viscosity supersolution of (3.2). From the definitions it is easy to check that

$$\mathcal{J}^{2,-}C(x) = -\mathcal{N}_K^2(x) \quad \forall\, x \in \partial K.$$

By plugging this formula into (3.2) we obtain exactly condition (3.1) of the viability theorem and complete the proof of (ii).

To prove (i) we choose $\overline{\mu} > 0$ small enough so that $K := \{x \in \mathcal{O} : V(x) \le \mu\}$, for $\mu \le \overline{\mu}$, is closed in $\mathbb{R}^N$ (for instance, $\overline{\mu} < \inf_{y \in \partial\mathcal{O}} \liminf_{x \to y} V(x)$). Then the preceding part of this proof gives the viability of K and the estimate (2.8) for all x such that $V(x) \le \overline{\mu}$. Therefore, for some $\delta_o > 0$, (2.8) holds for all x with $|x| \le \delta_o$, and this gives the almost sure stabilizability of the origin. □

Next we give the proof of Theorem 2.6 about asymptotic stability. It is obtained by first applying Theorem 2.5 to a new system with an extra variable and then using martingale inequalities as, e.g., in [18].

Proof of Theorem 2.6. We consider the differential system

$$\begin{cases} dX_t &= f(X_t, \alpha_t)dt &+ \sigma(X_t, \alpha_t)dB_t, \\ dZ_t &= l(X_t)dt \end{cases}$$

with initial data $X_0 = x$ and $Z_0 = 0$. We rewrite this system in $\mathbb{R}^{N+1}$ as

$$(CSDE2) \begin{cases} d(X_t, Z_t) = \overline{f}(X_t, Z_t, \alpha_t)dt + \overline{\sigma}(X_t, Z_t, \alpha_t)d(B_t, 0), \quad t > 0, \\ (X_0, Z_0) = (x, 0), \end{cases}$$

where $\overline{f}(x, z, \alpha) = (f(x, \alpha), l(x))$ and $\overline{\sigma}(x, z, \alpha) = (\sigma(x, \alpha), 0)$. Clearly it satisfies conditions (2.1) and (2.2). Let us consider the function

$$\begin{aligned} W(x, z) :& \mathcal{O} \times \mathbb{R} \to \mathbb{R}, \\ & (x, z) \longmapsto V(x) + |z|. \end{aligned}$$

We claim that it is a Lyapunov function for $(CSDE2)$. In fact, W is positive definite (because $W \geq 0$ and $W = 0$ only for $(x, z) = (0, 0)$); W is l.s.c., continuous at $(0, 0)$, and proper since V is so. We have only to prove that W satisfies condition (2.4). Fix $x \neq 0$ and (x, z) with $z > 0$ and a smooth function ϕ such that $W - \phi$ has a local minimum at (x, z), i.e.,

$$V(x) + z - \phi(x, z) \leq V(y) + w - \phi(y, w),$$

for every (y, w), $w > 0$ in a neighborhood of (x, z). If we choose $w = z$ we get a minimum in x for the function $V(\cdot) - \phi(\cdot, z)$; therefore $(D_x\phi(x, z), D^2_{xx}\phi(x, z)) \in J^{2,-}V(x)$. If we choose $y = x$ we find a minimum in z for the smooth function $w \longmapsto w - \phi(x, w)$, so $D_z\phi(x, z) = 1$. Then there exists $\overline{\alpha} \in A$ such that $(\sigma(x, \overline{\alpha}), 0)^T (D_x\phi(x, z), 1) = 0$ and

$$\left\{-D\phi(x, z) \cdot \overline{f}(x, z, \overline{\alpha}) - trace\left[\overline{a}(x, z, \overline{\alpha})D^2\phi(x, z)\right]\right\}$$

$$= \left\{-(D_x\phi(x, z), 1) \begin{pmatrix} f(x, \overline{\alpha}) \\ l(x) \end{pmatrix} - trace\left[\begin{pmatrix} a(x, \overline{\alpha}) & 0 \\ 0 & 0 \end{pmatrix} D^2\phi(x, z)\right]\right\}$$

$$= \left\{-D_x\phi(x, z) \cdot f(x, \overline{\alpha}) - trace\left[a(x, \overline{\alpha})D^2_{xx}\phi(x, z)\right]\right\} - l(x) \geq 0,$$

since V is a strict Lyapunov function. Now fix (x, z) with $z < 0$ and let ϕ be a smooth function such that

$$V(x) - z - \phi(x, z) \leq V(y) - w - \phi(y, w)$$

for every (y, w), $w < 0$ in a neighborhood of (x, z). We argue as before and now get that there exists $\overline{\alpha} \in A$ such that $(\sigma(x, \overline{\alpha}), 0)^T \cdot (D_x\phi(x, z), -1) = 0$ and

$$-D_x\phi(x, z) \cdot f(x, \overline{\alpha}) - trace\left[a(x, \overline{\alpha})D^2_{xx}\phi(x, z)\right] + l(x)$$

$$> -D_x\phi(x, z) \cdot f(x, \overline{\alpha}) - trace\left[a(x, \overline{\alpha})D^2_{xx}\phi(x, z)\right] - l(x) \geq 0$$

because l is positive and V is a Lyapunov function. Finally, we consider $(x, 0)$ and a smooth function ϕ such that

$$V(x) - \phi(x, 0) \leq V(y) - w - \phi(y, w)$$

for every (y, w), $w < 0$ in a neighborhood of $(x, 0)$ and

$$V(x) - \phi(x, 0) \leq V(y) + w - \phi(y, w)$$

for all (y, w), $w > 0$ in a neighborhood of $(x, 0)$. Then $(D_x\phi(x, z), D^2_{xx}\phi(x, z)) \in J^{2,-}V(x)$, $D_z\phi(x, 0) \geq -1$, and $D_z\phi(x, 0) \leq 1$. Therefore there exists $\overline{\alpha} \in A$ such that $(\sigma(x, \overline{\alpha}), 0)^T \cdot (D_x\phi(x, z), D_z\phi(x, z)) = 0$ and

$$\left\{-D\phi(x, z) \cdot \overline{f}(x, z, \overline{\alpha}) - trace\left[\overline{a}(x, z, \overline{\alpha})D^2\phi(x, z)\right]\right\}$$

$$= \left\{-D_x\phi(x, z) \cdot f(x, \overline{\alpha}) - trace\left[a(x, \overline{\alpha})D^2_{xx}\phi(x, z)\right]\right\} - D_z\phi(x, z)l(x)$$

$$= \left\{-D_x\phi(x, z) \cdot f(x, \overline{\alpha}) - trace\left[a(x, \overline{\alpha})D^2_{xx}\phi(x, z)\right]\right\} - l(x) \geq 0.$$

This completes the proof of the claim, so we can apply Theorem 2.5 to get for every $x \in \mathcal{O}$ an admissible control $\overline{\alpha}_\cdot \in \mathcal{A}_x$ such that the corresponding trajectory $(\overline{X}_\cdot, \overline{Z}_\cdot)$ of (*CSDE*2) with initial data $(x,0)$ remains a.s. in the level set $K = \{(y,w) \in \mathcal{O} \times \mathbb{R} \,|\, W(y,w) \le W(x,0)\}$. Then, for all $t \ge 0$ and a.s., $\overline{X}_t \in \mathcal{O}$,

$$W(\overline{X}_t, \overline{Z}_t) = V(\overline{X}_t) + \overline{Z}_t = V(\overline{X}_t) + \int_0^t l(\overline{X}_s)ds \le W(x,0) = V(x)$$

and

$$0 \le V(\overline{X}_t) \le V(x) - \int_0^t l(\overline{X}_s)ds. \tag{3.3}$$

In particular, since $l \ge 0$, for some $r > 0$, $|\overline{X}_t| \le r$ for all t a.s.

Next we claim that $l(\overline{X}_t) \to 0$ a.s. as $t \to +\infty$. Let us assume by contradiction that the claim is not true: then there exist $\varepsilon > 0$, a subset $\Omega_\varepsilon \subseteq \Omega$ with $\mathbf{P}(\Omega_\varepsilon) > 0$, and for every $\omega \in \Omega_\varepsilon$ a sequence $t_n(\omega) \to +\infty$ such that $l(\overline{X}_{t_n}(\omega)) > \varepsilon$. We define

$$F(r) := \max_{|x| \le r, \alpha \in A} |f(x,\alpha)|, \qquad \Sigma(r) := \max_{|x| \le r, \alpha \in A} \|\sigma(x,\alpha)\|.$$

We compute

$$\mathbf{E}\left\{\sup_{t \le s \le t+h} |\overline{X}_s - \overline{X}_t|^2\right\}$$

$$= \mathbf{E}\left\{\sup_{t \le s \le t+h} \left|\int_t^s f(\overline{X}_u, \overline{\alpha}_u)du + \int_t^s \sigma(\overline{X}_u, \overline{\alpha}_u)dB_u\right|^2\right\}$$

$$\le 2\mathbf{E}\left\{\sup_{t \le s \le t+h} \left|\int_t^s f(\overline{X}_u, \overline{\alpha}_u)du\right|^2\right\} + 2\mathbf{E}\left\{\sup_{t \le s \le t+h} \left|\int_t^s \sigma(\overline{X}_u, \overline{\alpha}_u)dB_u\right|^2\right\}$$

$$\le 2F^2(r)h^2 + 2\mathbf{E}\left\{\sup_{t \le s \le t+h} \left|\int_t^s \sigma(\overline{X}_u, \overline{\alpha}_u)dB_u\right|^2\right\} =: K.$$

By Theorem 3.4 in [19] (the process $|\int_t^s \sigma(\overline{X}_u, \overline{\alpha}_u)dB_u|$ is a positive semimartingale) we get

$$K \le 2F^2(r)h^2 + 8 \sup_{t \le s \le t+h} \mathbf{E}\left\{\left|\int_t^s \sigma(\overline{X}_u, \overline{\alpha}_u)dB_u\right|^2\right\}$$

and by the Ito isometry,

$$K \le 2F^2(r)h^2 + 8\mathbf{E}\left\{\int_t^{t+h} |\sigma(\overline{X}_u, \overline{\alpha}_u)|^2 du\right\} \le 2F^2(r)h^2 + 8\Sigma^2(r)h.$$

Then, the Chebyshev inequality gives

$$\mathbf{P}\left\{\sup_{t \le s \le t+h} |\overline{X}_s - \overline{X}_t| > k\right\} \le \frac{\mathbf{E}\left\{\sup_{t \le s \le t+h} |\overline{X}_s - \overline{X}_t|^2\right\}}{k^2}$$

$$\le \frac{2F^2(r)h^2 + 8\Sigma^2(r)h}{k^2}.$$

Since l is continuous, we can fix δ such that $|l(x) - l(y)| \le \frac{\varepsilon}{2}$ if $|x - y| \le \delta$ and $|x|, |y| \le r$. We define

$$C := \left\{ \omega \in \Omega \ : \ \sup_{0 \le s \le h} |\overline{X}_s - x| \le \delta \right\}$$

and choose $0 < k < \mathbf{P}(\Omega_\varepsilon)$ and $h > 0$ depending on δ and ε such that

$$\mathbf{P}_x(C) \ge 1 - \frac{2F^2(r)h^2 + 8\Sigma^2(r)h}{\delta^2} \ge 1 + k - \mathbf{P}(\Omega_\varepsilon).$$

By the uniform continuity of l, the set

$$B := \left\{ \omega \in \Omega \ : \ \sup_{0 \le s \le h} |l(\overline{X}_s) - l(x)| \le \varepsilon/2 \right\}$$

contains C and then

$$\mathbf{P}_x(B) \ge 1 + k - \mathbf{P}(\Omega_\varepsilon). \tag{3.4}$$

From inequality (3.3), letting $t \to \infty$, we get

$$V(x) \ge \mathbf{E}_x \int_0^{+\infty} l(\overline{X}_s) ds \ge \int_{\Omega_\varepsilon} \int_0^{+\infty} l(\overline{X}_s)\, ds\, d\mathbf{P} \ge \int_{\Omega_\varepsilon} \sum_n \int_{t_n(\omega)}^{t_n(\omega)+h} l(\overline{X}_s) ds\, d\mathbf{P}$$

$$\ge \int_{\Omega_\varepsilon} \sum_n h \inf_{[t_n(\omega), t_n(\omega)+h]} l(\overline{X}_t) \ge h \sum_n \int_{\Omega_\varepsilon} \inf_{[t_n(\omega), t_n(\omega)+h]} l(\overline{X}_t) d\mathbf{P}$$

$$\ge h \sum_n \frac{\varepsilon}{2} \mathbf{P}\left[\left(\sup_{0 \le s \le h} |l(\overline{X}_s) - l(x)| \le \varepsilon/2 \mid x = \overline{X}_{t_n} \right) \cap \Omega_\varepsilon \right].$$

By the properties of the solutions of (*CSDE*) estimate (3.4) gives $\mathbf{P}\left(\sup_{0 \le s \le h} |l(\overline{X}_s) - l(x)| \le \varepsilon/2 \| \ x = \overline{X}_{t_n}\right) \ge 1 + k - \mathbf{P}_x(\Omega_\varepsilon)$ for every n. Therefore $\mathbf{P}\left[\left(\sup_{0 \le s \le h} |l(\overline{X}_s) - l(x)| \le \varepsilon/2 \| \ x = \overline{X}_{t_n}\right) \cap \Omega_\varepsilon\right] \ge k$ for every n. Then by the previous inequality, we get

$$V(x) \ge h \sum_n \frac{\varepsilon}{2} k = +\infty.$$

This gives a contradiction; thus $\mathbf{P}(\Omega_\varepsilon) = 0$ for every $\varepsilon > 0$. We have proved that $l(\overline{X}_t) \to 0$ a.s. as $t \to +\infty$, now the positive definiteness of l implies that $|\overline{X}_t| \to 0$ a.s. as $t \to +\infty$. □

Remark 6. If the function l is only nonnegative semidefinite, the proof of the last theorem gives, for any x, a control $\overline{\alpha}_\cdot$ whose trajectory $\overline{X}_t$ satisfies a.s. $V(\overline{X}_t) \le V(x)$ and $l(\overline{X}_t) \to 0$ as $t \to +\infty$. Then the set $\mathcal{L} := \{y \mid l(y) = 0\}$ is an attractor, for a suitable choice of the control, in the sense that $dist(\overline{X}_t, \mathcal{L}) \to 0$ a.s. as $t \to +\infty$. For uncontrolled diffusion processes, results of this kind can be found in [37] and [18] and are considered stochastic versions of a theorem by La Salle. The earlier paper of Kushner [32] also studies a stochastic version of the La Salle invariance principle, namely, that the omega limit set of the process is an invariant subset of $\mathcal{L}$ in a suitable sense.

4. Almost sure feedback stabilization of affine systems. In this section we give a result on the *feedback stabilizability* of systems affine in the control in the case where there exists a smooth strict control Lyapunov function. It is an analogue for the almost sure stability of a celebrated theorem of Artstein [2] and Sontag [42] for deterministic systems, extended by Florchinger [21] to the stability of controlled diffusions in probability.

We begin with the simple case of a single-input affine system with uncontrolled diffusion, that is,

$$dX_t = (f(X_t) + \alpha_t g(X_t))\, dt + \sigma(X_t) dB_t, \tag{4.1}$$

where f, g, σ are vector fields in $\mathbb{R}^N$ with $f(0) = 0$ and $\sigma(0) = 0$, B_t is a one-dimensional Brownian motion, and the control α_t takes values in $\mathbb{R}$. We seek a function $k : \mathbb{R}^N \to \mathbb{R}$, at least continuous in $\mathbb{R}^N \setminus \{0\}$, such that the origin is a.s. asymptotically stable for the stochastic differential equation

$$dX_t = (f(X_t) + k(X_t) g(X_t))\, dt + \sigma(X_t) dB_t. \tag{4.2}$$

Then k is called an *a.s. asymptotically stabilizing feedback* for the control system (4.1).

If there are no constraints on the control, a smooth strict control Lyapunov function V satisfies, in $\mathbb{R}^N \setminus \{0\}$,

$$f \cdot DV + trace\left[\frac{1}{2}\sigma\sigma^T D^2 V\right] + \inf_{\alpha \in \mathbb{R}} \{\alpha g \cdot DV\} \le -l, \qquad \sigma \cdot DV = 0.$$

Set $\gamma(x) := f \cdot DV + trace\left[\sigma\sigma^T D^2 V\right]/2 + l/2$ and observe that the inequality for V means

$$g(x) \cdot DV(x) = 0 \quad \Rightarrow \quad \gamma(x) \le -l(x)/2 < 0.$$

It is clear that $k(x) := -\gamma(x)/g(x) \cdot DV(x)$, $k(x) := 0$ if $g(x) \cdot DV(x) = 0$ could be a stabilizing feedback, but it is discontinuous where $g(x) \cdot DV(x)$ vanishes. If this case occurs, we build a continuous feedback by means of Sontag's universal formula [42], i.e.,

$$k(x) := -\frac{\gamma(x) + \sqrt{\gamma^2(x) + (g(x) \cdot DV(x))^4}}{g(x) \cdot DV(x)} \quad \text{if } g(x) \cdot DV(x) \neq 0, \tag{4.3}$$

and $k(x) = 0$ if $g(x) \cdot DV(x) = 0$. By the argument in [42], $k \in C(\mathbb{R}^N \setminus \{0\})$ if $V \in C^2(\mathbb{R}^N \setminus \{0\})$ and $k \in C^1(\mathbb{R}^N \setminus \{0\})$ if f, g, l are of class C^1 and $V \in C^3(\mathbb{R}^N \setminus \{0\})$. Moreover

$$(f + kg) \cdot DV + trace\left[\frac{1}{2}\sigma\sigma^T D^2 V\right] \le -\frac{l}{2}, \qquad \sigma \cdot DV = 0,$$

in $\mathbb{R}^N \setminus \{0\}$, so V is a strict Lyapunov function for (4.2) and the origin is a.s. asymptotically stable. In conclusion, k is a stabilizing feedback for the affine control system (4.1).

If the control must satisfy a hard constraint, say $\alpha \in [-1, 1]$, it is not hard to check that $k(x)$ can be used in a neighborhood of the origin provided that DV and $D^2 V$ are bounded near 0 and either $g(x) \to 0$ or $DV(x) \to 0$ as $x \to 0$.

Next we use the same idea for the more general system with both the drift and the diffusion terms affine in the control

$$dX_t = \left(f(X_t) + \sum_{i=1}^{P-1} \alpha_t^i g_i(X_t) \right) dt + \left(\sigma(X_t) + \alpha_t^P \tau(X_t) \right) dB_t, \tag{4.4}$$

where f, g_i, σ, τ are vector fields in $\mathbb{R}^N$, B_t is a standard one-dimensional Brownian motion, and the controls α_t^i, $i = 1, \dots, P$, are $\mathbb{R}$-valued. The existence of a strict control Lyapunov function V implies that for some real number r the vector $\sigma + r\tau$ is orthogonal to DV, so $\tau \cdot DV \neq 0$ at all points where $\sigma \cdot DV \neq 0$, and we can define for all $x \in \mathbb{R}^N \setminus \{0\}$,

$$h(x) := \begin{cases} 0 & \text{if } \sigma(x) \cdot DV(x) = 0, \\ -\frac{\sigma(x) \cdot DV(x)}{\tau(x) \cdot DV(x)} & \text{if } \sigma(x) \cdot DV(x) \neq 0. \end{cases}$$

PROPOSITION 4.1. *Assume system* (4.4) *has a strict control Lyapunov function* $V \in C^2(\mathbb{R}^N \setminus \{0\})$ *and the function* h *is continuous in* $\mathbb{R}^N \setminus \{0\}$. *Then there exists continuous functions* $k_i : \mathbb{R}^N \setminus \{0\} \to \mathbb{R}$, $i = 1, \dots, P-1$, *such that* $(k_1(x), \dots, k_{P-1}(x), h(x))$ *is an a.s. asymptotically stabilizing feedback for system* (4.4).

Moreover, $k_i(x) \in [-1, 1]$ *for* x *in a neighborhood of 0 if* DV *and* D^2V *are bounded near* 0, *and either* $DV(x) \to 0$ *or* $g_i(x) \to 0$ *for all* i *as* $x \to 0$.

Proof. We recall from [42] that the function $\phi(a, 0) := 0$ for $a < 0$, $\phi(a, b) := (a + \sqrt{a^2 + b^2})/b$ is real-analytic in the set $S := \{(a, b) \in \mathbb{R}^2 : b > 0 \text{ or } a < 0\}$. We set

$$\gamma(x) := f(x) \cdot DV(x) + trace \left[(\sigma(x) + h(x)\tau(x))(\sigma(x) + h(x)\tau(x))^T \frac{D^2V(x)}{2} \right] + \frac{l(x)}{2},$$

$$\beta(x) := \sum_{i=1}^{P-1} (g_i(x) \cdot DV(x))^2.$$

Since V is a strict control Lyapunov function,

$$\gamma(x) + \inf_{\alpha_i \in \mathbb{R}} \sum_{i=1}^{P-1} \alpha_i g_i(x) \cdot DV(x) \leq -\frac{l(x)}{2},$$

so, for $x \neq 0$,

$$\beta(x) = 0 \quad \Rightarrow \quad \gamma(x) \leq -l(x)/2 < 0.$$

Therefore $(\gamma(x), \beta(x)) \in S$. Now we define, for $i = 1, \dots, P-1$,

$$k_i(x) := -\phi(\gamma(x), \beta(x)) g_i(x) \cdot DV(x), \quad x \neq 0,$$

and $k(0) = 0$. Then $(k_1(x), \dots, k_{P-1}(x), h(x))$ is continuous in $\mathbb{R}^N \setminus \{0\}$ and satisfies

$$\left(f + \sum_{i=1}^{P-1} k_i g_i \right) \cdot DV + trace \left[(\sigma + h\tau)(\sigma + h\tau)^T \frac{D^2V}{2} \right] + \frac{l}{2}$$
$$= \gamma - \beta\phi(\gamma, \beta) = -\sqrt{\gamma^2 + \beta^2} < 0.$$

Since $(\sigma + h\tau) \cdot DV = 0$ by definition of h, V is a strict Lyapunov function for the equation

$$dX_t = \left(f(X_t) + \sum_{i=1}^{P-1} k_i(X_t) g_i(X_t) \right) dt + (\sigma(X_t) + h(X_t)\tau(X_t))\, dB_t.$$

Therefore the origin is a.s. asymptotically stable for this equation.

Finally, we check the boundedness of k in a neighborhood of 0. This is trivial for $\beta(x) = 0$. If $\beta(x) \neq 0$, then

$$|k| \leq \frac{|\gamma + |\gamma| + \beta|}{\sqrt{\beta}}.$$

Since either $DV \to 0$ or $g_i \to 0$ for all i, $\beta(x) \to 0$ as $x \to 0$. We fix $\delta > 0$ such that $\beta(x) \leq \delta$ implies $\gamma(x) < 0$ and then choose a neighborhood of the origin where $\beta(x) \leq \delta$. In this set $|k(x)| \leq \sqrt{\beta(x)} \to 0$. □

Remark 7. The proof above gives an explicit formula for the stabilizing feedback in terms of the data and the Lyapunov function V only, which reduces to (4.3) if $\tau \equiv 0$ and $P = 2$. From the formula, one sees that the feedback is C^1 in $\mathbb{R}^N \setminus \{0\}$ if h, f, g, σ, τ, and l are C^1 in $\mathbb{R}^N \setminus \{0\}$ and $V \in C^3(\mathbb{R}^N \setminus \{0\})$.

Note also that the continuity assumption on h is automatically satisfied if $\tau \cdot DV$ is either always nonnull or identically 0.

Finally, it is straightforward to extend the proposition to the case of M-dimensional noise with independent Brownian components $B_t^1, \ldots, B_t^M$ and a diffusion term of the form $\sum_{i=P}^{P+M-1} (\sigma_i + \alpha_t^i \tau_i)\, dB_t^i$, with σ_i, τ_i vector fields and α_t^i scalar controls.

5. Some variants and extensions. In this section we collect several remarks on other applications of our methods. We begin with the *almost sure exponential stabilizability*. It means that there exists a positive rate λ and $\gamma \in \mathcal{K}$ such that for every initial data x there exists an admissible control $\overline{\alpha}_\cdot \in \mathcal{A}_x$ whose corresponding trajectory $\overline{X}_\cdot$ satisfies

$$|\overline{X}_t| \leq e^{-\lambda t} \gamma(|x|) \quad \text{a.s.}$$

Proposition 5.1 (almost sure exponential stabilizability). *Under assumptions* (2.1) *and* (2.2)*, the null state is a.s. exponentially stabilizable for (CSDE) if there exists a control Lyapunov function V satisfying conditions* (i), (ii), (iii) *of Definition* 2.3 *and, for some $\lambda > 0$,*

(iv)′ *for every $(p, Y) \in \mathcal{J}^{2,-}V(x)$ there exists $\overline{\alpha} \in A$ such that*

$$\sigma(x, \overline{\alpha})^T p = 0 \quad \textit{and} \quad -p \cdot f(x, \overline{\alpha}) - \textit{trace}\,[a(x, \overline{\alpha})Y] - \lambda V(x) \geq 0.$$

Proof. We consider the system

$$\begin{cases} dX_t = f(X_t, \alpha_t)dt + \sigma(X_t, \alpha_t)dB_t, \\ dY_t = dt \end{cases}$$

with initial data $X_0 = x$ and $Y_0 = 0$, and the Lyapunov function $W(x, y) = e^{\lambda y} V(x)$. By applying Theorem 2.5 we obtain the existence of a control $\overline{\alpha}_\cdot$ such that the corresponding trajectory a.s. satisfies $V(\overline{X}_t) \leq V(x)e^{-\lambda t}$, which is the desired inequality. □

Next we extend the results of section 2 to the stabilizability of a general closed set $M \subseteq \mathbb{R}^N$. We denote by $d(x, M)$ the distance between a point $x \in \mathbb{R}^N$ and M.

DEFINITION 5.2 (almost sure stabilizability at M). *The system* $(CSDE)$ *is* a.s. (stochastic open-loop) stabilizable *at* M *if there exists* $\gamma \in \mathcal{K}$ *such that, for every* x *in a neighborhood of* M*, there is an admissible control function* $\overline{\alpha}_\cdot \in \mathcal{A}_x$ *whose trajectory* $\overline{X}_\cdot$ *verifies*

$$d(\overline{X}_t, M) \le \gamma(d(x, M)) \quad \forall\, t \ge 0 \quad a.s.$$

If, in addition,

$$\lim_{t \to +\infty} d(\overline{X}_t, M) = 0 \quad a.s.,$$

the system is a.s. (stochastic open-loop) locally asymptotically stabilizable *at* M.

If these properties hold for all $x \in \mathbb{R}^N$*, the system is* a.s. (stochastic open-loop) globally asymptotically stabilizable *at* M.

Remark 8. If M is a.s. stabilizable, then it is viable for $(CSDE)$. In fact, the definition gives for $x \in M$ and $\varepsilon > 0$ an admissible control such that a.s. $d(X_t, M) \le \varepsilon$ for all $t \ge 0$.

Then for such control and any $\lambda > 0$ $\mathbf{E}_x \int_0^{+\infty} d(X_t, M) e^{-\lambda t} dt \le \frac{\varepsilon}{\lambda}$, and so

$$\inf_{\alpha_\cdot \in \mathcal{A}_x} \mathbf{E}_x \int_0^{+\infty} d(X_t, M) e^{-\lambda t} dt = 0.$$

The convexity assumption (2.2) and an existence theorem for optimal controls [27] imply that the inf is attained, and the minimizing control produces a trajectory remaining in M for all $t \ge 0$.

DEFINITION 5.3 (control Lyapunov functions at M). *Let* $\mathcal{O}$ *be an open neighborhood of the closed set* M*. A function* $V : \mathcal{O} \to [0, +\infty)$ *is a control Lyapunov function at* M *for* $(CSDE)$ *if*

(i) V *is lower semicontinuous;*

(ii) *there exists* $\gamma_1 \in \mathcal{K}_\infty$ *such that* $V(x) \le \gamma_1(d(x, M))$ *for all* $x \in \mathcal{O}$*;*

(iii) *there exists* $\gamma_2 \in \mathcal{K}_\infty$ *such that* $\gamma_2(d(x, M)) \le V(x)$ *for all* $x \in \mathcal{O}$*;*

(iv) *for all* $x \in \mathcal{O} \setminus M$ *and* $(p, Y) \in \mathcal{J}^{2,-}V(x)$ *there exists* $\overline{\alpha} \in A$ *such that condition* (2.4) *holds.*

The function V *is a* strict control Lyapunov function *at* M *if it satisfies conditions* (i)–(iii) *and*

(iv)$'$ *for some Lipschitz continuous* $l : \mathcal{O} \to \mathbb{R}$*,* $l > 0$ *for all* $x \in \mathcal{O} \setminus M$ *and* $(p, Y) \in \mathcal{J}^{2,-}V(x)$*, there exists* $\overline{\alpha} \in A$ *such that condition* (2.7) *holds.*

Now we can state the analogues of the first and second Lyapunov theorems for the almost sure stabilizability at M. Their proofs are easily obtained from the arguments of Theorems 2.5 and 2.6 by using $d(x, M)$ instead of $|x|$ and noting that conditions (ii) and (iii) in the Definition 5.3 say that the sublevel sets of the Lyapunov function form a basis of neighborhoods of M.

THEOREM 5.4. *Assume* (2.1), (2.2)*, and the existence of a control Lyapunov function* V *at* M*. Then*

(i) *the system* $(CSDE)$ *is a.s. stabilizable at* M*;*

(ii) *if, in addition, the domain* $\mathcal{O}$ *of* V *is all* $\mathbb{R}^N$*, for all* $x \notin M$ *there exists* $\overline{\alpha}_\cdot \in \mathcal{A}_x$ *such that the corresponding trajectory* $\overline{X}_\cdot$ *satisfies*

$$d(\overline{X}_t, M) \le \gamma_1^{-1}(\gamma_2(d(x, M))) \quad \forall\, t \ge 0 \quad a.s.$$

with $\gamma_1, \gamma_2 \in \mathcal{K}_\infty$ from Definition 5.3*; in particular, if M is bounded, the system is also a.s. Lagrange stabilizable.*

THEOREM 5.5. *Assume* (2.1), (2.2)*, and the existence of a strict control Lyapunov function V at M. Then*

(i) *the system $(CSDE)$ is a.s. locally asymptotically stabilizable at M;*

(ii) *if, in addition, the domain $\mathcal{O}$ of V is all $\mathbb{R}^N$, the system is a.s. globally asymptotically stabilizable at M.*

Remark 9 (stochastic target problems and absorbing sets). A stochastic target problem consists of steering the state of the system $(CSDE)$ in finite time into a given closed set $\mathcal{T}$ (the target) by an appropriate choice of the control. One of the objects of interest is the set of initial positions from which this goal can be achieved a.s. in a given time t. We define these reachability sets for $t > 0$ as

$$\mathcal{R}(t) = \{x \in \mathbb{R}^N \mid \exists \alpha_\cdot \in \mathcal{A}_x \,:\, X_t \in \mathcal{T} \text{ a.s.}\}.$$

We consider a target $\mathcal{T}$ containing 0 and being invariant for the stochastic system and we assume there exists a global strict control Lyapunov function V as defined in (2.4) such that

$$\inf_{\mathbb{R}^N \setminus \mathcal{T}} l(x) = L > 0.$$

We are going to show that each reachability set $\mathcal{R}(t)$ lies between two sublevel sets of the Lyapunov function V. The arguments in the proof of Theorem 2.6 show that for every initial point $x \notin \mathcal{T}$ there exists a control $\overline{\alpha}_\cdot \in \mathcal{A}_x$ such that the first entry time $\overline{\tau}_x$ of the corresponding trajectory in the target is a.s. bounded by

$$\overline{\tau}_x \le \left(V(x) - \inf_{\partial \mathcal{T}} V(y)\right) \Big/ L. \tag{5.1}$$

In particular, since the target $\mathcal{T}$ is invariant, it is reached a.s. in a finite time, and as such time is also uniformly bounded, $\mathcal{T}$ is an *absorbing set* for the system according to the terminology in [5]. Next, from the assumptions and inequality (5.1) we get

$$\left\{x \in \mathbb{R}^N \mid V(x) \le Lt + \inf_{\partial \mathcal{T}} V(y)\right\} \subseteq \mathcal{R}(t).$$

Using Chebyshev inequality and estimates of the same kind as in the proof of Theorem 2.6 we can find also for every $t > 0$ a positive number $k(t)$ depending continuously on t such that

$$\mathcal{R}(t) \subseteq \{x \in \mathbb{R}^N \mid V(x) \le k(t)\}.$$

Let us mention that Soner and Touzi [39] developed recently a PDE approach to stochastic target problems; see also [40] and the references therein for some interesting applications to geometric PDEs and front propagation problems.

6. Examples. We begin with an example of an uncontrolled system that does not have a continuous Lyapunov function but has an l.s.c. Lyapunov function and therefore is a.s. stable. It shows that allowing V to be merely l.s.c. in Theorem 2.5 really increases the range of the applications. Our example is a variant of a deterministic one by Krasowski [30], namely,

$$\begin{cases} \dot{X}_t = Y_t, \\ \dot{Y}_t = -X_t + Y_t (X_t^2 + Y_t^2)^3 \sin^2\left(\frac{\pi}{X_t^2 + Y_t^2}\right); \end{cases}$$

see [6] for a discussion of this and other deterministic examples.

Example 1. We transform the previous system into polar coordinates and perturb it with a white noise tangential to the circles $C_n := \{(x,y) : |(x,y)| = \frac{1}{\sqrt{n}}\}$ and nondegenerate between two consecutive circles:

$$\begin{cases} d\rho_t &= \left[\rho_t^7 \sin^2(\theta_t) \sin^2(\frac{\pi}{\rho_t^2})\right] dt + \left[\sigma(\rho_t,\theta_t) \sin^2(\frac{\pi}{\rho_t^2})\right] dB_t, \\ d\theta_t &= \left[-1 + \rho_t^6 \sin(\theta_t)\cos(\theta_t) \sin^2(\frac{\pi}{\rho_t^2})\right] dt, \end{cases}$$

where B_t is a one-dimensional Brownian motion and σ satisfies the hypotheses for the existence and uniqueness of the solution of the stochastic differential equation. As in the undisturbed case, the circles C_n are a.s. invariant and any point in C_n is eventually reached a.s. by any trajectory starting in C_n. Then any Lyapunov function V is constant on C_n because $V(\rho_t,\theta_t) \le V(\rho_0,\theta_0)$ a.s., and $c_n := V_{|C_n} \ne c_{n-1} := V_{|C_{n-1}}$ at least on a subsequence. By property (iv) in Definition 2.3 of the Lyapunov function, for every (ρ,θ) in the interior of $C_{n-1} \setminus C_n$ and every $(p,X) \in \mathcal{J}^{2,-}V(\rho,\theta)$, we get $(\sigma(\rho,\theta)\sin^2(\frac{\pi}{\rho^2}),0)\cdot p = 0$. Since the diffusion is nondegenerate in the ρ direction in the interior of $C_{n-1} \setminus C_n$, from the previous equality we deduce that, for such (ρ,θ), every element in $\mathcal{J}^{2,-}V(\rho,\theta)$ is of the form $((0,p_2),X)$. This implies that the function V is constant in the ρ direction in the interior of $C_{n-1}\setminus C_n$ and cannot be continuous.

Now we check that the Lyapunov function of the undisturbed system in the unit ball does the job also for our perturbed stochastic system. We take

$$V(\rho,\theta) := \frac{1}{\sqrt{n}} \quad \text{for } \frac{1}{\sqrt{n}} < \rho \le \frac{1}{\sqrt{n-1}} \quad \forall\theta.$$

This is a positive definite function, l.s.c. and continuous at 0. We calculate its second order subjets and plug them into (2.4). If $\rho \ne \frac{1}{\sqrt{n}}$ for all n, $(p,\mathbf{X}) \in J^{2,-}V(\rho,\theta)$ if and only if $p = 0$ and $\mathbf{X} \le 0$, so condition (2.4) is trivially satisfied. On the other hand, $(p,\mathbf{X}) \in J^{2,-}V(\frac{1}{\sqrt{n}},\theta)$ if and only if

$$p = \begin{pmatrix} s \\ 0 \end{pmatrix},\ s \ge 0, \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} a & b \\ b & c \end{pmatrix},\ c \le 0.$$

At the points with $\rho = \frac{1}{\sqrt{n}}$ the drift f of the system is $(0,-1)$ and the dispersion vector σ is $(0,0)$. Then

$$f\cdot p + \frac{1}{2} trace\left[\sigma\sigma^T\mathbf{X}\right] = 0, \qquad \sigma\cdot p = 0,$$

and condition (2.4) is satisfied. Therefore Theorem 2.5 applies and the system is a.s. Lyapunov stable at the origin.

The next two examples are about *stochastic perturbations of stabilizable systems*. We consider a deterministic controlled system in $\mathbb{R}^N$,

$$\dot{X}_t = f(X_t,\alpha_t), \tag{6.1}$$

globally asymptotically (open-loop) stabilizable at the origin, i.e., asymptotically controllable in the terminology of deterministic systems [43, 44]. By the converse Lyapunov theorem of Sontag [41, 44], there exists a strict continuous control Lyapunov

function for the system, i.e., for some positive definite continuous function L, a proper function V satisfying, in $\mathbb{R}^N \setminus \{0\}$,

$$\max_{\alpha \in A} \{-f(x,\alpha) \cdot DV\} - L(x) \geq 0 \tag{6.2}$$

in the viscosity sense. (This is perhaps not explicitly stated in the literature; the original result of Sontag [41] interprets this inequality in the sense of Dini derivatives of V along relaxed trajectories; the paper of Sontag and Sussmann [44] interprets it in the sense of directional Dini subderivatives; and both these senses are known to be equivalent to the viscosity one; see, e.g., [47, 7]).

In the following examples we perturb (6.1) in two different ways and give a condition under which V remains a control Lyapunov function for the almost sure stabilizability of the new stochastic system.

Example 2. Consider the controlled diffusion process

$$dX_t = f(X_t, \alpha)dt + \sigma(X_t)dB_t, \tag{6.3}$$

where B_t is an M-dimensional Brownian motion and σ a Lipschitzean $N \times M$ matrix. Then V *is a Lyapunov function for* (6.3) *if, for some open set* $\mathcal{O} \ni 0$ *and some continuous* $l : \mathcal{O} \to [0, +\infty)$, V *satisfies in viscosity sense in* $\mathcal{O} \setminus \{0\}$,

$$-trace\left[\frac{1}{2}\sigma\sigma^T D^2 V\right] + L - l \geq 0, \quad \sigma_i \cdot DV = 0 \quad \forall i, \tag{6.4}$$

and it is a strict Lyapunov function if l *is positive definite.*

In fact, this inequality and (6.2) give, for any $(p, \mathbf{X}) \in J^{2,-}V(x)$,

$$\max_{\alpha \in A} \{-f(x,\alpha) \cdot p\} - trace\left[\frac{1}{2}\sigma\sigma^T \mathbf{X}\right] - l \geq 0,$$

so V satisfies the inequality in condition (2.7), whereas the equality in condition (2.7) reduces to $\sigma_i \cdot p = 0$.

In the classical special case of $V(x) = |x|^2$ and $M = 1$, the sufficient condition (6.4) for V to be a Lyapunov function of (6.3) reads

$$l(x) := L(x) - |\sigma(x)|^2 \geq 0, \quad \sigma(x) \cdot x = 0.$$

For a noise of dimension $M = N$ an example of σ satisfying the orthogonality condition in (6.4) is

$$\sigma(x) = k\left(\mathbf{I} - \frac{DV(x) \otimes DV(x)}{|DV(x)|^2}\right)$$

for any constant k.

Example 3. Here we consider the perturbation of the deterministic system (6.1) by a function g of a K-dimensional diffusion process Y_t:

$$\begin{cases} \dot{X}_t = f(X_t, \alpha_t) + g(X_t, Y_t), \\ dY_t = b(Y_t, X_t, \alpha_t)dt + \tau(Y_t, X_t, \alpha_t)dB_t, \end{cases} \tag{6.5}$$

where the function $g : \mathbb{R}^n \times \mathbb{R}^K \to \mathbb{R}^n$ is Lipschitz continuous with $g(0, y) = 0$ for all y, B_t is a one-dimensional Brownian motion, and b, τ are vector fields in $\mathbb{R}^K$

with the usual assumptions. We are still assuming that (6.1) has a strict control Lyapunov function V, i.e., (6.2) holds with L positive definite. We are interested in the stabilizability of the perturbed system at the set $M := \{(x,y) \in \mathbb{R}^n \times \mathbb{R}^K \,:\, x = 0\}$, which corresponds to the origin of the unperturbed system (6.1); see Definition 5.2. Note that the assumption on g implies the viability of M for (6.5).

We claim that *the function V, defined by $V(x,y) := V(x)$ for all y, is a Lyapunov function at M for* (6.5) *(see Definition* 5.3*) if, for some open set $\mathcal{O} \ni 0$ and some continuous $l : \mathcal{O} \times \mathbb{R}^K \to [0,+\infty)$, V satisfies in viscosity sense in $\mathcal{O} \setminus \{0\}$,*

$$\inf_{y \in \mathbb{R}^K} \{-g(x,y) \cdot DV(x) - l(x,y)\} + L(x) \geq 0, \tag{6.6}$$

and V is a strict Lyapunov function if $l(x,y) > 0$ for all $x \neq 0$ and all y.

In fact, since $d((x,y),M) = |x|$, V satisfies conditions (i)–(iii) of Definition 5.3. By (6.2) and (6.6) V is also a viscosity supersolution in $\mathcal{O} \times \mathbb{R}^K \setminus M$ of

$$\sup_{a \in A} \{-f(x,a) \cdot DV(x)\} - g(x,y) \cdot DV(x) - l(x,y) \geq 0,$$

which is the inequality in (2.7) in this case, because V is constant in y. Finally, for the same reason, the condition in (2.7) of orthogonality of the diffusion vector to the level sets of V is trivially satisfied.

The inequality (6.6) is a smallness condition of the component of g in the direction of DV with respect to L in the set $\mathcal{O}$, uniformly in y. For $l \equiv 0$ and V smooth in $\mathcal{O} \setminus \{0\}$, it becomes

$$\sup_{y \in \mathbb{R}^K} g(x,y) \cdot DV(x) \leq L(x) \quad \text{in } \mathcal{O} \setminus \{0\}, \tag{6.7}$$

which is satisfied, in particular, if

$$\sup_{y \in \mathbb{R}^K} |g(x,y)| \leq L(x)/LipV,$$

where $LipV$ denotes the Lipschitz constant of V in $\mathcal{O}$. We recall that, under our assumption that the deterministic system (6.1) be asymptotically controllable, although V may not be smooth, it can be chosen semiconcave in $R^N \setminus \{0\}$ and therefore locally Lipschitz [38]. If we make this choice, it is enough that inequality (6.7) holds for all points $x \in \mathcal{O}$ where V is differentiable, and the last inequality is guaranteed for all perturbations g with small sup-norm with respect to y.

In the next two examples we give conditions on a radial function to be a Lyapunov function for almost sure stability.

Example 4. We consider as a candidate Lyapunov function for the general controlled system ($CSDE$) the function $V(x) = v(|x|)$, for some smooth $v : [0,+\infty) \to [0,+\infty)$ with $v'(r) > 0$ for $r > 0$. Since $DV(x) = xv'(|x|)/|x|$, in view of the orthogonality condition in (2.4), we restrict ourselves to controls $\alpha \in A$ such that

$$\sigma_i(x,\alpha) \cdot x = 0 \quad \forall\, i = 1, \dots, M. \tag{6.8}$$

We compute

$$trace\left[a(x,\alpha)D^2V(x)\right] = \frac{v'(|x|)}{|x|} trace\ a(x,\alpha) + \left(v''(|x|) - \frac{v'(|x|)}{|x|}\right) \frac{|\sigma(x,\alpha)^T x|^2}{|x|^2}$$

and use (6.8) to obtain that V is a Lyapunov function if and only if, in a neighborhood $\mathcal{O}$ of 0,

$$l(x) := \max_{\alpha\in A,\, \sigma(x,\alpha)^T x=0} [-f(x,\alpha)\cdot x - trace\, a(x,\alpha)]\, \frac{v'(|x|)}{|x|} \geq 0,$$

i.e.,

$$\min_{\alpha\in A,\, \sigma(x,\alpha)^T x=0} [f(x,\alpha)\cdot x + trace\, a(x,\alpha)] \leq 0. \tag{6.9}$$

This condition is independent of the choice of v. Moreover, if $l > 0$ and Lipschitz in $\mathcal{O}\setminus\{0\}$ and $l \to 0$ as $x \to 0$, then V is a strict Lyapunov function. Note that, although the radial component of the diffusion must be null by (6.8), its rotational component still plays a destabilizing role. In fact, $trace\, a(x,\alpha) \geq 0$ and whenever it is nonnull it must be compensated by a negative radial component of f.

In particular, a single-input affine system with uncontrolled diffusion and one-dimensional noise B_t,

$$dX_t = (f(X_t) + \alpha_t g(X_t))\, dt + \sigma(X_t) dB_t, \quad \alpha_t \in [-1,1],$$

has a radial Lyapunov function in $\mathcal{O}$ if and only if

$$\sigma(x)\cdot x = 0 \quad \text{and} \quad |g(x)\cdot x| \geq f(x)\cdot x + \frac{|\sigma(x)|^2}{2} \quad \text{in } \mathcal{O},$$

and $V(x) = |x|^2/2$ is a strict Lyapunov function in $\mathcal{O}$ if and only if

$$l(x) := |g(x)\cdot x| - f(x)\cdot x - \frac{|\sigma(x)|^2}{2} > 0 \quad \text{in } \mathcal{O}\setminus\{0\}.$$

Moreover, $k(x) := -sign(g(x)\cdot x)$ is a stabilizing feedback if $g(x)\cdot x$ does not change sign; if it does, k is discontinuous, and then a continuous stabilizing feedback in a neighborhood of 0 is given by the formula (4.3) in section 4.

Example 5. Here we study a system in $\mathbb{R}^2$ written in polar coordinates (ρ,θ) and look for radial Lyapunov functions, i.e., of the form $V(\rho,\theta) = v(\rho)$. Consider the stochastic controlled system:

$$(CSDE)\begin{cases} d\rho_t = f(\rho_t,\theta_t,\alpha_t)dt + \sigma(\rho_t,\theta_t,\alpha_t)dB_t, \\ d\theta_t = g(\rho_t,\theta_t,\alpha)dt + \tau(\rho_t,\theta_t,\alpha_t)dB_t, \end{cases}$$

where all functions f,σ,g,τ are 2π-periodic and B_t is (for simplicity) a one-dimensional Brownian motion. The conditions for a function $V = v(\rho)$ to be a Lyapunov function of this system at the set $M := \{(0,\theta) : \theta\in\mathbb{R}\}$ are the following. The orthogonality condition in (2.7) requires that for every (ρ,θ) there exists a subset $A(\rho,\theta) \neq \emptyset$ of the control set A such that

$$\sigma(\rho,\theta,\alpha) = 0 \qquad \forall \alpha \in A(\rho,\theta).$$

Then the condition (2.7) is satisfied if v is a viscosity supersolution of the ordinary differential inequality

$$\sup_{\alpha\in A(\rho,\theta)} \{-f(\rho,\theta,\alpha)\cdot v'(\rho)\} \geq 0$$

for $\rho > 0$ and for each fixed $\theta \in [0, 2\pi]$. Of course the same result can be obtained from the previous example with some calculations based on the Ito chain rule.

The last two examples are about the stabilization of systems to sets M different from the origin, namely, the complement of a ball and a periodic orbit.

Example 6. We consider the general system $(CSDE)$ and the set

$$M := \{x \mid |x| \geq R\} = \mathbb{R}^N \setminus B_R.$$

We assume M is viable for the system. We take the radial function V

$$V(x) := \begin{cases} R^2 - |x|^2 & |x| < R, \\ 0 & |x| \geq R \end{cases}$$

and use the calculations of Example 4 to see that V is a Lyapunov function at M if and only if for every x with $|x| < R$ there exists $\overline{\alpha} \in A$ such that

$$\sigma_i(x, \overline{\alpha}) \cdot x = 0 \quad \forall i \quad \text{and} \quad f(x, \overline{\alpha}) \cdot x + trace\ a(x, \overline{\alpha}) \geq 0.$$

Contrary to Example 4, here the rotational component of the diffusion has a stabilizing effect. In fact, the drift $f(x, a)$ is allowed also to point away from M if its negative radial component is compensated by the positive term $trace\ a(x, \overline{\alpha})$.

If $K \subset B_R$ is a compact set and

$$l(x) := \max_{\alpha \in A,\, \sigma(x,\alpha)^T x = 0} [f(x, \alpha) \cdot x + trace\, a(x, \alpha)] > 0 \quad \text{in } B_R \setminus K,$$

then M is locally asymptotically stable by Theorem 5.5, and for all initial points $x \notin K$ there is a control whose trajectories tend a.s. to M as $t \to +\infty$. In this case we can say that K can be made a.s. repulsive by a suitable choice of the controls. In particular, we have a criterion of instability of an equilibrium point.

Note also that if $l > 0$ on $\partial M = \partial B_R$, then for some control the trajectories starting in a suitable neighborhood of ∂M reach M in finite time a.s., as we observed in the last remark of section 5. In particular, if $l > 0$ in $\overline{B_R}$, then for every $x \in B_R$ there exists a control $\overline{\alpha}_\cdot$ such that the exit time of the corresponding trajectory $\overline{X}_\cdot$ from B_R is a.s. bounded by $(R^2 - |x|^2)/\min_{B_R} l$.

Example 7. Consider $(CSDE)$ in $\mathbb{R}^2$ and assume the circle $\gamma := \{x : |x| = R\}$ is a viable set. By the results of [11] this occurs if for all $x \in \gamma$ there exists $\overline{\alpha} \in A$ such that

$$\sigma(x, \overline{\alpha}) \cdot x = 0 \quad \text{and} \quad f(x, \overline{\alpha}) \cdot x + trace\, a(x, \overline{\alpha}) = 0.$$

Then γ is locally asymptotically stabilizable if, in a neighborhood $\{x : R - \varepsilon \leq |x| \leq R + \varepsilon\}$,

$$\max_{\alpha \in A,\, \sigma(x,\alpha)^T x = 0} [f(x, \alpha) \cdot x + trace\ a(x, \alpha)] > 0 \quad \text{if } |x| < R,$$

$$\min_{\alpha \in A,\, \sigma(x,\alpha)^T x = 0} [f(x, \alpha) \cdot x + trace\ a(x, \alpha)] < 0 \quad \text{if } |x| > R.$$

This follows immediately from the arguments of Examples 4 and 6.

REFERENCES

[1] L. Arnold and B. Schmalfuss, *Lyapunov's second method for random dynamical systems*, J. Differential Equations, 177 (2001), pp. 235–265.

[2] Z. Artstein, *Stabilization with relaxed controls*, Nonlinear Anal., 7 (1983), pp. 1163–1173.

[3] J.-P. Aubin, *Viability Theory*, Birkäuser, Boston, 1991.

[4] J.-P. Aubin and G. Da Prato, *The viability theorem for stochastic differential inclusions*, Stochastic Anal. Appl., 16 (1998), pp. 1–15.

[5] J.-P. Aubin and G. Da Prato, *Stochastic Lyapunov method*, NoDEA Nonlinear Differential Equations Appl., 2 (1995), pp. 511–525.

[6] A. Bacciotti and L. Rosier, *Liapunov Functions and Stability in Control Theory*, Lecture Notes in Control and Inform. Sci. 267, Springer-Verlag, London, 2001.

[7] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton–Jacobi–Bellman Equations*, Birkäuser, Boston, 1997.

[8] M. Bardi and A. Cesaroni, *Viscosity Lyapunov functions for almost sure stability of degenerate diffusions*, in Elliptic and Parabolic Problems (Rolduc/Gaeta, 2001), J. Bemelmans et al., eds., World Scientific, River Edge, NJ, 2002, pp. 322–331.

[9] M. Bardi, M. G. Crandall, L. C. Evans, M. H. Soner, and P. E. Souganidis, *Viscosity Solutions and Applications*, Lecture Notes in Math. 1660, Springer-Verlag, Berlin, 1997.

[10] M. Bardi and P. Goatin, *Invariant sets for controlled degenerate diffusions: A viscosity solutions approach*, in Stochastic Analysis, Control, Optimization, and Applications: A Volume in Honor of W. H. Fleming, W. M. McEneaney, G. G. Yin, and Q. Zhang, eds., Birkhäuser, Boston, 1999, pp. 191–208.

[11] M. Bardi and R. Jensen, *A geometric characterization of viable sets for controlled degenerate diffusions*, Set-Valued Anal., 10 (2002), pp. 129–141.

[12] A. Cesaroni, *A Converse Lyapunov Theorem for Almost Sure Stabilizability*, Preprint, Dipartimento di Matematica Pura e Applicata, University of Padova, Padova, Italy, 2004. Available online at http://cpde.iac.rm.cnr.it/preprint.php

[13] A. Cesaroni, *Lyapunov Stabilizability of Controlled Diffusions via a Superoptimality Principle for Viscosity Solutions*, Preprint, Dipartimento di Matematica Pura e Applicata, University of Padova, Padova, Italy, 2004. Available online at http://cpde.iac.rm.cnr.it/preprint.php

[14] A. Cesaroni, *Stability Properties of Controlled Diffusion Processes Via Viscosity Methods*, Ph.D. thesis, University of Padova, Padova, Italy, 2004.

[15] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth Analysis and Control Theory*, Springer-Verlag, New York, 1998.

[16] F. H. Clarke, Yu. S. Ledyaev, E. D. Sontag, and A. I. Subbotin, *Asymptotic controllability implies feedback stabilization*, IEEE Trans. Automat. Control, 42 (1997), pp. 1394–1407.

[17] M. C. Crandall, H. Ishii, and P. L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27 (1992), pp. 1–67.

[18] H. Deng, M. Krstić, and R. J. Williams, *Stabilization of stochastic nonlinear systems driven by noise of unknown covariance*, IEEE Trans. Automat. Control, 46 (2001), pp. 1237–1253.

[19] J. L. Doob, *Stochastic Processes*, John Wiley & Sons, New York, 1953.

[20] W. H. Fleming and H. M. Soner, *Controlled Markov Process and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[21] P. Florchinger, *Lyapunov-like techniques for stochastic stability*, SIAM J. Control Optim., 33 (1995), pp. 1151–1169.

[22] P. Florchinger, *Feedback stabilization of affine in the control stochastic differential systems by the control Lyapunov function method*, SIAM J. Control Optim., 35 (1997), pp. 500–511.

[23] P. Florchinger, *A stochastic Jurdjevic–Quinn theorem*, SIAM J. Control Optim., 41 (2002), pp. 83–88.

[24] L. Grüne, *Asymptotic behavior of dynamical and control systems under perturbation and discretization*, Lecture Notes in Mathematics 1783, Springer-Verlag, Berlin, 2002.

[25] W. Hahn, *Stability of Motion*, Springer-Verlag, New York, 1967.

[26] R. Z. Has'minskii, *Stochastic Stability of Differential Equations*, Sijthoff and Noordhoff, Alphen aan den Rijn—Germantown, MD, 1980.

[27] U. G. Haussmann and J. P. Lepeltier, *On the existence of optimal controls*, SIAM J. Control Optim., 28 (1990), pp. 851–902.

[28] M. Kocan and P. Soravia, *Lyapunov functions for infinite-dimensional systems*, J. Funct. Anal., 192 (2002), pp. 342–363.

[29] F. Kozin, *On almost sure asymptotic sample properties of diffusion processes defined by stochastic differential equation*, J. Math. Kyoto Univ., 4 (1964/1965), pp. 515–528.

[30] N. N. Krasowski, *The converse of the theorem of K. P. Persidskij on uniform stability*, Prik. Mat. i Meh., 19 (1955), pp. 273–278, in Russian.
[31] H. J. Kushner, *Stochastic Stability and Control*, Academic Press, New York, 1967.
[32] H. J. Kushner, *Stochastic stability*, in Stability of Stochastic Dynamical Systems, Proceedings of the International Symposium, University of Warwick (Coventry, 1972), Lecture Notes in Math. 294, Springer, Berlin, 1972, pp. 97–124.
[33] G. S. Ladde and V. Lakshmikantham, *Random Differential Inequalities*, Academic Press, New York, 1980.
[34] P.-L. Lions, *Optimal control of diffusion processes and Hamilton–Jacobi–Bellman equations. Part* 1: *The dynamic programming principle and applications*, Comm. Partial Differential Equations 8 (1983), pp. 1101–1174.
[35] P.-L. Lions, *Optimal control of diffusion processes and Hamilton–Jacobi–Bellman equations. Part* 2: *Viscosity solutions and uniqueness*, Comm. Partial Differential Equations 8 (1983), pp. 1229-1276.
[36] X. Mao, *Exponential Stability of Stochastic Differential Equations*, Marcel Dekker, New York, 1994.
[37] X. Mao, *Stochastic versions of the LaSalle theorem*, J. Differential Equations, 153 (1999), pp. 175–195.
[38] L. Rifford, *Existence of Lipschitz and semiconcave control-Lyapunov functions*, SIAM J. Control Optim., 39 (2000), pp. 1043–1064.
[39] H. M. Soner and N. Touzi, *Stochastic target problems, dynamic programming, and viscosity solutions*, SIAM J. Control Optim., 41 (2002), pp. 404–424.
[40] H. M. Soner and N. Touzi, *A stochastic representation for mean curvature type geometric flows*, Ann. Probab., 31 (2003), pp. 1145–1165.
[41] E. D. Sontag, *A Lyapunov-like characterization of asymptotic controllability*, SIAM J. Control Optim., 21 (1983), pp. 462–471.
[42] E. D. Sontag, *A "universal" construction of Artstein's theorem on nonlinear stabilization*, Systems Control Lett., 13 (1989), pp. 117–123.
[43] E. D. Sontag, *Stability and stabilization: Discontinuities and the effect of disturbances*, in Nonlinear Analysis, Differential Equations, and Control (Montreal, QC, 1998), F. H. Clarke and R. J. Stern, eds., Kluwer, Dordrecht, 1999, pp. 551–598.
[44] E. D. Sontag and H. J. Sussmann, *Non smooth control Lyapunov functions*, in Proceedings of the IEEE Conference on Decision and Control (New Orleans, 1995), IEEE Publications, Piscataway, NJ, 1995, pp. 2799–2805.
[45] P. Soravia, *Stability of dynamical systems with competitive controls: The degenerate case*, J. Math. Anal. Appl., 191 (1995), pp. 428–449.
[46] P. Soravia, *Feedback stabilization and H-infinity control of nonlinear systems affected by disturbances*, in Dynamics, Bifurcations, and Control (Kloster Irsee, 2001), Lecture Notes in Control and Inform. Sci. 273, Springer, Berlin, 2002, pp. 173–190.
[47] A. I. Subbotin, *Generalized Solutions of First-order PDEs*, Birkäuser, Boston, 1995.
[48] J. Yong and X. Y. Zhou, *Stochastic Controls, Hamiltonian Systems and HJB Equations*, Springer-Verlag, New York, 1999.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 1, pp. 99–110

THE APPROXIMATION OF HIGHER-ORDER INTEGRALS OF THE CALCULUS OF VARIATIONS AND THE LAVRENTIEV PHENOMENON*

ALESSANDRO FERRIERO†

Abstract. We prove the following approximation theorem: given a function $x : [a, b] \to \mathbb{R}^N$ in the Sobolev space $\mathbf{W}^{\nu+1,1}$, $\nu \geq 1$, and $\epsilon > 0$, there exists a function x_ϵ in $\mathbf{W}^{\nu+1,\infty}$ such that

$$\int_a^b \sum_{i=1}^m L_i(x_\epsilon^{(\nu)}, x_\epsilon^{(\nu+1)})\psi_i(t, x_\epsilon, x_\epsilon', \dots, x_\epsilon^{(\nu)}) < \int_a^b \sum_{i=1}^m L_i(x^{(\nu)}, x^{(\nu+1)})\psi_i(t, x, x', \dots, x^{(\nu)}) + \epsilon,$$

$$\begin{aligned} &x_\epsilon(a) = x(a), && x_\epsilon(b) = x(b),\\ &x_\epsilon'(a) = x'(a), && x_\epsilon'(b) = x'(b),\\ &\vdots\\ &x_\epsilon^{(\nu)}(a) = x^{(\nu)}(a), && x_\epsilon^{(\nu)}(b) = x^{(\nu)}(b), \end{aligned}$$

provided that, for every i in $\{1, \dots, m\}$, $L_i\psi_i$ is continuous in a neighborhood of x, L_i is convex in its second variable, and ψ_i evaluated along x has positive sign. We discuss the optimality of our assumptions comparing them with an example of Sarychev [*J. Dynam. Control Systems*, 3 (1997), pp. 565–588].

As a consequence, we obtain the nonoccurrence of the Lavrentiev phenomenon. In particular, the integral functional $\int_a^b L(x^{(\nu)}, x^{(\nu+1)})$ does not exhibit the Lavrentiev phenomenon for any given boundary values $x(a) = A$, $x(b) = B$, $x'(a) = A'$, $x'(b) = B'$, $\dots$, $x^{(\nu)}(a) = A^{(\nu)}$, $x^{(\nu)}(b) = B^{(\nu)}$.

Furthermore, we prove the following necessary condition: an action functional with Lagrangian of the form $\sum_{i=1}^m L_i(x^{(\nu)}, x^{(\nu+1)})\psi_i(t, x, x', \dots, x^{(\nu)})$, with $\nu \geq 0$, exhibiting the Lavrentiev phenomenon takes the value $+\infty$ in any neighborhood of a minimizer.

Key words. calculus of variations, Lavrentiev phenomenon, reparameterization

AMS subject classifications. 49J30, 49N45, 49N60

DOI. 10.1137/S0363012903437721

1. Introduction.

In 1926, Lavrentiev [11] proposed an example of a first-order integral functional of the calculus of variations, $\mathcal{I}(x) = \int_a^b L(t, x, x')$, whose infimum taken over the space of the absolutely continuous functions $\mathbf{W}^{1,1}(a, b)$ is strictly less than the infimum taken over the space of Lipschitz continuous functions $\mathbf{W}^{1,\infty}(a, b)$, with $x(a) = A$ and $x(b) = B$. Later, Manià [13] published a simpler example of the same phenomenon where the Lagrangian is

$$L_1(x')\psi_1(t, x) = |x'|^6(x^3 - t)^2.$$

Several papers have been devoted to the problem of finding conditions under which the Lavrentiev phenomenon does not occur: Angell [2], Clarke, Vinter [8], Ball, Mizel [3], Lowen [12], Alberti, Serra Cassano [1]. In a recent paper by Cellina, Ferriero, and Marchini [5] a large class of Lagrangians of the form $L_1(x, x')\psi_1(t, x)$ has been treated, including the autonomous and some nonautonomous cases, under no additional conditions besides the convexity of L_1 in x' and the positivity of ψ_1.

*Received by the editors November 14, 2003; accepted for publication (in revised form) October 9, 2004; published electronically June 27, 2005.

http://www.siam.org/journals/sicon/44-1/43772.html

†Dipartimento di Matematica e Applicazioni, Universitá degli Studi di Milano-Bicocca,Via R. Cozzi 53, 20126 Milano, Italy (ferriero@matapp.unimib.it).

Besides the first-order case, the Lavrentiev phenomenon occurs as well in the case with $(\nu+1)$-order derivatives, $\mathcal{I}(x) = \int_a^b L(t, x, x', \dots, x^{(\nu+1)})$. For $\nu = 1$, in 1994 Cheng and Mizel [7] described a restricted Lavrentiev phenomenon in which the gap occurs for a dense subset of the absolutely continuous nonnegative functions, and they proved that even autonomous Lagrangian $L(x, x', x'')$ can exhibit it. Some years later Sarychev [15] proved that a class of Lagrangians of the form

$$L_1(x'')\psi_1(x, x') + L_2(x'')$$

exhibits the Lavrentiev phenomenon provided that $\psi_1(x, x') = \phi(kx - k|x' - 1|^{k-1} - (k-1)|x' - 1|^k)$ for appropriate constants k, that L_1, L_2, ϕ satisfy certain growth conditions, and that $\phi(0) = 0$. For example, $L_1(x'') = |x''|^7$, $L_2(x'') = \alpha|x''|^{3/2}$, $\phi_1(\cdot) = (\cdot)^2$, $k = 3$, and $\alpha > 0$ sufficiently small yield a Lagrangian whose integral exhibits the Lavrentiev phenomenon when the boundary values are $x(0) = 0$, $x(1) = 5/3$, $x'(0) = 1$, $x'(1) = 2$.

The Lagrangians proposed by Manià and Sarychev have the property that L_1 evaluated along the minimizer x is not integrable (this is possible because there exists at least one point t in $[a, b]$ such that ψ_1 evaluated along x in t is 0). A condition avoiding the occurrence of this fact will turn out, in this paper, to be essential for the nonoccurrence of the Lavrentiev phenomenon.

We prove the following general approximation theorem: let $x : [a, b] \to \mathbb{R}^N$ be a function in $\mathbf{W}^{\nu+1,1}$ (independently on whether is a minimizer or not), then the integrability of L_i evaluated along x (or the assumption that $\psi_i > 0$), for every i, implies that, given $\epsilon > 0$, there exists a function x_ϵ in $\mathbf{W}^{\nu+1,\infty}$ with the same boundary values of x in a and in b, i.e., $x_\epsilon(a) = x(a)$, $x_\epsilon(b) = x(b)$, $x'_\epsilon(a) = x'(a)$, $x'_\epsilon(b) = x'(b)$, $\dots$, $x^{(\nu)}_\epsilon(a) = x^{(\nu)}(a)$, $x^{(\nu)}_\epsilon(b) = x^{(\nu)}(b)$, such that

$$\begin{aligned}&\int_a^b \sum_{i=1}^m L_i(x^{(\nu)}_\epsilon, x^{(\nu+1)}_\epsilon)\psi_i(t, x_\epsilon, x'_\epsilon, \dots, x^{(\nu)}_\epsilon)\\ &< \int_a^b \sum_{i=1}^m L_i(x^{(\nu)}, x^{(\nu+1)})\psi_i(t, x, x', \dots, x^{(\nu)}) + \epsilon.\end{aligned}$$

We underline that an application of this result is the nonoccurrence of the Lavrentiev phenomenon for a class of functionals of the calculus of variations with $(\nu+1)$-order derivatives, $\nu \geq 1$. (The case $\nu = 0$, $m = 1$ has already been treated in [5]. The case $\nu = 0$, $m > 1$ can be obtained modifying slightly the proof of the main result of [5]; see [10].) Moreover, we infer a necessary condition for the Lavrentiev phenomenon.

In section 2 we state our results, we discuss the optimality of the assumptions, and we infer the nonoccurrence of the Lavrentiev phenomenon. In section 3 we prove the main result. In section 4 we deal with a necessary condition for the Lavrentiev phenomenon: a functional

$$\mathcal{I}(x) = \int_a^b \sum_{i=1}^m L_i(x^{(\nu)}, x^{(\nu+1)})\psi_i(t, x, x', \dots, x^{(\nu)}),$$

with $\nu \geq 0$, exhibiting the Lavrentiev phenomenon takes the value $+\infty$ in any neighborhood of a minimizer $\bar{x}$.

2. The main result and the Lavrentiev phenomenon. For $\delta > 0$, $B[c, \delta]$ denotes the closed ball in $\mathbb{R}^N$ centered in c with radius δ. For a function x in $\mathbf{C}^\nu[a, b]$,

with values in $\mathbb{R}^N$, the closed δ-tube along $(x, \dots, x^{(\nu)})$

$$\mathsf{T}^{\nu}_{\delta}[x] = \{(t, z_0, \dots, z_\nu) \in [a,b] \times \mathbb{R}^{(\nu+1)N} : \\ (z_0, \dots, z_\nu) \in B[x(t), \delta] \times \cdots \times B[x^{(\nu)}(t), \delta],\ t \in [a,b]\}$$

and the closed δ-neighborhood of the image $\mathrm{Im}(x^{(\nu)})$ of $x^{(\nu)}$

$$\mathsf{I}_{\delta}[x^{(\nu)}] = \{z \in \mathbb{R}^N : \mathrm{dist}(z, \mathrm{Im}(x^{(\nu)})) \le \delta\}$$

are compact sets.

We recall that the space $\mathbf{W}^{\nu+1,p}(a,b)$ can be seen as the space of functions x in $\mathbf{C}^{\nu}[a,b]$ such that $x^{(\nu)}$ is absolutely continuous with derivative in $\mathbf{L}^p(a,b)$, $p \ge 1$.

The following approximation theorem is our main result.

THEOREM 2.1. *Let x be a function in $\mathbf{W}^{\nu+1,1}(a,b)$, $\nu \ge 1$, and let the real-valued functions $L_1, \dots, L_m$ and $\psi_1, \dots, \psi_m$ be continuous on $\mathsf{I}_{\delta}[x^{(\nu)}] \times \mathbb{R}^N$ and on $\mathsf{T}^{\nu}_{\delta}[x]$, respectively, for some $\delta > 0$.*

Assume that, for every i in $\{1, \dots, m\}$,

- *$L_i(\xi, \cdot)$ is convex, for every ξ in $\mathsf{I}_{\delta}[x^{(\nu)}]$,*
- *ψ_i is nonnegative, and $\psi_i(t, x(t), x'(t), \dots, x^{(\nu)}(t)) > 0$, for every t in $[a,b]$.*

Then

(i) $\mathcal{I}(x) = \displaystyle\int_a^b \sum_{i=1}^{m} L_i(x^{(\nu)}(t), x^{(\nu+1)}(t)) \psi_i(t, x(t), x'(t), \dots, x^{(\nu)}(t))\, dt > -\infty$;

(ii) *given any $\epsilon > 0$, there exists a function x_ϵ in $\mathbf{W}^{\nu+1,\infty}(a,b)$ such that*

$$\mathcal{I}(x_\epsilon) < \mathcal{I}(x) + \epsilon,$$

and

$$\begin{aligned} &x_\epsilon(a) = x(a), && x_\epsilon(b) = x(b), \\ &x'_\epsilon(a) = x'(a), && x'_\epsilon(b) = x'(b), \\ &\vdots \\ &x^{(\nu)}_\epsilon(a) = x^{(\nu)}(a), && x^{(\nu)}_\epsilon(b) = x^{(\nu)}(b). \end{aligned}$$

As a corollary we obtain the nonoccurrence of the Lavrentiev phenomenon.

THEOREM 2.2. *Let $\Omega_0, \dots, \Omega_\nu$ be open sets in $\mathbb{R}^N$, $\nu \ge 1$, such that the set $E = \{x \in \mathbf{W}^{\nu+1,1}(a,b) : x(t) \in \Omega_0, \dots, x^{(\nu)}(t) \in \Omega_\nu\ \forall t \in [a,b]\}$ is nonempty.*

Let $L_1, \dots, L_m : \Omega_\nu \times \mathbb{R}^N \to \mathbb{R}$ and $\psi_1, \dots, \psi_m : [a,b] \times \Omega_0 \times \cdots \times \Omega_\nu \to (0, +\infty)$ be continuous and such that $L_i(\xi, \cdot)$ is convex, for any ξ in Ω_ν, and any i in $\{1, \dots, m\}$.

Then, for all boundary values $A, B \in \Omega_0$, $A^{(1)}, B^{(1)} \in \Omega_1$, $\dots$, $A^{(\nu)}, B^{(\nu)} \in \Omega_\nu$, the infimum of

$$\mathcal{I}(x) = \int_a^b \sum_{i=1}^{m} L_i(x^{(\nu)}(t), x^{(\nu+1)}(t)) \psi_i(t, x(t), x'(t), \dots, x^{(\nu)}(t))\, dt$$

over the space $E_{a,b} = \{x \in E : x(a) = A,\ x(b) = B,\ x'(a) = A^{(1)},\ x'(b) = B^{(1)},\ \dots,\ x^{(\nu)}(a) = A^{(\nu)},\ x^{(\nu)}(b) = B^{(\nu)}\}$ is equal to the infimum of the same functional $\mathcal{I}$ over the space $E_{a,b} \cap \mathbf{W}^{\nu+1,\infty}(a,b)$.

Proof. Let $\{x_n\}_n \subset E_{a,b}$ be a minimizing sequence for $\mathcal{I}$: by the fact that $\psi_i > 0$, for every i, the theorem follows from Theorem 2.1 applied to any x_n, with $\epsilon = 1/n$. ∎

Setting $m = 1$, $\psi_1 = 1$, and $L_1 = L$, we obtain that a Lagrangian depending only on $x^{(\nu)}$ and $x^{(\nu+1)}$ satisfies the assumptions of Theorem 2.2. Hence, the integral functional

$$\int_a^b L(x^{(\nu)}(t), x^{(\nu+1)}(t))dt$$

does not exhibit the Lavrentiev phenomenon, for any boundary values

$$\begin{aligned} &x(a) = A, && x(b) = B,\\ &x'(a) = A^{(1)}, && x'(b) = B^{(1)},\\ &\vdots\\ &x^{(\nu)}(a) = A^{(\nu)}, && x^{(\nu)}(b) = B^{(\nu)}. \end{aligned}$$

This extends some previous results ([1], [4]), where functionals without boundary conditions, or with boundary conditions only in a, have been considered.

We point out that the assumption $\psi_i(t, x(t), x'(t), \dots, x^{(\nu)}(t)) \neq 0\ \forall t \in [a, b]$ in Theorem 2.1 will be used only to infer that $\int_a^b L_i(x^{(\nu)}, x^{(\nu+1)})$ is finite, provided that $\mathcal{I}(x)$ is finite (point (a) in the proof). The theorem holds under the weaker assumption $\int_a^b |L_i(x^{(\nu)}, x^{(\nu+1)})| < +\infty$, for every i.

To verify how sharp our assumptions are, consider the following example of A. V. Sarychev [15]: for $\nu = 1$, $m = 1$, minimize the functional

$$\int_0^1 |x''(t)|^7[3x(t) - 3|x'(t) - 1|^2 - 2|x'(t) - 1|^3]^2 dt,$$

with boundary conditions $x(0) = 0$, $x(1) = 5/3$, $x'(0) = 1$, $x'(1) = 2$. He proved that the infimum taken over the space $\mathbf{W}^{2,1}(0, 1)$, assumed in $\bar{x}(t) = (2/3)\sqrt[2]{t^3} + t$, is strictly lower than the infimum taken over the space $\mathbf{W}^{2,\infty}(0, 1)$.

The assumption $\int_a^b |L_1(x', x'')| < +\infty$ along $\bar{x}$ is not verified. Indeed, setting $\psi_1(t, x, \xi) = [3x - 3|\xi - 1|^2 - 2|\xi - 1|^3]^2$ and $L_1(\xi, w) = |w|^7$, we see that $\psi_1 \geq 0$ (but, for example, $\psi_1(0, x(0), x'(0)) = 0$) and that

$$\int_0^1 |\bar{x}''(t)|^7 dt = \int_0^1 \frac{1}{(2\sqrt{t})^7} dt = +\infty.$$

3. Proof of the main theorem. In what follows, $\mathbf{x}$ denotes the matrix $(x, \dots, x^{(\nu-1)})$ and $\mathsf{x} = x^{(\nu-1)}$, so that $\mathsf{x}' = x^{(\nu)}$, $\mathsf{x}'' = x^{(\nu+1)}$ (similarly, $\mathbf{z} = (z, \dots, z^{(\nu-1)})$, and $\mathsf{z} = z^{(\nu-1)}$). The Lagrangian we consider takes the form

$$\sum_{i=1}^m L_i(\mathsf{x}'(t), \mathsf{x}''(t))\psi_i(t, \mathbf{x}(t), \mathsf{x}'(t)).$$

(In case $\nu = 1$, $\mathbf{x}$, x', x'' coincide with x, x', x'', respectively.)

(i) For every $t \in [a, b]$, $L_i(\mathsf{x}'(t), \mathsf{x}''(t)) \geq L_i(\mathsf{x}'(t), 0) + \langle p_0(t), \mathsf{x}''(t)\rangle$, where $p_0(t)$ is any selection from the subdifferential $\partial_w L_i(\mathsf{x}'(t), 0)$ of L_i with respect to its second variable. Set $E_i = \{t \in [a, b] : [L_i(\mathsf{x}'(t), \mathsf{x}''(t))]^- \neq 0\}$, so that

$$\begin{aligned} &\int_a^b [L_i(\mathsf{x}'(t), \mathsf{x}''(t))\psi_i(t, \mathbf{x}(t), \mathsf{x}'(t))]^- dt\\ &\leq -\int_{E_i} [L_i(\mathsf{x}'(t), 0) + \langle p_0(t), \mathsf{x}''(t)\rangle]\psi_i(t, \mathbf{x}(t), \mathsf{x}'(t))dt, \end{aligned}$$

for any i. Since ψ_i is bounded and, by Proposition 2 in [5], $p_0(t)$ is bounded, the claim follows by Hölder's inequality.

(ii) Fix $\epsilon > 0$; set $\bar\epsilon = \epsilon/m$. Without loss of generality, we shall assume $\epsilon < 1$, and also $\delta < 1$.

In case $\int_a^b L_i(\mathsf{x}'(t), \mathsf{x}''(t))\psi_i(t, \mathbf{x}(t), \mathsf{x}'(t))dt = +\infty$, for some i, any Lipschitz function x_ϵ satisfying the boundary conditions is acceptable. Hence we can assume, for every i,

$$\int_a^b L_i(\mathsf{x}'(t), \mathsf{x}''(t))\psi_i(t, \mathbf{x}(t), \mathsf{x}'(t))dt < +\infty.$$

The proof is in three steps. In Step (1) of the proof we introduce the new functions $\tilde L_i$ such that $\tilde L_i = L_i + \text{const}$ and such that their polar functions $\tilde L_i^*$ (with respect to the second variable) are nonnegative. In Step (3) we define a variation z_n in $\mathbf{W}^{\infty,1}(a,b)$, with the same boundary values of x in a and in b, such that $\mathcal{I}(z_n) < \mathcal{I}(x) + \epsilon$. In order to define z_n, in Step (2) we define a sequence of reparameterizations s_n of $[a,b]$.

Step (1). We claim that there exists functions $\tilde L_i$ and a constant η such that $\tilde L_i = L_i + \eta$ and $\tilde L_i^* \geq 0$, for any i.

In fact, consider the set

$$V_i = \{(\xi, p) : \xi \in \mathsf{I}_\delta[x^{(\nu)}],\, p \in \partial_w L_i(\xi, w),\, |w| \leq 1\}.$$

By Proposition 2 in [5], arguing by contradiction, we obtain that V_i is compact. Let $L_i^*(\xi, p) = \sup_{w\in\mathbb{R}^N} \langle p, w\rangle - L_i(\xi, w)$ be the polar function of L_i with respect to its second variable. Then, $\min_{V_i} L_i^*$ is attained and is finite. Applying Proposition 3 in [5], we obtain that $L_i^*(\xi, p) \geq \min_{V_i} L_i^*$, for every $\xi \in \mathsf{I}_\delta[x^{(\nu)}]$, for every $p \in \partial_w L_i(\xi, w)$ and for every $w \in \mathbb{R}^N$. Set $\eta = \min\{\min_{V_1} L_1^*, \dots, \min_{V_m} L_m^*\}$.

Consider $\tilde L_i(\xi, w) = L_i(\xi, w) + \eta$. Since $\partial_w L_i(\xi, w) = \partial_w \tilde L_i(\xi, w)$, we have that $\tilde L_i^*(\xi, p) \geq 0$, for any i. (We denote $\tilde{\mathcal{I}}_i$ the functional $\int_a^b \tilde L_i \psi_i$.)

(a) We set some preliminary constants, depending on $\bar\epsilon$ fixed, that we shall use in the following steps.

By the condition on ψ_i, there exists $c > 0$ such that $\psi_i(t, \mathbf{x}(t), \mathsf{x}'(t)) \geq c$, for every t in $[a,b]$, and we obtain

$$\begin{aligned} +\infty &> \int_a^b |L_i(\mathsf{x}'(t), \mathsf{x}''(t))\psi_i(t, \mathbf{x}(t), \mathsf{x}'(t))|dt + \eta \int_a^b \psi_i(t, \mathbf{x}(t), \mathsf{x}'(t))dt \\ &\geq \int_a^b |\tilde L_i(\mathsf{x}'(t), \mathsf{x}''(t))\psi_i(t, \mathbf{x}(t), \mathsf{x}'(t))|dt \geq c \int_a^b |\tilde L_i(\mathsf{x}'(t), \mathsf{x}''(t))|dt. \end{aligned}$$

Set $\ell_i = \int_a^b |\tilde L_i(\mathsf{x}'(s), \mathsf{x}''(s))|ds$, $\ell = \max\{\ell_1, \dots, \ell_m\}$, and Ψ and $\tilde{\mathbf{L}}$ the maximum value of $|\psi_1|, \dots, |\psi_m|$ over $\mathsf{T}_\delta^\nu[x]$ and of $|\tilde L_1|, \dots, |\tilde L_m|$ over $\mathsf{I}_\delta[x^{(\nu)}] \times B[0, |\mathsf{x}''(\tau)| + \delta]$, respectively. Denote $\alpha = \max\{1, (b-a)^\nu\}$.

From the uniform continuity of $\psi_1, \dots, \psi_m$ on $\mathsf{T}_\delta^\nu[x]$, we infer that we can fix $h \in \mathbb{N}$, $1/2^h < \delta$, such that whenever $(t_1, \mathbf{x}_1, \xi_1)$, $(t_2, \mathbf{x}_2, \xi_2) \in \mathsf{T}_\delta^\nu[x]$ and

$$|t_1 - t_2| \leq \frac{b-a}{2^h}, \quad |\mathbf{x}_{1,j} - \mathbf{x}_{2,j}| \leq \frac{1}{2^h}\ \forall j \in \{0, \dots, \nu-1\}, \quad |\xi_1 - \xi_2| \leq \frac{1}{2^h},$$

we have

$$|\psi_i(t_1, \mathbf{x}_1, \xi_1) - \psi_i(t_2, \mathbf{x}_2, \xi_2)| < \min\left\{\frac{\bar\epsilon}{8(\ell + \tilde{\mathbf{L}} + 1)},\, \frac{\bar\epsilon}{2(|\eta| + 1)(b-a)}\right\},$$

for any i.

Let $\theta : \mathbb{R} \to [0,1]$ be a $\mathbf{C}^\infty$ increasing function with value 0 on $(-\infty, 0]$ and 1 on $[1, +\infty)$. Observe that $1 \leq ||\theta^{(j)}||_\infty \leq ||\theta^{(j+1)}||_\infty$, for any $j \geq 0$. Set $\Theta = ||\theta^{(\nu+1)}||_\infty$.

There exists a point τ in (a,b) that is a Lebesgue point for the functions $\tilde{L}_1(\mathsf{x}'(\cdot), \mathsf{x}''(\cdot))\psi_1(\cdot, \mathbf{x}(\cdot), \mathsf{x}'(\cdot))$, $\ldots$, $\tilde{L}_m(\mathsf{x}'(\cdot), \mathsf{x}''(\cdot))\psi_m(\cdot, \mathbf{x}(\cdot), \mathsf{x}'(\cdot))$ and x'', $\mathsf{x}''(\tau)$ in $\mathbb{R}^N$. By definition of Lebesgue point, there exists a positive number ρ less than

$$\min\left\{\frac{1}{2^{h+4}(\nu+2)(\nu+1)\nu\Theta\alpha^2}, \frac{\bar{\epsilon}}{32\tilde{\mathbf{L}}\Psi}\right\}$$

such that, for any λ^-, λ^+ in $(0, \rho)$,

$$\int_{\tau-\lambda^-}^{\tau+\lambda^+} |\tilde{L}_i(\mathsf{x}'(t), \mathsf{x}''(t))\psi_i(t, \mathbf{x}(t), \mathsf{x}'(t)) - \tilde{L}_i(\mathsf{x}'(\tau), \mathsf{x}''(\tau))\psi_i(\tau, \mathbf{x}(\tau), \mathsf{x}'(\tau))|dt \leq (\lambda^+ + \lambda^-)\bar{\epsilon}$$

for any i, and

$$\int_{\tau-\lambda^-}^{\tau+\lambda^+} |\mathsf{x}''(t) - \mathsf{x}''(\tau)|dt \leq (\lambda^+ + \lambda^-)\frac{1}{2^{h+4}(\nu+1)\nu\Theta\alpha}.$$

Fix $t_0^- = (b-a)v^-/2^\gamma$, $t_0^+ = (b-a)v^+/2^\gamma$, where $\gamma \in \mathbb{N}$, $v^-, v^+ \in \{0, 1, \ldots, 2^\gamma\}$, $v^- < v^+$, are such that $\tau \in (\tau^-, \tau^+) \subset (\tau - \rho, \tau + \rho)$.

We define the absolutely continuous function $\mathsf{z}' : [a,b] \to \mathbb{R}^N$ by $\mathsf{z}'(t) = x^{(\nu)}(a) + \int_a^t \mathsf{z}''$, where

$$\mathsf{z}''(t) = \begin{cases} x^{(\nu+1)}(\tau) + \dfrac{1}{\tau^+ - \tau^-}\displaystyle\int_{\tau^-}^{\tau^+} [\mathsf{x}'' - \mathsf{x}''(\tau)], & t \in [\tau^-, \tau^+], \\ x^{(\nu+1)}(t), & \text{otherwise.} \end{cases}$$

By definition, $\mathsf{z}''(t) = \mathsf{x}''(t)$, $\mathsf{z}'(t) = \mathsf{x}'(t)$, for any t in $[a, \tau^-] \cup [\tau^+, b]$. For any t in $[\tau^-, \tau^+]$, we have that $\mathsf{z}''(t) \in B[0, |\mathsf{x}''(\tau)| + \delta/2]$ and

$$|\mathsf{z}'(t) - \mathsf{x}'(t)| \leq 2\int_{\tau^-}^{\tau^+} |\mathsf{x}''(\tau) - \mathsf{x}''| < (\tau^+ - \tau^-)\frac{1}{2^{h+3}(\nu+1)\nu\Theta\alpha}.$$

Step (2). Our purpose is to show that there exists a sequence of reparameterizations s_n of $[a,b]$ into itself such that $\mathsf{z}' \circ s_n$ is Lipschitz continuous on $[a,b]$.

From the uniform continuity of $x, \ldots, x^{(\nu)}$ on $[a, \tau^-] \cup [\tau^+, b]$, we infer that we can fix $k \in \mathbb{N}$, such that whenever $|s_1 - s_2| \leq (b-a)/2^k$, we have $|x^{(j)}(s_1) - x^{(j)}(s_2)| < (\tau^+ - \tau^-)^{\nu+2}$, for any j in $\{1, \ldots, \nu\}$.

For $v = 0, \ldots, 2^k - 1$, set $I_v = [(b-a)v/2^k, (b-a)(v+1)/2^k]$, $H_v = \int_{I_v} |\mathsf{z}''(s)|ds$, $\mu = \max\{2^{k+1}H_v/(b-a) : v = 0, \ldots, 2^k - 1\}$, and

$$T_{H_v} = \left\{s \in I_v : |\mathsf{z}''(s)| \leq \frac{2^{k+1}H_v}{b-a}\right\};$$

we have that $|T_{H_v}| \geq (b-a)/2^{k+1}$.

Since $\{(\mathsf{z}'(s), \mathsf{z}''(s)) : s \in \bigcup_{v=0}^{2^k-1} T_{H_v}\}$ belongs to a compact set and $L_1, \dots, L_m$ are continuous, there exists a constant M, such that

$$\left| \tilde{L}_i(\mathsf{z}'(s) + \xi, 2\mathsf{z}''(s) + w)\frac{1}{2} - \tilde{L}_i\Big(\mathsf{z}'(s) + \xi, \mathsf{z}''(s) + \frac{w}{2}\Big)\right| \leq M,$$

for any $s \in \bigcup_{v=0}^{2^k-1} T_{H_v}$, any $|\xi| \leq \delta$, any $|w| \leq \delta$, and any i.

For every $n \in \mathbb{N}$, set $S_n^v = \{s \in I_v : |\mathsf{z}''(s)| > n\}$. From the integrability of z'' it follows that $\int_{S_n^v} (|\mathsf{z}''(s)|/n - 1)ds$ converges to 0, as n goes to ∞. Hence, we can fix a subset Σ_n^v of T_{H_v} such that $|\Sigma_n^v| = 2\int_{S_n^v} (|\mathsf{z}''(s)|/n - 1)ds$.

We define the absolutely continuous functions t_n by $t_n(s) = a + \int_a^s t_n'$, where

$$t_n'(s) = \begin{cases} 1 + (|\mathsf{z}''(s)|/n - 1), & s \in S_n = \bigcup_{v=0}^{2^k-1} S_n^v, \\ 1 - 1/2, & s \in \Sigma_n = \bigcup_{v=0}^{2^k-1} \Sigma_n^v, \\ 1, & \text{otherwise.} \end{cases}$$

One verifies that t_n admits inverse function s_n on the interval $[a, b]$. Furthermore, for any v in $\{0, \dots, 2^k - 1\}$, the restriction of t_n to I_v maps I_v onto itself. Hence, $|t_n(s) - s| \leq (b - a)/2^k$, for any s in $[a, b]$. If n is greater than $|\mathsf{x}''(\tau)| + \delta/2$, the restriction of t_n to $[\tau^-, \tau^+]$ is the identity.

The function $\mathsf{z}' \circ s_n$ is Lipschitz continuous on $[a, b]$. In fact, fix t where $s_n'(t)$ exists: we obtain

$$\left|\frac{d(\mathsf{z}' \circ s_n)}{dt}(t)\right| = |\mathsf{z}''(s_n(t))s_n'(t)| \begin{cases} = n, & t \in S_n, \\ \leq \mu, & t \in \Sigma_n, \\ \leq n, & \text{otherwise.} \end{cases}$$

Step (3). We construct a function $z_n : [a, b] \to \mathbb{R}^N$, with the same boundary values of x in a and b, such that z_n belongs to $\mathbf{W}^{\nu+1,\infty}(a, b)$ and $\tilde{\mathcal{I}}_i(z_n) < \tilde{\mathcal{I}}_i(x) + \bar{\epsilon}/2$.

Set $f'(t) = \theta((t - \tau^-)/(\tau^+ - \tau^-))$, for any t in $[a, b]$ (the function θ as defined in point (a)): then f' is identically 0 on $[a, \tau^-]$, it is identically 1 on $[\tau^+, b]$, and $||f^{(j+1)}||_\infty = ||\theta^{(j)}||_\infty/(\tau^+ - \tau^-)^j$, for any $j \geq 0$.

We define ν absolutely continuous functions $z_{n,\nu-1}, \dots, z_{n,0} : [a, b] \to \mathbb{R}^N$ by

$$\begin{aligned} z_{n,\nu-1}(t) &= x^{(\nu-1)}(a) + \int_a^t \mathsf{z}' \circ s_n + f'(t)D_{\nu-1}, \\ z_{n,\nu-2}(t) &= x^{(\nu-2)}(a) + \int_a^t z_{n,\nu-1} + f'(t)D_{\nu-2}, \\ &\vdots \\ z_{n,0}(t) &= x(a) + \int_a^t z_{n,1} + f'(t)D_0, \end{aligned}$$

where, for any j in $\{0, \cdots, \nu - 2\}$,

$$D_j = x^{(j)}(b) - x^{(j)}(a) - \int_a^b z_{n,j+1}, \qquad D_{\nu-1} = x^{(\nu-1)}(b) - x^{(\nu-1)}(a) - \int_a^b \mathsf{z}' \circ s_n.$$

Set $z_n = z_{n,0}$. The derivatives of z_n up to the order $\nu+1$ are

$$\begin{aligned}
&z_n'(t) = z_{n,1}(t) + f''(t)D_0,\\
&z_n''(t) = z_{n,2}(t) + f'''(t)D_0 + f''(t)D_1,\\
&\vdots\\
&z_n^{(\nu-1)}(t) = z_{n,\nu-1}(t) + \textstyle\sum_{j=0}^{\nu-2} f^{(\nu-j)}(t)D_j,\\
&z_n^{(\nu)}(t) = \mathsf{z}'(s_n(t)) + \textstyle\sum_{j=0}^{\nu-1} f^{(\nu-j+1)}(t)D_j,\\
&z_n^{(\nu+1)}(t) = \mathsf{z}''(s_n(t))s_n'(t) + \textstyle\sum_{j=0}^{\nu-1} f^{(\nu-j+2)}(t)D_j.
\end{aligned}$$

We denote by H' the function $\sum_{j=0}^{\nu-1} f^{(\nu-j+1)}D_j$. By the properties of $f^{(j)}$ and s_n, we have that z_n belongs to $\mathbf{W}^{\nu+1,\infty}(a,b)$, with $||z_n^{(\nu+1)}||_\infty \le n + ||\mathsf{H}''||_\infty$ (where $||\cdot||_\infty$ is the essential supremum on (a,b)), and it has the same boundary values of x in a and b.

(b) We claim that $||z_n^{(j)} - x^{(j)}||_\infty \le 1/2^h$ and $||z_n^{(j)} \circ t_n - x^{(j)}||_\infty \le 1/2^h$, for any j in $\{0,\dots,\nu\}$, eventually in n.

In fact, for any n greater than $|\mathsf{x}''(\tau)| + \delta/2$, we have

$$\begin{aligned}
|D_{\nu-1}| &\le \int_a^{\tau^-} |\mathsf{x}' - \mathsf{x}'\circ s_n| + \int_{\tau^-}^{\tau^+} |\mathsf{x}' - \mathsf{z}'| + \int_{\tau^+}^b |\mathsf{x}' - \mathsf{x}'\circ s_n|\\
&\le (\tau^+ - \tau^-)^2\left[3\alpha(\tau^+-\tau^-)^\nu + \frac{1}{2^{h+3}(\nu+1)\nu\Theta\alpha}\right]\\
&\le (\tau^+ - \tau^-)^2 \frac{1}{2^{h+2}(\nu+1)\nu\Theta\alpha},\\
|D_{\nu-2}| &\le \int_a^{\tau^+}\left|\mathsf{x}'(t) - \mathsf{x}'(a) - \int_a^t \mathsf{z}'\circ s_n - f'(t)D_{n,\nu-1}\right| dt\\
&\quad + \int_{\tau^+}^b \left|\mathsf{x}'(t) - \mathsf{x}'(b) + \int_t^b \mathsf{z}'\circ s_n - [1-f'(t)]D_{n,\nu-1}\right| dt\\
&\le \int_a^{\tau^+}\int_a^t |\mathsf{x}' - \mathsf{z}'\circ s_n| dt + (\tau^+-\tau^-)|D_{n,\nu-1}| + \int_{\tau^+}^b\int_t^b |\mathsf{x}' - \mathsf{z}'\circ s_n| dt\\
&\le (\tau^+-\tau^-)^3\left[4\alpha(\tau^+-\tau^-)^{\nu-1} + \frac{1}{2^{h+3}(\nu+1)\nu\Theta\alpha}\right] + (\tau^+-\tau^-)|D_{\nu-1}|\\
&\le (\tau^+-\tau^-)^3 \frac{2}{2^{h+2}(\nu+1)\nu\Theta\alpha},\\
&\vdots\\
|D_j| &\le (\tau^+-\tau^-)^{\nu-j+1}\frac{\nu-j}{2^{h+2}(\nu+1)\nu\Theta\alpha} \le (\tau^+-\tau^-)^{\nu-j+1}\frac{1}{2^{h+2}(\nu+1)\Theta}\\
&\qquad\qquad \forall j\in\{0,\dots,\nu-1\},
\end{aligned}$$

so that $||\mathsf{H}'||_\infty \le \sum_{j=0}^{\nu-1} ||f^{(\nu-j+1)}||_\infty(\tau^+-\tau^-)^{\nu-j+1}/[2^{h+2}(\nu+1)\Theta] \le (\tau^+-\tau^-)/2^{h+2}$,

$||\mathsf{H}''||_\infty \le 1/2^{h+2}$, and

$$\begin{aligned}
|\mathsf{z}'(s_n(t)) - \mathsf{x}'(t)| &\le (\tau^+ - \tau^-)\left[3\alpha(\tau^+ - \tau^-)^{\nu+1} + \frac{1}{2^{h+3}(\nu+1)\nu\Theta\alpha}\right] \\
&\le \frac{1}{2^{h+2}(\nu+1)\nu\Theta\alpha}, \\
|z_{n,\nu-1}(t) - x^{(\nu-1)}(t)| &\le \int_a^b |\mathsf{z}' \circ s_n - \mathsf{x}'| + (b-a)|D_{\nu-1}| \le (1+b-a)|D_{\nu-1}| \\
&\le \frac{2\alpha}{2^{h+2}(\nu+1)\nu\Theta\alpha}, \\
&\vdots \\
|z_{n,j}(t) - x^{(j)}(t)| &\le \frac{(\nu-j+1)\alpha}{2^{h+2}(\nu+1)\nu\Theta\alpha} \le \frac{1}{2^{h+2}} \qquad \forall j \in \{0,\dots,\nu-1\}.
\end{aligned}$$

Hence, we can fix n such that $M\Psi|\Sigma_n| < \bar{\epsilon}/8$, $||z_n^{(j)} - x^{(j)}||_\infty \le 1/2^{h+1}$, and

$$||z_n^{(j)} \circ t_n - x^{(j)}||_\infty \le ||z_n^{(j)} \circ t_n - x^{(j)} \circ t_n||_\infty + ||x^{(j)} \circ t_n - x^{(j)}||_\infty \le 1/2^h,$$

for any j in $\{0,\dots,\nu\}$. The graph of the function $(\mathbf{z}_n, \mathsf{z}_n')$ is included in $\mathsf{T}_\delta^\nu[x]$, and $\mathsf{z}_n''(t) \in B[0, |\mathsf{x}''(\tau)| + \delta]$, for any t in $[\tau^-, \tau^+]$. (From what follows, it turns out that z_n is the sought variation x_ϵ.)

(c) We show that $\tilde{\mathcal{I}}_i(z_n) < \tilde{\mathcal{I}}_i(x) + \bar{\epsilon}/2$, for any i.

Using the change of variable formula [16], we compute $\tilde{\mathcal{I}}_i(z_n) - \tilde{\mathcal{I}}_i(x)$ as the sum of the following three appropriate terms:

$$\begin{aligned}
&\int_a^b \tilde{L}_i(\mathsf{z}_n'(t_n(s)), \mathsf{z}_n''(t_n(s)))\psi_i(t_n(s), \mathbf{z}_n(t_n(s)), \mathsf{z}_n'(t_n(s)))t_n'(s)ds \\
&- \int_a^b \tilde{L}_i(\mathsf{x}'(s), \mathsf{x}''(s))\psi_i(s, \mathbf{x}(s), \mathsf{x}'(s))ds \\
&= \int_a^b \left[\tilde{L}_i\left(\mathsf{z}_n'(t_n(s)), \mathsf{z}_n''(t_n(s))\right)t_n'(s) - \tilde{L}_i(\mathsf{z}_n'(t_n(s)), \mathsf{z}''(s) + t_n'(s)\mathsf{H}''(t_n(s)))\right] \\
&\qquad \times \psi_i(t_n(s), \mathbf{z}_n(t_n(s)), \mathsf{z}_n'(t_n(s)))ds \\
&\quad + \int_a^b \tilde{L}_i(\mathsf{z}_n'(t_n(s)), \mathsf{z}''(s) + t_n'(s)\mathsf{H}''(t_n(s))) \\
&\qquad \times [\psi_i(t_n(s), \mathbf{z}_n(t_n(s)), \mathsf{z}_n'(t_n(s))) - \psi_i(s, \mathbf{x}(s), \mathsf{x}'(s))]ds \\
&\quad + \int_a^b \left[\tilde{L}_i\left(\mathsf{z}_n'(t_n(s)), \mathsf{z}''(s) + t_n'(s)\mathsf{H}''(t_n(s))\right) - \tilde{L}_i(\mathsf{x}'(s), \mathsf{x}''(s))\right]\psi_i(s, \mathbf{x}(s), \mathsf{x}'(s))ds \\
&= I_i^1 + I_i^2 + I_i^3.
\end{aligned}$$

To estimate I_i^1, it is enough to estimate its integrand over the sets S_n and Σ_n (because it is identically 0 elsewhere). Since $\Sigma_n \subset T$ and $||\mathsf{H}''||_\infty \le \delta$, we obtain that

$$\begin{aligned}
&\tilde{L}_i\left(\mathsf{z}'(s) + \mathsf{H}'(t_n(s)), 2\mathsf{z}''(s) + \mathsf{H}''(t_n(s))\right)\frac{1}{2} \\
&- \tilde{L}_i\left(\mathsf{z}'(s) + \mathsf{H}'(t_n(s)), \mathsf{z}''(s) + \frac{\mathsf{H}''(t_n(s))}{2}\right) \le M,
\end{aligned}$$

for every s in Σ_n. By Propositions 3 and 4 in [5], for every s in S_n,

$$\tilde{L}_i\left(\mathsf{z}_n'(t_n(s)), n\frac{\mathsf{z}''(s)+t_n'(s)\mathsf{H}''(t_n(s))}{|\mathsf{z}''(s)|}\right)\frac{|\mathsf{z}''(s)|}{n}$$
$$-\tilde{L}_i(\mathsf{z}_n'(t_n(s)), \mathsf{z}''(s)+t_n'(s)\mathsf{H}''(t_n(s)))\le -\left(\frac{|\mathsf{z}''(s)|}{n}-1\right)\tilde{L}_i^*(\mathsf{z}_n'(t_n(s)),p)\le 0,$$

where $p\in\partial_w L_i(\mathsf{z}_n'(t_n(s)), n(\mathsf{z}''(s)+t_n'(s)\mathsf{H}''(t_n(s)))/|\mathsf{z}''(s)|)$. Using the fact that ψ_i is positive and bounded by Ψ, we have $I_i^1\le M\Psi|\Sigma_n|<\bar\epsilon/8$.

To estimate I_i^2, we observe that

$$\tilde{L}_i(\mathsf{z}_n'(t_n(s)), \mathsf{z}''(s)+t_n'(s)\mathsf{H}''(t_n(s)))=\begin{cases}\tilde{L}_i(\mathsf{z}_n'(s), \mathsf{z}''(s)+\mathsf{H}''(s)), & s\in[\tau^-,\tau^+],\\ \tilde{L}_i(\mathsf{x}'(s),\mathsf{x}''(s)), & \text{otherwise.}\end{cases}$$

By the fact that $|\psi_i(t_n(s),\mathbf{z}_n(t_n(s)),\mathsf{z}_n'(t_n(s)))-\psi_i(s,\mathbf{x}(s),\mathsf{x}'(s))|\le\bar\epsilon/[8(\ell+\tilde{\mathbf{L}}+1)]$, for any s in $[a,b]$, and that $\mathsf{z}''+\mathsf{H}''\in B[0,|\mathsf{x}''(\tau)|+\delta]$ on $[\tau^-,\tau^+]$, we have $I_i^2\le\bar\epsilon/8$.

To estimate I_i^3, it is enough to estimate the integrals over $[\tau^-,\tau^+]$ (because it is identically 0 elsewhere). Recalling that τ is a Lebesgue point for $\tilde{L}_i(\mathsf{x}'(\cdot),\mathsf{x}''(\cdot))\psi_i(\cdot,\mathbf{x}(\cdot),\mathsf{x}'(\cdot))$, we have

$$\begin{aligned}I_i^3&\le\int_{\tau^-}^{\tau^+}[\tilde{L}_i(\mathsf{z}_n'(s),\mathsf{z}''(s)+\mathsf{H}''(s))\psi_i(s,\mathbf{x}(s),\mathsf{x}'(s))\\&\qquad-\tilde{L}_i(\mathsf{x}'(\tau),\mathsf{x}''(\tau))\psi_i(\tau,\mathbf{x}(\tau),\mathsf{x}'(\tau))]ds+\frac{\bar\epsilon}{8}\\&\le 4\rho\tilde{\mathbf{L}}\Psi+\frac{\bar\epsilon}{8}<\frac{\bar\epsilon}{4}.\end{aligned}$$

Hence, $I_i^1+I_i^2+I_i^3<\bar\epsilon/2$, for any i.

Conclusion. We have obtained

$$\begin{aligned}&\int_a^b L_i(\mathsf{z}_n'(t),\mathsf{y}_n''(t))\psi_i(t,\mathbf{z}_n(t),\mathsf{z}_n'(t))dt-\int_a^b L_i(\mathsf{x}'(t),\mathsf{x}''(t))\psi_i(t,\mathbf{x}(t),\mathsf{x}'(t))dt\\&<\int_a^b[L_i(\mathsf{z}_n'(t),\mathsf{z}_n''(t))+\eta]\psi_i(t,\mathbf{z}_n(t),\mathsf{z}_n'(t))dt\\&\quad-\int_a^b[L_i(\mathsf{x}'(t),\mathsf{x}''(t))+\eta]\psi_i(t,\mathbf{x}(t),\mathsf{x}'(t))dt+\frac{\bar\epsilon}{2}\\&=\int_a^b\tilde{L}_i(\mathsf{z}_n'(t),\mathsf{z}_n''(t))\psi_i(t,\mathbf{z}_n(t),\mathsf{z}_n'(t))dt-\int_a^b\tilde{L}_i(\mathsf{x}'(s),\mathsf{x}''(s))\psi_i(s,\mathbf{x}(s),\mathsf{x}'(s))ds+\frac{\bar\epsilon}{2}\\&<\bar\epsilon.\end{aligned}$$

Hence, $\mathcal{I}(z_n)-\mathcal{I}(x)<\sum_{i=1}^m\bar\epsilon=\epsilon$.

So, setting $x_\epsilon=z_n$, we have proved the theorem.

4. A necessary condition for the Lavrentiev phenomenon. The content of this section is provided to show the following necessary condition: a functional

$$\mathcal{I}(x)=\int_a^b\sum_{i=1}^m L_i(x^{(\nu)},x^{(\nu+1)})\psi_i(t,x,x',\dots,x^{(\nu)}),$$

with $\nu \geq 0$, exhibiting the Lavrentiev phenomenon takes the value $+\infty$ in any neighborhood of a minimizer $\bar{x}$; or equivalently if $\mathcal{I}$ assumes only finite values on a neighborhood of $\bar{x}$, then $\mathcal{I}$ does not exhibit the Lavrentiev phenomenon.

This is proved in the following corollary to Theorem 2.1 and Theorem 1 in [5].

COROLLARY 4.1. *Let $\Omega_0, \ldots, \Omega_\nu$ be open sets in $\mathbb{R}^N$, $\nu \geq 0$, such that the set $E = \{x \in \mathbf{W}^{\nu+1,1}(a,b) : x(t) \in \Omega_0, \ \ldots, \ x^{(\nu)}(t) \in \Omega_\nu \ \forall t \in [a,b]\}$ is nonempty. Let $A, B \in \Omega_0$, $A^{(1)}, B^{(1)} \in \Omega_1, \ \ldots, \ A^{(\nu)}, B^{(\nu)} \in \Omega_\nu$ be given boundary values.*

Let $L_1, \ldots, L_m : \Omega_\nu \times \mathbb{R}^N \to \mathbb{R}$ and $\psi_1, \ldots, \psi_m : [a,b] \times \Omega_0 \times \cdots \times \Omega_\nu \to [0, +\infty)$ be continuous and such that $L_i(\xi, \cdot)$ is convex, for any ξ in Ω_ν, any i in $\{1, \ldots, m\}$.

Let

$$\mathcal{I}(x) = \int_a^b \sum_{i=1}^m L_i(x^{(\nu)}(t), x^{(\nu+1)}(t))\psi_i(t, x(t), x'(t), \ldots, x^{(\nu)}(t))dt$$

be a functional exhibiting the Lavrentiev phenomenon, and let $\bar{x}$ be a minimum of $\mathcal{I}$ over $E_{a,b} = \{x \in E : x(a) = A, \ x(b) = B, \ x'(a) = A^{(1)}, \ x'(b) = B^{(1)}, \ \ldots, \ x^{(\nu)}(a) = A^{(\nu)}, \ x^{(\nu)}(b) = B^{(\nu)}\}$.

Assume that for any $\delta > 0$ there exists $\sigma_\delta > 0$ such that $\sigma_\delta \to 0$, for $\delta \to 0$, and that ψ_i restricted to $\mathsf{T}^\nu_\delta[\bar{x}]$ may vanish only on the graph of $(\bar{x}, \bar{x}', \ldots, \bar{x}^{(\nu)})$ or on a σ_δ-neighborhood of $(a, A, \ldots, A^{(\nu)})$ or on a σ_δ-neighborhood of $(b, B, \ldots, B^{(\nu)})$, for any i in $\{1, \ldots, m\}$.

Then, for any $\epsilon > 0$, there exists x_ϵ in $E_{a,b}$ such that the graph of $(x_\epsilon, x'_\epsilon, \ldots, x^{(\nu)}_\epsilon)$ is included in $\mathsf{T}^\nu_\epsilon[\bar{x}]$ and $\mathcal{I}(x_\epsilon) = +\infty$.

Proof. Fix $\epsilon > 0$. From Theorem 2.1 and Theorem 1 in [5], it follows that $\int_a^b |L_i(\bar{x}^{(\nu)}, \bar{x}^{(\nu+1)})| = +\infty$, for at least one i in $\{1, \ldots, m\}$.

Without loss of generality, we suppose that $\int_a^{(a+b)/2} |L_i(\bar{x}^{(\nu)}, \bar{x}^{(\nu+1)})| = +\infty$.

Let $g : (-\infty, +\infty) \to [0,1]$ be a $\mathbf{C}^\infty$ increasing function with value 1 on $[b, +\infty)$ and 0 on $(-\infty, (a+b)3/4]$. We define the integrable function $x_{\delta,\nu+1} : [a,b] \to \mathbb{R}^N$ by

$$x_{\delta,\nu+1}(t) = \begin{cases} 0, & t \in [a, a+\sigma_\delta), \\ \bar{x}^{(\nu+1)}(t - \sigma_\delta), & \text{otherwise}, \end{cases}$$

and ν absolutely continuous functions $x_{\delta,j}(t) = A^{(j)} + \int_a^t x_{\delta,j+1} + g(t)D_{\delta,j}$, for any t in $[a,b]$, where $D_{\delta,j} = B^{(j)} - A^{(j)} - \int_a^b x_{\delta,j+1}$, for any j in $\{0, \ldots, \nu\}$.

Set $x_\delta = x_{\delta,0}$. The derivatives of x_δ up to the order $\nu + 1$ are

$$\begin{aligned} x'_\delta(t) &= x_{\delta,1}(t) + g'(t)D_{\delta,0}, \\ x''_\delta(t) &= x_{\delta,2}(t) + g''(t)D_{\delta,0} + g'(t)D_{\delta,1}, \\ &\vdots \\ x^{(\nu+1)}_\delta(t) &= x_{\delta,\nu+1}(t) + \textstyle\sum_{j=0}^\nu g^{(\nu-j+1)}(t)D_{\delta,j}. \end{aligned}$$

By definition, x_δ belongs to $\mathbf{W}^{\nu+1,1}(a,b)$, it has the same boundary values of $\bar{x}$ in a and in b, and, for j in $\{\nu, \nu+1\}$, for any t in $[a+\sigma_\delta, (a+b)3/4]$, we have $x^{(j)}_\delta(t) = \bar{x}^{(j)}(t - \sigma_\delta)$. Furthermore, there exist constants c_j, d_j, independent on δ, such that $|D_{\delta,j}| \leq c_j \int_{b-\sigma_\delta}^b |\bar{x}^{(\nu+1)}|$ and $||x_{\delta,j} - x^{(j)}_\delta||_\infty \leq d_j \int_{b-\sigma_\delta}^b |\bar{x}^{(\nu+1)}|$. Hence, for any j in $\{0, \ldots, \nu\}$,

$$||\bar{x}^{(j)} - x^{(j)}_\delta||_\infty \leq \left(c_j + ||g^{(\nu+1)}||_\infty \sum_{j=0}^\nu d_j \right) \int_{b-\sigma_\delta}^b |\bar{x}^{(\nu+1)}|.$$

By hypothesis, we can choose $\bar\delta > 0$ such that $(c_j + ||g^{(\nu+1)}||_\infty \sum_{j=0}^{\nu} d_j) \int_{b-\sigma_{\bar\delta}}^{b} |\bar x^{(\nu+1)}| < \epsilon$ and $\sigma_{\bar\delta} < (b-a)/4$.

Set $\Psi_i = \min\{\psi_i(t, x_{\bar\delta}(t), \dots, x_{\bar\delta}^{(\nu)}(t)) : t \in [a+\sigma_{\bar\delta}, (a+b)3/4]\}$: by hypothesis, Ψ_i is positive. We have obtained that the graph of $(x_{\bar\delta}, x'_{\bar\delta}, \dots, x_{\bar\delta}^{(\nu)})$ belongs to $\mathsf{T}_\epsilon^\nu[\bar x]$ and

$$\begin{aligned}
&\int_a^b |L_i(x_{\bar\delta}^{(\nu)}(t), x_{\bar\delta}^{(\nu+1)}(t))\psi_i(t, x_{\bar\delta}(t), \dots, x_{\bar\delta}^{(\nu)}(t))|dt \\
&\geq \int_{a+\sigma_{\bar\delta}}^{(a+b)3/4} |L_i(\bar x^{(\nu)}(t-\sigma_{\bar\delta}), \bar x^{(\nu+1)}(t-\sigma_{\bar\delta}))|\psi_i(t, x_{\bar\delta}(t), \dots, x_{\bar\delta}^{(\nu)}(t))dt \\
&\geq \Psi_i \int_a^{(a+b)/2} |L_i(\bar x^{(\nu)}(t), \bar x^{(\nu+1)}(t))|dt = +\infty.
\end{aligned}$$

From (i) in the proof of Theorem 2.1 and Theorem 1 in [5], we infer that $\mathcal{I}(x_{\bar\delta}) = +\infty$.

So, setting $x_\epsilon = x_{\bar\delta}$, we have proved the corollary. □

The corollary above applies to the functionals of Manià and Sarychev, for instance, and to the examples of functionals exhibiting the Lavrentiev phenomenon proposed in [3], [4], [11], [12], [13], [14], and [15].

Acknowledgment. We thank the anonymous referees for many interesting comments.

REFERENCES

[1] G. Alberti and F. Serra Cassano, *Non-occurrence of gap for one-dimensional autonomous functionals*, in Calculus of Variations, Homogenization and Continuum Mechanics, Ser. Adv. Math. Appl. Sci. 18, World Scientific, River Edge, NJ, 1994, pp. 1–17.

[2] T. S. Angell, *A note on approximation of optimal solutions of free problems of the calculus of variations*, Rend. Circ. Mat. Palermo (2), 28 (1979), pp. 258–272.

[3] J. Ball and V. J. Mizel, *One-dimensional variational problems whose minimizers do not satisfy the Euler-Lagrange equation*, Arch. Rational Mech. Anal., 90 (1985), pp. 325–388.

[4] M. Belloni, *Interpretation of Lavrentiev phenomenon by relaxation: The higher order case*, Trans. Amer. Math. Soc., 347 (1995), pp. 2011–2023.

[5] A. Cellina, A. Ferriero, and E. M. Marchini, *Reparametrizations and approximate values of integrals of the calculus of variations*, J. Differential Equations, 193 (2003), pp. 374–384.

[6] L. Cesari, *Optimization—Theory and Applications*, Springer-Verlag, New York, 1983.

[7] C. W. Cheng and V. J. Mizel, *On the Lavrentiev phenomenon for autonomous second-order integrands*, Arch. Rational Mech. Anal., 126 (1994), pp. 21–33.

[8] F. H. Clarke and R. B. Vinter, *Regularity properties of solutions to the basic problem in the calculus of variations*, Trans. Amer. Math. Soc., 289 (1985), pp. 73–98.

[9] F. H. Clarke and R. B. Vinter, *A regularity theory for variational problems with higher order derivatives*, Trans. Amer. Math. Soc., 320 (1990), pp. 227–251.

[10] A. Ferriero, *The Lavrentiev Phenomenon in the Calculus of Variations*, Ph.D. thesis, Universitá degli Studi di Milano-Bicocca, Milano, Italy, 2004.

[11] M. Lavrentiev, *Sur quelques problemes du calcul des variations*, Ann. Mat. Pura Appl., 4 (1926), pp. 7–28.

[12] P. D. Lowen, *On the Lavrentiev phenomenon*, Canad. Math. Bull., 30 (1987), pp. 102–108.

[13] B. Manià, *Sopra un esempio di Lavrentieff*, Boll. Unione Mat. Ital., 13 (1934), pp. 147–153.

[14] V. J. Mizel, *Recent progress on the Lavrentiev phenomenon, with applications*, in Differential Equations and Control Theory, Lecture Notes in Pure and Appl. Math. 225, Dekker, New York, 2002, pp. 257–261.

[15] A. V. Sarychev, *First- and second-order integral functionals of the calculus of variations which exhibit the Lavrentiev phenomenon*, J. Dynam. Control Systems, 3 (1997), pp. 565–588.

[16] J. Serrin and D. E. Varberg, *A general chain rule for derivatives and the change of variable formula for the Lebesgue integral*, Amer. Math. Monthly, 76 (1969), pp. 514–520.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 1, pp. 111–139

TIME-OPTIMAL SYNTHESIS FOR LEFT-INVARIANT CONTROL SYSTEMS ON $SO(3)$*

UGO BOSCAIN† AND YACINE CHITOUR‡

Abstract. Consider the control system (Σ) given by $\dot{x} = x(f + ug)$, where $x \in SO(3)$, $|u| \leq 1$, and $f, g \in so(3)$ define two perpendicular left-invariant vector fields normalized so that $\|f\| = \cos(\alpha)$ and $\|g\| = \sin(\alpha)$, $\alpha \in]0, \pi/4[$. In this paper, we provide an upper bound and a lower bound for $N(\alpha)$, the maximum number of switchings for time-optimal trajectories of (Σ). More precisely, we show that $N_S(\alpha) \leq N(\alpha) \leq N_S(\alpha) + 4$, where $N_S(\alpha)$ is a suitable integer function of α such that $N_S(\alpha) \underset{\alpha \to 0}{\sim} \pi/(4\alpha)$. The result is obtained by studying the time-optimal synthesis of a projected control problem on $\mathbb{R}P^2$, where the projection is defined by an appropriate Hopf fibration. Finally, we study the projected control problem on the unit sphere S^2. It exhibits interesting features which will be partly rigorously derived and partially described by numerical simulations.

Key words. optimal control, optimal synthesis, minimum time, $SO(3)$

AMS subject classifications. 49k15

DOI. 10.1137/S0363012904441532

1. Introduction.

Let (Σ) be the control system given by

$$\dot{x} = x(f + ug), \tag{1.1}$$

where $x \in SO(3)$, $|u| \leq 1$, and $f, g \in so(3)$ give rise to two nonzero perpendicular left-invariant vector fields on $SO(3)$. In this paper, we consider the following problem: given any pair of points x_1, x_2 of $SO(3)$, find a trajectory of (1.1) steering x_1 to x_2 in minimum time. That issue is known as the problem of determining the time-optimal synthesis (TOS) for (Σ). The strategy to determine a TOS usually consists of two steps:

1. Reduction procedure: it is based on the Pontryagin maximum principle (PMP) which is a first-order necessary condition for optimality. Roughly speaking, the PMP reduces the candidates for time optimality to the so-called extremals, which are solutions of a pseudo-Hamiltonian system. This reduction procedure may be refined using higher-order conditions, such as Clebsch–Legendre conditions, higher-order maximum principle, envelopes, conjugate points, and index theory (cf. for instance [2, 3, 5, 12, 18, 19, 21, 26, 27, 28, 30, 32, 33, 34, 35]).
2. Selection procedure: it consists of selecting the time-optimal trajectories among the extremals that passed the test of Step 1 (see for instance [8, 12, 16, 25]).

Step 1 is already nontrivial and, in general, the second one is extremely difficult: if the state space is two dimensional, the problem of determining the TOS for single-input control systems is now well understood [9, 11, 12, 14, 15, 23, 24, 32, 33]. However, for higher dimensions, very few examples of complete TOS for a nonlinear control system are available (see for instance [29]). Intermediate issues were thus deeply investigated:

*Received by the editors March 1, 2004; accepted for publication (in revised form) February 18, 2005; published electronically, June 27, 2005.
http://www.siam.org/journals/sicon/44-1/44153.html
†SISSA-ISAS, Via Beirut 2-4, 34014 Trieste, Italy (boscain@sissa.it).
‡Université Paris XI, Laboratoiré des signaux et systèmes (L2S) Supélec, 3 rue Joliot-Cure, 91190 Gif-sur-Yvette, France (chitour@lss.supelec.fr).

determining estimates for the number of switchings of optimal trajectories, describing the local structure of optimal trajectories, finding families of trajectories sufficient for optimality (cf. [6, 13, 19, 22, 26, 28, 36]), etc.

For the control system (Σ), we normalize the two perpendicular vector fields induced by f and g in such a way that $\|f\| = \cos(\alpha)$, $\|g\| = \sin(\alpha)$, with $\alpha \in]0, \pi/2[$ (for the precise meaning of "perpendicular" and of the symbol $\|.\|$, we refer to section 2.1). Defining $X_+ := f + g$ and $X_- := f - g$, we have $\|X_+\| = \|X_-\| = 1$ and α is the angle between f and X_+.

By a standard argument (see section 2 below), one can show that every time optimal trajectory is a finite concatenation of bang arcs (i.e., $u \equiv \pm 1$) or singular arcs ($u = 0$) the Fuller phenomenon (i.e., existence of a trajectory of a control system joining two points in (finite) minimum time, with an infinite number of switchings, cf. [20,37]) never occurs. (A switching time—or simply a switching—along an extremal is a time t_0 so that the control u is not constant in any open neighborhood of t_0.) Moreover, one can easily show that the supremum $N(\alpha)$ of the number of switchings over all time optimal trajectories of (Σ) is finite.

By using the index theory developed by Agrachev, it is proved in [3] that

$$N(\alpha) \leq N_A := \left[\frac{\pi}{\alpha}\right], \tag{1.2}$$

where $[\cdot]$ stands for the integer part. That result was not only an indirect indication that $N(\alpha)$ would tend to ∞ as α tends to zero, but it also provided a hint on the asymptotic of $N(\alpha)$ as α tends to zero.

A related line of work regards the study of the distributional version of (Σ), which is the driftless control system given by $\dot{x} = x(u_1 f_1 + u_2 f_2)$, $|u_1|, |u_2| \leq 1$, and $f_1, f_2 \in so(3)$ linearly independent. Indeed, assuming that $\|f_1\| = \|f_2\|$, Sussmann and Tang [36] showed that time-optimal trajectories have at most four switchings and they provided a finitely parametrized family of trajectories sufficient for optimality. That result was extended to the general case (f_1 and f_2 just linearly independent; cf. [13]): time-optimal trajectories have at most five switchings. For both works, the elimination from optimality of extremals with, respectively, five or six bangs relies on the envelope theory developed in the context of control theory by Sussmann (cf. [35]).

In light of the previous results, there was strong evidence for two radically situations as α tends to zero: for (Σ), $N(\alpha)$ is expected to go to infinity, and as for the distributional control system, there exists a universal bound on the number of switchings. The main result of the present paper confirms that difference, i.e., $N(\alpha)$ tends to ∞ as α tends to zero. More precisely, we complete the inequality (1.2) as follows.

THEOREM 1. *Let (Σ) be the control system defined in (1.1) with f, g perpendicular so that $\|f\| = \cos(\alpha)$ and $\|g\| = \sin(\alpha)$, $\alpha \in]0, \pi/4[$. Then, if $N(\alpha)$ is the maximum number of switchings along a time-optimal trajectory of (Σ), we have*

$$N_S(\alpha) \leq N(\alpha) \leq N_S(\alpha) + 4, \text{ where } N_S(\alpha) := 2\left[\frac{\pi}{8\alpha}\right] - \left[2\left[\frac{\pi}{8\alpha}\right] - \frac{\pi}{4\alpha}\right]. \tag{1.3}$$

The above theorem improves (1.2) in two ways: (i) for α small, it (essentially) divides the upper bound of $N(\alpha)$ by four with respect to (1.2); and (ii) it provides a lower bound of $N(\alpha)$ differing from the upper bound by a constant.

The lower bound is in fact our main contribution and, to get it, one must prove the existence of time-optimal trajectories of (Σ) admitting at least a number of switchings equal to that lower bound. Our strategy consists of projecting the control problem

onto another $(\Sigma)_S$ defined next. First, let $\mathbb{R}P^2$ be the two-dimensional real projective space (i.e., the two-dimensional manifold made of the directions of $\mathbb{R}^3$) and fix a point $x_0 \in SO(3)$. Consider the Hopf fibration $\Pi : SO(3) \to \mathbb{R}P^2$ defined by $\mathrm{Ker} d\Pi(x_0) = Span\{x_0 f\}$, which means, roughly speaking, that Π annihilates the drift term f at x_0. Then, we project (Σ) by Π and obtain a single-input $SO(3)$-equivariant control system $(\Sigma)_S$ on $\mathbb{R}P^2$ given by $\dot{y} = y(f_S + u g_S)$, with $f_S = d\Pi(f)$ and $g_S = d\Pi(g)$, that is locally controllable. We then consider the minimum time problem for connecting $\Pi(x_0)$ to any other point of $\mathbb{R}P^2$.

In fact, we study a slightly different time optimal problem by lifting $(\Sigma)_S$ to the unit sphere S^2. By an abuse of notation, we still denote by $(\Sigma)_S$ the control system obtained in that way. Hence $\Pi(x_0)$ is identified with the north pole and $\mathbb{R}P^2$ is identified with $\underline{\mathcal{NH}}$, the subset of the sphere made of the union of $\mathcal{NH}$, the (open) top hemisphere of S^2, together with half of the equator.

The time optimal problem now consists of connecting, in minimum time, the north pole to any point of $\underline{\mathcal{NH}}$. Thanks to the suitable choice of the Hopf projection and since α belongs to the interval $]0, \pi/4[$, all extremals of the projected problem are bang-bang (i.e., they are a finite concatenation of trajectories corresponding to controls $+1$ or -1). Let $N_S(\alpha)$ be the supremum of the number of switchings for time-optimal trajectories of $(\Sigma)_S$ starting at the north pole and ending in $\underline{\mathcal{NH}}$ (such trajectories of $(\Sigma)_S$ are actually entirely contained in $\underline{\mathcal{NH}}$, see Lemma 7).

The use of the Hopf fibration Π is motivated by two facts: first, every time-optimal trajectory for the time optimal problem on $(\Sigma)_S$ staying in $\underline{\mathcal{NH}}$ is the projection by Π of a time-optimal trajectory for the time optimal problem on (Σ) and thus $N_S(\alpha) \leq N(\alpha)$. Taking full advantage of the theory developed in [12], we will actually compute *exactly* $N_S(\alpha)$ as given in (1.3). Second, using the fact that the fiber above $\Pi(x_0)$ is the support of a singular arc (for this problem singular arcs are integral curves of the drift xf), we show that every regular bang-bang trajectory with at least $N_S(\alpha) + 5$ cannot be optimal and thus, the upper bound.

It is then clear, by now, that the most delicate part of the argument relies on the exact determination of $N_S(\alpha)$. This is done by studying the TOS for the time optimal problem on $(\Sigma)_S$. Such a TOS is usually constructed, following the theory developed in [9, 11, 12, 14, 15, 23, 24, 32, 33], recursively on the number of extremals arcs, and by checking at each step whether they are optimal or not. For the problem on $\mathbb{R}P^2$, we are not able to complete all the steps of the above construction, which would imply as a by-product the existence of the TOS. In particular, we cannot show the optimality of all the extremals (i.e., the trajectories candidate for time optimality), but, from their study, we can demonstrate enough partial results in order to compute $N_S(\alpha)$ precisely and thus to conclude the proof of Theorem 1.

The complete time-optimal synthesis is then studied numerically (actually on the whole S^2) and is shown on the top of Figure 5.2. In particular, due to the compactness of S^2, one of the main issues is to understand the singularities developed by the minimum time wave front as it approaches the south pole. We provide numerical simulations that describe the evolution of the extremal front. As $\alpha \to 0$, these numerical simulations suggest the emergence of three cyclically alternating patterns of optimal synthesis, each of them depending on an arithmetic property of α.

The rest of the paper is organized as follows. Section 2 collects basic facts relative to the time-optimal trajectories of (Σ), and in section 3, the Hopf fibration is described and the proof of Theorem 1 is provided, assuming some facts about the time-optimal synthesis of $(\Sigma)_S$, whose arguments are deferred to the next section. In particular

we use the expression for $N_S(\alpha)$ and the relation between the length of interior bang arcs for the problem on $\mathbb{R}P^2$. The construction of the time-optimal synthesis of $(\Sigma)_S$ is investigated in section 4, where an exact computation of $N_S(\alpha)$ is established. We conclude the section with two remarks, the first one explaining the relation between the TOS on the sphere and the TOS of a controlled linear pendulum, and the second one establishing a link with an optimal control problem on $SU(2)$. Finally, in section 5, we provide the results of the numerical simulations, completing the study of the time-optimal synthesis (in particular of the possible behavior in a neighborhood of the south pole), and we propose some open problems stated as conjectures.

2. Statement of the problem and properties of optimal trajectories.

2.1. Basic facts. In this paper, we consider the control (Σ) given by (1.1), where $x \in SO(3)$, $|u| \leq 1$, and $f, g \in so(3)$. An admissible control u is a measurable function $u : [a, b] \to [-1, 1]$, where a, b depend (in general) on u; cf. [18]. A trajectory γ of (Σ) is an absolutely continuous curve $\gamma : J \to SO(3)$, where $J = [a, b]$ is a compact segment of $\mathbb{R}$ such that there exists an admissible control u for which $\dot\gamma(t) = \gamma(t)(f + u(t)g)$ holds a.e. in J. We then say that (γ, u), defined as before, is an admissible pair for (Σ).

Definition 1. *A trajectory γ of (Σ), defined on $[a, b]$, is time optimal if for every trajectory γ' of (Σ) defined on $[a', b']$ with $\gamma(a) = \gamma'(a')$ and $\gamma(b) = \gamma'(b')$, we have $b - a \leq b' - a'$.*

The Lie algebra $(so(3), [.,.])$ is isomorphic to the Lie algebra $(\mathbb{R}^3, \times)$, where $\times$ denotes the vector product. This isomorphism is realized by the map

$$\phi_L : so(3) \to \mathbb{R}^3 \tag{2.1}$$
$$\phi_L\left(\begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix}\right) := \begin{pmatrix} a \\ b \\ c \end{pmatrix},$$

and provides an inner product on $so(3)$ given by $\langle z_1, z_2\rangle := \langle \phi_L(z_1), \phi_L(z_2)\rangle$, where $z_1, z_2 \in so(3)$. The symbol $\langle .,.\rangle$ on the right-hand side of the above equation stands for the Euclidean inner product of $\mathbb{R}^3$. With this definition, it follows that $\langle z_1, z_2\rangle := -\frac{1}{2}Tr(z_1 z_2)$. In other words, this scalar product is the opposite of the Killing form on $so(3)$. In the following, $\|z\| := \sqrt{\langle z, z\rangle}$ and Id is the 3×3 identity matrix. We will sometimes consider the 2×2 matrix corresponding to the planar rotation of angle β and we use R_β to denote it.

In this paper, we will assume that f and g are perpendicular and normalized so that $\|f\| = \cos(\alpha)$ and $\|g\| = \sin(\alpha)$, $\alpha \in]0, \pi/2[$. Here, we adopt the following notation used throughout the paper, $c_\alpha := \cos(\alpha)$, $c_\alpha^2 := \cos^2(\alpha)$, $s_\alpha := \sin(\alpha)$, and $s_\alpha^2 := \sin^2(\alpha)$. We define $h := [f, g] = fg - gf$ and

$$X_+ := f + g, \quad X_- := f - g.$$

Note that $\|X_\varepsilon\| = 1$, with $\varepsilon = +, -$. For a vector field $z \in so(3)$, we use e^{tz} to denote the flow of z, acting on the right, so that $t \mapsto pe^{tz} \in SO(3)$ is the integral curve of z starting at p at time 0. Since z is linear, we have $e^{tz} = \sum_{n=0}^{\infty} \frac{(tz)^n}{n!}$. We use ad_z to denote the operator $w \mapsto [z, w]$, acting on vector fields. If z, w are vector fields, then $e^{t \operatorname{ad}_z}(w) := e^{tz} w e^{-tz}$. The Lie bracket relations between f, g, h are

$$[f, g] = h, \quad [g, h] = s_\alpha^2 f, \quad [h, f] = c_\alpha^2 g.$$

From them, one deduces the following classical relations that will be useful later:

$$e^{t\ \mathrm{ad}X_\varepsilon}(f) = (c_\alpha^2 + s_\alpha^2\cos(t))f + \varepsilon c_\alpha^2(1-\cos(t))g - \varepsilon\sin(t)h, \tag{2.2}$$

$$e^{t\ \mathrm{ad}X_\varepsilon}(g) = \varepsilon s_\alpha^2(1-\cos(t))f + (s_\alpha^2 + c_\alpha^2\cos(t))g + \sin(t)h, \tag{2.3}$$

$$e^{t\ \mathrm{ad}X_\varepsilon}(h) = \varepsilon s_\alpha^2\sin(t)f - c_\alpha^2\sin(t)g + \cos(t)h, \tag{2.4}$$

$$e^{t\ \mathrm{ad}X_\varepsilon}(X_{-\varepsilon}) = (\cos(2\alpha) + 2s_\alpha^2\cos(t))f + \varepsilon(\cos(2\alpha) - 2c_\alpha^2\cos(t))g - 2\varepsilon\sin(t)h, \tag{2.5}$$

$$e^{tX_\varepsilon} = Id + \sin(t)X_\varepsilon + (1-\cos(t))X_\varepsilon^2, \tag{2.6}$$

$$e^{tadf}(g) = \cos(tc_\alpha)g + \frac{\sin(tc_\alpha)}{c_\alpha}h. \tag{2.7}$$

2.2. Existence of optimal trajectories. A control system is *complete* if, for every measurable control function $u : [a,b] \to [-1,1]$ and every initial state p, there exists a trajectory γ corresponding to u, which is defined on the whole interval $[a,b]$ and satisfies $\gamma(a) = p$. Since $SO(3)$ is compact and the function $\mathcal{F}(x,u) := x(f+ug)$ is regular enough, the system (1.1) is complete. Note that (f,g) satisfies the strong bracket generating condition (cf. [31]) and the set of velocities $V(x) := \{x(f+ug),\ u \in [-1,1]\}$ is compact and convex. Then (cf. for instance [36]) we have the following proposition.

PROPOSITION 1. *For each pair of points p and q belonging to $SO(3)$, there exists a time-optimal trajectory joining p to q.*

2.3. Pontryagin maximum principle and switching functions. We next state the PMP (cf. [26]) for our minimum time problem on $SO(3)$. Define the following maps called, respectively, *Hamiltonian* and *minimized Hamiltonian*:

$$\mathcal{H} : T^*SO(3) \times [-1,1] \to \mathbb{R}, \quad \mathcal{H}(p,x,u) := \langle p, x(f+ug)\rangle, \tag{2.8}$$

$$H : T^*SO(3) \to \mathbb{R}, \quad H(p,x) := \min_{v\in[-1,1]} \mathcal{H}(p,x,v). \tag{2.9}$$

The PMP asserts that if $\gamma : [a,b] \to SO(3)$ is a time-optimal trajectory corresponding to a control $u : [a,b] \to [-1,1]$, then there exists a *nontrivial field of covectors along* γ, that is an absolutely continuous function $\lambda : t \in [a,b] \mapsto \lambda(t) \in T^*_{\gamma(t)}SO(3)$ (identified with $so(3)$) never vanishing and a constant $\lambda_0 \geq 0$ such that, for a.e. $t \in Dom(\gamma)$, we have

(i) $\dot\lambda(t) = -\frac{\partial\mathcal{H}}{\partial x}(\lambda(t),\gamma(t),u(t)) = -\lambda(t)(f+u(t)g)$,
(ii) $\mathcal{H}(\gamma(t),\lambda(t),u(t)) + \lambda_0 = 0$,
(iii) $\mathcal{H}(\gamma(t),\lambda(t),u(t)) = H(\gamma(t),\lambda(t))$.

Remark 1. The PMP is just a necessary condition for optimality. A trajectory γ (resp., a couple (γ,λ)) satisfying the conditions given by the PMP is said to be an *extremal* (resp., an *extremal pair*). An extremal corresponding to $\lambda_0 = 0$ is said to be an *abnormal extremal*, otherwise we call it a *normal extremal*. For a normal extremal, we can always normalize $\lambda_0 = 1$, and we do this all through the paper. Note that in general an extremal corresponds to more than one covector. For this reason, usually, one distinguishes between abnormal extremals that are *strict* (i.e., they correspond only to covectors satisfying $\lambda_0 = 0$) and abnormal extremals that are *nonstrict* (i.e., they correspond to covectors with $\lambda_0 = 0$ and to covectors with $\lambda_0 \neq 0$).

A control $u : [a,b] \to [-1,1]$ is said to be *bang-bang* if $u(t) \in \{-1,1\}$ a.e. in $[a,b]$. Moreover, if $u(t) \in \{-1,1\}$ and $u(t)$ is constant for almost every $t \in [a,b]$, then u is called a bang control. A *switching* time of u is a time $t \in [a,b]$ such that, for

every $\varepsilon > 0$, u is not bang on $(t-\varepsilon, t+\varepsilon) \cap [a,b]$. A control with a finite number of switchings is called *regular bang-bang*. A trajectory of Σ is a bang trajectory, bang-bang trajectory, or regular bang-bang trajectory, respectively, if it corresponds to a bang control, bang-bang control, or regular bang-bang control. The *switching functions*, associated with an extremal pair (γ, λ), are the three "components" of the covector $\lambda(t)$ on the basis $\{f, g, h\}$ transported to the point $\gamma(t)$. More precisely we have the following definition.

DEFINITION 2 (switching functions). *Let $\Phi_i(x,p)$ $(i = 1,2,3)$ be the Hamiltonian functions corresponding, respectively, to the vector fields f, g, h (cf.* [18]*), i.e., $\Phi_1(x,p) := \langle p, xf\rangle$, $\Phi_2(x,p) := \langle p, xg\rangle$, $\Phi_3(x,p) := \langle p, xh\rangle$ and (γ, λ) be an extremal pair. The switching functions associated with (γ, λ) are the evaluations of $\Phi_i(x,p)$ along the extremal, i.e.,*

$$\varphi_1(t) := \Phi_1(\gamma(t), \lambda(t)) = \langle \lambda(t), \gamma(t) f\rangle, \tag{2.10}$$

$$\varphi_2(t) := \Phi_2(\gamma(t), \lambda(t)) = \langle \lambda(t), \gamma(t) g\rangle, \tag{2.11}$$

$$\varphi_3(t) := \Phi_3(\gamma(t), \lambda(t)) = \langle \lambda(t), \gamma(t) h\rangle. \tag{2.12}$$

Remark 2. Note that the φ_i's are at least continuous and since λ never vanishes, the three switching functions cannot be all zero at the same time t. Moreover, using the switching functions, (ii) of PMP reads

$$\mathcal{H}(\lambda(t), \gamma(t), u(t)) = \varphi_1(t) + u(t)\varphi_2(t) + \lambda_0 = 0 \text{ a.e.} \tag{2.13}$$

The switching functions are important because they determine where the controls may switch. In fact, using the PMP, one easily gets the following proposition.

PROPOSITION 2. *A necessary condition for a time t to be a switching is that $\varphi_2(t) = 0$. Therefore, on any interval where φ_2 has no zeros (resp., finitely many zeros), the corresponding control is bang (resp., bang-bang). In particular, $\varphi_2 > 0$ (resp., $\varphi_2 < 0$) on $[a,b]$ implies $u = -1$ (resp., $u = +1$) a.e. on $[a,b]$. On the other hand, if φ_2 has a zero at t and $\dot\varphi_2(t)$ exists and is different from zero, then t is an isolated switching.*

As a corollary, it holds a.e. along an extremal trajectory that

$$u(t)\varphi_2(t) = -|\varphi_2(t)|. \tag{2.14}$$

An extremal trajectory γ of Σ defined on $[c,d]$ is said to be *singular* if the switching function φ_2 vanishes on $[c,d]$. To compute the control corresponding to a singular trajectory, one should compute the derivatives of the φ_i's. Using the Lie bracket relations between f, g, h, one gets the system of differential equations (called the *adjoint system*) satisfied a.e.:

$$\dot\varphi_1 = -u\varphi_3, \tag{2.15}$$

$$\dot\varphi_2 = \varphi_3, \tag{2.16}$$

$$\dot\varphi_3 = s_\alpha^2 u\varphi_1 - c_\alpha^2\varphi_2. \tag{2.17}$$

From (2.12) and (2.16), one immediately gets that φ_2 is at least a $\mathcal{C}^1$ function. Moreover, if γ is singular in $[a,b]$, then $\varphi_2 = 0$ and, from (2.16), we get $\varphi_3 = 0$ a.e. From (2.13) (cf. PMP (ii)), we get $\varphi_1 \equiv -1$ a.e. on $[a,b]$. From (2.17) we get $u = 0$ a.e., i.e., we have the following proposition.

PROPOSITION 3. *For the minimum time problem for (Σ), singular trajectories are integral curves of the drift, i.e., they correspond to a control a.e. vanishing.*

In what follows, we will use the following convention. The letter B refers to a bang trajectory and the letter S refers to a singular extremal trajectory. A concatenation of bang and singular trajectories will be labeled by the corresponding letter sequence, written in order from left to right. Sometimes, we will use a subscript to indicate the time duration of a trajectory so that we use B_t to refer to a bang trajectory defined on an interval of length t and, similarly, S_t for a singular trajectory defined on an interval of length t.

If we fix $u \in [-1, 1]$, then the integral curves of $x(f + ug)$ are periodic. In particular, the integral curves of xX_ε are periodic with period 2π while the integral curves of the drift xf are periodic with period $2\pi/c_\alpha$, which means the following proposition.

PROPOSITION 4. *If γ is an extremal trajectory of type B_t (resp., S_t), then $t < 2\pi$ (resp., $t < 2\pi/c_\alpha$).*

There are two quantities that remain constant along an extremal trajectory. The first one comes from the fact that the minimized Hamiltonian H is constant along the extremal pairs (γ, λ) (cf. (2.13) and (2.14)):

$$I_1 := -\varphi_1(t) + |\varphi_2(t)| = \lambda_0, \tag{2.18}$$

with λ_0 equal to 0 or 1 (cf. Remark 1). The second conserved quantity is

$$I_2 := c_\alpha^2\varphi_2^2 + s_\alpha^2\varphi_1^2 + \varphi_3^2 = K^2, \quad \text{for some } K \in \mathbb{R}. \tag{2.19}$$

Remark 3. Equations (2.15)–(2.17) are Hamiltonian equations on the dual of $so(3)$, with respect to the canonical Poisson structure induced by the brackets of $f, g, h \in so(3)$, and corresponding to the left-invariant Hamiltonian (2.8). The conserved quantity I_2 is the Casimir function (see for instance [1]).

There is a geometric interpretation of the above equations. Let (γ, λ) be a normal extremal lift of the time-optimal control problem. Then, the adjoint vector λ with coordinates $(\varphi_i)_{i=1,2,3}$ lies in the intersection of the region defined by (2.18) and the ellipsoid defined by (2.19).

2.4. Classification of optimal trajectories. In this section, we investigate the structure of time-optimal trajectories by analyzing the extremal flow defined in (2.15)–(2.17), subject to (2.18) and (2.19). First we study abnormal extremals (we prove that they are regular bang-bang and we establish a relation between the interior bang times). Then we study normal extremals that are bang-bang (again we find a relation between the interior bang times). Finally we study optimal trajectories containing a singular arc. The results presented in this section are well known, and some of them already contained in [3, 4], although in many cases without proof. To have a self-contained paper, we provide an argument for all of them.

2.4.1. Abnormal extremals. The following proposition describes the switching behavior of abnormal extremals.

PROPOSITION 5. *Let γ be an abnormal extremal. Then, it is regular bang-bang and the time duration between two consecutive switchings is always equal to π. In other words, γ is of kind $B_\pi B_\pi \cdots B_\pi B_t$ with $t \leq \pi$.*

Proof of Proposition 5. By definition, $\lambda_0 = 0$. Then (2.13) becomes

$$\varphi_1(t) = -u(t)\varphi_2(t) \quad \text{for a.e. } t \in Dom(\gamma). \tag{2.20}$$

If γ is singular on some interval $[c, d]$, then $\varphi_2 \equiv 0$ and from (2.16) $\varphi_3 \equiv 0$ on $[c, d]$. Equation (2.20) gives $\varphi_1 \equiv 0$, contradicting the nontriviality of λ (cf. Remark 2). Then γ cannot contain a singular arc. Therefore, $u^2 = 1$ a.e. $t \in Dom(\gamma)$.

From (2.17) and (2.20), we get a.e. $\dot{\varphi}_3(t) = (-s_\alpha^2 u(t)^2 - c_\alpha^2)\varphi_2(t) = -\varphi_2(t)$. This means that, in the (φ_3, φ_2) plane, the vector $z(t) := (\varphi_3(t), \varphi_2(t))$ rotates with angular velocity equal to 1 (cf. (2.16)). This implies γ is a regular bang-bang trajectory and the time duration between two consecutive switchings along γ is always equal to π. □

2.4.2. Normal bang-bang extremals. Let γ be a bang-bang trajectory starting at p_0 and ending at $p_0 e^{(t_0 X_+)} e^{(t_1 X_-)} e^{(t_2 X_+)} e^{(t_3 X_-)}$. The case in which the first bang is of kind X_- is similar. We have $\varphi_2(t_0) = \varphi_2(t_0 + t_1) = \varphi_2(t_0 + t_1 + t_2) = 0$ which implies

$$\begin{aligned} \langle \lambda(t_0 + t_1), p_2 e^{-t_1 adX_-}(g) \rangle &= \langle \lambda(t_0 + t_1), p_2 g \rangle \\ &= \langle \lambda(t_0 + t_1), p_2 e^{t_2 adX_+}(g) \rangle = 0, \end{aligned} \tag{2.21}$$

where $p_2 = p_0 e^{(t_0 X_+)} e^{(t_1 X_-)}$. We need the following definition. If z_1, z_2, z_3 are (possibly time-varying) vector fields of $SO(3)$, the application $q \mapsto q(z_1 \wedge z_2 \wedge z_3)$ is the *field of* 3*-vectors* associated with the z_i's, where $q(z_1 \wedge z_2 \wedge z_3)$ is an element of $\bigwedge^3 T_q SO(3)$, the 3-fold exterior power of $T_q SO(3)$. We now rewrite (2.21) by using fields of 3-vectors. We obtain

$$g \wedge e^{-t_1 adX_-}(g) \wedge e^{t_2 adX_+}(g) = 0. \tag{2.22}$$

Thanks to (2.3), (2.22) is equivalent to $r(t_1, t_2) f \wedge g \wedge h = 0$ for an appropriate real-valued function r. After computations, we get $r(t_1, t_2) = \sin(\frac{t_1 - t_2}{2})$. This implies that $t_1 = t_2 = t_3$.

Similar to what we did in the proof of Proposition 5, consider now a time-optimal trajectory of the form BB_TB, where B_T is a nontrivial interior bang arc associated with a normal extremal and $T \in]0, 2\pi[$. From (2.17) and (2.13) (with $\lambda_0 = 1$), we get $\dot{\varphi}_3 = -(\varphi_2 + s_\alpha^2 u)$. Using (2.16), this means that the vector $z = (\varphi_3, \varphi_2 + u s_\alpha^2)^T$ satisfies the differential equation

$$\dot{z} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} z, \quad t \in]0, T[,$$

with boundary conditions (the switching conditions imply $\varphi_2(0) = \varphi_2(T) = 0$) $z(0) = (\varphi_3(0), u s_\alpha^2)$ and $z(T) = (\varphi_3(T), u s_\alpha^2)$. Using $\varphi_2(0) = 0$ and the fact that $u > 0$ (resp. $u < 0$) implies $0 > \dot{\varphi}_2(0) = \varphi_3(0)$ (resp. $0 < \dot{\varphi}_2(0) = \varphi_3(0)$), one easily gets $\tan(T/2) = \varphi_3(0)/(u s_\alpha^2) < 0$. It follows that $T \in (\pi, 2\pi)$. In summary, we have proved the following proposition.

PROPOSITION 6. *Let γ be a bang-bang normal extremal. Then the time duration T along an interior bang arc is the same for all interior bang arcs and verifies $\pi < T < 2\pi$.*

Remark 4. From Propositions 5 and 6, we get that, for an extremal bang-bang trajectory (normal or abnormal), the time duration T along an interior bang arc is the same for all interior bang arcs and verifies $\pi \leq T < 2\pi$.

2.4.3. Optimal trajectories containing a singular arc. The purpose of this section is to describe the structure of time-optimal trajectories containing singular arcs.

PROPOSITION 7. *Let γ be a time-optimal trajectory containing a singular arc. Then γ is of the type $B_t S_s B_{t'}$, with $s \leq \frac{\pi}{c_\alpha}$ if $t > 0$ or $t' > 0$ and $s < 2\frac{\pi}{c_\alpha}$ otherwise.*

Proof of Proposition 7. Let γ be a time-optimal trajectory containing a singular arc S_t, $t > 0$. From Proposition 3, we know that $t < 2\pi/c_\alpha$.

Assume now that γ contains a singular arc and a nontrivial interior bang arc. Then, we may assume that γ contains a piece of the type S_sB_t or B_tS_s (say the first), with B_t a complete bang arc. Then we have $\varphi_2(s) = \varphi_3(s) = \varphi_2(s+t) = 0$. This translates to $g \wedge h \wedge e^{tadX_\varepsilon}(g) = 0$. Using (2.3), it implies that $\cos(t) = 1$, i.e., $t = 2\pi$. This contradicts the time optimality of γ. Finally, from (2.7), we get $e^{\frac{\pi}{c_\alpha}adf}(g) = -g$. From this, we deduce for $t \geq 0$, $e^{\frac{\pi}{c_\alpha}f}e^{tX_\varepsilon} = e^{tX_{-\varepsilon}}e^{\frac{\pi}{c_\alpha}f}$. Therefore, for $t, s \geq 0$, we have $e^{sf}e^{tX_\varepsilon} = e^{(s-\frac{\pi}{c_\alpha})f}e^{tX_{-\varepsilon}}e^{\frac{\pi}{c_\alpha}f}$. Then, if $s > \frac{\pi}{c_\alpha}$ and $t > 0$ and taking into account what precedes, $e^{sf}e^{tX_\varepsilon}$ cannot be optimal. □

2.5. Uniform bound on the number of switchings for time-optimal trajectories. For $\alpha \in]0, \pi/2[$, let $N(\alpha)$ be the supremum of the number of switchings of any time-optimal trajectory on $SO(3)$. Thanks to the left invariance of the control system (1.1), we may assume that the supremum is taken over any time optimal trajectory starting at Id. In this section, we prove the following proposition.

PROPOSITION 8. *For $\alpha \in]0, \pi/2[$, $N(\alpha)$ is finite (and thus achieved).*

Proof of Proposition 8. Let us first prove the following claim.

CLAIM. Every optimal trajectory of (Σ) is a finite concatenation of bang and singular arcs.

Let $\gamma : [a, b] \to SO(3)$ be a time-optimal trajectory of (Σ). Let S be the set of zeros of φ_2 such that if $t \in S$, then φ_2 does not vanish identically in some neighborhood of t. Clearly, S is the set of times t such that $\gamma(t)$ is the junction of two bang arcs or the junction of a singular arc and a bang arc. The conclusion follows if S is finite. Reasoning by contradiction, S must have a limit point $\bar{t}$. Moreover $\bar{t} \in S$, otherwise φ_2 would vanish identically in a neighborhood of $\bar{t}$, contradicting the fact that $\bar{t}$ is a limit point of S. Note also that $\dot{\varphi}_2$ is continuous in an open (in $[a, b]$) neighborhood N of $\bar{t}$ (see Remark 2). By definitions of S and $\bar{t}$, there exists a sequence (t_n) in N converging to $\bar{t}$ such that $\varphi_2(t_n) \neq 0$. Choose n large enough so that if $[t'_n, t''_n]$ is the maximal subinterval containing t_n with $\varphi_2 \neq 0$ on (t'_n, t''_n), then $[t'_n, t''_n] \subset N$. Clearly, $\varphi_2(t'_n) = \varphi_2(t''_n) = 0$, γ is a bang arc on $[t'_n, t''_n]$, and $t''_n - t'_n$ tends to zero as n goes to infinity since t_n tends to $\bar{t}$. But, by Proposition 6, $t''_n - t'_n \geq \pi$ for n large enough. So we have reached a contradiction and S is finite. The claim is proved. □

To finish the proof of Proposition 8, it remains to show that the (finite) number of switchings for any time-optimal trajectory is uniformly bounded over $SO(3)$. The argument goes by contradiction: there would then exist a sequence of regular bang-bang time-optimal trajectories $B_{s_n}B_{T_n}\cdots B_{T_n}B_{t_n}$, where $s_n, t_n < 2\pi$, $\pi < T_n < 2\pi$, and the number of switchings m_n goes to infinity as n goes to infinity. Therefore, there exists a sequence of points (x_n) of $SO(3)$ such that the minimum time τ_n needed to connect Id to x_n by a trajectory of (Σ) goes to infinity as n goes to infinity.

To reach a contradiction, it is enough to show that there exists a time $\mathcal{T}$ so that, for every point $x \in SO(3)$, there exists a trajectory γ of (Σ) connecting Id to x with $T(\gamma) \leq \mathcal{T}$. By a compactness argument and thanks to the $SO(3)$-invariance of (Σ), that would result from the following fact: there exists $\bar{t} > 0$ and an open neighborhood $U \subset SO(3)$ of Id such that every point $x \in U$ can be reached from Id in time less than or equal to $\bar{t}$. The latter simply results from the facts that (Σ) has the accessibility property and e^{tf} is periodic. □

Remark 5. Since the degree of nonholonomy of the distribution generated by (f, g) is equal to 2, by standard controllability arguments, one can quantitatively relate the size of U and $\bar{t}$ as follows: U contains a ball of radius $\frac{C\bar{t}}{\alpha^2}$ for some positive

constant C. Therefore, $N(\alpha)$ can be bounded above by $\frac{C'}{\alpha^2}$, for some positive constant C'.

3. The Hopf fibration and proof of Theorem 1.

3.1. The Hopf projection. In this section, we describe explicitly the Hopf projection from $SO(3)$ to $\mathbb{R}P^2$. This projection provides $SO(3)$ with a structure of fiber bundle with base $\mathbb{R}P^2$ and fiber S^1. In what follows, we use the identification of $\mathbb{R}P^2$ with $S^2\backslash\sim$, where $\sim$ is the antipodal map, that is $\mathbb{R}P^2$ is the set of rows (y_1, y_2, y_3), $\sum y_i^2 = 1$, where $(y_1, y_2, y_3) \sim (-y_1, -y_2, -y_3)$. In what follows, $\mathbb{R}P^2$ is identified with the subset $\underline{\mathcal{NH}}$, made of the (open) top hemisphere together with half of the equator. Fix a point $y_0 \in \mathbb{R}P^2$. The Hopf projection is defined as

$$\begin{aligned}\Pi : SO(3) &\to \mathbb{R}P^2,\\ x &\mapsto y = y_0 x,\end{aligned}$$

where $y_0 x$ is the standard matrix product. Then, any left-invariant vector field $V : x \mapsto xv$ on $SO(3)$, $v \in so(3)$, is transformed by $d\Pi$ into the (left-equivariant) vector field $V_S = d\Pi(V) : y \mapsto yv$. In the following, we call, respectively, the control systems $\dot{x} = x(f + ug)$, $x \in SO(3)$, and $\dot{y} = y(f + ug)$, $y \in \mathbb{R}P^2$, the control systems *upstairs* and *downstairs*. As explained next, it is crucial to choose y_0 so that the drift term at the initial point, Idf, vanishes downstairs, so we require $\Pi(Id) = y_0$. Indeed, if this is the case, then (i) from the point y_0, we have local controllability (see next section) and this greatly helps in the construction of the optimal synthesis; and (ii) if a trajectory upstairs starts with a singular arc (that is, with $u \equiv 0$, i.e., it is an integral curve of the drift f), then its projection is a point. That suitable choice of y_0 is made possible by the following normalizations:

$$f = c_\alpha \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \text{ and } g = s_\alpha \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}, \quad y_0 = (0, 0, 1).$$

In that way, the control system downstairs reads $\dot{y} = F_S(y) + uG_S(y)$, where $F_S(y) = yf$, $G_S(y) = yg$. In order to respect the convention in control theory where states are represented by column vectors and costates by row vectors, we will consider the transposed control system and, with the change of notations $y^T \to y$, $f^T = -f \to f$, $g^T = -g \to g$, we obtain downstairs

$$\left\{\begin{array}{rcl} \dot{y} &=& F_S(y) + uG_S(y), \quad |u| \le 1, \qquad \text{where} \\ F_S(y) &=& fy = f \times y = c_\alpha \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} -y_2 \\ y_1 \\ 0 \end{pmatrix}, \\ G_S(y) &=& gy = g \times y = s_\alpha \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 0 \\ -y_3 \\ y_2 \end{pmatrix}. \end{array}\right. \tag{3.1}$$

3.2. Proof of Theorem 1. For the rest of the paper, we assume $\alpha \in]0, \pi/4[$. In this section we prove Theorem 1, using a lemma describing the structure of time-optimal trajectories of $(\Sigma)_S$ connecting the north pole to any point of $\underline{\mathcal{NH}}$, which will be studied in the next section.

LEMMA 1. *Consider the control system $(\Sigma)_S$ and the time-optimal trajectories connecting the north pole to any point of $\underline{\mathcal{NH}}$. Then*

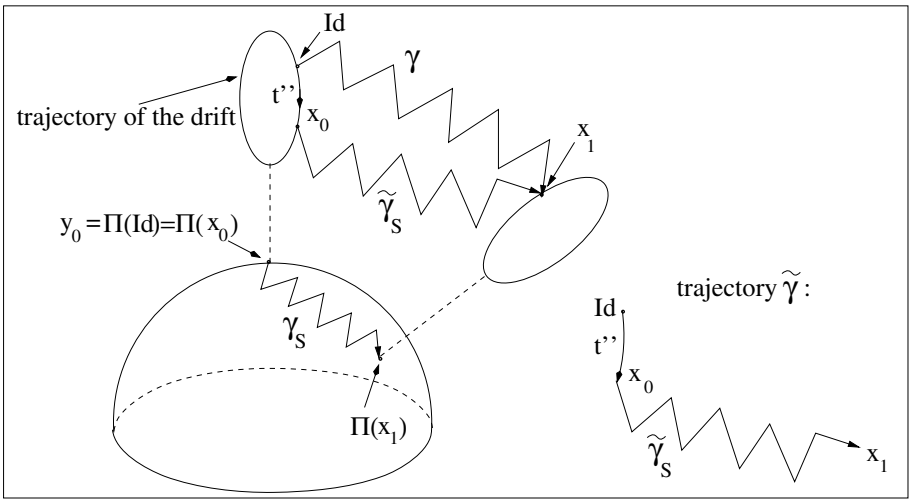


FIG. 3.1. *Proof of Theorem 1.*

(i) *They are regular bang-bang with same time durations for the interior bang arcs, i.e., they are of the type* $B_s B_{v(s)} \cdots B_{v(s)} B_t$, *where* $s \in [0, \pi]$, $t \in [0, v(s)]$ *and*

$$v(s) = \pi + 2 \arctan\left(\frac{s_s}{c_s + \cot^2(\alpha)}\right). \tag{3.2}$$

(ii) *Set* $N_0(\alpha) := 2\left[\frac{\pi}{8\alpha}\right]$. *Then the maximum number of switching of these trajectories is*

$$N_S(\alpha) = N_0(\alpha) - \left[N_0(\alpha) - \frac{\pi}{4\alpha}\right]. \tag{3.3}$$

The above formula means that $N_S(\alpha)$ *can take the values* $N_0(\alpha)$, $N_0(\alpha) + 1$, *or* $N_0(\alpha) + 2$.

(iii) *They are projections, through the Hopf map* Π, *of time-optimal trajectories of* (Σ) *starting at* Id.

Proof of Lemma 1. For the proof of (i), see Proposition 10 and Proposition 11. For the proof of (iii), see Lemma 7 and Lemma 8. For the proof of (ii) see Proposition 13. □

We now prove separately the two inequalities of Theorem 1.

Proof of the inequality $N_S(\alpha) \leq N(\alpha)$. From (iii) of Lemma 1, every time-optimal trajectory of $(\Sigma)_S$ connecting the north pole to any point of $\underline{\mathcal{NH}}$ is the projection by Π of a time-optimal trajectory of (Σ) with the same time duration (in particular, of the time-optimal trajectory connecting the two fibers). Therefore, $N_S(\alpha) \leq N(\alpha)$. □

Proof of the inequality $N(\alpha) \leq N_S(\alpha) + 4$. We refer to Figure 3.1. Consider a time-optimal trajectory γ of (Σ) containing $N(\alpha)$ switchings. With no loss of generality, we may assume that $N(\alpha) > 2$. By Propositions 6 and 7, we deduce that γ is regular bang-bang and is of the type $B_s B_T \cdots B_T B_t$, with $s, t \geq 0$, $\pi \leq T \leq 2\pi$. Since every subarc of a time-optimal trajectory is also time optimal, we may assume that $s = t = 0$. Let Id and x_1 be the initial and terminal points of γ and consider γ_S, a time-optimal trajectory for $(\Sigma)_S$ connecting $\Pi(Id)$ and $\Pi(x_1)$. From (i) of Lemma 1,

γ_S is of the type $B_{s'}B_{v(s')}\cdots B_{v(s')}B_{t'}$ with $s' \leq \pi$, $t' < v(s')$ and m interior bangs. We thus have $m \leq N_S(\alpha) - 1$. We now build, from γ_S, a suboptimal trajectory connecting Id and x_1 as follows: we can lift γ_S to $SO(3)$ to an admissible trajectory $\tilde{\gamma}_S$ of (Σ) connecting x_0 and x_1, with x_0 in the fiber of $\Pi(Id)$. It is also clear that γ_S and $\tilde{\gamma}_S$ have same time durations. By construction of the fiber of $\Pi(Id)$, we get $x_0 = e^{t''f}$ with $t'' \leq 2\pi$. Finally, the curve $\tilde{\gamma}$ obtained as the concatenation of $e^{t''f}$ and $\tilde{\gamma}_S$ is an admissible trajectory of (Σ) connecting Id and x_1. Its time duration is equal to

$$T(\tilde{\gamma}) = t'' + T(\tilde{\gamma}_S) = t'' + T(\gamma_S) = t'' + mv(s') + s' + t',$$

with $m \leq N_S(\alpha) - 1$. Since γ is time optimal, we have $T(\gamma) \leq T(\tilde{\gamma})$, which implies that

$$\Big(N(\alpha) - 1\Big)T \leq t'' + mv(s') + s' + t'.$$

Using all the estimates on T, s', t', t'' (i.e., $T \in [\pi, 2\pi)$ (cf. Remark 4), $s' \leq \pi,\ t' \leq \max_{s'\in[0,\pi]} v(s'),\ t'' < 2\pi$), we deduce that

$$\Big(N(\alpha) - 1\Big)\pi < N_S(\alpha)V(\alpha) + 3\pi, \text{ where } V(\alpha) := \max_{s\in[0,\pi]} v(s),$$

from which we have

$$N(\alpha) - N_S(\alpha) < N_S(\alpha)\frac{V(\alpha) - \pi}{\pi} + 4. \tag{3.4}$$

Set $r(\alpha) := N_S(\alpha)\frac{V(\alpha)-\pi}{\pi}$. A simple computation shows that

$$r(\alpha) = N_S(\alpha)\frac{2}{\pi}\arcsin(\tan^2(\alpha)).$$

Using (3.3), it is easy to see that $r(\alpha) \in]0, 1[$ on $]0, \pi/4[$. Since $N(\alpha)$ and $N_S(\alpha)$ are integers, we get $N(\alpha) - N_S(\alpha) \leq 4$. □

4. The time-optimal synthesis downstairs. In this section, to compute $N_S(\alpha)$, we study the time-optimal synthesis for the problem downstairs (3.1), starting from the point y_0.

DEFINITION 3. *A time-optimal synthesis for the problem downstairs* (3.1)*, starting from the point y_0, is a family of time-optimal trajectories* $\Gamma = \{\gamma_y : [0, b_y] \mapsto \mathbb{R}P^2,\ y \in \mathbb{R}P^2 :\ \gamma_y(0) = y_0,\ \gamma_y(b_y) = y\}$.

For that purpose, we use the theory of optimal syntheses on two-dimensional manifolds developed by Sussmann, Bressan, Piccoli and the first author in [9, 10, 11, 14, 15, 23, 24, 32, 33] and recently rewritten in [12]. The core of the theory consists of an explicit algorithmic construction (by induction on the number of switchings) of the optimal synthesis.

Note that the previous theory uses a more elaborated concept of synthesis, namely, that of *regular synthesis* (see for instance [8, 12, 16, 25] and cf. section 4.1.1).

In the following, in order to compute $N_S(\alpha)$ we just need to follow the steps of the algorithmic construction mentioned above, without requiring the existence of a regular synthesis. In what follows, by *time-optimal synthesis*, we refer to one in the sense of Definition 3, whose existence is simply guaranteed by Proposition 1.

Consider a two-dimensional smooth manifold M and the problem of computing the time-optimal synthesis from a fixed point $y_0 \in M$ for the control system:

$$\dot{y} = F(y) + uG(y),\quad y \in M,\quad |u| \leq 1, \tag{4.1}$$

where F and G are $\mathcal{C}^\infty$ vector fields. We introduce three functions:

$$\Delta_A(y) := \mathrm{Det}(F(y), G(y)) = F_1(y)G_2(y) - F_2(y)G_1(y), \tag{4.2}$$

$$\Delta_B(y) := \mathrm{Det}(G(y), [F,G](y)) = G_1(y)[F,G]_2(y) - G_2(y)[F,G]_1(y), \tag{4.3}$$

$$f_S(y) := -\Delta_B(y)/\Delta_A(y). \tag{4.4}$$

The sets $\Delta_A^{-1}(0), \Delta_B^{-1}(0)$ of zeros of Δ_A, Δ_B are, respectively, the set of points where F and G are parallel and the set of points where G is parallel to $[F,G]$. These loci are fundamental in the construction of the optimal synthesis. In fact, assuming that they are smooth embedded one-dimensional submanifold of M, we have the following:

- In each connected region of $M\backslash(\Delta_A^{-1}(0) \cup \Delta_B^{-1}(0))$, every extremal trajectory is bang-bang with at most one switching. Moreover, if the trajectory is switching, then the value of the control switches from -1 to $+1$ if $f_S > 0$ and from $+1$ to -1 if $f_S < 0$.
- The support of singular trajectories (that are trajectories for which the switching function identically vanishes, see Definition 4 below) is always contained in the set $\Delta_B^{-1}(0)$.
- A trajectory not switching on the set of zeros of G is an abnormal extremal (i.e., a trajectory with vanishing Hamiltonian) if and only if it switches on the locus $\Delta_A^{-1}(0)$.

Then the synthesis is built recursively on the number of switchings of extremal trajectories, canceling at each step the nonoptimal trajectories (see [12, Chapter 1]).

Remark 6. As we will see later (see Proposition 10), the condition $\alpha < \pi/4$ guarantees that *there are no singular trajectories* for the problem downstairs.

4.1. Basic definitions and facts on optimal synthesis on two-dimensional manifolds. Consider the minimum time problem for the control system (4.1). In this section, we recall some key facts for the construction of time-optimal synthesis following [12].

The first ingredient is, as usual, the PMP that, on a two-dimensional manifold, has exactly the same form as described in section 2.3 but with the following change of notation: $x \in SO(3) \to y \in M$, $\lambda(t) \in T_{\gamma(t)}SO(3) \to \lambda(t) \in T_{\gamma(t)}M$. As for the problem upstairs, switchings are described by the switching function given by Definition 4.

Definition 4 (switching function). *Let (γ, λ) be an extremal pair. The corresponding switching function is defined as $\phi(t) := \langle \lambda(t), G(\gamma(t))\rangle$.*

Again, ϕ is at least continuously differentiable ($\dot\phi(t) = \langle \lambda(t), [F,G](\gamma(t))\rangle$, cf. discussion in (2.16), and it determines the switching rule, according to Proposition 2 with the change of notation $\varphi_2 \to \phi$. Also, an extremal trajectory γ, defined on $[a,b]$, is called singular if $\phi \equiv 0$ in $[a,b]$. The following three lemmas illustrate the role of the two functions defined in (4.2) and (4.3). The proofs can be found in [9, 12, 24].

Lemma 2. *Let γ be an extremal trajectory that is singular in $[a,b] \subset Dom(\gamma)$. Then $\gamma|_{[a,b]}$ is associated with the so-called singular control $\varphi(\gamma(t))$, where*

$$\varphi(y) = -\frac{\nabla\Delta_B(y)\cdot F(y)}{\nabla\Delta_B(y)\cdot G(y)}, \tag{4.5}$$

with Δ_A and Δ_B defined in (4.2) *and* (4.3). *Moreover, on* $\mathrm{Supp}(\gamma)$, *$\varphi(y)$ is always well defined and its absolute value is less than or equal to* 1. *Finally* $\mathrm{Supp}(\gamma|_{[a,b]}) \subset \Delta_B^{-1}(0)$.

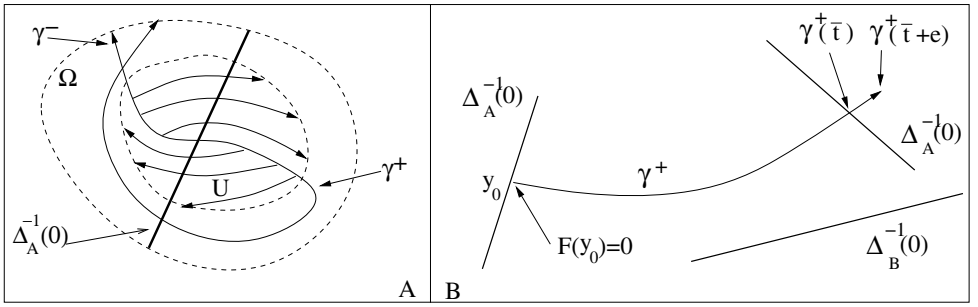


FIG. 4.1. *Lemma* 5 *and Lemma* 6.

LEMMA 3. *Let γ be an extremal bang-bang trajectory for the control problem* (4.1)*, $t_0 \in Dom(\gamma)$ be a time such that $\phi(t_0) = 0$ and $G(\gamma(t_0)) \neq 0$. Then, the following conditions are equivalent:* (i) *γ is an abnormal extremal;* (ii) *$\gamma(t_0) \in \Delta_A^{-1}(0)$; and* (iii) *$\gamma(t) \in \Delta_A^{-1}(0)$ for every time $t \in Dom(\gamma)$ such that $\phi(t) = 0$.*

The following lemma describes what happens when Δ_A and Δ_B are different from zero.

LEMMA 4. *Let $\Omega \subset M$ be an open set such that $\Omega \cap (\Delta_A^{-1}(0) \cup \Delta_B^{-1}(0)) = \emptyset$. Then all connected components of* $\mathrm{Supp}(\gamma) \cap \Omega$*, where γ is an extremal trajectory of* (4.1)*, are bang-bang with at most one switching. Moreover, if $f_S > 0$ throughout Ω, then $\gamma|_\Omega$ is associated with a constant control equal to $+1$ or -1 or has a switching from -1 to $+1$. If $f_S < 0$ throughout Ω, then $\gamma|_\Omega$ is associated with a constant control equal to $+1$ or -1 or has a switching from $+1$ to -1.*

DEFINITION 5. *Let $\gamma^+ : [0, \tau] \to M$ (resp. $\gamma^- : [0, \tau] \to M$) be the trajectory of* (4.1) *starting at y_0 and corresponding to the constant control $u \equiv 1$ (resp., $u \equiv -1$). For $t \in]0, \tau]$, let $\Gamma^+(t)$ (resp., $\Gamma^-(t)$) be the support of the curve $\gamma^+|_{[0,t]}$ (resp., $\gamma^-_{[0,t]}$).*

Under the assumption $F(y_0) = 0$ and $\Delta_B(y_0) \neq 0$, the next lemma (for a proof, see for instance [12, 26]) describes the shape of the optimal synthesis in a neighborhood of y_0. That local behavior of the optimal synthesis remains actually the same as long as $\Gamma^+(t)$ and $\Gamma^-(t)$ do not intersect $\Delta_B^{-1}(0)$ and $\Delta_A^{-1}(0)$ (except of course at y_0).

LEMMA 5. *Consider the control system* (4.1)*. Assume that $F(y_0) = 0$ and $\Delta_B(y_0) \neq 0$. Let Ω be an open neighborhood of y_0 such that $\Omega \cap \Delta_B^{-1}(0) = \emptyset$ and $\Omega \cap \Delta_A^{-1}(0)$ is an embedded one-dimensional submanifold of Ω. Let $\gamma^+ : [0, \tau] \to M$ (resp., $\gamma^- : [0, \tau] \to M$) be the trajectory of* (4.1) *starting at y_0 and corresponding to the constant control $u \equiv 1$ (resp., $u \equiv -1$). Then, for every $t_+, t_- \in]0, \tau[$ such that* (a) $\Gamma^+(t_+), \Gamma^-(t_-) \subset \Omega$, (b) $\Gamma^+(t_+) \cap \Delta_A^{-1}(0) = \Gamma^-(t_-) \cap \Delta_A^{-1}(0) = \{y_0\}$, *and* (c) $\Gamma^+(t_+) \cap \Gamma^-(t_-) = \{y_0\}$*, we have the following. There exists an open neighborhood U of $\Gamma^+(t_+) \cup \Gamma^-(t_-)$ contained in Ω such that, for every $y \in U$, there exists a unique extremal trajectory of* (4.1) *of the type $B_s B_t$ contained in U, which is time optimal and steers y_0 to y. In particular, the system* (4.1) *is controllable in U and γ^+ (resp., γ^-) is time optimal up to t_+ (resp., t_-); see Figure* 4.1(a).

Finally, we need one more lemma, related to Lemma 3 and whose hypotheses are illustrated in Figure 4.1(b).

LEMMA 6. *Consider the control system* (4.1)*. Assume that* (i) *$F(y_0) = 0$, $\Delta_B(y_0) \neq 0$,* (ii) *there exists $\bar{t}_+ > 0$ such that $\Gamma^+(\bar{t}_+) \cap \Delta_A^{-1}(0) = \{y_0, \gamma^+(\bar{t}_+)\}$, and* (iii) *there exists $\varepsilon > 0$ such that $\Gamma^+(\bar{t}_+ + \varepsilon) \cap \Delta_B^{-1}(0) = \emptyset$. Then γ^+ is extremal exactly up to time $\bar{t}_+$. Moreover, any extremal trajectory γ defined on $[0, T]$ with $T > \bar{t}_+$ and*

coinciding with γ^+ *on* $[0, \bar{t}_+]$ *switches at* $\bar{t}_+$ *to the constant control* $u \equiv -1$ *and thus* γ *is an abnormal extremal (cf. Lemma* 3*). A similar statement holds for* γ^-*.*

Remark 7. Under the hypotheses of Lemma 6, one can prove that the abnormal extremal γ restricted to an interval $[0, \bar{T}]$ is a *nonstrict abnormal extremal* if $\bar{T} < \bar{t}^+$, while it becomes a *strict abnormal extremal* if $\bar{T} \geq \bar{t}^+$ (cf. section 2.3). In other words, γ becomes a strict abnormal extremal after the first switching. These facts are analyzed in details in [11] and [12] (see Chapter 4, and in particular section 4.3, where strict abnormal extremals are called nontrivial abnormal extremals).

4.1.1. Frame curves and frame points. In this section, we briefly recall, for the sake of completeness, the main results of the theory developed in [23, 24] (see also [12]). That material is only used here and in section 5, where some numerical simulations and conjectures are presented. In [23, 24] (see also [12]), it was proved that the control system (4.1), under generic conditions on F and G (with the additional assumption $F(y_0) = 0$), admits a time-optimal *regular synthesis* in finite time T, starting from y_0. By generic conditions, we mean conditions verified on an open and dense subset of the set of $\mathcal{C}^\infty$ vector fields endowed with the $\mathcal{C}^3$ topology (see [12, formula 2.6, p. 39]). More precisely, let $\mathcal{R}(T)$ be the reachable set in time $T > 0$ given by

$$\begin{aligned}\mathcal{R}(T) := \{y \in M :\ &\exists\ b_y \in [0, T] \text{ and a trajectory} \\ &\gamma_y : [0, b_y] \to M \text{ of (4.1) such that } \gamma_y(0) = y_0,\ \gamma_y(b_y) = y\}.\end{aligned}$$

Then a *time-optimal regular synthesis* is defined by (i) a family of time-optimal trajectories $\Gamma = \{\gamma_y : [0, b_y] \to M,\ y \in \mathcal{R}(T) :\ \ \gamma_y(0) = y_0,\ \gamma_y(b_y) = y\}$ such that if $\gamma_y \in \Gamma$ and $\bar{y} = \gamma_{y(t)}$ for some $t \in [0, b_y]$, then $\gamma_{\bar{y}} = \gamma_y|_{[0,t]}$; and (ii) a stratification of $\mathcal{R}(T)$ (roughly speaking a partition of $\mathcal{R}(T)$ in manifolds of different dimensions; see [12, Definition 27, p. 56]) such that the optimal trajectories of Γ can be obtained from a feedback $u(y)$ satisfying

- on strata of dimension 2, $u(y) = \pm 1$;
- on strata of dimension 1, called *frame curves* (FC), $u(y) = \pm 1$ or $u(y) = \varphi(y)$, where $\varphi(y)$ is defined by (4.5).

The strata of dimension 0 are called *frame points* (FP). Every FP is an intersection of two FCs. In [24] (see also [12]), a complete classification of all types of FPs and FCs, under generic conditions, is provided. All the possible FCs are

- FCs of kind Y (resp., X), corresponding to subsets of the trajectories γ^+ (resp., γ^-) defined as the trajectory exiting y_0 with constant control $+1$ (resp., constant control -1);
- FCs of kind C, called *switching curves*, i.e., curves made of switching points;
- FCs of kind S, i.e., singular trajectories;
- FCs of kind K, called overlaps and reached optimally by two trajectories coming from different directions;
- FCs which are arcs of optimal trajectories starting at FPs. These trajectories "transport" special information.

The FCs of kind Y, C, S, K are depicted in Figure 4.2. There are 18 topological equivalence classes of FPs. A detailed description can be found in [10, 12, 24].

Remark 8. The proof of the existence of a regular synthesis is shown by means of a constructive algorithm (working recursively on the number of switchings) that builds explicitly the optimal trajectories (see [12, section 2.5, p. 56]). We stress the fact that the existence of a regular synthesis cannot be guaranteed before the complete

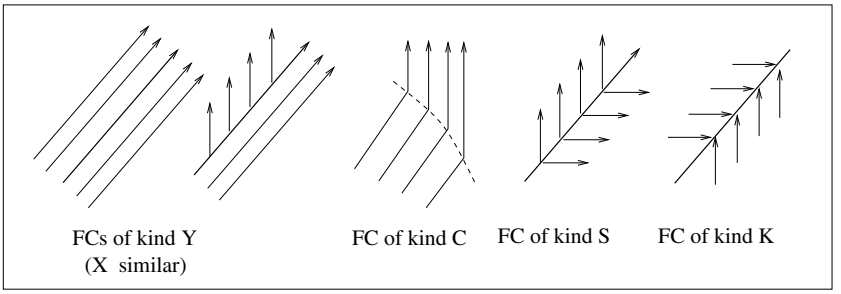


FIG. 4.2.

execution of the algorithm. Since for our system (3.1), we do not reach the end of that construction, we cannot conclude that such a regular synthesis exists. However, we conjecture that last fact (see also section 5).

4.2. The problem downstairs. In this section, we apply the theory recalled in section 4.1 to the control system (3.1) on S^2 in order to compute $N_S(\alpha)$, the maximum number of switchings for time-optimal trajectories connecting the north pole to any point of $\underline{NH}$. First we need some notations.

DEFINITION 6. *Set*

$$\begin{aligned} X_S^+(y) &= F_S(y) + G_S(y) = X^+ y = X^+ \times y \\ &= \begin{pmatrix} 0 & -c_\alpha & 0 \\ c_\alpha & 0 & -s_\alpha \\ 0 & s_\alpha & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} -c_\alpha y_2 \\ c_\alpha y_1 - s_\alpha y_3 \\ s_\alpha y_2 \end{pmatrix}, \\ X_S^-(y) &= F_S(y) - G_S(y) = X^- y = X^- \times y \\ &= \begin{pmatrix} 0 & -c_\alpha & 0 \\ c_\alpha & 0 & s_\alpha \\ 0 & -s_\alpha & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} -c_\alpha y_2 \\ c_\alpha y_1 + s_\alpha y_3 \\ -s_\alpha y_2 \end{pmatrix}. \end{aligned}$$

Let $\gamma : [t_1, t_2] \to S^2$ be a trajectory of (4.1). If γ corresponds to the constant control $+1$ (resp., -1) in $[t_1, t_2]$, we say that $\gamma|_{[t_1,t_2]}$ is a X^+–trajectory (resp., X^-–trajectory). Moreover, we call $\gamma^\pm$ the trajectories exiting the point x_0 with, respectively, constant control $+1$ and -1. Let $t_{op}^\pm$ be the last times for which $\gamma^\pm$ are optimal. We define $\gamma_{op}^\pm := \gamma^\pm|_{[0,t_{op}^\pm]}$. If $\gamma_1 : [a, b] \to S^2$ and $\gamma_2 : [b, c] \to S^2$ are trajectories of (3.1) such that $\gamma_1(b) = \gamma_2(b)$, then the concatenation $\gamma_2 * \gamma_1$ is the trajectory

$$(\gamma_2 * \gamma_1)(t) := \begin{cases} \gamma_1(t) \text{ for } t \in [a, b], \\ \gamma_2(t) \text{ for } t \in [b, c]. \end{cases}$$

Note that in the notation $\gamma_2 * \gamma_1$, γ_1 comes first.

The first quantities to be computed are $\Delta_A^{-1}(0), \Delta_B^{-1}(0)$ and the sign of f_S. Referring to Figure 4.4, we have for the system (3.1)

$$\begin{aligned} &\Delta_A^{-1}(0) = \{(y_1, y_2, y_3)^T \in S^2 : y_2 = 0\}, \\ &\Delta_B^{-1}(0) = \{(y_1, y_2, y_3)^T \in S^2 : y_3 = 0\}, \\ &f_S(y) > 0, \ \forall\, y \in \{(y_1, y_2, y_3)^T \in S^2 : y_2 y_3 > 0\}, \\ &f_S(y) < 0, \ \forall\, y \in \{(y_1, y_2, y_3)^T \in S^2 : y_2 y_3 < 0\}. \end{aligned}$$

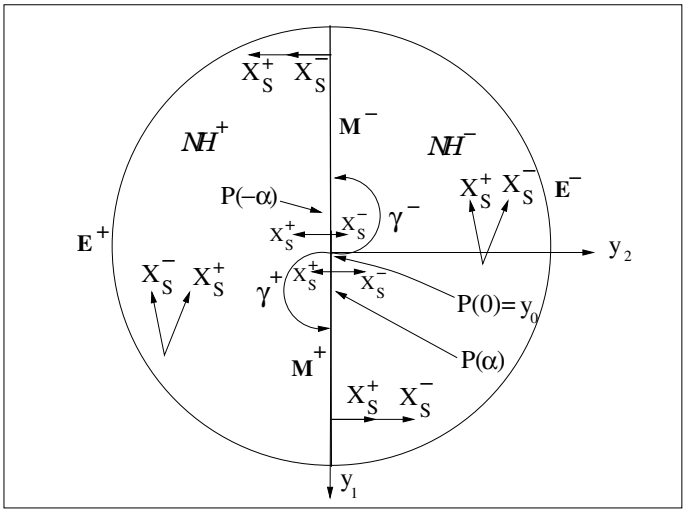


FIG. 4.3.

The set $\Delta_B^{-1}(0)$ is called the *equator* and $\Delta_A^{-1}(0)$ the *meridian*. Moreover, let $\mathcal{NH}$ be the (open) top hemisphere, i.e., the set of points $(y_1, y_2, y_3)^T$ so that $y_3 > 0$ and (see Figure 4.3)

$$\begin{aligned} \mathcal{NH}^+ &:= \{y \in \mathcal{NH} : y_2 < 0\}, \\ M^+ &= \{y \in \mathcal{NH} : y_1 > 0, y_2 = 0\}, \\ E^+ &= \{y \in S^2 : y_2 < 0, y_3 = 0\}. \end{aligned}$$

Similarly

$$\begin{aligned} \mathcal{NH}^- &= \{y \in \mathcal{NH} : y_2 > 0\}, \\ M^- &= \{y \in \mathcal{NH} : y_1 < 0, y_2 = 0\}, \\ E^- &= \{y \in S^2 : y_2 > 0, y_3 = 0\}. \end{aligned}$$

We also parametrize points y of the meridian by the oriented angle between $\overrightarrow{0y_0}$ and $\overrightarrow{0y}$. We use $P(\xi)$, $\xi \in [-\pi, \pi]$, to denote the point of the meridian defined by the angle ξ. Then $P(0) = y_0$ and $P(\alpha)$ (resp., $P(-\alpha)$) is the center of rotation in the north hemisphere of X_S^+ (resp. X_S^-). We also have that γ^+ (resp., γ^-), up to time π, is a half-circle with diameter $[y_0, P(2\alpha)]$ (resp., $[y_0, P(-2\alpha)]$); see Figure 4.3. From Lemma 4, Proposition 9 follows.

PROPOSITION 9. *Let $\gamma : [0, T] \to S^2$, $\gamma(0) = y_0$ be an optimal trajectory for the control system* (3.1). *Then*

- *γ has at most a $X^+ * X^-$ switching in $\mathcal{NH}^-$, that is, if* $\mathrm{Supp}(\gamma|_{[a,b]}) \subset \mathcal{NH}^-$, *then $\gamma|_{[a,b]}$ corresponds to one of the three following controls:*
  - (-) *$u = +1$ in $[a, b]$,*
  - (-) *$u = -1$ in $[a, b]$,*
  - (-) *there exists $c \in]a, b[$ such that $u = -1$ in $[a, c[$ and $u = +1$ in $]c, b]$;*
- *γ has at most an $X_S^- * X_S^+$ switching in $\mathcal{NH}^+$;*
- *γ has at most an $X_S^- * X_S^+$ switching in the region $\{x \in S^2 : y_2 > 0, y_3 < 0\}$;*
- *γ has at most an $X_S^+ * X_S^-$ switching in the region $\{x \in S^2 : y_2 < 0, y_3 < 0\}$.*

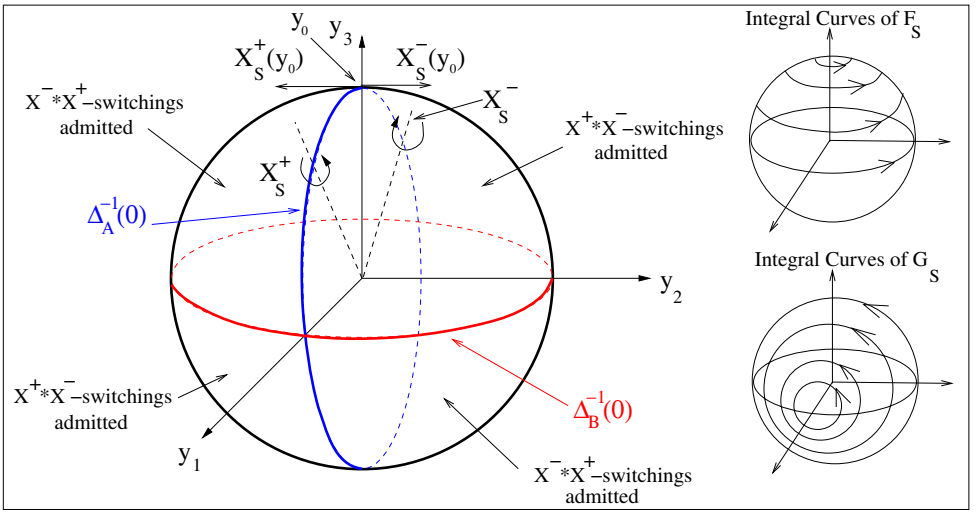


FIG. 4.4.

In Figure 4.4, the integral curves of F_S, G_S, X_S^+, X_S^- and the loci $\Delta_A^{-1}(0), \Delta_B^{-1}(0)$ are depicted. Moreover, the allowed switchings are indicated.

Remark 9. Note that, in $\mathcal{NH}^+$ (resp., $\mathcal{NH}^-$), X_S^+ points on the right (resp., on the left) of X_S^-, while, on the meridian, X_S^+ and X_S^- are parallel (see Figure 4.3). More precisely, X_S^+ and X_S^- point in the same direction on $\{P(\xi), \xi \in]\alpha, \pi-\alpha[\bigcup]-\pi+\alpha, -\alpha[\}$ and in opposite directions on $\{P(\xi), \xi \in]-\alpha, \alpha[\bigcup]\pi-\alpha, \pi]\ \bigcup\ [-\pi, -\pi+\alpha[\}$.

4.2.1. Two properties of extremal trajectories. The following two propositions are essential in the construction of the optimal synthesis.

PROPOSITION 10. *Every time-optimal trajectory of* (3.1)*, starting at the north pole, is regular bang-bang.*

Proof of Proposition 10. Since $\alpha < \frac{\pi}{4}$, by taking into account Lemmas 5 and 6, the curves γ^+ and γ^- defined in Definition 6 do not intersect the equator and are time optimal until the first time they meet the meridian, i.e., exactly up to time π. Moreover, since singular arcs are contained in the equator and thanks to Lemma 6, any time optimal trajectory γ of (3.1), with at least one switching, is of the form $B_s B_t \cdots$, with $s \in]0, \pi]$ and $t > 0$. Finally, since γ is the projection of a time-optimal trajectory $\tilde{\gamma}$ of (1.1), the latter is also of the type $B_s B_t \cdots$. Therefore, by Proposition 7, $\tilde{\gamma}$ and so γ, cannot contain any singular arc. □

PROPOSITION 11. *Let $\gamma : [0, T] \to S^2$ be a time-optimal trajectory for the control system* (3.1) *of the type $B_s B_{t_1} B_{t_2} \cdots$. Then, all time durations of interior bang arcs are equal to $v(s)$, where*

$$v(s) := \pi + 2 \arctan\left(\frac{s_s}{c_s + \cot^2(\alpha)}\right). \tag{4.6}$$

Proof of Proposition 11. Consider $\tilde{\gamma} : [0, T] \to SO(3)$, an optimal trajectory that projects on γ through the Hopf fibration Π. Thanks to Proposition 6 (see also Remark 4), we have $\tilde{\gamma} = B_s B_{t_1} B_{t_1} \cdots$, where $t_1 \in [\pi, 2\pi[$. Moreover, since that curve projects on a time-optimal trajectory for (3.1), we will establish a relation between s and t_1.

We start from the relations $\varphi_2(s) = \varphi_2(s + t_1) = 0$, which can be written as

$$\langle \lambda(s), G_S(\gamma(s)) \rangle = \langle \lambda(s + t_1), G_S(\gamma(s + t_1)) \rangle = 0. \tag{4.7}$$

Recall that $\lambda(s) = \lambda(0)e^{-sX_\varepsilon}$, $\lambda(s + t_1) = \lambda(0)e^{-sX_\varepsilon}e^{-t_1X_{-\varepsilon}}$ and $\gamma(s) = e^{sX_\varepsilon}\gamma(0)$, $\gamma(s + t_1) = e^{t_1X_{-\varepsilon}}e^{sX_\varepsilon}\gamma(0)$. Since γ is nontrivial, $\lambda(0)$ is a nonzero line vector of $\mathbb{R}^3$. Moreover, $\gamma(0) = y_0 = (0, 0, 1)^T$. Equation (4.7) can be written as

$$\lambda(0)e^{-sX_\varepsilon}(g \times e^{sX_\varepsilon}\gamma(0)) = 0, \quad \lambda(0)e^{-sX_\varepsilon}e^{-t_1X_{-\varepsilon}}(g \times e^{t_1X_{-\varepsilon}}e^{sX_\varepsilon}\gamma(0)) = 0.$$

The previous equations can be transformed to

$$\det(e^{sX_\varepsilon}\lambda(0)^T, g, e^{sX_\varepsilon}\gamma(0)) = 0, \quad \det(e^{t_1X_{-\varepsilon}}e^{sX_\varepsilon}\lambda(0)^T, g, e^{t_1X_{-\varepsilon}}e^{sX_\varepsilon}\gamma(0)) = 0$$

and then to

$$\det(e^{sX_\varepsilon}\lambda(0)^T, g, e^{sX_\varepsilon}\gamma(0)) = 0, \quad \det(e^{sX_\varepsilon}\lambda(0)^T, e^{-t_1X_{-\varepsilon}}g, e^{sX_\varepsilon}\gamma(0)) = 0.$$

Since $e^{sX_\varepsilon}\lambda(0)^T$ is not zero, we deduce that

$$\det(g, e^{sX_\varepsilon}\gamma(0), e^{-t_1X_{-\varepsilon}}g) = 0. \tag{4.8}$$

We end up with the relation

$$-s_\alpha^2 \cos(s - t_1/2) = c_\alpha^2 \cos(t_1/2). \tag{4.9}$$

Taking into account that $\pi \le t_1 < 2\pi$, we can simplify the previous equation to get (4.6). □

4.3. Construction of the time-optimal synthesis. In this section, we present, step by step, the construction of the TOS for (3.1). We will not complete that construction, but only provide here the steps for which the outcome is justified by a rigorous argument. For the other steps of the construction, we refer to the last section where we propose conjectures on their outcomes, which are supported by numerical simulation.

Step 1. By Lemmas 5 and 6, for every $\varepsilon > 0$, there exists an open neighborhood U of $\Gamma^+(\pi - \varepsilon) \cup \Gamma^-(\pi - \varepsilon)$ (recall Definition 5) where the time-optimal synthesis is described in Figure 4.5(a). Moreover, $t_{op}^+ = t_{op}^- = \pi$ (recall Definition 6).

Step 2. Taking into account the analysis of Sections 4.1 and 4.2, the time-optimal trajectories for the problem downstairs are described by the following proposition.

PROPOSITION 12. *Every time-optimal trajectory for the system* (3.1)*, starting from the north pole, is contained in the following two sets of extremals, which are parametrized by the length of the first bang arc, the one of the last bang arc, and the number of arcs:*

$$\Xi^+(s, t) = \overbrace{e^{X_S^\varepsilon t}e^{X_S^{-\varepsilon}v(s)} \cdots e^{X_S^- v(s)}e^{X_S^+ s}}^{m \quad terms} y_0, \tag{4.10}$$

$$\Xi^-(s, t) = \underbrace{e^{X_S^{\varepsilon'} t}e^{X_S^{-\varepsilon'}v(s)} \cdots e^{X_S^+ v(s)}e^{X_S^- s}}_{m' \quad terms} y_0, \tag{4.11}$$

where $s \in [0, \pi]$, $t \in [0, v(s)]$, the number of bang arcs (m and m', respectively) is an integer, and

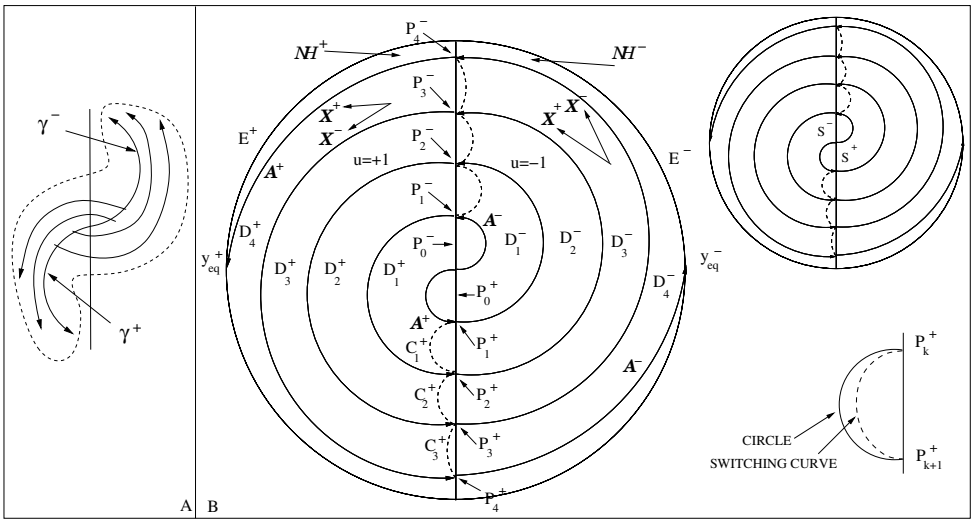


FIG. 4.5.

(-) $\varepsilon = +1$ *(resp.,* $\varepsilon = -1$*) if* m *is odd (resp., even);*
(-) $\varepsilon' = +1$ *(resp.,* $\varepsilon' = -1$*) if* m' *is even (resp., odd).*

Step 3. Let $\mathcal{A}^+$ and $\mathcal{A}^-$ be the two extremal trajectories starting, respectively, with controls $u \equiv 1$ and $u \equiv -1$, and switching after time π, i.e., corresponding, respectively, to $\Xi^+(\pi, \cdot)$ and $\Xi^-(\pi, \cdot)$. These two curves are abnormal extremals and their respective first bang arcs coincide with γ^+_{op} and γ^-_{op}. As explained in Remark 7, these two curves become strict abnormal extremals after time π.

To describe them, consider, for $\varepsilon = \pm$ and $0 \le k \le \tilde{k}$ ($\tilde{k}$ defined below), the half-circles $L^\varepsilon_k \subset Clos(\mathcal{NH}^\varepsilon)$, whose centers lie on $\overrightarrow{0P(\varepsilon\alpha)}$ and pass through the points $P^{-\varepsilon}_k$ and P^ε_{k+1}, where

$$P^+_n := P(2n\alpha), \quad P^-_n := P(-2n\alpha)$$

for the integers n so that $2n\alpha \le \frac{\pi}{2} + 2\alpha$. Note that $\frac{\pi}{2} + 2\alpha < \pi$ for $\alpha < \frac{\pi}{4}$ and, in fact, the last P^ε_n belongs to the bottom-half hemisphere, i.e., $n \le \tilde{k}$, where $\tilde{k} := 2 + \left[\frac{\pi}{4\alpha}\right]$. It is easy to see that $\mathcal{A}^+$ intersects the top half-meridian according to the following ordered sequence of points: y_0, P^+_1, P^-_2, $P^+_3, \ldots$. Similarly, $\mathcal{A}^-$ intersects the top half-meridian at y_0, P^-_1, P^+_2, $P^-_3, \ldots$. Moreover, let y^+_{eq} and y^-_{eq} be the antipodal points of the equator which are the respective first intersections of $\mathcal{A}^+$ and $\mathcal{A}^-$ with the equator. Note that they are reached at the same time T_{eq}. Finally, consider the open subset of the top hemisphere bounded below by the equator and obtained by removing the supports of $\mathcal{A}^+$ and $\mathcal{A}^-$ up to time T_{eq}, i.e., all the L^ε_k. That set is the disconnected union of the two "snake-shaped" simply connected regions S^+ and S^- (defined so that each S^ε contains the center of rotation of X^ε_S). Clearly S^+ and S^- are made of open segments of the meridian and open simply connected regions $D^\varepsilon_k \subset \mathcal{NH}^\varepsilon$ defined as follows. For $k = 0$, D^ε_0 is delimited by $\mathrm{Supp}(\gamma^\varepsilon_{op})$ and the segment $[P_0, P^\varepsilon_1]$ and, for $k \ge 1$, D^ε_k is delimited by L^ε_{k-1} on the top, L^ε_k on the bottoms, and by the segments $[P^{-\varepsilon}_{k-1}, P^{-\varepsilon}_k]$ and $[P^\varepsilon_k, P^\varepsilon_{k+1}]$ on the sides; see Figure 4.5(b).

In what follows, if A, B are two subsets of points of S^ε, we say that A is *above* B (or equivalently B is *below* A) if $A \subset D^{\varepsilon'}_k$ and $B \subset D^{\varepsilon''}_{k'}$ with $k < k'$, for some $\varepsilon', \varepsilon''$.

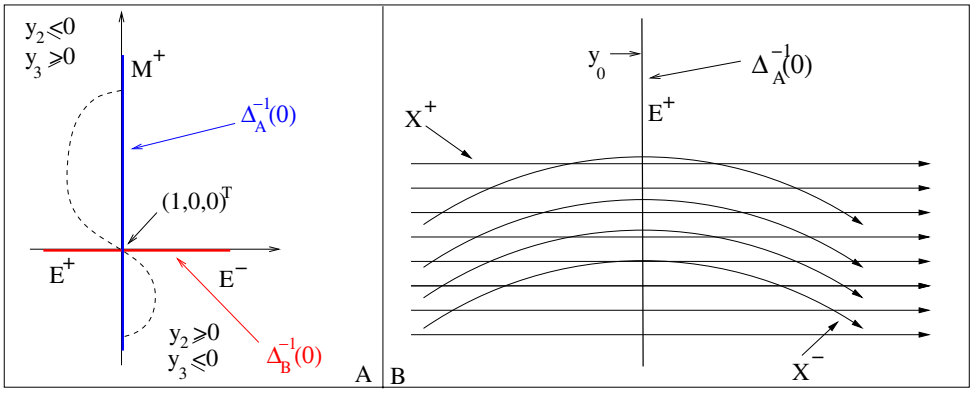


FIG. 4.6.

Step 4. The switching curves (SC), associated with the set of extremals given in (4.10) and (4.11), are defined as follows: they can be divided into two families, (C_k^+) and (C_k^-). If $\varepsilon = \pm$, $1 \le k \le N_0 - 1$, and $s \in [0, \pi]$, then

$$C_1^\varepsilon(s) = e^{X_S^\varepsilon v(s)} e^{X_S^{-\varepsilon} s} y_0, \quad C_{k+1}^\varepsilon(s) = e^{X_S^\varepsilon v(s)} C_k^{-\varepsilon}(s).$$

The boundary points of C_k^ε are $C_k^\varepsilon(0) = P_k^\varepsilon$ and $C_k^\varepsilon(\pi) = P_{k+1}^\varepsilon$. By using Proposition 9 and since $v(s) \ge \pi$, $s \in [0, \pi]$, the support of C_k^ε is contained in the subset of $Clos(\mathcal{NH}^\varepsilon)$, delimited by the half-circle centered on $\overrightarrow{0P(\alpha(2k+1))}$ and passing through the points $P_k^\varepsilon, P_{k+1}^\varepsilon$, and the segment of the meridian $[P_k^\varepsilon, P_{k+1}^\varepsilon]$; see Figure 4.5(b). In particular, a SC with boundary points in the top hemisphere is entirely contained in the top hemisphere and the intersection of its support with the top meridian reduces to its boundary points (see Lemma 3).

We next describe the shape of the first SC intersecting the equator. By symmetry, we may assume $\varepsilon = +$. We claim that its intersection with the equator reduces to the point $P(\frac{\pi}{2}) = (1, 0, 0)^T$. Indeed, by the switching rules established in Proposition 9, the SC intersecting the equator is contained in $\{y \in S^2 : y_2 \le 0, y_3 \ge 0\} \cup \{y \in S^2 : y_2 \ge 0, y_3 \le 0\}$. Taking into account the regularity of the SC and the values of its boundary points, the claim is proved; see Figure 4.6(a).

4.4. Computation of $N_S(\alpha)$. In the previous section, we provided detailed information about extremal trajectories and switching curves but we did not show that every extremal of (4.10) and (4.11) is in fact time optimal. Anyway, a rigorous derivation of $N_S(\alpha)$ is possible with the available knowledge of time-optimal trajectories combined with the subsequent lemmas.

LEMMA 7. *Every time-optimal trajectory γ starting at y_0 intersects the equator at most once.*

Proof of Lemma 7. We argue by contradiction. There would exist two distinct points of the equator q_i, q_f so that $\gamma(t_i) = q_i$, $\gamma(t_f) = q_f$ and $\gamma|_{(t_i,t_f)}$ is entirely contained in the (closed) bottom hemisphere. Let γ_{sing} be the integral curve of F_S (contained in the equator) connecting q_i to q_f. Consider now the region of the bottom hemisphere bounded by γ_{sing} and $\gamma|_{(t_i,t_f)}$. Taking into account, first, the relative positions of X_S^+, X_S^-, F_S, and G_S along the equator and, second, the sign of f_S in the bottom hemisphere, one can check that $T(\gamma_{sing}) \le T(\gamma|_{(t_i,t_f)})$. The argument is similar to that of [32] (see also [12]) and is based on the use of Stokes theorem. Since

time-optimal trajectories starting at y_0 do not contain a singular arc, it follows that γ cannot be time optimal. We have reached a contradiction. □

LEMMA 8. *Every time-optimal trajectory γ, starting at y_0 and remaining in $\underline{\mathcal{NH}}$, is the projection of a time-optimal trajectory of (Σ) starting at Id.*

Proof of Lemma 8. From the definition of the Hopf fibration, every trajectory γ of $(\Sigma)_S$, starting at y_0, associated with an admissible control u and staying in $\underline{\mathcal{NH}}$, is the projection of the trajectory $\bar{\gamma}$ of (Σ) starting at Id with the same control u. In particular, γ and $\bar{\gamma}$ have same time duration. It is clear that if γ is time optimal, then $\bar{\gamma}$ is also time optimal. □

LEMMA 9. *Recall that $S^\varepsilon \subset \mathcal{NH}$. With the notations above, choose any point y in the region S^ε and let γ_y be a time-optimal trajectory connecting the north pole y_0 to y. If $s \in]0, \pi[$ is the time duration of the first bang arc and $T(y)$ the total time duration of γ_y, then $\gamma_y|_{(s,T(y)]}$ is entirely contained in S^ε.*

Proof of Lemma 9. By the switching rules of Proposition 9, along every time-optimal trajectory contained in $\mathcal{NH}^\varepsilon$, the control must switch from ε to $-\varepsilon$, when arriving at a switching curve C_k^ε. In addition, the time-optimal trajectory switches from being an arc of a circle (integral curve of X_S^ε) to another arc of a circle of bigger radius (integral curve of $X_S^{-\varepsilon}$). After rectification of the flow of X_S^ε (i.e., the one entering the SC C_k^ε), and then by taking into account Remark 9, one gets the situation depicted in Figure 4.6(b).

By contradiction, we assume that there exists a time-optimal trajectory γ with time duration T and first bang arc time duration $s < T$ such that γ connects y_0 to $y \in S^\varepsilon$ and $\gamma|_{(s,T]}$ exits from S^ε. Let t' be the smallest time (in $[0, T]$) so that $\gamma|_{(t',T]}$ is entirely contained in S^ε. Then $\gamma(t')$ belongs to $\mathrm{Supp}(\mathcal{A}^+_{[0,T_{eq}]}) \cup \mathrm{Supp}(\mathcal{A}^-_{[0,T_{eq}]})$ (see step 3 of section 4.3 for the definition of T_{eq}). If $\gamma(t')$ is on the (top) meridian, then it has to switch so that the interior bang time duration is constant, equal to π. Therefore, $\gamma|_{(t',T]}$ will never re-enter S^+. We thus deduce that $\gamma(t')$ is not on the meridian and, with no loss of generality, we will assume that $\gamma(t')$ belongs to the (one-dimensional) interior of some L_k^+, $k \geq 1$. Now we make the following two claims.

CLAIM 1. *With the notations above, there exist $t'' < t' < t'''$ such that*

$$\gamma|_{(t',t''')} \subset D_k^{\varepsilon'} \subset S^\varepsilon \text{ and } \gamma|_{(t'',t')} \subset D_{k+1}^{\varepsilon'} \subset S^{-\varepsilon}, \text{ for some } \varepsilon' \in \{+,-\},$$

i.e., γ passes (backward in time) from S^ε to $S^{-\varepsilon}$ at time t' by going "down."

Proof of Claim 1. It is clear that there exists a neighborhood U of t' so that $\gamma|_U$ is an integral curve of $X_S^{-\varepsilon}$. Thanks to Remark 9 and to the argument above, $\gamma|_U$ intersects $Int(L_k^+)$ transversally (see Figure 4.7) in such a way that γ, run backward in time, goes from D_k^+ to D_{k+1}^+. Claim 1 is proved. □

Now, by definition of t', $\gamma(t') \in \mathcal{A}^\varepsilon|_{[0,T_{eq}]}$. Let γ_{ab} be the restriction of $\mathcal{A}^\varepsilon$ between y_0 and $\gamma(t')$. Consider $\tilde{\gamma}$, the concatenation of γ_{ab} and $\gamma|_{(t',T]}$. The conclusion of Lemma 9 will follow if one can show that the time duration T' of $\tilde{\gamma}$ is less than T, the time duration of γ. This, in turn, amounts to showing that T' (the time duration of γ_{ab}) is less than t' (the time duration of $\gamma|_{[0,t']}$). This is the object of the next claim.

CLAIM 2. *With the notations above, we have $T' < t'$.*

Proof of Claim 2. The trajectory γ, run backward in time from t', is an $X_S^{-\varepsilon'}$-integral curve until it hits a SC in some $D_L^\varepsilon \in \mathcal{NH}^{\varepsilon'}$, for some integer $L \geq k+1$, at a point $C_L^{\varepsilon'}(\bar{s})$, $\bar{s} \in]0, \pi[$. One can easily conclude that the only possibility is $L = k+1$. By Claim 1, a time-optimal trajectory can pass (backward in time) from S^ε to $S^{-\varepsilon}$ only by going down, i.e., by passing from some $D_k^{\varepsilon'}$ to $D_{k+1}^{\varepsilon'}$. Therefore, by

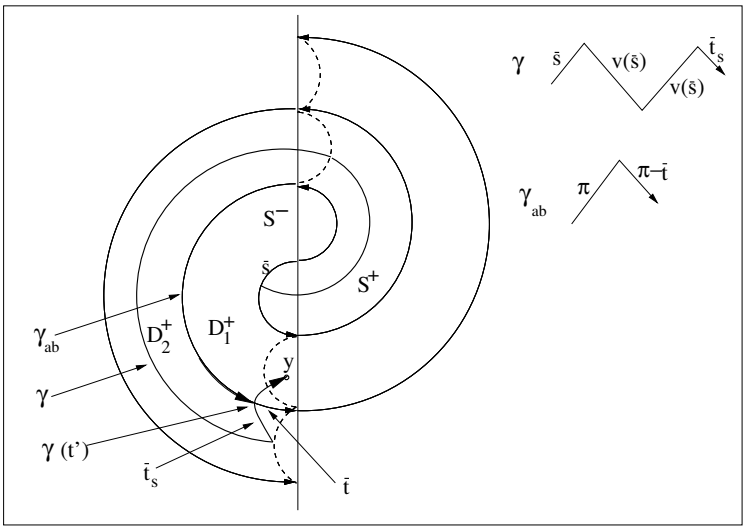


FIG. 4.7. *Proof of Lemma* 9. *Here to fix the ideas we set* $\varepsilon = -,\ \varepsilon' = +,\ k = 1$.

an elementary counting argument, one gets

$$t' = \bar{s} + t_{\bar{s}} + (k+1)v(\bar{s}),$$

where $t_{\bar{s}}$ is the time needed to go from $\gamma(t')$ to $C^+_{k+1}(\bar{s})$. On the other hand,

$$T' = (K+1)\pi - \tilde{t},$$

where $\tilde{t} \in]0, \pi[$ is the time needed for $\mathcal{A}^\varepsilon$ to go from $\gamma(t')$ to P^+_{k+1}. Since $v(\bar{s}) > \pi$, $T' < t'$. The proof of Lemma 9 is finished. □

Remark 10. Coupled with the proof of Claim 2, a simple continuity argument implies that $\mathcal{A}^+$ and $\mathcal{A}^-$ are time-optimal trajectories in the top hemisphere.

Gathering all the information on time-optimal trajectories, we are now able to compute $N_S(\alpha)$.

PROPOSITION 13. *For* $\alpha \in]0, \pi/4[$, *we have*

$$N_S(\alpha) := 2\left[\frac{\pi}{8\alpha}\right] - \left[2\left[\frac{\pi}{8\alpha}\right] - \frac{\pi}{4\alpha}\right]. \tag{4.12}$$

Proof of Proposition 13. Let $y \in S^+$ and γ be a time-optimal trajectory connecting y_0 to y. The point y belongs to some D^+_k, $k \le N_0$ and, by Lemma 9, γ remains in S^+. Since the function v takes values in $[\pi, \pi + \pi/2]$, it is easy to see from (4.10) and (4.11) and Remark 10 that γ, run backward in time, will go through the ordered sequence of regions D^+_k, D^-_{k-1}, D^+_{k-2}, etc. until hitting one of the two curves γ^+_{op} or γ^-_{op}. Moreover, in each of the regions D^ε_l, γ will switch *exactly* once, thanks to Proposition 9. Therefore, the number of times an optimal trajectory γ starting at y_0 switches is exactly equal to the number of times γ crosses the subset of the meridian

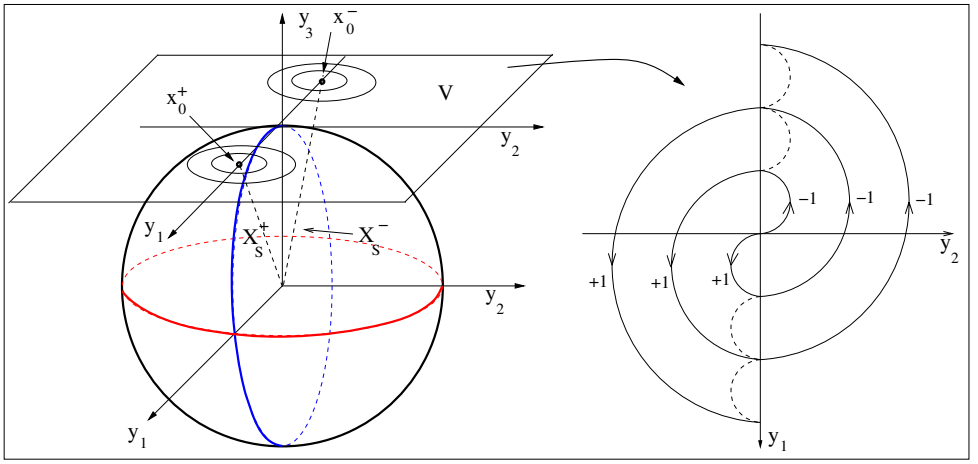


FIG. 4.8. *Stereographic projection and synthesis of the linear pendulum.*

contained in $\mathcal{NH}$. The same conclusion holds for points belonging to S^-, $\mathcal{A}^+$, and $\mathcal{A}^-$. By a systematic examination of all the possible cases, we end up with (4.12). Note that $N_S(\alpha)$ is the number of switchings for a time-optimal trajectory ending on the equator. □

4.5. Geometric remarks.

4.5.1. Relations with the linear pendulum. In a fixed neighborhood of the north pole, the control system on the sphere (3.1) behaves, when $\alpha > 0$ is small enough, as a controlled linear pendulum. More precisely, let us consider the stereographic projection of the sphere from the south pole $(0, 0, -1)$ on V, the tangent plane to the sphere at the north pole. If y_1, y_2, y_3 are the coordinates of the three-dimensional Euclidean space where the sphere is embedded, a system of coordinates on V is (y_1, y_2); see Figure 4.8. Let x_0^+ and x_0^- be the projections of the equilibrium points of X_S^+ and X_S^- in $\mathcal{NH}$.

An alternative way of parametrizing this problem (instead of fixing the radius of the sphere and varying the axes of rotations) consists of fixing the points $x_0^+ = (1, 0)^T$, $x_0^- = (-1, 0)^T$ and varying the radius r of the sphere. The relation between α and r is $\tan(\alpha) = 1/r$. The range $\alpha \in]0, \pi/4[$ becomes $r \in]1, \infty[$ and $\alpha \to 0$ corresponds to $r \to \infty$. In V, fix a ball $B(0, r_0)$ of radius $r_0 > 0$ centered in the origin, and consider the stereographic projection of the integral curves of X_0^+ and X_0^-. For $r \to \infty$, they become circles centered at the points $x_0^\pm$. Then, one easily sees that, in $B(0, r_0)$, the limit system (and the associated synthesis) corresponds to a controlled linear pendulum (with the associated synthesis) given by the equation $\dot{y}_1 = -y_2$, $\dot{y}_2 = y_1 - u$, $|u| \le 1$. Note that $\lim_{\alpha \to \infty} v(s) = \pi$, that is exactly the time duration of interior bang arcs for the linear pendulum.

4.5.2. The time-optimal problem on $SU(2)$. The optimal control problem on $\underline{NH}$ is the projection (by a Hopf fibration) of an optimal control problem on $SO(3)$. Similarly, the corresponding problem on the whole sphere S^2 is the projection (by an appropriate Hopf fibration) of an optimal control problem on $SU(2)$. Indeed, $SU(2)$ is the universal (double) covering of $SO(3)$ and they have the same Lie algebra $so(3)$. The existence of that double covering justifies, by a factor 2, the difference

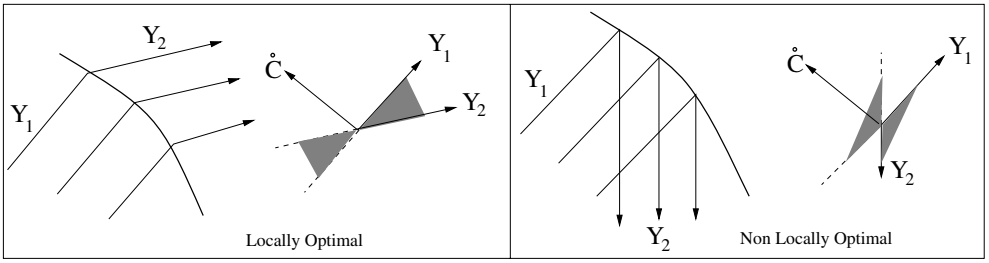


FIG. 5.1. *Definition 7.*

between our bound and the bound (1.2), on the maximal number of switchings for the control problem on $SO(3)$. Indeed, the index theory developed by Agrachev and Gamkrelidze in [3, 5] provides a bound on the number of switchings by proving that a certain extremal is not optimal because it loses *local optimality* working at the Lie algebraic level. This is why the upper bound in (1.2) corresponds (essentially) to a control problem on $SU(2)$, and thus, after projection, on a control problem on the whole sphere S^2, and not just on $\underline{NH}$. The other factor 2, of the difference between our bound and the bound given in (1.2), comes from the fact that in [3] the index of the second variation was estimated up to an additive factor 1 (see [3, p. 275]).

5. Conclusion and open problems. In the previous section, we derived a set of properties of the optimal synthesis that were sufficient to compute the maximum number of switchings of a time-optimal trajectory joining y_0 to any point of the north hemisphere. This enabled us to provide a precise estimate for $N(\alpha)$, $\alpha \in]0, \pi/4[$. However, the following questions remain unsolved.

Question 1. Are all the extremal trajectories (4.10) and (4.11) optimal in the north hemisphere?

The answer to this question depends on the answer to the next question.

Question 1′. In the north hemisphere, are the switching curves $C_k^\varepsilon(s)$, $s \in]0, \pi[$, locally optimal? (The points $s = 0, s = \pi$ are not included since we already know that the two abnormal extremal $\mathcal{A}^\pm$ are optimal in $\mathcal{NH}$.)

Roughly speaking we say that a switching curve is locally optimal if it never "reflects" the trajectories. More precisely, we have the following definition (clarified by Figure 5.1).

DEFINITION 7. *Consider a smooth switching curve C between two smooth vector fields Y_1 and Y_2 on a smooth two-dimensional manifold. Let $C(s)$ be a smooth parametrization of C. We say that C is locally optimal if, for every $s \in Dom(C)$, we have*

$$\dot{C}(s) \neq \alpha_1 Y_1(C(s)) + \alpha_2 Y_2(C(s)), \text{ for every } \alpha_1, \alpha_2 \text{ such that } \alpha_1 \alpha_2 \geq 0. \tag{5.1}$$

The points of a switching curve on which relation (5.1) is not satisfied are usually called "conjugate points."

Remark 11. Note that, if all the switching curves are locally optimal in the north hemisphere, it follows that the set of extremals (4.10) and (4.11) (restricted to $\underline{\mathcal{NH}}$) is an optimal synthesis for problem (3.1) on $\mathbb{R}P^2$. In this case, on $\mathbb{R}P^2$, the extremals (4.10) and (4.11) lose global optimality before losing local optimality.

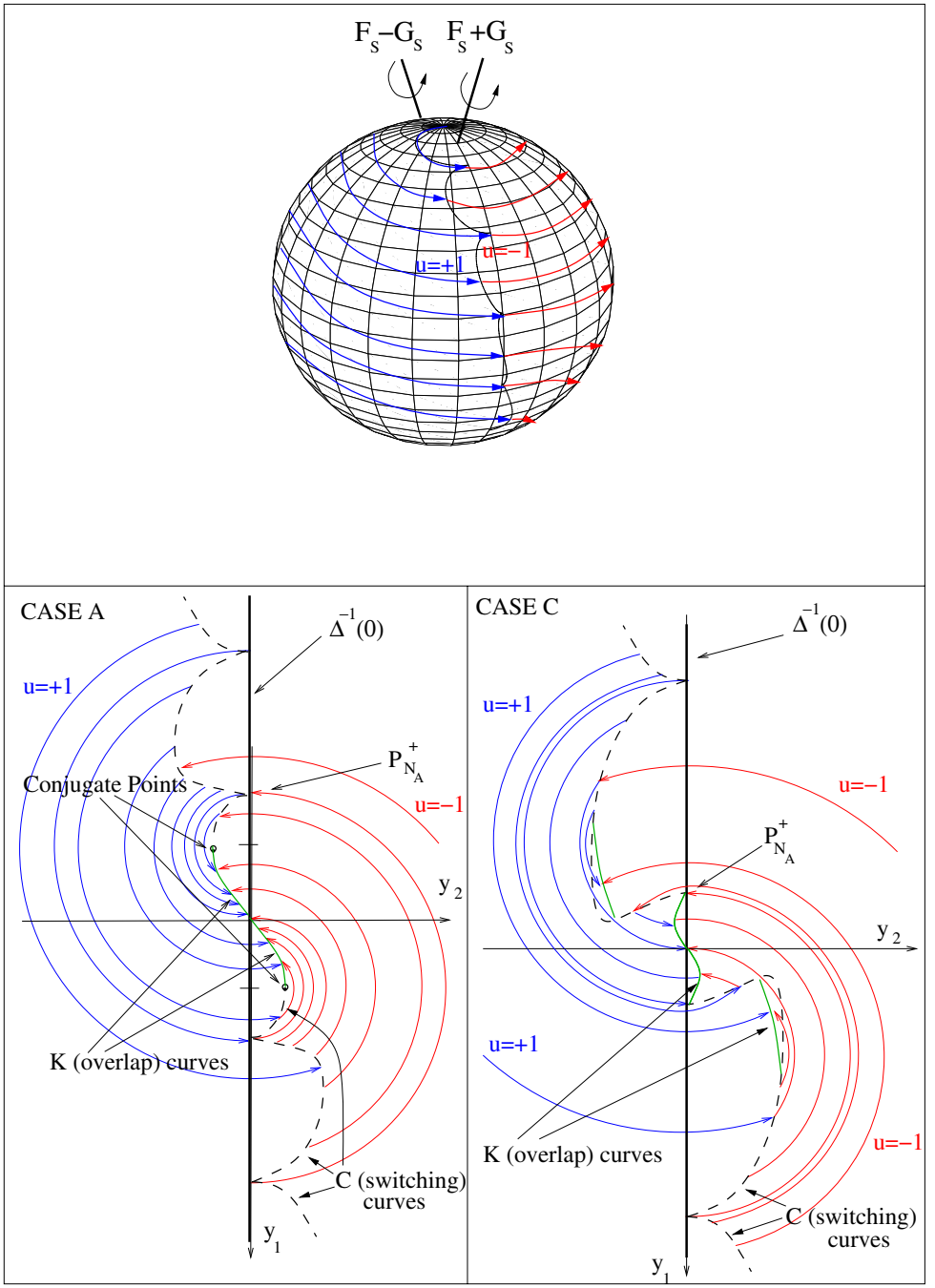


FIG. 5.2. *The time-optimal synthesis on the sphere (top) and optimal synthesis in a neighborhood of the south pole, Case A and Case C (bottom).*

Question 2. If the answer to Question 1′ is yes, what about the same question for the optimal control problem on S^2? More precisely, one would like to understand how the extremal trajectories (4.10) and (4.11) are going to lose optimality in a neighborhood of the south pole (i.e., if the loss of optimality is local or just global).

Question 3. What is the shape of the optimal synthesis in a neighborhood of the south pole?

In this section, we present the results of some numerical simulations which provide some hints regarding the above questions. More precisely we make the following observations.

- There is strong numerical evidence for a positive answer to Question 1′. This means that the switching curves in the north hemisphere never reflect trajectories. In other words, situations like those considered in the proof of Lemma 9 (cf. Figure 4.7) are not possible.
- As regards Question 2, we conjecture the following:

  C1. The curves $C_k^\varepsilon(s)$, $s \in]0, \pi[$ are locally optimal if and only if $X_S^+(C_k^\varepsilon(0)) = \alpha_1 X_S^- C_k^\varepsilon(0)$ and $X_S^+(C_k^\varepsilon(\pi)) = \alpha_2 X_S^- C_k^\varepsilon(\pi)$ with $\alpha_1, \alpha_2 \geq 0$ but not both vanishing.

  This condition is verified if and only if $k \leq \left[\frac{\pi-\alpha}{2\alpha}\right] - 1$, which simply follows from Remark 9.

  Set $N_A := \left[\frac{\pi}{2\alpha}\right]$. Analyzing the evolution of the minimum time wave front in a neighborhood of the south pole, it is reasonable to conjecture the following.

  C2. For $T \leq (N_A - 1)\pi$, the synthesis built above is optimal. Every $x \in S^2$ is reached in time $T \leq (N_A + 1)\pi$. sxsEvery optimal trajectory has at most N_A switchings and there exists an optimal trajectory having $N_A - 1$ switchings.

  On the top of Figure 5.2, the optimal synthesis is plotted.
- Regarding Question 3, numerical simulations suggest that the shape of the optimal synthesis for time $T > (N_A - 1)\pi$ depends on the remainder

$$r := \pi - 2\alpha N_A = \pi - 2\alpha \left[\frac{\pi}{2\alpha}\right].$$

  Note that r belongs to the interval $[0, 2\alpha[$. More precisely, we conjecture the following

  C3. For $\alpha \in]0, \pi/4[$, there exist two positive numbers α_1 and α_2 such that $0 < \alpha_1 < \alpha < \alpha_2 < 2\alpha$ and

    Case A: $r \in]\alpha_2, 2\alpha[$. The switching curve starting at $P_{N_A}^+$ glues to an overlap curve that passes through the origin (see the bottom of Figure 5.2, Case A).

    Case B: $r \in [\alpha_1, \alpha_2]$. An overlap curve starts exactly at $P_{N_A}^+$ and passes through the origin.

    Case C: $r \in]0, \alpha_1[$. The situation is more complicated and it is depicted in the bottom of Figure 5.2, Case C.

  For $r = 0$, the situation is the same as in Case A, but for the switching curve starting at $P_{N_A-1}^+$.

Acknowledgments. The authors are grateful to Andrei Agrachev for suggesting the problem and for many geometric hints. The authors would also like to thank Benedetto Piccoli, Gregorio Falqui, Mario Sigalotti, and Paolo Mason for helpful discussions.

REFERENCES

[1] R. Abraham and J. E. Marsden, *Foundations of Mechanics*, Benjamin-Cummings, London, 1978.

[2] A. A. Agrachev, *Methods of control theory in nonholonomic geometry*, in Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zurich, 1994), Birkhauser, Basel, 1995, pp. 1473–1483.

[3] A. A. Agrachev and R. V. Gamkrelidze, *Symplectic geometry for optimal control,* in Nonlinear Controllability and Optimal Control, Monogr. Textbooks Pure Appl. Math. 133, Dekker, New York, 1990, pp. 263–277.

[4] A. A. Agrachev and Y. L. Sachkov, *Control Theory from the Geometric Viewpoint*, Encyclopaedia Math. Sci. 87, Control Theory and Optimization II, Springer-Verlag, Berlin, 2004.

[5] A. A. Agrachev and R. V. Gamkrelidze, *Symplectic methods for optimization and control*, in Geometry of Feedback and Optimal Control, Monogr. Textbooks Pure Appl. Math. 207, Dekker, New York, 1998, pp. 19–77.

[6] A. A Agrachev and M. Sigalotti, *On the local structure of optimal trajectories in* $\mathbb{R}^3$, SIAM J. Control Optim., 42, (2004), pp. 513–531.

[7] A. A. Agrachev, G. Stefani and P. Zezza, *Strong optimality of a bang-bang trajectory*, SIAM J. Control Optim., 41 (2002), pp. 991–1014.

[8] V. Boltyanskii, *Sufficient condition for optimality and the justification of the dynamic programming principle*, SIAM J. Control Optim., 4 (1966), pp. 326–361.

[9] U. Boscain and B. Piccoli, *Extremal syntheses for generic planar systems*, J. Dynam. Control Systems, 7 (2001), pp. 209–258.

[10] U. Boscain and B. Piccoli, *Morse properties for the minimum time function on* 2*-D manifolds*, J. Dynam. Control Systems, 7 (2001), pp. 385–423.

[11] U. Boscain and B. Piccoli, *On automaton recognizability of abnormal extremals*, SIAM J. Control Optim., 40 (2002), pp. 1333–1357.

[12] U. Boscain and B. Piccoli, *Optimal Syntheses for Control Systems on* 2*-D Manifolds*, Math. Appl. (Berlin) 43, Springer, Berlin, 2004.

[13] U. Boscain and Y. Chitour, *On the minimum time problem for driftless left-invariant control systems on* $SO(3)$*,* Commun. Pure Appl. Anal., 1 (2002), pp. 285–312.

[14] A. Bressan and B. Piccoli, *Structural stability for time-optimal planar syntheses*, Dyn. Continuous Discrete Impuls. Syst., 3 (1997), pp. 335–371.

[15] A. Bressan and B. Piccoli, *A generic classification of time optimal planar stabilizing feedbacks*, SIAM J. Control Optim., 36 (1998), pp. 12–32.

[16] P. Brunovský, *On the structure of optimal feedback systems*, in Proceedings of the International Congress of Mathematicians (Helsinki, 1978), Acad. Sci. Fennica, Helsinki, 1980, pp. 841–846.

[17] L. E. Dubins, *On curves of minimal length with a constraint on average curvature and with prescribed initial and terminal positions and tangents*, Am. J. Math, 79 (1957), pp. 497–516.

[18] V. Jurdjevic, *Geometric Control Theory,* Cambridge Stud. Adv. Math., Cambridge University Press, Cambridge, 1997.

[19] A. J. Krener and H. Schattler, *The structure of small-time reachable sets in low dimensions*, SIAM J. Control Optim., 27 (1989), pp. 120–147.

[20] I. Kupka, *Ubiquity of the Fuller phenomenon,* in Nonlinear Controllability and Optimal Control, Monogr. Textbooks Pure Appl. Math. 133, Dekker, New York, 1990.

[21] E. B. Lee and L. Markus, *Foundations of Optimal Control Theory,* John Wiley, New York, London, Sydney, 1967.

[22] A. Marigo and B. Piccoli, *Regular syntheses and solutions to discontinuous ODEs*, ESAIM Control Optim. Calc. Var., 7 (2002), pp. 291–308.

[23] B. Piccoli, *Regular time-optimal syntheses for smooth planar systems,* Rend. Sem. Mat. Univ. Padova, 95 (1996), pp. 59–79.

[24] B. Piccoli, *Classifications of generic singularities for the planar time-optimal synthesis*, SIAM J. Control Optim., 34 (1996), pp. 1914–1946.

[25] B. Piccoli and H. J. Sussmann, *Regular synthesis and sufficiency conditions for optimality*, SIAM J. Control Optim., 39 (2000), pp. 359–410.

[26] L. S. Pontryagin, V. Boltianski, R. Gamkrelidze, and E. Mitchtchenko, *The Mathematical Theory of Optimal Processes*, John Wiley, New York, 1961.

[27] A. V. Sarychev, *Index of second variation of a control system*, (Russian) Mat. Sb. (N.S.) 113(155), (1980), pp. 464–486, 496.

[28] H. Schattler, *Regularity properties of optimal trajectories: Recently developed techniques*, in Nonlinear Controllability and Optimal Control, Monogr. Textbooks Pure Appl. Math. 133, Dekker, New York, 1990, pp. 351–381.

[29] P. Souères and J. P. Laumond, *Shortest paths synthesis for a car-like robot*, IEEE Trans. Automat. Control, 41 (1996), pp. 672–688.

[30] G. Stefani, *Higher order variations: How can they be defined in order to have good properties?* in Nonsmooth Analysis and Geometric Methods in Deterministic Optimal Control (Minneapolis, MN, 1993), IMA Vol. Math. Appl. 78, Springer, New York, 1996, pp. 227–237.

[31] R. Strichartz, *Sub-Riemannian geometry,* J. Differ. Geom., 24 (1983), pp. 221–263.

[32] H. J. Sussmann, *The structure of time-optimal trajectories for single-input systems in the plane: The general real-analytic case,* SIAM J. Control Optim., 25 (1987), pp. 868–904.

[33] H. J. Sussmann, *Regular synthesis for time optimal control of single-input real-analytic systems in the plane,* SIAM J. Control Optim., 25 (1987), pp. 1145–1162.

[34] H. J. Sussmann, *Envelopes, conjugate points, and optimal bang-bang extremals*, in Algebraic and Geometric Methods in Nonlinear Control Theory, Math. Appl. 29, Reidel, Dordrecht, 1986, pp. 325–346.

[35] H. J. Sussmann, *Envelopes, higher-order optimality conditions and Lie brackets,* in Proceedings of the 28th IEEE Conference on Decision and Control, Tampa, 1989, pp. 1107–1112.

[36] H. J. Sussmann and G. Tang, *Time-Optimal Control of a Satellite with Two Rotors Attached along Its Two Fixed Axes,* Preprint, 1995.

[37] M. I. Zelikin and V. F. Borisov, *Theory of Chattering Control. With Applications to Astronautics, Robotics, Economics, and Engineering*, Systems Control Found. Appl., Birkhäuser, Boston, 1994.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 1, pp. 140–158

GLOBAL EXACT BOUNDARY CONTROLLABILITY OF A CLASS OF QUASILINEAR HYPERBOLIC SYSTEMS OF CONSERVATION LAWS II*

DE-XING KONG† AND HUI YAO‡

Abstract. In this paper, by a new constructive method, the authors reprove the global exact boundary controllability of a class of quasilinear hyperbolic systems of conservation laws with linearly degenerate characteristics. It is shown that the system with nonlinear boundary conditions is globally exactly boundary controllable in the class of piecewise C^1 functions. In particular, the authors give the optimal control time of the system. Finally, a new example belonging to this kind of system is also provided.

Key words. quasilinear hyperbolic system, conservation laws, global exact boundary controllability, Cauchy problem, Goursat problem, classical discontinuous solution, contact discontinuity, optimal control time

AMS subject classifications. 93C20, 49J20

DOI. 10.1137/S0363012903432651

1. Introduction.

Consider the following quasilinear system in a form of conservation laws:

$$\begin{cases} u_t + f(u,v)_x = 0, \\ v_t + g(u,v)_x = 0, \end{cases} \tag{1.1}$$

where $u = u(t,x)$ and $v = v(t,x)$ are unknown functions and f, $g \in C^2(\mathcal{N})$ for some closed bounded domain $\mathcal{N}$ in $\mathbb{R}^2$. Let $F = (f,g)^T$ and

$$\nabla F(U) = \begin{pmatrix} f_u & f_v \\ g_u & g_v \end{pmatrix},$$

where $U = (u,v)$. We assume that

(H_1) On the domain $\mathcal{N}$ under consideration, system (1.1) is *strongly strictly hyperbolic*, i.e., for any given $U \in \mathcal{N}$, $\nabla F(U)$ has two distinct real eigenvalues $\lambda_1(U)$, $\lambda_2(U)$:

$$\lambda_1(U) < 0 < \lambda_2(U) \quad \forall\, U \in \mathcal{N}. \tag{1.2}$$

Let $\vec{l}_i(U) = (l_{i1}(U), l_{i2}(U))$ (resp., $\vec{r}_i(U) = (r_{i1}(U), r_{i2}(U))^T$) be a left (resp., right) eigenvector corresponding to $\lambda_i(U)$ $(i = 1,2)$:

$$\vec{l}_i(U)\nabla F(U) = \lambda_i(U)\vec{l}_i(U) \qquad (\text{resp., } \nabla F(U)\vec{r}_i(U) = \lambda_i(U)\vec{r}_i(U)).$$

(H_2) System (1.1) is *linearly degenerate*:

$$\nabla\lambda_i(U)\cdot\vec{r}_i(U) \equiv 0 \quad (i = 1,2) \quad \forall\, U \in \mathcal{N}. \tag{1.3}$$

*Received by the editors August 4, 2003; accepted for publication (in revised form) April 5, 2005; published electronically June 27, 2005. This work was supported in part by the National Science Foundation of China under grant 10371073, the Deutsche Forschungsgemeinschaft (DFG), and the Qi Ming Xing programme of Shanghai Government.
http://www.siam.org/journals/sicon/44-1/43265.html
†Department of Mathematics, Shanghai Jiao Tong University, Shanghai 200030, China (makong@cityu.edu.hk).
‡Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong.

(H_3) For any given real number a, let $H_i^a \triangleq \{U \,|\, \lambda_i(U) = a\}$ $(i = 1, 2)$. For any given $U_1, U_2 \in H_i^a$, there exists a C^1 curve segment $U = U(\tau)$ $(\tau \in [\tau_1, \tau_2])$ in $\mathcal{N}$ such that

$$U(\tau_j) = U_j \quad (j = 1, 2) \quad \text{and} \quad U(\tau) \in H_i^a \quad \forall\, \tau \in [\tau_1, \tau_2] \tag{1.4}$$

and

$$\nabla \lambda_i (U(\tau)) \neq 0 \qquad \forall\, \tau \in [\tau_1, \tau_2] \quad (i = 1, 2). \tag{1.5}$$

(H_4) There are *global* Riemann invariants for the system (1.1)

$$R_1 = R_1(U), \quad R_2 = R_2(U). \tag{1.6}$$

Remark 1.1. *Any quasilinear hyperbolic system with two unknown functions*

$$\frac{\partial u_i}{\partial t} + \sum_{j=1}^{2} a_{ij}(u_1, u_2) \frac{\partial u_j}{\partial x} = 0 \qquad (i = 1, 2)$$

can always be reduced to a system with the diagonal form at least in a local domain. This means that for any quasilinear hyperbolic system with two unknown functions, the local Riemann invariants always exist. On the other hand, many physical systems (for example, the system of isentropic gas) always possess global Riemann invariants.

By the assumptions (H_2) and (H_4), in the Riemann invariants, (1.1) can be rewritten as

$$\begin{cases} \dfrac{\partial R_1}{\partial t} + \mu_1(R_2) \dfrac{\partial R_1}{\partial x} = 0, \\ \dfrac{\partial R_2}{\partial t} + \mu_2(R_1) \dfrac{\partial R_2}{\partial x} = 0, \end{cases} \tag{1.7}$$

where

$$\mu_1(R_2(U)) = \lambda_1(U) \quad \text{and} \quad \mu_2(R_1(U)) = \lambda_2(U). \tag{1.8}$$

Recently, Kong [5] investigates the following *exact boundary control problem* for the system (1.1). Consider system (1.1) posed on the domain

$$\mathcal{D} = \{(t, x) |\, t \geq 0, \ -1 \leq x \leq 1\}$$

with the nonlinear boundary conditions

$$\begin{array}{ll} B_1(u, v, t) + h_1(t) = 0 & \text{at } x = -1, \\ B_2(u, v, t) + h_2(t) = 0 & \text{at } x = 1 \end{array} \tag{1.9}$$

and the initial data

$$t = 0 : (u, v) = \begin{cases} \left(u_0^-(x), v_0^-(x)\right) & \forall\, x \in [-1, 0], \\ \left(u_0^+(x), v_0^+(x)\right) & \forall\, x \in [0, 1], \end{cases} \tag{1.10}$$

where $B_i(u, v, t)$ are given smooth functions, $(u_0^-(x), v_0^-(x)) \in \mathcal{N}$ and $(u_0^+(x), v_0^+(x)) \in \mathcal{N}$ are C^1 vector functions, defined for $x \in [-1, 0]$ and $x \in [0, 1]$, respectively, satisfying

$$\left(u_0^-(0), v_0^-(0)\right) \neq \left(u_0^+(0), v_0^+(0)\right). \tag{1.11}$$

EXACT BOUNDARY CONTROL PROBLEM. *Given*

$$U_z(x) = \begin{cases} (u_z^-(x), v_z^-(x)) \in C^1([-1,0]) \times C^1([-1,0]), \\ (u_z^+(x), v_z^+(x)) \in C^1([0,1]) \times C^1([0,1]), \end{cases} \qquad z = 0,\ T, \tag{1.12}$$

can we find a time $T > 0$ and control inputs $h_1(t)$, $h_2(t)$ in the class of piecewise C^1 functions defined on $[0,\ T]$, such that the boundary control system (1.1), (1.9) *has a piecewise C^1 solution $U = U(t,x)$ containing contact discontinuities and satisfying the initial condition* (1.10) *and the terminal condition*

$$U(T,x) = U_T(x)? \tag{1.13}$$

Kong [5] proves the following theorem.

THEOREM A. *Under the hypotheses (H_1)–(H_4), for given $U_z(x)$ $(z = 0, T)$ (see* (1.12)*) and for any $T > \bar{T}_0$, there exist piecewise C^1 control inputs $h_1(t)$ and $h_2(t)$ defined for $t \in [0,T]$ such that system* (1.1), (1.9) *possesses a piecewise C^1 solution $U = U(t,x)$ on the domain*

$$\mathcal{D}(T) = \{(t,x) |\ 0 \le t \le T,\ \ -1 \le x \le 1\} \tag{1.14}$$

containing four contact discontinuities and satisfying

$$U(0,x) = U_0(x), \quad U(T,x) = U_T(x) \quad \forall\ x \in [-1,1], \tag{1.15}$$

where $\bar{T}_0$ is defined by

$$\bar{T}_0 = \max\left\{-\frac{2}{\bar{\lambda}_1},\ \frac{2}{\underline{\lambda}_2}\right\} + \max\left\{-\frac{2}{\bar{\lambda}_1},\ \frac{2}{\underline{\lambda}_2},\ \frac{4}{\underline{\lambda}_2 - \bar{\lambda}_1}\right\}, \tag{1.16}$$

in which

$$\bar{\lambda}_1 = \max_{|R_2| \le M} \mu_1(R_2), \qquad \underline{\lambda}_2 = \min_{|R_1| \le M} \mu_2(R_1). \tag{1.17}$$

Here M is given by

$$M = \max_{\substack{z=0,T \\ i=1,2}} \left\{ \|R_i(u_z^-(x), v_z^-(x)\|_{C^0([-1,0])},\ \|R_i(u_z^+(x), v_z^+(x)\|_{C^0([0,1])} \right\}. \tag{1.18}$$

REMARK 1.2. *Theorem* A *shows that the system* (1.1) *with nonlinear boundary conditions* (1.9) *is globally exactly boundary controllable in the class of piecewise C^1 functions. However, the control time $\bar{T}_0$, defined by* (1.16), *is not optimal.*

In this paper, by a new constructive method, we reprove the global exact boundary controllability of the system (1.1) with the optimal control time. The main result is the following theorem.

THEOREM 1.1. *Under the hypotheses (H_1)–(H_4), for given $U_z(x)$ $(z = 0, T)$ (see* (1.12)*) and for any $T > T_0$, there exist piecewise C^1 control inputs $h_1(t)$ and $h_2(t)$ defined for $t \in [0,T]$ such that system* (1.1), (1.9) *possesses a piecewise C^1 solution $U = U(t,x)$, containing four contact discontinuities and satisfying* (1.15), *on the domain $\mathcal{D}(T)$, where T_0 is defined by*

$$T_0 = \max\left\{-\frac{2}{\bar{\lambda}_1},\ \frac{2}{\underline{\lambda}_2},\ \frac{4}{\underline{\lambda}_2 - \bar{\lambda}_1}\right\}. \tag{1.19}$$

REMARK 1.3. *Comparing* (1.16) *with* (1.19)*, we observe that* $\bar{T}_0 > T_0$*. The condition* $T > T_0$ *in Theorem* 1.1 *is sharp and* T_0 *is optimal in the sense that, if* $T \le T_0$*, we may find a pair of initial and terminal states such that no matter what control inputs we choose, the system will not go from the given initial state to the desired terminal state during the time interval* $[0, T]$*. In this sense,* T_0 *defined by* (1.19) *is called the optimal control time of the system* (1.1)*.*

In order to illustrate the assertion claimed in Remark 1.3, i.e., the affirmation that the condition $T > T_0$ in Theorem 1.1 is sharp and T_0 is optimal, for simplicity we only consider the case that the characteristics λ_1, λ_2 are constants. In the present situation, there are two cases

Case I: $0 < -\lambda_1 \le \lambda_2$;

Case II: $0 < \lambda_2 < -\lambda_1$.

For Case I, we find that $T_0 = -2/\lambda_1$. If $T \le T_0$, by the characteristic method, it is easy to show that, there exists a pair of initial and terminal states such that no matter what control inputs we choose, the system will not go from the given initial state to the desired terminal state during the time interval $[0, T]$. In fact, it suffices to take the pair of initial and terminal states satisfying

$$R_1\left(u_0^+(1), v_0^+(1)\right) \neq \begin{cases} R_1\left(u_T^-(\lambda_1 T + 1), v_T^-(\lambda_1 T + 1)\right) & \text{if } \lambda_1 T + 1 \le 0, \\ R_1\left(u_T^+(\lambda_1 T + 1), v_T^+(\lambda_1 T + 1)\right) & \text{if } \lambda_1 T + 1 \ge 0. \end{cases} \tag{1.20}$$

Noting that the Riemann invariant R_1 keeps continuous across the contact discontinuities corresponding to the characteristic filed λ_2, we observe that, for such a pair of initial and terminal states satisfying (1.20), no matter what control inputs we choose, the system will not go from the given initial state to the desired terminal state during the time interval $[0, T]$.

For Case II, we have a similar discussion.

REMARK 1.4. *The hypothesis* (H_3) *is a technical assumption only for constructing contact discontinuities, Kong* [5] *gives two examples to show that some physical systems always satisfy it. Moreover, if the initial and terminal data are* C^1 *smooth, then the hypothesis* (H_3) *is not needed. In this case, Theorem* 1.1 *is the result given in Li and Zhang* [9]*.*

REMARK 1.5. *As in other works (e.g.,* [9]*), the solution of our exact boundary control problem does not possess uniqueness. In fact, even if the initial and terminal data are smooth, the solution of the system does not have uniqueness either (see* [8]*,* [9]*). In particular, the control inputs are not unique. This is due to the fact that we have some freedom to choose the control inputs, because the waiting time* $T > T_0$*. This fact can be observed from the proof of* Theorem 1.1*.*

This paper is organized as follows. Theorem 1.1 is proved in section 2 by a new constructive method. Section 3 gives some important supplementary remarks, while section 4 provides a new example of system such as described in section 3, namely, the system for time-like extremal surfaces in the $(1+n)$-dimensional Minkowski space $\mathbb{R}^{1+n}$.

2. Proof of Theorem 1.1. For readers' convenience, before starting the proof of Theorem 1.1, we first recall the definition of contact discontinuity.

DEFINITION 2.1. *$U = U(t, x)$ is called a piecewise C^1 solution containing a kth $(k = 1, 2)$ contact discontinuity $x = x_k(t)$ if $U = U(t, x)$ satisfies the system* (1.1) *out of $x = x_k(t)$ in the classical sense and satisfies the Rankine–Hugoniot condition on*

$x = x_k(t)$, *i.e.*,

$$\sigma \cdot [U] = [F], \tag{2.1}$$

$$\sigma = \lambda_k\left(U^+\right) = \lambda_k\left(U^-\right), \tag{2.2}$$

where $U^{\pm} = U(t, x_k(t) \pm 0)$ *and* $\sigma = x_k'(t)$.

Let

$$\begin{array}{llll} A = (0,-1), & B = (0,1), & C = (T,1), & D = (T,-1), \\ O = (0,0), & N = (T,0), & E = (t_e, x_e), & F = (t_f, x_f), \end{array} \tag{2.3}$$

where E is the intersection point of the lines

$$L_1: \ x = \bar{\lambda}_1 t + 1 \quad \text{and} \quad L_2: \ x = \underline{\lambda}_2 t - 1 \tag{2.4}$$

and F is the intersection point of the lines

$$\bar{L}_1: \ x = \bar{\lambda}_1 (t - T) - 1 \quad \text{and} \quad \bar{L}_2: \ x = \underline{\lambda}_2 (t - T) + 1. \tag{2.5}$$

Noting (1.19) and $T > T_0$, we observe that

$$t_e < t_f. \tag{2.6}$$

Step 1: Generalized Riemann problem for the system (1.1). First, we solve the generalized Riemann problem for the system (1.1) with discontinuous initial data (1.10). The following lemma comes from [7].

LEMMA 2.1. *Under the hypotheses* (H_1)*–*(H_4)*, the generalized Riemann problem* (1.1), (1.10) *has a unique piecewise* C^1 *solution* $(u, v) = (u(t,x), v(t,x))$*, containing two* C^1 *contact discontinuities starting from* O*, on the maximum determined domain* Ω_1 *enclosed by the characteristics* $x = x_1(t)$, $x = x_2(t)$ *and the* x*-axis:*

$$\Omega_1 = \{(t,x) |\, 0 \le t \le t_p,\ x_2(t) \le x \le x_1(t)\}, \tag{2.7}$$

where $x = x_1(t)$ *satisfies*

$$\frac{dx_1(t)}{dt} = \lambda_1(u, v), \qquad x_1(0) = 1, \tag{2.8}$$

where $x = x_2(t)$ *satisfies*

$$\frac{dx_2(t)}{dt} = \lambda_2(u, v), \qquad x_2(0) = -1, \tag{2.9}$$

and while t_p *is the time coordinate of the intersection point, denoted by* $P = (t_p, x_p)$*, of the characteristic* $x = x_1(t)$ *with the characteristic* $x = x_2(t)$*. See Figure* 1.

The solution (u, v) of the generalized Riemann problem (1.1), (1.10) is denoted by $U = U_1(t,x)$ on the domain Ω_1, its two contact discontinuities are denoted by $x = \xi_1(t)$ and $x = \xi_2(t)$, respectively. By the definition of contact discontinuity, $x = \xi_i(t)$ $(i = 1, 2)$ satisfy

$$\frac{\xi_i(t)}{dt} = \lambda_i\left(U_1^{\pm}\right), \quad \xi_i(0) = 0 \quad (i = 1, 2). \tag{2.10}$$

Moreover, let $P_1 = (t_{p_1}, x_{p_1})$ (resp. $P_2 = (t_{p_2}, x_{p_2})$) be the the intersection point of the characteristic $x = x_2(t)$ (resp. $x = x_1(t)$) with the contact discontinuity $x = \xi_1(t)$ (resp. $x = \xi_2(t)$). See Figure 1.

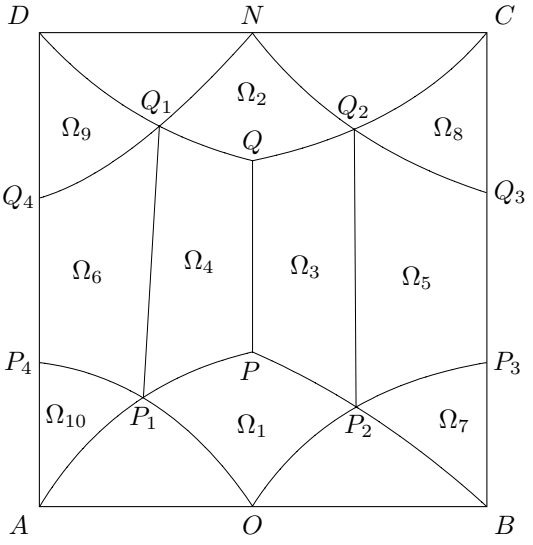


FIG. 1. *Domains* Ω_i $(i = 1, \ldots, 10)$*; characteristics:* $AP:\ x = x_2(t)$, $BP:\ x = x_1(t)$, $CQ:\ x = \tilde{x}_2(t)$, $DQ:\ x = \tilde{x}_1(t)$*; contact discontinuities:* $OP_1:\ x = \xi_1(t)$, $P_1P_4:\ x = \zeta_1(t)$, $OP_2:\ x = \xi_2(t)$, $P_2P_3:\ x = \zeta_2(t)$, $NQ_1:\ x = \tilde{\xi}_2(t)$, $Q_1Q_4:\ x = \tilde{\zeta}_2(t)$, $NQ_2:\ x = \tilde{\xi}_1(t)$, $Q_2Q_3:\ x = \tilde{\zeta}_1(t)$.

Similarly, we can solve the backward generalized Riemann problem for the system (1.1) with discontinuous initial data (1.13). We have the following lemma.

LEMMA 2.2. *Under the hypotheses* (H_1)*–*(H_4)*, the backward generalized Riemann problem* (1.1), (1.13) *has a unique piecewise* C^1 *solution* $(u, v) = (u(t, x), v(t, x))$*, containing two* C^1 *contact discontinuities starting from* N*, on the maximum determined domain* Ω_2 *enclosed by the characteristics* $x = \tilde{x}_1(t)$, $x = \tilde{x}_2(t)$ *and the line* $t = T$*:*

$$\Omega_2 = \{(t, x) |\, t_q \le t \le T,\ \tilde{x}_1(t) \le x \le \tilde{x}_2(t)\}, \tag{2.11}$$

where $x = \tilde{x}_1(t)$ *satisfies*

$$\frac{d\tilde{x}_1(t)}{dt} = \lambda_1\left(u, v\right), \qquad \tilde{x}_1(T) = -1, \tag{2.12}$$

where $x = x_2(t)$ *satisfies*

$$\frac{d\tilde{x}_2(t)}{dt} = \lambda_2\left(u, v\right), \qquad \tilde{x}_2(T) = 1 \tag{2.13}$$

and while t_q *is the time coordinate of the intersection point, denoted by* $Q = (t_q, x_q)$*, of the characteristic* $x = \tilde{x}_1(t)$ *with the characteristic* $x = \tilde{x}_2(t)$*. See Figure* 1.

The solution (u, v) of the backward generalized Riemann problem (1.1), (1.13) is denoted by $U = U_2(t, x)$ on the domain Ω_2, its two contact discontinuities are denoted by $x = \tilde{\xi}_1(t)$ and $x = \tilde{\xi}_2(t)$, respectively. By the definition of contact discontinuity, $x = \tilde{\xi}_i(t)$ $(i = 1, 2)$ satisfy

$$\frac{\tilde{\xi}_i(t)}{dt} = \lambda_i\left(U_2^{\pm}\right), \quad \tilde{\xi}_i(0) = 0 \quad (i = 1, 2). \tag{2.14}$$

Moreover, let $Q_1 = (t_{q_1}, x_{q_1})$ (resp. $Q_2 = (t_{q_2}, x_{q_2})$) be the the intersection point of the characteristic $x = \tilde{x}_1(t)$ (resp. $x = \tilde{x}_2(t)$) with the contact discontinuity $x = \tilde{\xi}_2(t)$ (resp. $x = \tilde{\xi}_1(t)$). See Figure 1.

REMARK 2.1. *From the argument mentioned above, we observe that*

$$0 \le t_p \le t_e, \qquad t_f \le t_q \le T. \tag{2.15}$$

Step 2: Mixed initial-boundary value problem for the system (1.1). Let Ω_3 be the domain enclosed by the characteristics PP_2 and QQ_2, and the straight line segments PQ and P_2Q_2. It is easy to see that the straight line PQ can be expressed by

$$x = x_p + \alpha(t - t_p) \triangleq c(t), \qquad t \in [t_p, t_q], \tag{2.16}$$

where α is the slope of the line PQ

$$\alpha = \frac{x_q - x_p}{t_q - t_p}. \tag{2.17}$$

Similarly, the straight line P_2Q_2 can be expressed by

$$x = x_{p_2} + \alpha_2(t - t_{p_2}) \triangleq c_2(t), \qquad t \in [t_{p_2}, t_{q_2}], \tag{2.18}$$

where α_2 is the slope of the line P_2Q_2

$$\alpha_2 = (x_{q_2} - x_{p_2})/(t_{q_2} - t_{p_2}). \tag{2.19}$$

Noting (1.7), the system (1.1) can be equivalently rewritten as

$$\begin{cases} \dfrac{\partial R_1}{\partial x} + \dfrac{1}{\mu_1(R_2)} \dfrac{\partial R_1}{\partial t} = 0, \\ \dfrac{\partial R_2}{\partial x} + \dfrac{1}{\mu_2(R_1)} \dfrac{\partial R_2}{\partial x} = 0 \end{cases} \tag{2.20}$$

for smooth solutions. We next consider the mixed initial-boundary value problem for the system (2.20) (equivalently, (1.1)) on the domain Ω_3 with the following boundary conditions:

on the characteristic PP_2: $x = x_1(t)$ $(t \in [t_{p_2}, t_p])$

$$R_2 = R_2(U_1(t, x_1(t))) \triangleq r_2(t), \qquad t \in [t_{p_2}, t_p], \tag{2.21}$$

on the characteristic QQ_2: $x = \tilde{x}_2$ $(t \in [t_q, t_{q_2}])$

$$R_1 = R_1(U_2(t, \tilde{x}_2(t))) \triangleq \tilde{r}_1(t), \qquad t \in [t_q, t_{q_2}] \tag{2.22}$$

and the initial condition on the line segment PQ

$$R_1 = s_1(t), \qquad R_2 = s_2(t), \quad t \in [t_p, t_q], \tag{2.23}$$

where $s_1(t)$, $s_2(t)$ are C^1 functions of $t \in [t_p, t_q]$. Here we have interchanged the role of x and t variables. In order to ensure that the mixed initial-boundary value problem (2.20)–(2.23) has a C^1 solution on Ω_3, the initial data $(s_1(t), s_2(t))$ must satisfy certain compatibility conditions. First of all, it is required that

$$s_1(t_p) = R_1(U_1(t_p, x_p)), \qquad s_2(t_q) = R_2(U_2(t_q, x_q)). \tag{2.24}$$

Moreover, it is also required that

$$s_1(t_q) = R_1(U_2(t_q, x_q)) = \tilde{r}_1(t_q), \qquad s_2(t_p) = R_2(U_1(t_p, x_p)) = r_2(t_p). \tag{2.25}$$

Notice that along the characteristic QQ_2: $x = \tilde{x}_2(t)$

$$\tilde{r}_1'(t) = \frac{\partial R_1}{\partial t} + \mu_2(R_1)\frac{\partial R_1}{\partial x} = (\mu_2(R_1) - \mu_1(R_2))\frac{\partial R_1}{\partial x}.$$

Then,

$$\tilde{r}_1'(t_q) = (\mu_2(R_1(U_2(t_q, x_q))) - \mu_1(R_2(U_2(t_q, x_q))))\,\frac{\partial R_1}{\partial x}(t_q, x_q). \tag{2.26}$$

On the other hand, along the line PQ: $x = c(t)$

$$s_1'(t) = \frac{\partial R_1}{\partial t} + \alpha\frac{\partial R_1}{\partial x} = (\alpha - \mu_1(s_2))\frac{\partial R_1}{\partial x}.$$

Then,

$$s_1'(t_q) = (\alpha - \mu_1(s_2(t_q)))\,\frac{\partial R_1}{\partial x}(t_q, x_q). \tag{2.27}$$

Therefore, we need that

$$s_1'(t_q) = \frac{\alpha - \mu_1(R_2(U_2(t_q, x_q)))}{\mu_2(R_1(U_2(t_q, x_q))) - \mu_1(R_2(U_2(t_q, x_q)))}\tilde{r}_1'(t_q). \tag{2.28}$$

Similarly, at (t_p, x_p) we require that

$$s_2'(t_p) = \frac{\alpha - \mu_2(R_1(U_1(t_p, x_p)))}{\mu_1(R_2(U_1(t_p, x_p))) - \mu_2(R_1(U_1(t_p, x_p)))}r_2'(t_p). \tag{2.29}$$

We have the following proposition.

PROPOSITION 2.1. *The angles formed by the line segment PQ and the characteristic QC, by the line segment PQ and the characteristic PB are less than π.*

Proof. Consider the angle between PQ and QC. The worst case is given by

$$P = \left(\frac{2}{\underline{\lambda}_2 - \underline{\lambda}_1}, \frac{\underline{\lambda}_1 + \underline{\lambda}_2}{\underline{\lambda}_2 - \underline{\lambda}_1}\right), \qquad Q = \left(T - \frac{2}{\underline{\lambda}_2 - \underline{\lambda}_1}, -\frac{\underline{\lambda}_1 + \underline{\lambda}_2}{\underline{\lambda}_2 - \underline{\lambda}_1}\right).$$

In the worst case, the slope of PQ is

$$\frac{dx}{dt} = \frac{2\,(\underline{\lambda}_1 + \underline{\lambda}_2)}{4 - T\,(\underline{\lambda}_2 - \underline{\lambda}_1)}.$$

In order to ensure the angle formed by the line segment PQ and the characteristic QC is less than π, it is sufficient to require that the slope of the line segment PQ is less than the slope of the characteristic QC, i.e.,

$$\frac{2\,(\underline{\lambda}_1 + \underline{\lambda}_2)}{4 - T\,(\underline{\lambda}_2 - \underline{\lambda}_1)} < \underline{\lambda}_2. \tag{2.30}$$

Noting (1.19) and $T > T_0$, we have

$$T > \frac{4}{\underline{\lambda}_2 - \bar{\lambda}_1}.$$

Hence, in order to guarantee the validity of inequality (2.30), it suffices to require that

$$T > \frac{2}{\underline{\lambda}_2}.$$

This is true because of (1.19) and the fact $T > T_0$.

Similarly, consider the angle between PQ and PB. The worst case is given by

$$P = \left(\frac{2}{\bar{\lambda}_2 - \bar{\lambda}_1}, \frac{\bar{\lambda}_1 + \bar{\lambda}_2}{\bar{\lambda}_2 - \bar{\lambda}_1}\right), \qquad Q = \left(T - \frac{2}{\bar{\lambda}_2 - \bar{\lambda}_1}, -\frac{\bar{\lambda}_1 + \bar{\lambda}_2}{\bar{\lambda}_2 - \bar{\lambda}_1}\right).$$

A similar argument yields

$$T > -\frac{2}{\bar{\lambda}_1}.$$

This is also true because of (1.19) and the fact $T > T_0$. Thus, the proof of Proposition 2.1 is completed. □

Therefore, using Lemma 2.3 and Remark 2.1 in [9], we obtain the following lemma.

LEMMA 2.3. *Under the hypotheses* (H_1)–(H_4), *the mixed initial-boundary value problem* (2.20)–(2.23) *admits a unique* C^1 *solution* $(R_1, R_2) = (R_1(t,x), R_2(t,x))$ *on* Ω_3, *provided that the compatibility conditions* (2.24)–(2.25) *and* (2.28)–(2.29) *hold.*

By Lemma 2.3, let $U = U_3(t,x)$ be the solution of the system (1.1) corresponding to $(R_1(t,x), R_2(t,x))$ given by Lemma 2.3 on the domain Ω_3.

On the other hand, let Ω_4 be the domain enclosed by the characteristics PP_1 and QQ_1, and the straight line segment PQ: $x = c(t) = x_p + \alpha(t - t_p)$, $t_p \le t \le t_q$ and the straight line segment P_1Q_1:

$$x = x_{p_1} + \alpha_1(t - t_{p_1}) \triangleq c_1(t), \qquad t \in [t_{p_1}, t_{q_1}], \tag{2.31}$$

where α_1 is the slope of the line P_1Q_1

$$\alpha_1 = \frac{x_{q_1} - x_{p_1}}{t_{q_1} - t_{p_1}}.$$

On the domain Ω_4, we consider the mixed initial-boundary value problem for the system (2.20) (equivalently, (1.1)) with the following boundary conditions:

on the characteristic PP_1: $x = x_2(t)$ $(t \in [t_{p_1}, t_p])$

$$R_1 = R_1(U_1(t, x_2(t))) \triangleq r_1(t), \qquad t \in [t_{p_1}, t_p], \tag{2.32}$$

on the characteristic QQ_1: $x = \tilde{x}_1(t)$ $t \in [t_q, t_{q_1}]$

$$R_2 = R_2(U_2(t, \tilde{x}_1(t))) \triangleq \tilde{r}_2(t), \qquad t \in [t_q, t_{q_1}] \tag{2.33}$$

and the initial condition on the line segment PQ

$$R_1 = s_1(t), \qquad R_2 = s_2(t), \quad t \in [t_p, t_q]. \tag{2.34}$$

As before, we choose $s_1(t)$, $s_2(t)$ to satisfy the compatibility conditions (2.28)–(2.29) and

$$\begin{aligned} s_1'(t_p) &= \frac{\alpha - \mu_1(R_2(U_1(t_p, x_p)))}{\mu_2(R_1(U_1(t_p, x_p))) - \mu_1(R_2(U_1(t_p, x_p)))} r_1'(t_p), \\ s_2'(t_q) &= \frac{\alpha - \mu_2(R_1(U_2(t_q, x_q)))}{\mu_1(R_2(U_2(t_q, x_q))) - \mu_2(R_1(U_2(t_q, x_q)))} \tilde{r}_2'(t_q). \end{aligned} \tag{2.35}$$

Similar to Proposition 2.1, we have the following.

PROPOSITION 2.2. *The angles formed by the line segment PQ and the characteristic QD and by the line segment PQ and the characteristic PA are less than π.*

Using Lemma 2.3 and Remark 2.1 in [9] again, we obtain the following lemma.

LEMMA 2.4. *Under the hypotheses (H_1)–(H_4), the mixed initial-boundary value problem* (2.20), (2.32)–(2.34) *admits a unique C^1 solution $(R_1, R_2) = (R_1(t,x), R_2(t,x))$ on Ω_4, provided that the compatibility conditions* (2.24)–(2.25), (2.28)–(2.29) *and* (2.35) *hold.*

We denote the solution of the system (1.1) corresponding to $(R_1(t,x), R_2(t,x))$, given by Lemma 2.4, on the domain Ω_4 by $U = U_4(t,x)$.

Step 3: Cauchy problem for the system (2.20). Consider the Cauchy problem in the x-direction for the system (2.20) with the following initial condition on the line segment P_2Q_2:

$$\begin{aligned} R_1(t, c_2(t)) &= R_1(U_3(t, c_2(t))) \triangleq \sigma_1(t), \\ R_2(t, c_2(t)) &= R_2(U_3(t, c_2(t))) \triangleq \sigma_2(t), \end{aligned} \qquad t \in [t_{p_2}, t_{q_2}]. \tag{2.36}$$

Similar to Proposition 2.1, we have the following.

PROPOSITION 2.3. *The angles formed by the line segment P_2Q_2 and the characteristic Q_2C, by the line segment P_2Q_2 and the characteristic P_2B are less than π.*

Proof. Consider the angle between P_2Q_2 and Q_2C. The worst case is given by

$$P_2 = \left(\frac{1}{\underline{\lambda}_2 - \underline{\lambda}_1}, \frac{\underline{\lambda}_2}{\underline{\lambda}_2 - \underline{\lambda}_1}\right), \qquad Q_2 = \left(T - \frac{1}{\underline{\lambda}_2 - \underline{\lambda}_1}, -\frac{\underline{\lambda}_1}{\underline{\lambda}_2 - \underline{\lambda}_1}\right).$$

In the worst case, the slope of P_2Q_2 is

$$\frac{dx}{dt} = \frac{\underline{\lambda}_1 + \underline{\lambda}_2}{2 - T(\underline{\lambda}_2 - \underline{\lambda}_1)}.$$

In order to ensure the angle formed by the line P_2Q_2 and the characteristic Q_2C is less than π, it is sufficient to require that the slope of the line P_2Q_2 is less than the slope of the characteristic Q_2C, i.e.,

$$\frac{\underline{\lambda}_1 + \underline{\lambda}_2}{2 - T(\underline{\lambda}_2 - \underline{\lambda}_1)} < \underline{\lambda}_2. \tag{2.37}$$

Noting (1.19) and $T > T_0$, we have

$$T > \frac{4}{\underline{\lambda}_2 - \bar{\lambda}_1}.$$

Hence, in order to guarantee the validity of inequality (2.37), it suffices to require that

$$T > \frac{1}{\underline{\lambda}_2}.$$

Of course, this is true because of (1.19) and the fact $T > T_0$.

Similarly, consider the angle between P_2Q_2 and P_2B. The worst case is given by

$$P_2 = \left(\frac{1}{\bar{\lambda}_2 - \bar{\lambda}_1}, \frac{\bar{\lambda}_2}{\bar{\lambda}_2 - \bar{\lambda}_1}\right), \qquad Q_2 = \left(T - \frac{1}{\bar{\lambda}_2 - \bar{\lambda}_1}, -\frac{\bar{\lambda}_1}{\bar{\lambda}_2 - \bar{\lambda}_1}\right).$$

A similar argument gives

$$T > -\frac{1}{\overline{\lambda}_1}.$$

Obviously, this is true because of (1.19) and the fact $T > T_0$. Thus, the proof of Proposition 2.3 is finished. □

Using the Corollary 2.1 in [9], we have the following lemma.

LEMMA 2.5. *Under the hypotheses* (H_1)–(H_4), *the Cauchy problem* (2.20), (2.36) *admits a unique* C^1 *solution* $(R_1, R_2) = (R_1(t,x), R_2(t,x))$ *on the maximum determined domain* Ω_5 *enclosed by the straight line segment* P_2Q_2, *the straight line segment* P_3Q_3, *the characteristic* P_2P_3: $x = \zeta_2(t)$ $(t \in [t_{p_2}, t_{p_3}])$ *and the characteristic* Q_2Q_3: $x = \tilde{\zeta}_1(t)$ $(t \in [t_{q_3}, t_{q_2}])$, *where* $P_3 = (t_{p_3}, 1)$ *is the intersection point of the characteristic* $x = \zeta_2(t)$ *with the line* BC, *and* $Q_3 = (t_{q_3}, 1)$ *is the intersection point of the characteristic* $x = \tilde{\zeta}_1(t)$ *with the line* BC, *while the characteristic* $\zeta = \tilde{\zeta}_1(t)$ *satisfies*

$$\frac{d\tilde{\zeta}_1(t)}{dt} = \mu_1(R_2), \qquad \tilde{\zeta}_1(t_{q_2}) = x_{q_2} \tag{2.38}$$

and the characteristic $\zeta = \zeta_2(t)$ *satisfies*

$$\frac{d\zeta_2(t)}{dt} = \mu_2(R_1), \qquad \zeta_2(t_{p_2}) = x_{p_2}. \tag{2.39}$$

See Figure 1.

We denote the solution of the system (1.1) corresponding to $(R_1(t,x), R_2(t,x))$, given by Lemma 2.5, on the domain Ω_5 by $U = U_5(t,x)$.

Similarly, we consider the Cauchy problem in the anti-x-direction for the system (2.20) with the following initial condition on the line segment P_1Q_1:

$$\begin{aligned} R_1(t, c_1(t)) &= R_1(U_4(t, c_1(t))) \triangleq \theta_1(t), \\ R_2(t, c_1(t)) &= R_2(U_4(t, c_1(t))) \triangleq \theta_2(t), \end{aligned} \qquad t \in [t_{p_1}, t_{q_1}]. \tag{2.40}$$

Similar to Proposition 2.3, we have the following.

PROPOSITION 2.4. *The angles formed by the line segment* P_1Q_1 *and the characteristic* Q_1D, *by the line segment* P_1Q_1 *and the characteristic* P_1A *are less than* π.

By the Corollary 2.1 in [9], we have the following lemma.

LEMMA 2.6. *Under the hypotheses* (H_1)–(H_4), *the Cauchy problem* (2.20), (2.40) *admits a unique* C^1 *solution* $(R_1, R_2) = (R_1(t,x), R_2(t,x))$ *on the maximum determined domain* Ω_6 *enclosed by the straight line segment* P_1Q_1, *the straight line segment* P_4Q_4, *the characteristic* P_1P_4: $x = \zeta_1(t)$ $(t \in [t_{p_1}, t_{p_4}])$ *and the characteristic* Q_1Q_4: $x = \tilde{\zeta}_2(t)$ $(t \in [t_{q_4}, t_{q_1}])$ (*see Figure* 1), *where* $P_4 = (t_{p_4}, 1)$ *is the intersection point of the characteristic* $\zeta = \zeta_1(t)$ *with the line* AD, *and* $Q_4 = (t_{q_4}, 1)$ *is the intersection point of the characteristic* $\zeta = \tilde{\zeta}_2(t)$ *with the line* AD, *while the characteristic* $\zeta = \zeta_1(t)$ *satisfies*

$$\frac{d\zeta_1(t)}{dt} = \mu_1(R_2), \qquad \zeta_1(t_{p_1}) = x_{p_1} \tag{2.41}$$

and the characteristic $\zeta = \tilde{\zeta}_2(t)$ *satisfies*

$$\frac{d\tilde{\zeta}_2(t)}{dt} = \mu_2(R_1), \qquad \tilde{\zeta}_2(t_{q_1}) = x_{q_1}. \tag{2.42}$$

See Figure 1.

Let $U = U_6(t,x)$ be the solution of the system (1.1) corresponding to $(R_1(t,x), R_2(t,x))$, given by Lemma 2.6, on the domain Ω_6.

Step 4: Goursat problem for the system (2.20). We next consider the Goursat problem in the x-direction for the system (2.20) with the following characteristic boundary conditions:

on the characteristic P_2B: $x = x_1(t)$ $(t \in [0, t_{p_2}])$

$$R_2 = R_2(U_1(t, x_1(t))) \triangleq \eta_2(t), \qquad t \in [0, t_{p_2}], \tag{2.43}$$

on the characteristic P_2P_3: $x = \zeta_2(t)$ $(t \in [t_{p_2}, t_{p_3}])$

$$R_1 = R_1(U_5(t, \zeta_2(t))) \triangleq \eta_1(t), \qquad t \in [t_{p_2}, t_{p_3}]. \tag{2.44}$$

By the Lemma 2.2 in [9], we have the following.

LEMMA 2.7. *Under the hypotheses* (H_1)*–*(H_4)*, the Goursat problem* (2.20), (2.43)*–*(2.44) *has a unique* C^1 *solution* $(R_1, R_2) = (R_1(t,x), R_2(t,x))$ *on the domain* Ω_7 *enclosed by the straight line segment* BP_3*, the characteristic* BP_2*:* $x = x_1(t)$ $(t \in [0, t_{p_2}])$ *and the characteristic* P_2P_3*:* $x = \zeta_2(t)$ $(t \in [t_{p_2}, t_{p_3}])$*. See Figure* 1*. Moreover, it holds that*

$$\begin{aligned} &R_1(t, \zeta_2(t)) = \eta_1(t) \qquad \forall\, t \in [t_{p_2}, t_{p_3}], \\ &R_2(t, x_1(t)) = \eta_2(t) \qquad \forall\, t \in [0, t_{p_2}]. \end{aligned} \tag{2.45}$$

Let $U = U_7(t,x)$ be the solution of the system (1.1) corresponding to $(R_1(t,x), R_2(t,x))$, given by Lemma 2.7, on the domain Ω_7.

Similarly, consider the Goursat problem in the x-direction for the system (2.20) with the following characteristic boundary conditions:

on the characteristic Q_2C: $x = \tilde{x}_2(t)$ $(t \in [t_{q_2}, T])$

$$R_1 = R_1(U_2(t, \tilde{x}_2(t))) \triangleq \tilde{\eta}_1(t), \quad t \in [t_{q_2}, T], \tag{2.46}$$

on the characteristic Q_2Q_3: $x = \tilde{\zeta}_1(t)$ $(t \in [t_{q_3}, t_{q_2}])$

$$R_2 = R_2(U_5(t, \tilde{\zeta}_1(t))) \triangleq \tilde{\eta}_2(t), \quad t \in [t_{q_3}, t_{q_2}]. \tag{2.47}$$

Similar to Lemma 2.7, we have the following.

LEMMA 2.8. *Under the hypotheses* (H_1)*–*(H_4)*, the Goursat problem* (2.20), (2.46)*–*(2.47) *has a unique* C^1 *solution* $(R_1, R_2) = (R_1(t,x), R_2(t,x))$ *on the domain* Ω_8 *enclosed by the straight line segment* Q_3C*, the characteristic* Q_2C*:* $x = \tilde{x}_2(t)$ $(t \in [t_{q_2}, T])$ *and the characteristic* Q_2Q_3*:* $x = \tilde{\zeta}_1(t)$ $(t \in [t_{q_3}, t_{q_2}])$*. See Figure* 1*. Moreover, it holds that*

$$\begin{aligned} &R_1(t, \tilde{x}_2(t)) = \tilde{\eta}_1(t) \quad \forall\, t \in [t_{q_2}, T], \\ &R_2(t, \tilde{\zeta}_1(t)) = \tilde{\eta}_2(t) \quad \forall\, t \in [t_{q_3}, t_{q_2}]. \end{aligned} \tag{2.48}$$

Let $U = U_8(t,x)$ be the solution of the system (1.1) corresponding to $(R_1(t,x), R_2(t,x))$, given by Lemma 2.8, on the domain Ω_8.

On the other hand, we consider the Goursat problem in the anti-x-direction for the system (2.20) with the following characteristic boundary conditions:

on the characteristic Q_1D: $x = \tilde{x}_1(t)$ $(t \in [t_{q_1}, T])$

$$R_2 = R_2(U_2(t, \tilde{x}_1(t))) \triangleq \tilde{\gamma}_2(t), \qquad t \in [t_{q_1}, T], \tag{2.49}$$

on the characteristic Q_1Q_4: $x = \tilde{\zeta}_2(t)$ $(t \in [t_{q_4}, t_{q_1}])$

$$R_1 = R_1(U_6(t, \tilde{\zeta}_2(t))) \triangleq \tilde{\gamma}_1(t), \qquad t \in [t_{q_4}, t_{q_1}]. \tag{2.50}$$

Similar to Lemma 2.8, we have the following.

LEMMA 2.9. *Under the hypotheses* (H_1)*–*(H_4)*, the Goursat problem* (2.20), (2.49)*–*(2.50) *has a unique* C^1 *solution* $(R_1, R_2) = (R_1(t,x), R_2(t,x))$ *on the domain* Ω_9 *enclosed by the straight line segment* Q_4D*, the characteristic* Q_1D*:* $x = \tilde{x}_1(t)$ $(t \in [t_{q_1}, T])$ *and the characteristic* Q_4Q_1*:* $x = \tilde{\zeta}_2(t)$ $(t \in [t_{q_4}, t_{q_1}])$*. See Figure* 1*. Moreover, it holds that*

$$\begin{aligned} &R_1(t, \tilde{\zeta}_2(t)) = \tilde{\gamma}_1(t) \quad \forall\, t \in [t_{q_4}, t_{q_1}], \\ &R_2(t, \tilde{x}_1(t)) = \tilde{\gamma}_2(t) \quad \forall\, t \in [t_{q_1}, T]. \end{aligned} \tag{2.51}$$

Let $U = U_9(t,x)$ be the solution of the system (1.1) corresponding to $(R_1(t,x), R_2(t,x))$, given by Lemma 2.9, on the domain Ω_9.

Finally, we consider the Goursat problem in the anti-x-direction for the system (2.20) with the following characteristic boundary conditions:

on the characteristic AP_1: $x = x_2(t)$ $(t \in [0, t_{p_1}])$

$$R_1 = R_1(U_1(t, x_2(t))) \triangleq \gamma_1(t), \qquad t \in [0, t_{p_1}], \tag{2.52}$$

on the characteristic P_1P_4: $x = \zeta_1(t)$ $(t \in [t_{p_1}, t_{p_4}])$

$$R_2 = R_2(U_6(t, \zeta_1(t))) \triangleq \gamma_1(t), \qquad t \in [t_{p_1}, t_{p_4}]. \tag{2.53}$$

Similar to Lemma 2.9, we have the following.

LEMMA 2.10. *Under the hypotheses* (H_1)*–*(H_4)*, the Goursat problem* (2.20), (2.52)*–*(2.53) *has a unique* C^1 *solution* $(R_1, R_2) = (R_1(t,x), R_2(t,x))$ *on the domain* Ω_{10} *enclosed by the straight line segment* AP_4*, the characteristic* AP_1*:* $x = x_2(t)$ $(t \in [0, t_{p_1}])$ *and the characteristic* P_1P_4*:* $x = \zeta_1(t)$ $(t \in [t_{p_1}, t_{p_4}])$*. See Figure* 1*. Moreover, it holds that*

$$\begin{aligned} &R_1(t, x_2(t)) = \gamma_1(t) \quad \forall\, t \in [0, t_{p_1}], \\ &R_2(t, \zeta_1(t)) = \gamma_2(t) \quad \forall\, t \in [t_{p_1}, t_{p_4}]. \end{aligned} \tag{2.54}$$

Let $U = U_{10}(t,x)$ be the solution of the system (1.1) corresponding to $(R_1(t,x), R_2(t,x))$, given by Lemma 2.10, on the domain Ω_{10}.

Step 5: Piecewise $\mathbf{C^1}$ solution with four $\mathbf{C^1}$ contact discontinuities. Now we choose $s_1(t)$, $s_2(t)$ in (2.23) from the space $C^1[t_p, t_q]$ so that the compatibility conditions (2.24)–(2.25), (2.28)–(2.29) and (2.35) are all satisfied and the C^0 norms of $s_1(t)$, $s_2(t)$ are bounded by M.

Define the piecewise C^1 function

$$U = U(t,x) = \begin{cases} U_1(t,x) & \text{for } (t,x) \in \Omega_1, \\ \cdots & \cdots \\ U_{10}(t,x) & \text{for } (t,x) \in \Omega_{10} \end{cases} \tag{2.55}$$

and the piecewise C^1 curves

$$
\begin{aligned}
&\ell_1 = \{\,(t,x)\,|\,x=\xi_1(t) \quad \text{for } t\in[0,t_{p_1}]; \qquad x=\zeta_1(t) \quad \text{for } t\in[t_{p_1},t_{p_4}]\},\\
&\ell_2 = \{\,(t,x)\,|\,x=\xi_2(t) \quad \text{for } t\in[0,t_{p_2}]; \qquad x=\zeta_2(t) \quad \text{for } t\in[t_{p_2},t_{p_3}]\},\\
&\tilde{\ell}_1 = \{\,(t,x)\,|\,x=\tilde{\xi}_1(t) \quad \text{for } t\in[t_{q_2},T]; \qquad x=\tilde{\zeta}_1(t) \quad \text{for } t\in[t_{q_3},t_{q_2}]\},\\
&\tilde{\ell}_2 = \{\,(t,x)\,|\,x=\tilde{\xi}_2(t) \quad \text{for } t\in[t_{q_1},T]; \qquad x=\tilde{\zeta}_2(t) \quad \text{for } t\in[t_{q_4},t_{q_1}]\}.
\end{aligned}
\tag{2.56}
$$

By the construction of U_i $(i=1,\dots,10)$, we observe that $U=U(t,x)$ defined by (2.55) is a C^1 function out of curves ℓ_1, ℓ_2, $\tilde{\ell}_1$ and $\tilde{\ell}_2$; meanwhile ℓ_1 (i.e., OP_4), ℓ_2 (i.e., OP_3), $\tilde{\ell}_1$ (i.e., NQ_3) and $\tilde{\ell}_2$ (i.e., NQ_4) are C^1 smooth curves. See Figure 1.

Obviously, $U=U(t,x)$, defined by (2.55), satisfies the system (1.1) on the domain $\mathcal{D}(T)$, but out of the curves ℓ_1, ℓ_2, $\tilde{\ell}_1$ and $\tilde{\ell}_2$, in the class sense. On the other hand, it is clear that $U=U(t,x)$ satisfies the condition (1.15).

In what follows, we show that the curves ℓ_1 and $\tilde{\ell}_1$ (resp., ℓ_2 and $\tilde{\ell}_2$) are contact discontinuities corresponding to $\lambda_1(U)$ (resp., $\lambda_1(U)$).

In fact, we only need to prove that the Rankine–Hugoniot conditions (2.1)–(2.2) hold on the curves $x=\zeta_1(t)$ $(t\in[t_{p_1},t_{p_4}])$, $x=\zeta_2(t)$ $(t\in[t_{p_2},t_{p_3}])$, $x=\tilde{\zeta}_1(t)$ $(t\in[t_{q_3},t_{q_2}])$ and $x=\tilde{\zeta}_2(t)$ $(t\in[t_{q_4},t_{q_1}])$.

It follows from (1.8) and (2.53) that

$$
\lambda_1\left(U_6\left(t,\zeta_1(t)\right)\right)=\lambda_1\left(U_{10}\left(t,\zeta_1(t)\right)\right)\triangleq\sigma(t)\quad\forall\ t\in[t_{p_1},t_{p_4}].
\tag{2.57}
$$

This is just the desired (2.2) for the case $k=1$ and $x=\zeta_1(t)$.

We next show that the Rankine–Hugoniot condition (2.1) holds on $x=\zeta_1(t)$.

By (H_3), we know that, for any fixed $t\in[t_{p_1},t_{p_4}]$, there exists a C^1 curve segment $U=U(\tau)$ $(\tau\in[\tau_1,\tau_2])$ in $\mathcal{N}$ such that

$$
U(\tau_1)=U_6(t,\zeta_1(t)),\qquad U(\tau_2)=U_{10}\left(t,\zeta_1(t)\right),
$$

and

$$
\lambda_1\left(U(\tau)\right)=\sigma(t)\quad\forall\ \tau\in[\tau_1,\tau_2]\,.
\tag{2.58}
$$

Differentiating (2.58) with respect to τ gives

$$
\nabla\lambda_1\left(U(\tau)\right)\cdot\frac{dU}{d\tau}(\tau)=0\quad\forall\ \tau\in[\tau_1,\tau_2]\,.
\tag{2.59}
$$

By (1.3), (1.5), and (2.59), we observe that $\frac{dU}{d\tau}(\tau)$ is proportional to $\vec{r}_2(U(\tau))$. Then we have

$$
\sigma(t)\frac{dU}{d\tau}(\tau)=\nabla F\left(U(\tau)\right)\frac{dU}{d\tau}(\tau)\quad\forall\ \tau\in[\tau_1,\tau_2]\,.
$$

Integrating this yields the Rankine–Hugoniot condition (2.1) on $x=\zeta_1(t)$.

Others are similar.

Step 6: Control inputs $h_1(t)$ and $h_2(t)$. Finally, we define $h_1(t)$ and $h_2(t)$ as follows:

$$
h_1(t)=\begin{cases}
-B_1(U_{10}(t,-1),t) & \text{as } t\in[0,t_{p_4}],\\
-B_1(U_6(t,-1),t) & \text{as } t\in[t_{p_4},t_{q_4}],\\
-B_1(U_9(t,-1),t) & \text{as } t\in[t_{q_4},T]
\end{cases}
\tag{2.60}
$$

and

$$h_2(t) = \begin{cases} -B_2(U_7(t,1),t) & \text{as } t \in [0, t_{p_3}], \\ -B_2(U_5(t,1),t) & \text{as } t \in [t_{p_3}, t_{q_3}], \\ -B_2(U_8(t,1),t) & \text{as } t \in [t_{q_3}, T]. \end{cases} \tag{2.61}$$

$h_1(t)$ and $h_2(t)$ are just the desired control inputs such that the system (1.1), (1.9) possesses a piecewise C^1 solution $U = U(t,x)$ containing four contact discontinuities and satisfying (1.15) on the domain $\mathcal{D}(T)$, provided that $T > T_0$. Thus, the proof of Theorem 1.1 is completed. □

3. Some remarks. In this section, we give some important supplementary remarks.

REMARK 3.1. *Some physical systems always satisfy the hypotheses* (H_1)*–*(H_4)*, for example, the system of isentropic gas with the Von Kármán–Tsien pressure law, the system of relativistic gas dynamics with the relativistic counterpart of the Chaplygin pressure law, etc. (see* [5]*).*

REMARK 3.2. *Consider the general quasilinear system of conservation laws*

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0, \tag{3.1}$$

where $u = (u_1, \dots, u_n)^T$ *is the unknown vector function of* (t,x)*,* $f\colon \mathbb{R}^n \to \mathbb{R}^n$ *is a given* C^2 *vector function of* u*. Suppose that on the domain under consideration,* (3.1) *is a nonstrictly hyperbolic system with two characteristics, and each characteristic has a constant multiplicity, say, on the domain under consideration,*

$$\lambda_1(u) \equiv \cdots \equiv \lambda_m(u) \triangleq \lambda(u) < \mu(u) \triangleq \lambda_{m+1}(u) \equiv \cdots \equiv \lambda_n(u),$$

where $1 \le m \le n-1$*. When* $m > 1$ *or* $m < n-1$*, the system* (3.1) *is nonstrictly hyperbolic. In particular, when* $n \ge 4$ *and* $1 < m < n-1$*, the system* (3.1) *is rich, and* $\lambda(u)$*,* $\mu(u)$ *must be linearly degenerate (see* [1]*,* [4]*, or* [10]*). In this case, the system* (3.1) *can be rewritten as*

$$\begin{aligned} &\frac{\partial R_i}{\partial t} + \lambda(u)\frac{\partial R_i}{\partial x} = 0 \qquad (i = 1, \dots, m), \\ &\frac{\partial R_j}{\partial t} + \mu(u)\frac{\partial R_j}{\partial x} = 0 \qquad (j = m+1, \dots, n), \end{aligned} \tag{3.2}$$

where R_i $(i = 1, \dots, n)$ *are the Riemann invariants. For the present situation, we have a similar result.*

In fact, the rich systems generalize the class of 2×2 systems while preserving their essential properties:

(1) diagonalization with the help of the strict Riemann invariants;

(2) the infinite dimension of the entropy space.

See [10]. Therefore, noting the above properties and the linear degeneracy of the characteristic fields $\lambda(u)$, $\mu(u)$, in a manner completely similar to the proof of Theorem 1.1 we can obtain the conclusion stated in Remark 3.2. In this sense, we may call the system (3.1) (equivalently, (3.2)) a *generalized* 2×2 *system with linearly degenerate characteristics.*

REMARK 3.3. *Hypothesis* (H_3) *is a geometric assumption for constructing contact discontinuities with nonsmall jumps. It is needed since we do not require that the*

oscillations of $u_0^{\pm}(x)$, $v_0^{\pm}(x)$, $u_T^{\pm}(x)$, $v_T^{\pm}(x)$ *and the jumps* $|u_0^+(0) - u_0^+(0)|$, $|v_0^+(0) - v_0^+(0)|$, $|u_T^+(0) - u_T^+(0)|$, $|v_T^+(0) - v_T^+(0)|$ *are small. If the above jumps are small, then the hypothesis* (H_3) *is not needed. Moreover, in Theorem* 1.1*, a part of contact discontinuities in the solution* $U = U(t,x)$ *may disappear. If the initial and terminal functions are continuous at* $x = 0$*, then the contact discontinuities in* $U = U(t,x)$ *degenerate weak discontinuities. If the initial and terminal functions are* C^1 *smooth on* $[-1,1]$*, then the solution* $U = U(t,x)$ *is also* C^1 *smooth, in this case Theorem* 1.1 *is only the result given in* [9].

REMARK 3.4. *The hypothesis* (H_4) *is also needful since we do not require that the oscillations and jumps of the initial and terminal functions are small. If the oscillations and jumps of the initial and terminal functions are small, then the hypothesis* (H_4) *is not needed. Moreover, it is also required in the proof of Theorem* 1.1 *that the mapping defined by* (1.6) *is inverse, i.e., we can solve* U *from* (1.6).

4. A new example — the system for time-like extremal surfaces in the $(1+n)$-dimensional Minkowski space $\mathbb{R}^{1+n}$. In this section, we provide a new example of system such as described in Remark 3.2, namely, the system for time-like extremal surfaces in the $(1+n)$-dimensional Minkowski space $\mathbb{R}^{1+n}$.

We first give some notations and definitions.

Let $X = (t, x_1, \dots, x_n)^T$ be the position vector of a point in the $(1+n)$-dimensional Minkowski space $\mathbb{R}^{1+n}$. The scalar product of two vectors X and $\bar{X}$ is

$$X \cdot \bar{X} = \sum_{i=1}^{n} x_i \bar{x}_i - t\bar{t}; \tag{4.1}$$

in particular,

$$X \cdot X = \sum_{i=1}^{n} x_i^2 - t^2. \tag{4.2}$$

The Lorentz metric of the space $\mathbb{R}^{1+n}$ reads as

$$ds^2 = \sum_{i=1}^{n} dx_i^2 - dt^2. \tag{4.3}$$

A nonzero vector $X \in \mathbb{R}^{1+n}$ is called *space-like* (resp., *time-like or light-like*) if

$$X \cdot X > 0 \quad (\text{resp.}, \ < 0 \ \text{ or } \ = 0). \tag{4.4}$$

We now consider a smooth surface Σ in the space $\mathbb{R}^{1+n}$. Let a point $\mathbf{p}$ be in Σ, i.e., $\mathbf{p} \in \Sigma$. The surface Σ is said to be *time-like* (resp., *space-like or light-like*) at $\mathbf{p}$ if the unit normal vector of Σ at $\mathbf{p}$, denoted by $\vec{n}_p$, is space-like (resp., time-like or light-like). If Σ is time-like (resp., space-like or light-like) at every point $\mathbf{p} \in \Sigma$, then the surface Σ is called to be *time-like* (resp., *space-like or light-like*).

Let $z = (x_2, \dots, x_n)^T$. The local equation of a surface Σ in $\mathbb{R}^{1+n}$ in a suitable coordinate system can be written as

$$z = f(t, x_1) \quad \text{or} \quad z_j = f_j(t, x_1) \quad (j = 2, \dots, n). \tag{4.5}$$

In what follows, we use the notation

$$x = x_1$$

and we only consider the time-like[1] surfaces. By the definition, it is easy to prove the following.

PROPOSITION 4.1. *The surface Σ is time-like if and only if*

$$1+|f_x|^2-|f_t|^2-|f_t|^2|f_x|^2+\langle f_t,f_x\rangle^2>0.$$

DEFINITION 4.1. *The surface Σ is called an extremal surface if f is the critical point of the area functional*

$$I=\iint\sqrt{1+|f_x|^2-|f_t|^2-|f_t|^2|f_x|^2+\langle f_t,f_x\rangle^2}dxdt, \tag{4.6}$$

where $\langle\cdot,\cdot\rangle$ stands for the inner product.

More generally, we consider a vector function $\phi=(\phi_1,\dots,\phi_n)^T$, which is the critical point of the area functional

$$I=\iint\sqrt{1+|\phi_x|^2-|\phi_t|^2-|\phi_t|^2|\phi_x|^2+\langle\phi_t,\phi_x\rangle^2}dxdt, \tag{4.7}$$

where $\langle\cdot,\cdot\rangle$ stands for the inner product. The corresponding Euler–Lagrange equation is as follows:

$$\begin{aligned}&\left(\frac{\phi_t}{\sqrt{1+|\phi_x|^2-|\phi_t|^2-|\phi_t|^2|\phi_x|^2+\langle\phi_t,\phi_x\rangle^2}}\right)_t\\&+\left(\frac{\phi_x}{\sqrt{1+|\phi_x|^2-|\phi_t|^2-|\phi_t|^2|\phi_x|^2+\langle\phi_t,\phi_x\rangle^2}}\right)_x\\&+\left(\frac{|\phi_x|^2\phi_t-\langle\phi_t,\phi_x\rangle\phi_x}{\sqrt{1+|\phi_x|^2-|\phi_t|^2-|\phi_t|^2|\phi_x|^2+\langle\phi_t,\phi_x\rangle^2}}\right)_t\\&-\left(\frac{\langle\phi_t,\phi_x\rangle\phi_t-|\phi_t|^2\phi_x}{\sqrt{1+|\phi_x|^2-|\phi_t|^2-|\phi_t|^2|\phi_x|^2+\langle\phi_t,\phi_x\rangle^2}}\right)_x=0.\end{aligned} \tag{4.8}$$

REMARK 4.1. *In fact, $\phi=\phi(t,x)$ stands for a time-like extremal surface in the $(1+(1+n))$-dimensional Minkowski space $\mathbb{R}^{1+(1+n)}$.*

REMARK 4.2. *In particular, let (t,x,y) be points in the $(1+2)$-dimensional Minkowski space $\mathbb{R}^{1+2}$. We now consider a time-like extremal surface Σ_{1+2} taking the form*

$$y=\phi(t,x). \tag{4.9}$$

Corresponding to (4.8)*, the Euler–Lagrange equation reads as*

$$\left(\frac{\phi_t}{\sqrt{1+\phi_x^2-\phi_t^2}}\right)_t-\left(\frac{\phi_x}{\sqrt{1+\phi_x^2-\phi_t^2}}\right)_x=0, \tag{4.10}$$

which is the Born–Infeld equation (see [2]*).*

[1]The *time-like* corresponds to the relation of *cause and effect* in physics.

REMARK 4.3. *Recently, Brenier* [3] *suggests an equation for extremal surfaces in the* $(1+4)$*-dimensional Minkowski space, which is related to classical electrodynamics. By prescribing* $(t,s)\to(t,s,Y(t,s))$ *to be an extremal surface in the* $(1+4)$*-dimensional Minkowski space* (t,s,x_1,x_2,x_3) *with the signature* $(-,+,+,+,+)$*, we know that the area functional is*

$$I=\iint\sqrt{1+|\partial_s Y|^2-|\partial_t Y|^2-|\partial_s Y\times\partial_t Y|^2}dsdt. \tag{4.11}$$

The corresponding Euler–Lagrange equation is

(4.12)
$$\left(\frac{\phi_t}{\sqrt{1+|\phi_x|^2-|\phi_t|^2-|\phi_t\times\phi_x|^2}}\right)_t-\left(\frac{\phi_x}{\sqrt{1+|\phi_x|^2-|\phi_t|^2-|\phi_t\times\phi_x|^2}}\right)_x$$
$$+\left(\frac{\phi_x\times(\phi_t\times\phi_x)}{\sqrt{1+|\phi_x|^2-|\phi_t|^2-|\phi_t\times\phi_x|^2}}\right)_t-\left(\frac{\phi_t\times(\phi_t\times\phi_x)}{\sqrt{1+|\phi_x|^2-|\phi_t|^2-|\phi_t\times\phi_x|^2}}\right)_x=0,$$

where x *stands for* s *in* (4.11)*, and* $\phi=(\phi_1,\phi_2,\phi_3)^T$ *represents* Y *in* (4.11).

Let

$$u=\phi_x,\qquad v=\phi_t. \tag{4.13}$$

Then (4.8) can be equivalently rewritten as

$$\begin{cases} u_t-v_x=0,\\ \left(\dfrac{v}{\sqrt{1+|u|^2-|v|^2-|v|^2|u|^2+\langle u,v\rangle^2}}\right)_t\\ -\left(\dfrac{u}{\sqrt{1+|u|^2-|v|^2-|v|^2|u|^2+\langle u,v\rangle^2}}\right)_x\\ +\left(\dfrac{|u|^2v-\langle u,v\rangle u}{\sqrt{1+|u|^2-|v|^2-|v|^2|u|^2+\langle u,v\rangle^2}}\right)_t\\ -\left(\dfrac{\langle u,v\rangle v-|v|^2u}{\sqrt{1+|u|^2-|v|^2-|v|^2|u|^2+\langle u,v\rangle^2}}\right)_x=0 \end{cases} \tag{4.14}$$

for smooth solutions. The following lemma comes from [6].

LEMMA 4.1. *If the surface is time-like, that is,*

$$\triangle(u,v)\overset{\triangle}{=}1+|u|^2-|v|^2-|v|^2|u|^2+\langle u,v\rangle^2>0, \tag{4.15}$$

then (4.14) *is a nonstrictly hyperbolic system with two* n*-constant multiple eigenvalues:*

$$\lambda_1\equiv\cdots\equiv\lambda_n\overset{\triangle}{=}\lambda_-<\lambda_+\overset{\triangle}{=}\lambda_{n+1}\equiv\cdots\equiv\lambda_{2n}, \tag{4.16}$$

where

$$\lambda_\pm=\frac{1}{1+|u|^2}\left(-\langle u,v\rangle\pm\sqrt{1+|u|^2-|v|^2-|v|^2|u|^2+\langle u,v\rangle^2}\right);$$

moreover, the system (4.14) *is linear degenerate.*

Noting Remark 3.2, under suitable assumptions we can obtain the global exact boundary controllability for the system (4.14) (equivalently, (4.8)) in the class of piecewise C^1 functions. Here we omit the details.

Acknowledgments. The authors would like to thank the referee for pertinent comments and valuable suggestions. The first author (Kong) thanks Prof. Ta-Tsien Li for his constant encouragement and valuable suggestions. This work was completed while Kong was visiting University of Potsdam during the summer of 2003. Kong thanks Prof. B.-W. Schulze for his invitation and hospitality.

REFERENCES

[1] G. Boillat, *Chocs caractéristiques*, C. R. Acad. Sci. Paris, Sér. A, 274 (1972), pp. 1018–1021.
[2] M. Born and L. Infeld, *Foundation of the new field theory*, Proc. R. Soc. Lond., A144 (1934), pp. 425–451.
[3] Y. Brenier, *Some geometric PDEs related to hydrodynamics and electrodynamics*, Proc. ICM, 3 (2002), pp. 761–772.
[4] H. Freistühler, *Linear degeneracy and shock waves*, Math. Zeit., 207 (1991), pp. 583–596.
[5] D.-X. Kong, *Global exact boundary controllability of a class of quasilinear hyperbolic systems of conservation laws*, Systems Control Lett., 47 (2002), pp. 287–298.
[6] D.-X. Kong, Q.-Y. Sun, and Y. Zhou, *The equation for time-like extremal surfaces in Minkowski space* $\mathbb{R}^{2+n}$, to appear.
[7] D.-X. Kong and M. Tsuji, *Global solutions for* 2×2 *hyperbolic systems with linearly degenerate characteristics*, Funkcialaj Ekvacioj, 42 (1999), pp. 129–155.
[8] T.-T. Li and B.-P. Rao, *Exact boundary controllability for quasilinear hyperbolic systems*, SIAM J. Control Optim., 41 (2003), pp. 1748–1755.
[9] T.-T. Li and B.-Y. Zhang, *Global exact controllability of a class of quasilinear hyperbolic systems*, J. Math. Anal. Appl., 225 (1998), pp. 289–311.
[10] D. Serre, *Systems of Conservation Laws* 2: *Geometric Structures, Oscillations, and Initial-Boundary Value Problems*, Cambridge University Press, Cambridge, 2000.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 1, pp. 159–194

ON THE BACKWARD STOCHASTIC RICCATI EQUATION IN INFINITE DIMENSIONS*

GIUSEPPINA GUATTERI† AND GIANMARIO TESSITORE‡

Abstract. We study backward stochastic Riccati equations (BSREs) arising in quadratic optimal control problems with infinite dimensional stochastic differential state equations. We allow the coefficients, both in the state equation and in the cost, to be random. In such a context BSREs are backward stochastic differential equations existing in a non-Hilbert space and involving quadratic nonlinearities. We propose two different notions of solutions to BSREs and prove, for both of them, existence and uniqueness results. We also show that such solutions allow us to perform the synthesis of the optimal control. Finally we apply our results to the optimal control of a delay equation and of a wave equation with random damping.

Key words. backward stochastic differential equations, Riccati equation, linear quadratic optimal control, Hilbert spaces, stochastic coefficients

AMS subject classifications. 93E20, 60H10

DOI. 10.1137/S0363012903425507

1. Introduction.

Backward stochastic Riccati differential equations (BSREs) naturally arise in the study of stochastic optimal linear quadratic control problems with stochastic coefficients.

The interest of proving existence and uniqueness results for such a class of equations was first addressed by Bismut in [2]. It was clear from the beginning that to study those highly nonlinear backward stochastic differential equations was already a challenging task in the finite dimensional case (see [3], [20], or the historical review in [12]). The difficulty comes essentially from the fact that, in its general formulation, the BSRE involves quadratic terms in both the unknowns (in particular in the so-called martingale term). Moreover the nonlinearity can be well defined only in a subset of the space of nonnegative matrices (where the equation naturally exists).

Several works followed the pioneering paper [2] (see [19], [12], [13], [14], [15]). In particular only very recently, in [21], the proof of the existence and uniqueness of a solution of the BSRE was given in the general case corresponding to a finite dimensional, linear quadratic problem with random coefficients and state- and control-dependent noise. This last result, somehow, completes the theory of finite dimensional BSREs. We remark that in all the above literature it is clear that the treatment of the equation cannot be solely based on general backward stochastic differential equation techniques but needs to exploit the interplay between the Riccati equation and its control theoretic interpretation (for results on general backward stochastic differential equations with quadratic nonlinearities see [11] and [17]).

On the other hand several works, motivated by control of stochastic partial differential equations, have been devoted to linear quadratic optimal control problems for infinite dimensional stochastic differential equations with deterministic coefficients

*Received by the editors March 31, 2003; accepted for publication (in revised form) August 17, 2004; published electronically June 27, 2005.
http://www.siam.org/journals/sicon/44-1/42550.html

†Dipartimento di Matematica, Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milano, Italia (giuseppina.guatteri@mate.polimi.it).

‡Dipartimento di Matematica, Università di Parma, via D'Azeglio 85, 43100 Parma, Italia (gianmario.tessitore@unipr.it).

(see, for instance, [22] and references within). The corresponding Riccati equation is a deterministic nonlinear ODE in a suitable space of symmetric, nonnegative, Hilbert valued operators.

The present paper is, as far as we know, the first attempt to consider infinite dimensional BSREs. Such equations naturally arise in several models; namely they appear in all the situations in which one has to perform the synthesis of the optimal control for a linear quadratic problem having, as state equation, an infinite dimensional stochastic evolution equation with random coefficients (see examples in sections 9 and 10). We also emphasize that the study of infinite dimensional BSREs introduces specific new difficulties in the theory of backward stochastic differential equations. Specifically these are nonlinear backward stochastic differential equations that involve unbounded linear terms and quadratic nonlinearities. Moreover, and this is the main difficulty, they naturally exist in a non-Hilbertian infinite dimensional space.

In order to separate difficulties we consider here only the case in which the nonlinearity does not depend on the "martingale term" of the backward equation. In other words we consider the infinite dimensional analogue of the equation considered, in the finite dimensional case in [19]. We believe that, as we explain in the following, this case already presents serious new difficulties.

To be more precise: in this paper we consider a quadratic optimal control problem for a system governed by the following *state equation*:

$$\begin{cases} dy(t) = (Ay(t) + A_\sharp(t)y(t) + B(t)u(t))\,dt + C(t)y(t)\,dW(t), \quad t \in [0,T], \\ y(0) = x. \end{cases} \tag{1.1}$$

In the above equation y is the *state* of the system and u is the *control*; y has values in a Hilbert space H and u has values in another Hilbert space U; W is a cylindrical Ξ-valued Wiener process defined on a probability space $\{\Omega, \mathcal{F}, \mathbb{P}\}$, where Ξ is a third Hilbert space. Expanding notation with respect to an orthonormal basis $\{f_i : i \in \mathbb{N}\}$ in Ξ we have $C(t)y(t)\,dW(t) = \sum_{i=1}^{\infty} C_i(t)y(t)d\beta_i(t)$, where $\{\beta_i : i \in \mathbb{N}\} := \{(f_i, W)_\Xi : i \in \mathbb{N}\}$ is a family of standard independent Brownian motions.

We assume that the unbounded operator $A : \mathcal{D}(A) \subset H \to H$ is independent of $\omega \in \Omega$ and $t \in [0,T]$ and is the infinitesimal generator of a C_0-semigroup. On the contrary $A_\sharp$, B, and C are allowed to be random; namely they are bounded, operator valued, stochastic processes that we assume to be predictable relatively to the filtration $\mathcal{F} = \{\mathcal{F}_t : t \geq 0\}$ generated by W (this last condition is not restrictive; see Remark 2.4).

Our purpose is to minimize, over all predictable controls u, the quadratic cost functional

$$\mathbb{E}\int_0^T \left(|\sqrt{S}(s)y(s)|_H^2 + |u(s)|_U^2\right) ds + \mathbb{E}(P_T y(T), y(T))_H, \tag{1.2}$$

where S is a predictable stochastic process and P_T is a random variable, both taking values in the set of linear, symmetric, nonnegative, and bounded operators from H into H.

If we define the *stochastic value function* by

$$(P(t)x,x)_H \doteq \inf_u \mathbb{E}^{\mathcal{F}_t,\, y(t)=x}\left[\int_t^T \left(|\sqrt{S}(s)y(s)|_H^2 + |u(s)|_U^2\right) ds + (P_T y(T), y(T))_H\right], \tag{1.3}$$

then P solves, at least in a formal way, the following backward stochastic differential equation:

(1.4)
$$\begin{cases} -dP(t) = \Big(A^*P(t)+P(t)A+A_\sharp^*(t)P(t)+P(t)A_\sharp(t)-P(t)B(t)B^*(t)P(t)+S(t)\Big)\,dt \\ \qquad\qquad + \operatorname{Tr}[C^*(t)P(t)C(t)+C^*(t)Q(t)+Q(t)C(t)]\,dt+Q(t)\,dW(t), \quad t\in[0,T], \\ P(T) = P_T. \end{cases}$$

We notice that the unknowns in (1.4) are the two processes P and Q (the second one is sometimes referred to as a *martingale* term). Process P has values in the cone $\Sigma^+(H)$ of bounded, nonnegative, linear symmetric operators in H and process Q in the space $L_2(\Xi,\Sigma(H))$ of Hilbert–Schmidt operators from Ξ to the space $\Sigma(H)$ of bounded, linear symmetric operators in H. Moreover, again making the notation explicit, we have

$$\operatorname{Tr}[C^*(t)PC(t)+C^*(t)Q+QC(t)] = \sum_{i=1}^{\infty}[C_i^*(t)PC_i(t)+C_i^*(t)(Qf_i)+(Qf_i)C_i(t)].$$

The specificity of our situation resides in the fact that the above equation involves both the unbounded term A^*P+PA and the quadratic term PBB^*P. Moreover $\Sigma(H)$ is not a Hilbert space; thus some essential tools in stochastic calculus commonly used in the theory of backward stochastic differential equations, such as the Kunita–Watanabe martingale representation theorem, fail to hold. To overcome this difficulty one could try to compute the operator valued random variables on the vectors of a basis and then apply classical representation results to each component, but this procedure does not seem to allow the reconstruction of a suitable operator valued process Q. The point is that, due to the presence of an unbounded term, we cannot consider (1.4) in its classical sense. Normally this leads to a mild formulation of the equations. Here, due to the difficulty of handling the martingale representation term Q, this approach causes problems. As a matter of fact mild formulation requires defining the process like $s\to e^{(s-t)A^*}Q(s)e^{(s-t)A}h$, $h\in H$, while only the processes $Q(\cdot)h$ with h independent on t are well defined.

For the same reason, in the generality considered here, it seems difficult to show uniqueness of weak solutions of BSREs.

To cope with such a roadblock we propose the following strategy inspired by the notion of "strong solution" for partial differential equations; see [1] or [16] and references therein. Roughly speaking the method consists of first considering equations with more regular data and then defining the solution in the general case by a limiting procedure.

To continue with this program we devote the first part of the paper (up to section 5) to the case in which the process S and the random variable P_T (corresponding, respectively, to the running and final cost) take values in the Hilbert space $L_2(H)$ of Hilbert–Schmidt operators $H\to H$ (see assumption (A5)). To begin we prove existence, uniqueness, and stability with respect to approximations of the solution to a class of infinite dimensional backward stochastic differential equations with unbounded linear term and lipschitz nonlinearity; see Theorem 4.4. This result is essentially included in [10] in sofar as existence and uniqueness are concerned (except that we find a slightly more regular solution) while the part dealing with stability seems to be new and of independent interest.

The above general result is then applied to the affine Lyapunov equation

$$\begin{cases} -dP(t) = (A^*P(t) + P(t)A + \mathrm{Tr}[C^*(t)P(t)C(t) + C^*(t)Q(t) + Q(t)C(t)])\,dt \\ \qquad\qquad +(A^*_\sharp(t)P(t) + P(t)A_\sharp(t) + L(t))\,dt + Q(t)\,dW(t), \quad t \in [0,T], \\ P(T) = P_T \end{cases} \tag{1.5}$$

when L is a given Hilbert–Schmidt valued predictable process.

Then by fixed point technique and a priori estimates (see also [19]) we are able to show that if S and P_T take values in the Hilbert space $L_2(H)$, then (1.4) has, in $L_2(H)$, a unique *mild* solution (P,Q). By that we mean a pair of processes verifying $\mathbb{P}$-a.s. for all $t \in [0,T]$:

$$\begin{aligned} P(t) = &\int_t^T e^{(s-t)A^*}\mathrm{Tr}[C^*(s)P(s)C(s) + C^*(s)Q(s) + Q(s)C(s)]e^{(s-t)A}\,ds \\ &+e^{(T-t)A^*}P_Te^{(T-t)A} + \int_t^T e^{(s-t)A^*}Q(s)e^{(s-t)A}\,dW(s) \\ &+ \int_t^T e^{(s-t)A^*}S(s)e^{(s-t)A}\,ds + \int_t^T e^{(s-t)A^*}(A^*_\sharp(s)P(s) \\ &+P(s)A_\sharp(s) - P(s)B(s)B^*(s)P(s))e^{(s-t)A}\,ds. \end{aligned} \tag{1.6}$$

Moreover, we prove that such a solution can be approximated by the classical solutions of the equations obtained replacing A by its Yosida approximations. Once we have a solution to the Riccati equation it is easy to perform in this Hilbertian framework, the standard synthesis of the optimal control: that is, to verify that $(P(0)x,x)_H$ is the optimal cost and that the unique optimal control $\overline{u}$ verifies the *feedback law* $\overline{u}(t) = -B^*(t)P(t)\overline{y}(t)$ (see Theorem 5.14).

Hilbert–Schmidt Assumption (A5) is too restrictive in many of the concrete applications (see the example in section 10 and Remark 10.1) so it is necessary to complete the above mentioned programm in order to include in the theory general running costs S and final conditions P_T. In section 6 we introduce the concept of *generalized solutions* of (1.4). By this we mean limits (in a suitable sense) of solutions corresponding to Hilbert–Schmidt data S and P_T. We are able to prove, under fairly general assumptions, that a generalized solution, in the above sense, exists and is unique (see Theorem 6.6). Notice that if existence of a generalized solution is somehow expected, uniqueness seems a more interesting result; its proof is largely based on the control-theoretic interpretation of (1.4). Moreover, we show that such a solution still allows to perform the synthesis of the optimal control as in the Hilbert–Schmidt case (see again Theorem 6.6). We also notice that their control theoretic interpretation imply that generalized solutions enjoy "strong continuity" property (see Lemma 6.5).

In section 7 we prove that generalized solutions verify the following *variation of constants* formula:

$$\begin{aligned} (P(t)x,x)_H = &(L_{t,T}P_Tx,x)_H + \int_t^T (L_{t,s}S(s)x,x)_H\,ds \\ &- \int_t^T (L_{t,s}P(s)B(s)B^*(s)P(s)x,x)_H\,ds \quad \mathbb{P}\text{-a.s.}, \end{aligned} \tag{1.7}$$

where $L_{t,s}$ is the evolution operator corresponding to the Lyapunov equation (1.5) with $L = 0$. We are also able to show that there exists a unique process P verifying (1.7). Thus (1.7) can be regarded as an alternative definition of solution to the BSRE (1.4). We notice that in both the definitions of solution we propose that only the P term in the BSRE is characterized. This is natural from the point of view of control theory and, in any case, is enough to complete the synthesis of the optimal control, see also Remark 6.3.

In sections 9 and 10 we show that our general results can be applied to a variety of concrete examples. The first example is a minimization of variance problem for a delay equation with a stochastic coefficient. The interest of such an example is that on one side it is extremely simple (and consequently applicable to a wide range of concrete situations) on the other it is connected with financial applications. Namely it is a firs step towards a mean variance hedging problem for a market with stochastic variance and memory effects. The second example is an optimal control problem for a wave equation in random media. In this case a stochastic coefficient is introduced, in a realistic way, assuming that the equation is subject to a stochastic damping due to the media. We notice that for the example in section 9 the Hilbert–Schmidt assumption (A5) is verified and we obtain mild solutions of the corresponding Riccati equation. On the contrary, for the example in section 9 the Hilbert–Schmidt assumption (A5) is never verified and we have to use the concept of generalized solution of the Riccati equation.

2. Main notation and assumptions. By H, U, and Ξ we will always indicate real separable Hilbert spaces.

If K is a Hilbert space, its inner scalar product and norm will be denoted by $(\cdot,\cdot)_K$ and $|\cdot|_K$, omitting the K when no confusion is possible.

For any Banach space E by $\mathcal{B}(E)$ we denote its Borel σ-field.

For any pair K_1 and K_2 of separable real Hilbert spaces we denote by $L(K_1, K_2)$ the Banach space of linear and bounded operators from K_1 to K_2 endowed by the norm $|T|_{L(K_1,K_2)} = \sup_{\{x\in K_1, |x|_{K_1}=1\}} |Tx|_{K_2}$ (as usual $L(H) = L(H, H)$).

By $\Sigma(H)$ we denote the subspace of all symmetric and bounded operators, and by $\Sigma^+(H)$ the cone of $\Sigma(H)$ that contains all positive semidefinite operators.

$L_2(K, H)$ denotes the Hilbert space of Hilbert–Schmidt operators from K to H, endowed with the Hilbert–Schmidt norm $|T|^2_{L_2(K,H)} = \sum_{i=1}^{\infty} |Te_i|^2_H$ ($\{e_i : i \in \mathbb{N}\}$ being an orthonormal basis in K), and we set $L_2(H, H) = L_2(H)$. $\Sigma_2(H)$ is the subset of $L_2(H)$ that consists of all linear and symmetric operators, and $\Sigma_2^+(H)$ is the cone of $\Sigma_2(H)$ that consists of all nonnegative operators.

The cylindrical Wiener process. We fix a probability basis $(\Omega, \mathcal{F}, \mathbb{P})$. A cylindrical Wiener process with value in Ξ is a family $W(t)$, $t \geq 0$, of linear mappings $\Xi \to L^2(\Omega)$ such that

(i) for every $h \in \Xi$, $\{W(t)h,\, t \geq 0\}$ is a real (continuous) Wiener process;

(ii) for every $h, k \in \Xi$ and $t, s \geq 0$, $\mathbb{E}(W(t)h \cdot W(s)k) = (t \wedge s)(h, k)_\Xi$.

We denote by $\mathcal{F}_t$ its natural filtration augmented with the set $\mathcal{N}$ of $\mathbb{P}$-null sets of $\mathcal{F}$. As is well known, the filtration $\mathcal{F}_t$ satisfies the usual conditions. By $\mathbb{E}^{\mathcal{F}_t}$ we denote the conditional expectation with respect to $\mathcal{F}_t$.

Finally by $\mathcal{P}$ we denote the predictable σ-field on $\Omega \times [0, T]$.

Some classes of stochastic process. Let K be any separable Hilbert space and let $\mathcal{B}(K)$ be its Borel σ-field on K. The following classes of processes will be used in this work.

- $L^p_{\mathcal{P}}(\Omega\times[0,T];K)$, $p\in[1,+\infty]$ denotes the subset of $L^p(\Omega\times[0,T];K)$, given by all equivalence classes admitting a predictable version. This space is endowed with the natural norm

$$|Y|^p_{L^p_{\mathcal{P}}(\Omega\times[0,T];K)}=\mathbb{E}\int_0^T|Y_s|^p_K\,ds$$

  Elements of this space are defined up to modification.
- $L^p_{\mathcal{P}}(\Omega;L^2([0,T];K))$ denotes the space of equivalence classes of processes Y, admitting a predictable version such that the norm

$$|Y|^p_{L^p_{\mathcal{P}}(\Omega;L^2([0,T];K))}=\mathbb{E}\bigg(\int_0^T|Y_s|^2_K\,ds\bigg)^{p/2}$$

  is finite. Elements of this space are defined up to modification.
- $C_{\mathcal{P}}([0,T];L^p(\Omega;K))$ denotes the space of K-valued processes Y such that $Y:[0,T]\to L^p(\Omega,K)$ is continuous and Y has a predictable modification, endowed with the norm

$$|Y|^p_{C_{\mathcal{P}}([0,T];L^p(\Omega;K))}=\sup_{t\in[0,T]}\mathbb{E}|Y_t|^p_K$$

  Elements of $C_{\mathcal{P}}([0,T];L^p(\Omega;K))$ are identified up to modification.
- $L^p_{\mathcal{P}}(\Omega;C([0,T];K))$ denotes the space of predictable processes Y with continuous paths in K such that the norm

$$|Y|^p_{L^p_{\mathcal{P}}(\Omega;C([0,T];K))}=\mathbb{E}\sup_{t\in[0,T]}|Y_t|^p_K$$

  is finite. Elements of this space are defined up to indistinguishability.

Now let us consider the space $L(H)$ of linear and bounded operators from a separable Hilbert space H to H. Moreover, it turns out that the σ-field generated by the operator norm in $L(H)$ is too large. For instance if A generates a C_0 semigroup, the map $t\to e^{tA}$ is not even measurable with respect to such σ-field, see [5, pp. 23–24]. We are, therefore, led to introduce the σ-field

$$\mathcal{L}_S=\sigma\{\{T\in L(H):Tu\in A\},\ \text{where } u\in H \text{ and } A\in\mathcal{B}(H)\}.$$

Again following [5] the elements of $\mathcal{L}_S$ are called *strongly measurable*.

We notice that the maps $P\to|P|_{L(H)}$ and $(P,u)\to Pu$ are measurable from $(L(H),\mathcal{L}_S)$ to $(\mathbb{R},\mathcal{B}(\mathbb{R}))$ and from $(L(H)\times H,\mathcal{L}_S\otimes\mathcal{B}(H))$ to $(H,\mathcal{B}(H))$, respectively.

Moreover, $\mathcal{L}_S$ is identical to the weak σ-field

$$\mathcal{L}_S=\sigma\{\{T\in L(H):(Tu,x)_H\in A\},\ \text{where } u,x\in H \text{ and } A\in\mathcal{B}(\mathbb{R})\}.$$

We define the following spaces.

- $L^\infty_{\mathcal{P},S}(\Omega\times[0,T];L(H))$ the space of essentially bounded, strongly measurable predictable processes $Y:\Omega\times[0,T]\to L(H)$. That is, Y is measurable from $(\Omega\times[0,T],\mathcal{P})$ to $(L(H),\mathcal{L}_S)$ and the real valued random valued $|Y|_{L(H)}$ is in $L^\infty(\Omega\times[0,T];\mathbb{R})$. By $|Y|_{L^\infty_{\mathcal{P},S}(\Omega\times[0,T];L(H))}$ we indicate the norm of $|Y|_{L(H)}$ in $L^\infty(\Omega\times[0,T];\mathbb{R})$. Elements of this space are identified up to modification.
- $L^\infty_S(\Omega,\mathcal{F}_t;L(H))$ is the space of measurable maps $Y:(\Omega,\mathcal{F}_t)\to(L(H),\mathcal{L}_S)$ such that $|Y|_{L(H)}$ is in $L^\infty(\Omega;\mathbb{R})$. By $|Y|_{L^\infty_S(\Omega;L(H))}$ we indicate the norm of $|Y|_{L(H)}$ in $L^\infty(\Omega;\mathbb{R})$.

- $L^1_{\mathcal{P},S}([0,T];L^\infty(\Omega,L(H)))$ is the space of predictable, strongly measurable processes such that $|Y|_{L(H)}$ is in $L^1([0,T];L^\infty(\Omega;\mathbb{R}))$. By $|Y|_{L^1_{\mathcal{P},S}([0,T];L^\infty(\Omega,L(H)))}$ we indicate the norm of $|Y|_{L(H)}$ in $L^1([0,T];L^\infty(\Omega;\mathbb{R}))$. Elements of this space are identified up to modification.

We identically define, with trivial changes the spaces: $L^\infty_{\mathcal{P},S}(\Omega\times[0,T];\Sigma^+(H))$, $L^\infty_{\mathcal{P},S}(\Omega\times[0,T];L(U,H))$, $L^1_{\mathcal{P},S}([0,T];L^\infty(\Omega,\Sigma^+(H)))$, and $L^\infty_S(\Omega,\mathcal{F}_t;\Sigma^+(H))$. Elements of these spaces are identified up to modification.

Statement of the problem and general assumptions on the coefficients. We consider the following infinite dimensional stochastic differential equation:

$$\begin{cases} dy(s) = (Ay(s) + A_\sharp(s)y(s) + B(s)u(s))\,ds + C(s)y(s)\,dW(s), & s\in[t,T],\\ y(t) = x, \end{cases} \tag{2.1}$$

where y is an H valued process that represents the *state* of the system and is our unknown, u is the *control* and the initial data x is in H. To stress its dependence on u, t, and x we will denote the (mild; see Definition 3.1) solution of (2.1) by $y^{t,x,u}$ when needed.

Our purpose is to minimize with respect to u the cost functional

$$\begin{aligned} J(0,x,u) = \mathbb{E}\Bigg[\int_0^T \left((S(s)y^{0,x,u}(s), y^{0,x,u}(s))_H + |u(s)|^2_U\right)\,ds \\ +(P_T y^{0,x,u}(T), y^{0,x,u}(T))_H\Bigg]. \end{aligned} \tag{2.2}$$

We also introduce the following random variables for $t\in[0,T]$:

$$\begin{aligned} J(t,x,u) = \mathbb{E}^{\mathcal{F}_t}\Bigg[\int_t^T \left((S(s)y^{t,x,u}(s), y^{t,x,u}(s))_H + |u(s)|^2_U\right)\,ds \\ +(P_T y^{t,x,u}(T), y^{t,x,u}(T))_H\Bigg]. \end{aligned}$$

We will work under the following general assumptions on A, B, and C that will hold throughout the paper.

Hypothesis 2.1.

(A1) $A : D(A)\subset H\to H$ is the infinitesimal generator of a C_0 semigroup $e^{tA} : H\to H$.

(A2) We assume that $A_\sharp\in L^\infty_{\mathcal{P},S}(\Omega\times[0,T];L(H))$. We denote by $M_{A_\sharp}$ a positive constant such that

$$|A_\sharp(t,\omega)|_{L(U,H)}\le M_{A_\sharp},\quad \mathbb{P}\text{-a.s. and for a.e. } t\in(0,T).$$

Moreover, $B\in L^\infty_{\mathcal{P},S}(\Omega\times[0,T];L(U,H))$. We denote by M_B a positive constant such that

$$|B(t,\omega)|_{L(U,H)}\le M_B,\quad \mathbb{P}\text{-a.s. and for a.e. } t\in(0,T).$$

(A3) We assume that C is of the form: $C=\sum_{i=1}^\infty C_i(\cdot,f_i)_\Xi$, where $\{f_i : i\in\mathbb{N}\}$ is an orthonormal basis in Ξ. Moreover, we suppose that

$$C_i\in L^\infty_{\mathcal{P},S}(\Omega\times[0,T];L(H)) \quad\text{and}\quad \left(\sum_{i=1}^\infty |C_i(t,\omega)|^2_{L(H)}\right)^{1/2}\le M_C,$$
$$\mathbb{P}\text{-a.s. for a.e. } t\in(0,T)$$

for a suitable positive constant M_C.

On S and P_T we will need to play with two different sets of assumptions. We introduce both of them here

(A4) $S \in L^1_{\mathcal{P},S}([0,T];L^\infty(\Omega;\Sigma^+(H)))$ and $P_T \in L^\infty_S(\Omega,\mathcal{F}_T;\Sigma^+(H))$;

(A5) $S \in L^2_{\mathcal{P}}(\Omega\times[0,T];\Sigma^+_2(H))$ and $P_T \in L^2(\Omega,\mathcal{F}_T;\Sigma^+_2(H))$.

We introduce, for later use, the Yosida approximants of the unbounded operator A, letting

$$A_h = AJ(h,A), \quad \text{where} \quad J(h,A) = h(hI-A)^{-1}, \quad h:1,2,\dots.$$

We denote by M_A a positive constant such that

$$\sup_{t\in[0,T]} |e^{tA_h}|_{L(H)} \le M_A \quad \forall h\in\mathbb{N} \quad \text{and} \quad \sup_{t\in[0,T]} |e^{tA}|_{L(H)} \le M_A \tag{2.3}$$

Remark 2.2. If we set $\beta_i(t) := (f_i, W(t))_\Xi$, then $\{\beta_i : i\in\mathbb{N}\}$ is a family of independent standard (real valued) Brownian motions. Moreover, the term $C(t)y(t)dW(t)$ can be rewritten as $\sum_{i=1}^\infty C_i(t)y(t)\,d\beta_i(t)$.

Remark 2.3. In section 9 we show that assumptions (A1)–(A5) are satisfied by a general class of controlled stochastic delay equations. In section 10 we point out that for stochastic controlled partial differential equations, assumptions (A1)–(A4) are satisfied while (A5) typically fails. We also notice that when H is finite dimensional, (A5) and (A4) reduce to the requirements $S \in L^2_{\mathcal{P},S}([0,T];L^\infty(\Omega;\Sigma^+(H)))$ *and* $P_T \in L^\infty_S(\Omega,\mathcal{F}_T;\Sigma^+(H))$ which slightly generalize the assumptions in [19] and [12], [13], [14], where S is uniformly bounded.

Remark 2.4. The fact that in the previous assumptions measurability and predictability has always been required with respect to the filtration $\{\mathcal{F}_t : t\ge 0\}$ generated by the noise $\{W_t : t\ge 0\}$ is not restrictive. Such a condition can in fact be easily weakened by the following standard procedure.

Let $\widehat{\Xi}\supset\Xi$ be a larger separable Hilbert space and let $\{\widehat{W}_t : t\ge 0\}$ be a cylindrical Wiener process with values in $\widehat{\Xi}$. Moreover, let $\{\widehat{f}_i : i\in\mathbb{N}\}$ an orthonormal basis in $\widehat{\Xi}$ with $\{\widehat{f}_i : i\in\mathbb{N}\}\supset\{f_i : i\in\mathbb{N}\}$. Finally let $\widehat{C}_i = C_i$ if $f_i\in\Xi$, $\widehat{C}_i = 0$ if $f_i\notin\Xi$ and $\widehat{C} = \sum_{i=1}^\infty \widehat{C}_i(\cdot,\widehat{f}_i)_{\widehat{\Xi}}$. If now we replace Ξ by $\widehat{\Xi}$, W by $\widehat{W}$, and C by $\widehat{C}$, (2.1) is unchanged while in all the assumptions filtration $\{\mathcal{F}_t : t\ge 0\}$ can be replaced by filtration $\{\widehat{\mathcal{F}}_t : t\ge 0\}$ generated by $\widehat{W}$. In addition, in order to allow $\mathcal{F}_0$ to be nontrivial there are no difficulties in letting the noise W to be defined in $[-\rho,+\infty[$, for some $\rho>0$, instead that in $[0,+\infty[$.

3. The state equation. This section is devoted to the state equation (2.1). We recall the well known notion of *mild solution*.

DEFINITION 3.1. *Given $x\in H$ and $u\in L^2_{\mathcal{P}}(\Omega\times[t,T];U)$, a mild solution of* (2.1) *is a process $y\in L^2_{\mathcal{P}}(\Omega\times[t,T];H)$ such that, almost surely in $\Omega\times[t,T]$,*

$$y(s) = e^{(s-t)A}x + \int_t^s e^{(s-\sigma)A}\left[A_\sharp(\sigma)y(\sigma)+B(\sigma)u(\sigma)\right]d\sigma + \int_t^s e^{(s-\sigma)A}C(\sigma)y(\sigma)\,dW(\sigma).$$

The following existence and uniqueness result is now well known.

THEOREM 3.2. *Assume* (A1)–(A3). *Given any $x\in H$ and $u\in L^2_{\mathcal{P}}(\Omega\times[t,T];U)$ problem* (2.1) *has a unique mild solution $y\in C_{\mathcal{P}}([t,T];L^2(\Omega;H))$. Moreover,*

$$\sup_{s\in[t,T]}\mathbb{E}|y(s)|^2 \le C_2\Big[|x|^2 + \mathbb{E}\int_t^T |u(s)|^2\,ds\Big] \tag{3.1}$$

for a suitable constant C_2 depending on T, M_B, M_C $M_{A_\sharp}$, and M_A.

Finally if $p > 2$ and

$$\mathbb{E}\Big(\int_t^T |u(s)|^2\,ds\Big)^{\frac{p}{2}} < \infty,$$

then we have that $y \in L^p_{\mathcal{P}}(\Omega; C([t,T];H))$ and

$$\mathbb{E}\sup_{s\in[t,T]}|y(s)|^p \le C_p\left[|x|^p + \mathbb{E}\Big(\int_t^T |u(s)|^2\,ds\Big)^{\frac{p}{2}}\right] \tag{3.2}$$

for some positive constant C_p depending on p, T, M_B, M_C, M_A, and $M_{A_\sharp}$.

Proof. The argument is identical to the one included in [5, Theorem 7.4] and [7, Proposition 3.2]. The only difference is that here the operators B and C are stochastic processes. Anyway, thanks to their boundedness stated in Hypotheses 2.1, one can proceed exactly as in the above mentioned papers. □

To stress dependence on the initial data and on the control we will, when necessary, denote the above solution by $y^{t,x,u}$.

For all $x \in H$ and $u \in L^p_{\mathcal{P}}(\Omega; L^2([t,T];U))$, $p \ge 2$ we also introduce the following family of approximating problems, $h \in \mathbb{N}$:

$$\begin{cases} dy_h(s) = (A_h y_h(s) + A_\sharp(s)y_h(s) + B(s)u(s))\,dt + C(s)y_h(s)\,dW(s), \quad s \in [t,T], \\ y(t) = x. \end{cases} \tag{3.3}$$

It is well known (see [5]) that, under the same hypotheses of Theorem 3.2, problem (3.3) has, for every $h \in \mathbb{N}$, a unique *classical* solution $y_h \in L^p_{\mathcal{P}}(\Omega; C([t,T];H))$ that, when necessary, we will denote by $y_h^{t,x,u}$.

The following stability result for the approximated problems holds.

THEOREM 3.3. *Assume that $x_h \to x$ in H and $u_h \to u$ in $L^p_{\mathcal{P}}(\Omega; L^2([t,T];U))$, as $h \to \infty$. If $p = 2$, $y_h^{t,x_h,u_h} \to y^{t,x,u}$ in $C_{\mathcal{P}}([t,T]; L^2(\Omega;H))$. If $p > 2$, $y_h^{t,x_h,u_h} \to y^{t,x,u}$ in $L^p_{\mathcal{P}}(\Omega; C([t,T];H))$.*

Proof. The proof consists in a straightforward application of the parameter depending contraction argument (see, for instance, [24, Theorem 10.1]). The case with $p = 2$ is treated also in [22, Theorem 1.1]. For the case $p > 2$ it is enough to proceed as in [7, Proposition 3.2]. □

4. Backward stochastic equations: Stability with respect to approximations. In this section we prove, for later use, a result on the stability of a generic backward stochastic equation with value in an real and separable Hilbert space K and Lipschitz nonlinearity. Beside the same hypotheses on the noise introduced in the previous section, we are given

(i) a positive number $T > 0$;
(ii) an unbounded operator $G : D(G) \subset K \to K$ and a sequence of bounded operators $G_h : K \to K$;
(iii) a map $\psi : [0,T] \times \Omega \times K \times L_2(\Xi, K) \to K$;
(iv) a final data $\eta \in L^2(\Omega, \mathcal{F}_T, \mathbb{P}; K)$.

We assume the following.

Hypothesis 4.1.

1. G generates a C_0-semigroup $\{e^{tG} : t \ge 0\}$ in K.

2. There exists a constant M_G such that

$$\sup_{t\in[0,T]} |e^{tG_h}|_{L(H)} \le M_G \quad \forall h\in\mathbb{N} \quad \text{and} \quad \sup_{t\in[0,T]} |e^{tG}|_{L(H)} \le M_G. \tag{4.1}$$

3. $\sup_{t\in[0,T]} \left|e^{tG_h}x - e^{tG}x\right| \to 0$ for all $x\in K$.
4. ψ is measurable from $\mathcal{P}\otimes\mathcal{B}(K)\otimes\mathcal{B}(L_2(\Xi,K))$ to $\mathcal{B}(K)$ and $\mathbb{E}\int_0^T |\psi(s,0,0)|_K^2 ds < +\infty$
5. There exists a constant M_ψ such that, $\mathbb{P}$ almost surely for almost every $t\in[0,T]$, the following holds for all $Y_1,Y_2\in K$, $Z_1,Z_2\in L_2(\Xi,K)$:

$$|\psi(t,Y_1,Z_1)-\psi(t,Y_2,Z_2)|_K \le M_\psi\left(|Y_1-Y_2|_K + |Z_1-Z_2|_{L_2(\Xi,K)}\right). \tag{4.2}$$

We consider the following backward stochastic equation:

$$\begin{cases} dY(s) = -GY(s)\,ds - \psi(s,Y(s),Z(s))\,ds - Z(s)\,dW(s), \quad s\in[0,T], \\ Y(T)=\eta \end{cases} \tag{4.3}$$

and the following sequence of approximating problems:

$$\begin{cases} dY_h(s) = -G_hY_h(s)\,ds - \psi(s,Y_h(s),Z_h(s))\,ds - Z_h(s)\,dW(s), \quad s\in[0y,T], \\ Y_h(T)=\eta. \end{cases} \tag{4.4}$$

Definition 4.2. *A mild solution of* (4.3) *is a couple of predictable processes* (Y,Z) *such that* Y *belongs to* $L^2_{\mathcal{P}}(\Omega,C([0,T];K))$, Z *belongs to* $L^2_{\mathcal{P}}(\Omega\times[0,T];L_2(\Xi;K))$, *and they verify for all* $t\in[0,T]$

$$Y(t) = e^{(T-t)G}\eta + \int_t^T e^{(s-t)G}\psi(s,Y(s),Z(s))\,ds + \int_t^T e^{(s-t)G}Z(s)\,dW(s) \quad \mathbb{P}\text{-a.s.} \tag{4.5}$$

An identical definition is given for a mild solution of (4.4).

Remark 4.3. G_h being bounded, it is immediate to check that the couple (Y_h,Z_h) is a mild solution of (4.4) if and only if it is a classical solution of (4.4); that is, it verifies, for all $t\in[0,T]$,

$$Y_h(t) = \eta_h + \int_t^T (G_hY_h(s) + \psi(s,Y_h(s),Z_h(s)))\,ds + \int_t^T Z_h(s)\,dW(s) \qquad \mathbb{P}\text{-a.s.} \tag{4.6}$$

The following result will be used in several occasions in what follows. As far as the existence and uniqueness part is concerned, is very similar to the one included in [10] (except from the fact that we obtain a more regular solution). On the contrary the part dealing with stability with respect to approximations is new.

Theorem 4.4. *Under Hypothesis* 4.1 *problem* (4.3) *has a unique mild solution* (Y,Z). *Moreover,* for all $h\in\mathbb{N}$, *problem* (4.4) *has a unique classical (equivalently mild) solution* (Y_h,Z_h).

Finally

$$\lim_{h\to\infty}\mathbb{E}\Big(\sup_{t\in[0,T]}|Y_h(t)-Y(t)|_K^2\Big)=0, \quad \lim_{h\to\infty}\mathbb{E}\int_0^T |Z_h(s)-Z(s)|^2_{L_2(\Xi;K)}ds = 0. \tag{4.7}$$

Proof. Part I. *Existence and uniqueness for a simplified equation.* We consider the simplified equation

(4.8)
$$Y(t) = e^{(T-t)G}\eta + \int_t^T e^{(s-t)G}F(s)\,ds + \int_t^T e^{(s-t)G}Z(s)\,dW(s), \quad t \in [0,T],$$

with $F \in L^2_{\mathcal{P}}(\Omega\times[0,T];K)$. In [10, Proposition 2.1] it is shown that the above equation admits a unique solution $(Y,Z) \in L^2_{\mathcal{P}}(\Omega\times[0,T];K) \times L^2_{\mathcal{P}}(\Omega\times[0,T];L_2(\Xi;K))$ given explicitly by

$$Y(t) = e^{(T-t)G}(\mathbb{E}^{\mathcal{F}_t}\eta) + \int_t^T e^{(s-t)G}(\mathbb{E}^{\mathcal{F}_t}F(s))\,ds, \tag{4.9}$$

$$Z(t) = -e^{(T-t)G}V(t) - \int_t^T e^{(s-t)G}L(t,s)\,ds, \tag{4.10}$$

where V and L verify

$$\mathbb{E}^{\mathcal{F}_t}\eta = \eta - \int_t^T V(\sigma)\,dW(\sigma), \quad 0 \le t \le T, \tag{4.11}$$

$$\mathbb{E}^{\mathcal{F}_t}F(s) = F(s) - \int_t^s L(\sigma,s)\,dW(\sigma), \quad 0 \le t \le s \le T \tag{4.12}$$

(existence and uniqueness of V and L are given by the Kunita–Watanabe martingale representation result applied in the Hilbert space K; again see [10]).

We now estimate such a solution in a suitable norm. For every $\beta > 0$,

$$\mathbb{E}\sup_{t\in[0,T]} e^{2\beta t}|Y(t)|_K^2 \le 2M_G^2\left[\mathbb{E}\sup_{t\in[0,T]} e^{2\beta t}\left(\int_t^T \mathbb{E}^{\mathcal{F}_t}|F(\sigma)|_K\,d\sigma\right)^2 + \mathbb{E}\sup_{t\in[0,T]} e^{2\beta t}|\mathbb{E}^{\mathcal{F}_t}\eta|_K^2\right].$$

Since

$$\left(\int_t^T |F(\sigma)|_K\,d\sigma\right)^2 \le \int_t^T e^{-2\beta s}\,ds \int_t^T e^{2\beta s}|F(s)|_K^2\,ds \le \frac{e^{-2\beta t}}{2\beta}\int_t^T e^{2\beta s}|F(s)|_K^2\,ds,$$

one gets that, thanks to Jensen and Doob inequalities,

$$\begin{aligned}\mathbb{E}\sup_{t\in[0,T]} e^{2\beta t}\left(\int_t^T \mathbb{E}^{\mathcal{F}_t}|F(s)|_K\,ds\right)^2 &\le \mathbb{E}\sup_{t\in[0,T]}\left(\mathbb{E}^{\mathcal{F}_t}\sup_{t\in[0,T]} e^{\beta t}\int_t^T |F(s)|_K\,ds\right)^2\\ &\le 4\mathbb{E}\sup_{t\in[0,T]} e^{2\beta t}\left(\int_t^T |F(s)|_K\,ds\right)^2 \le \frac{4}{2\beta}\mathbb{E}\int_0^T e^{2\beta s}|F(\sigma)|_K^2\,ds.\end{aligned}$$

Thus we have, using again Doob inequality,

$$\mathbb{E}\sup_{t\in[0,T]} e^{2\beta t}|Y(t)|_K^2 \le \frac{4M_G^2}{\beta}\mathbb{E}\int_0^T e^{2\beta s}|F(\sigma)|_K^2\,ds + 8M_G^2 e^{2\beta T}\mathbb{E}|\eta|_K^2. \tag{4.13}$$

As far as Z is concerned we have

$$|Z(t)|^2_{L_2(\Xi;K)} \le 2M_G^2\bigg[|V(t)|^2_{L_2(\Xi;K)} + \frac{e^{-2\beta t}}{2\beta}\int_t^T e^{2\beta s}|L(t,s)|^2_{L_2(\Xi;K)}\,ds\bigg].$$

Therefore,

$$\begin{aligned}\mathbb{E}\int_0^T e^{2\beta t}|Z(t)|^2_{L_2(\Xi;K)}\,dt &\le 2M_G^2\bigg[\mathbb{E}\int_0^T e^{2\beta t}|V(t)|^2_{L_2(\Xi;K)}\,dt\\ &+\frac{1}{2\beta}\mathbb{E}\int_0^T\int_t^T e^{2\beta s}|L(t,s)|^2_{L_2(\Xi;K)}\,ds\,dt\bigg] \le 2M_G^2\bigg[4e^{2\beta T}\mathbb{E}|\eta|^2_K\\ &+\frac{1}{2\beta}\mathbb{E}\int_0^T e^{2\beta s}\int_0^s |L(t,s)|^2_{L_2(\Xi;K)}\,dt\,ds\bigg]\end{aligned}$$

and we can conclude

$$\mathbb{E}\int_0^T e^{2\beta t}|Z(t)|^2_{L_2(\Xi;K)}\,dt \le 2M_G^2\bigg[4e^{2\beta T}\mathbb{E}|\eta|^2_K + \frac{2}{\beta}\int_0^T e^{2\beta s}\mathbb{E}|F(s)|^2_K\,ds\bigg]. \tag{4.14}$$

In an identical way we can prove that for all $h \in \mathbb{N}$ there exists a unique couple of processes (Y_h, Z_h) that belongs to $L^2_{\mathcal{P}}(\Omega; C([0,T];K)) \times L^2_{\mathcal{P}}(\Omega\times[0,T];L_2(\Xi;K))$ verifying, for all $t \in [0,T]$,

$$Y_h(t) = e^{(T-t)G_h}\eta + \int_t^T e^{(s-t)G_h}F(s)\,ds + \int_t^T e^{(s-t)G_h}Z_h(s)\,dW(s) \qquad \mathbb{P}\text{-a.s.} \tag{4.15}$$

with $F \in L^2_{\mathcal{P}}(\Omega\times[0,T];K)$.

Moreover, Y_h and Z_h verify (4.13) and (4.14).

Part II. *Stability with respect to approximations of the simplified equation.* By (4.10), we have, for a.e. $t \in [0,T]$,

$$\begin{aligned}Z_h(t) - Z(t) &= -e^{(T-t)G_h}V(t) + e^{(T-t)G}V(t) - \int_t^T e^{(s-t)G_h}L(t,s)\,ds\\ &+\int_t^T e^{(s-t)G}L(t,s)\,ds \quad \mathbb{P}\text{-a.s.}\end{aligned} \tag{4.16}$$

with $V \in L^2_{\mathcal{P}}(\Omega\times[0,T];L_2(\Xi;K))$ and $L \in L^2_{\mathcal{P}}(\Omega\times[0,T]\times[0,T];L_2(\Xi;K))$.

By the dominated convergence theorem, we immediately have that

$$\lim_{h\to+\infty}\mathbb{E}\int_0^T |Z_h(t) - Z(t)|^2_{L_2(\Xi;K)}\,dt = 0. \tag{4.17}$$

Now we consider the term $Y_h - Y$. We have

$$Y_h(t) - Y(t) \quad \big[e^{(T-t)G_h}\eta - e^{(T-t)G}\eta\big] + \mathbb{E}^{\mathcal{F}_t}\int_t^T \big[e^{(s-t)G_h}F(s) - e^{(s-t)G}F(s)\big]\,ds.$$

To estimate the first term of the right-hand side we can proceed as follows:

$$\begin{aligned}&\mathbb{E}\sup_{t\in[0,T]}\big|\mathbb{E}^{\mathcal{F}_t}[e^{(T-t)G_h}\eta - e^{(T-t)G}\eta]\big|^2_K \le \mathbb{E}\sup_{t\in[0,T]}\big(\mathbb{E}^{\mathcal{F}_t}\big|e^{(T-t)G_h}\eta - e^{(T-t)G}\eta\big|_K\big)^2\\ &\le \mathbb{E}\sup_{t\in[0,T]}\Big(\mathbb{E}^{\mathcal{F}_t}\sup_{t\in[0,T]}\big|e^{(T-t)G_h}\eta - e^{(T-t)G}\eta\big|_K\Big)^2\\ &\le 4\mathbb{E}\Big(\sup_{t\in[0,T]}\big|\big[e^{(T-t)G_h}\eta - e^{(T-t)G}\eta\big]\big|_K\Big)^2 \le 4\mathbb{E}\Big(\sup_{t\in[0,T]}\big|\big[e^{(T-t)G_h}\eta - e^{(T-t)G}\eta\big]\big|^2_K\Big).\end{aligned}$$

Similarly, for the second,

$$\begin{aligned}
&\mathbb{E}\sup_{t\in[0,T]}\left|\mathbb{E}^{\mathcal{F}_t}\left[\int_t^T e^{(s-t)G_h}F(s)-e^{(s-t)G}F(s)\,ds\right]\right|_K^2\\
&\quad\le \mathbb{E}\sup_{t\in[0,T]}\left(\mathbb{E}^{\mathcal{F}_t}\int_t^T |e^{(s-t)G_h}F(s)-e^{(s-t)G}F(s)|_K\,ds\right)^2\\
&\quad\le \mathbb{E}\sup_{t\in[0,T]}\left(\mathbb{E}^{\mathcal{F}_t}\int_0^T \sup_{\sigma\in[0,T]}|e^{\sigma G_h}F(s)-e^{\sigma G}F(s)|_K\,ds\right)^2\\
&\quad\le 4\mathbb{E}\left(\int_0^T \sup_{\sigma\in[0,T]}|e^{\sigma G_h}F(s)-e^{\sigma G}F(s)|_K\,ds\right)^2\\
&\quad\le 4\,T\mathbb{E}\int_0^T \sup_{\sigma\in[0,T]}|e^{\sigma G_h}F(s)-e^{\sigma G}F(s)|_K^2\,ds.
\end{aligned}$$

Therefore, we get that

$$\begin{aligned}
\mathbb{E}\sup_{t\in[0,T]}|Y_h(t)-Y(t)|_K^2 &\le 8\mathbb{E}\left(\sup_{t\in[0,T]}|e^{(T-t)G_h}\eta-e^{(T-t)G}\eta|_K^2\right)\\
&\quad+8T\,\mathbb{E}\int_0^T \sup_{\sigma\in[0,T]}|e^{\sigma G_h}F(s)-e^{\sigma G}F(s)|_K^2\,ds.
\end{aligned}$$

By point (iii) in Hypothesis 4.1 and the dominated convergence theorem we can conclude

$$\mathbb{E}\sup_{t\in[0,T]}|Y_h(t)-Y(t)|_K^2\to 0. \tag{4.18}$$

Part III. *Conclusion.* We let, for $\beta>0$, $\mathbb{K}(\beta)=L^2_{\mathcal{P}}(\Omega,C([0,T];K))\times L^2_{\mathcal{P}}(\Omega\times[0,T];L_2(\Xi;K))$ endowed with the norm (equivalent to the natural one)

$$|(Y,Z)|^2_{\mathbb{K}(\beta)}=\mathbb{E}\sup_{t\in[0,T]}e^{2\beta t}|Y(t)|^2+\mathbb{E}\int_0^T e^{2\beta s}|Z(s)|^2ds.$$

Moreover, we define a map $\Gamma:\mathbb{K}(\beta)\to\mathbb{K}(\beta)$ and a sequence of maps $\Gamma_h:\mathbb{K}(\beta)\to\mathbb{K}(\beta)$, $h\in\mathbb{N}$, letting $\Gamma(\widehat{Y},\widehat{Z})=(Y,Z)$ (respectively, $\Gamma_h(\widehat{Y},\widehat{Z})=(Y_h,Z_h)$), where (Y,Z) (respectively, (Y_h,Z_h)) is the solution of (4.8) (respectively, (4.15)) with $F(s)=\psi(s,\widehat{Y}(s),\widehat{Z}(s))$.

We notice that F belongs to $L^2_{\mathcal{P}}(\Omega\times[0,T];K)$ thus the above definition is justified by part I of the present proof.

Moreover, (4.13) and (4.14) immediately yield the following inequality, holding for all $(\widehat{Y},\widehat{Z})$, $(\widetilde{Y},\widetilde{Z})$ in $\mathbb{K}(\beta)$,

$$|\Gamma(\widehat{Y},\widehat{Z})-\Gamma(\widetilde{Y},\widetilde{Z})|^2_{\mathbb{K}(\beta)}\le\frac{4M_G^2M_\psi^2}{\beta}|(\widehat{Y},\widehat{Z})-(\widetilde{Y},\widetilde{Z})|^2_{\mathbb{K}(\beta)},$$

and an identical formula holds (with the same constant) for Γ_h.

So we can conclude that, for β large enough, Γ and Γ_h are contractions in $\mathbb{K}(\beta)$. Clearly the unique fixed point of Γ (respectively, Γ_h) is the unique mild solution of (4.3) (respectively, (4.4)).

Finally by the parameter depending contraction principle (see [24, Theorem 10.1]), relation (4.7) follows immediately if we prove that for all fixed $(\widehat{Y},\widehat{Z}) \in L^2_{\mathcal{P}}(\Omega, C([0,T];K)) \times L^2_{\mathcal{P}}(\Omega\times[0,T];L_2(\Xi;K))$, letting $(Y,Z)=\Gamma(\widehat{Y},\widehat{Z})$ and $(Y_h,Z_h)=\Gamma_h(\widehat{Y},\widehat{Z})$, then

$$\mathbb{E}\sup_{t\in[0,T]}|Y(s)-Y_h(s)|^2+\mathbb{E}\int_0^T|Z(s)-Z_h(s)|^2ds\to 0 \quad \text{as } h\to\infty.$$

The above relation is an immediate consequence of (4.18) and (4.17), letting $F(s)=\psi(s,\widehat{Y}(s),\widehat{Z}(s))$ in part II of the present proof. □

Remark 4.5. As a byproduct of the previous argument we have the following estimate for the solution (Y,Z) of (4.3):

$$|(Y,Z)|^2_{\mathcal{K}(\beta)}\le\hat{C}\left[e^{2\beta T}\mathbb{E}|P_T|^2_{\Sigma_2(H)}+\frac{1}{\beta}\int_0^T e^{2\beta s}\mathbb{E}|\psi(s,0,0)|^2_{\Sigma_2(H)}ds\right], \tag{4.19}$$

holding for β large enough, depending on T, M_G, M_ψ, and for a suitable constant $\hat{C}$, depending on T, M_G.

To prove it *just* notice that, for β large enough, Γ is a $1/2$ contraction in $\mathcal{K}(\beta)$. Since $(Y,Z)=\lim_{n\to\infty}\Gamma^n(0,0)$ we have $|(Y,Z)|_{\mathcal{K}(\beta)}\le 2|\Gamma(0,0)|_{\mathcal{K}(\beta)}$ and the claim follows by (4.14) and (4.13).

An identical estimate holds (with the same constant) for the solution (Y_h,Z_h) of the approximating equation (4.4).

Remark 4.6. Notice that although the semigroup generated by G is not, in general, a contraction semigroup and $\psi(\cdot,0,0)$ is only in $L^2_{\mathcal{P}}(\Omega\times[0,T];K)$, Y nevertheless has continuous trajectories. This is not true for standard (forward) stochastic differential equations (that is when the initial datum is specified rather the final one). For instance, in Theorem 3.2, if u is in $L^2_{\mathcal{P}}(\Omega\times[0,T];K)$, then y is only mean-square continuous.

The reason for such extra regularity of Y can be found in relation (4.9), at least for the simplified equation. Indeed in (4.9) it is clear that Y can be represented only by conditional expectations and deterministic convolutions. In particular, no stochastic convolution is involved in (4.9).

5. The Riccati equation in the Hilbert–Schmidt case. The natural space in which the deterministic Riccati equation is studied is the space $\Sigma(H)$ that is not an Hilbert space. Thus (see the introduction) we initially consider the Riccati equation in the Hilbert space $\Sigma_2(H)$ of symmetric and Hilbert–Schmidt linear operators in H.

5.1. The Lyapunov equation. We start from the linear part of the Riccati equation. Namely we consider the Lyapunov equation

$$\begin{cases} -dP(t)=(A^*P(t)+P(t)A+A^*_\sharp(t)P(t)+P(t)A_\sharp(t)+L(t))\,dt+Q(t)\,dW(t)\\ \qquad\qquad +\mathrm{Tr}[C^*(t)P(t)C(t)+C^*(t)Q(t)+Q(t)C(t)]\,dt, \quad t\in[0,T],\\ P(T)=P_T, \end{cases} \tag{5.1}$$

where, defining the notation with respect to the basis $\{f_i : i \in \mathbb{N}\}$ of Ξ, for all $P \in \Sigma_2(H)$, and $Q \in L_2(\Xi, \Sigma_2(H))$,

$$\mathrm{Tr}[C^*(t)PC(t) + C^*(t)Q + QC(t)] = \sum_{i=1}^{\infty} [C_i^*(t)PC_i(t) + C_i^*(t)(Qf_i) + (Qf_i)C_i(t)].$$

In order to give a precise definition of the *mild solution* of (5.1) we introduce the family $\{e^{t\mathcal{A}} : t \geq 0\}$ of linear operators $\Sigma(H) \to \Sigma(H)$, letting

$$e^{t\mathcal{A}}X := e^{tA^*}Xe^{tA}, \quad t \geq 0,\ X \in \Sigma(H),$$

We notice that the above family is a semigroup of bounded operators in the sense that

$$e^{t\mathcal{A}}e^{s\mathcal{A}}X = e^{(t+s)\mathcal{A}}X, \quad X \in \Sigma(H),\ t, s \geq 0,$$

but is not necessarily strongly continuous, in $\Sigma(H)$; see also [1, Chapter 1]. On the other hand, if we restrict it to $\Sigma_2(H)$, then it becomes a strongly continuous semigroup. Namely we have the following result (also concerning approximations) that will considerably simplify our work.

LEMMA 5.1. *Under hypothesis* (A1) *the family of linear operators* $\{e^{t\mathcal{A}} : t \geq 0\}$ *is a strongly continuous semigroup of bounded operators in* $\Sigma_2(H)$.

Moreover, for all $X \in L_2(H)$,

$$\lim_{h \to \infty} \sup_{t \in [0,T]} |e^{tA_h^*}Xe^{tA_h} - e^{tA^*}Xe^{tA}|_{L_2(H)} = 0. \tag{5.2}$$

Proof. We prove only continuity for $t = 0$. The proof of continuity in a generic t follows by semigroup law. Moreover, (5.2) is proved by an identical argument. We fix $X \in L_2(H)$ and a basis $\{e_i : i \in \mathbb{N}\}$ in H. Clearly

$$\sum_{i=1}^{\infty} |e^{tA^*}Xe^{tA}e_i - Xe_i|_H^2 \leq 2\sum_{i=1}^{\infty} |e^{tA^*}Xe^{tA}e_i - e^{tA^*}Xe_i|_H^2 + 2\sum_{i=1}^{\infty} |e^{tA^*}Xe_i - Xe_i|_H^2.$$

We have to prove that both the above terms converge to 0 as $t \downarrow 0$. As far as the second is concerned, since

$$|e^{tA^*}Xe_i - Xe_i|_H^2 \leq 2(M_A^2 + 1)|Xe_i|_H^2 \quad \text{and} \quad \sum_{i=1}^{\infty} |Xe_i|_H^2 = |X|_{L_2(H)}^2 < +\infty,$$

the claim follows by the dominated convergence theorem. As far as the first is concerned, we have

$$\sum_{i=1}^{\infty} |e^{tA^*}Xe^{tA}e_i - e^{tA^*}Xe_i|_H^2 \leq M_A^2 \sum_{i=1}^{\infty} |Xe^{tA}e_i - Xe_i|_H^2 = M_A^2 \sum_{i=1}^{\infty} |X^*e^{tA^*}e_i - X^*e_i|_H^2,$$

and the claim follows again by the dominated convergence theorem. ∎

Let us denote by $\mathcal{A} : \mathcal{D}(\mathcal{A}) \subset \Sigma_2(H) \to \Sigma_2(H)$ the infinitesimal generator of the semigroup $\{e^{t\mathcal{A}} : t \geq 0\}$ in $\Sigma_2(H)$. Notice that

$$(\mathcal{A}Xx, y)_H = (Xx, Ay)_H + (XAx, y)_H, \quad X \in \mathcal{D}(\mathcal{A}),\ x, y \in \mathcal{D}(A).$$

We now assume that $P_T \in L^2(\Omega, \mathcal{F}_T; \Sigma_2(H))$ and $L \in L^2_{\mathcal{P}}(\Omega \times [0,T]; \Sigma_2(H))$ and give the following definition of a mild solution (P,Q) of (5.1) with values in Hilbert–Schmidt case. We need also the following approximations to $\mathcal{A}$.

DEFINITION 5.2. *We define a sequence of bounded operators* $\mathcal{A}_n : \Sigma_2(H) \to \Sigma_2(H)$ *as follows:*

$$\mathcal{A}_h X \doteq A_h^* X + X A_h, \quad X \in \Sigma_2(H), \quad h = 1, 2, \dots$$

DEFINITION 5.3. *A* mild solution *of problem* (5.1) *is a pair of processes*

$$(P,Q) \in L^2_{\mathcal{P}}(\Omega, C([0,T]; \Sigma_2(H))) \times L^2_{\mathcal{P}}(\Omega \times [0,T]; L_2(\Xi; \Sigma_2(H)))$$

that verifies, for all $t \in [0,T]$,

$$\begin{aligned} P(t) =& e^{(T-t)A^*} P_T e^{(T-t)A} + \int_t^T e^{(s-t)A^*} \left[L(s) + A_\sharp^*(s) P(s) + P(s) A_\sharp(s) \right] e^{(s-t)A} \, ds \\ &+ \int_t^T e^{(s-t)A^*} \mathrm{Tr}[C^*(s) P(s) C(s) + C^*(s) Q(s) + Q(s) C(s)] e^{(s-t)A} \, ds + \\ &+ \int_t^T e^{(s-t)A^*} Q(s) e^{(s-t)A} \, dW(s) \qquad \mathbb{P}\text{-a.s.} \end{aligned} \tag{5.3}$$

We also introduce the regularized versions of (5.1) corresponding to the ones we have introduced for the state equation. Namely we consider

$$\begin{cases} -dP_h(t) = \left(A_h^* P_h(t) + P_h(t) A_h + A_\sharp^*(t) P_h(t) + P_h(t) A_\sharp(t) + L(t) \right) dt \\ \qquad\quad + Q_h(t) \, dW(t) + \mathrm{Tr}[C^*(t) P_h(t) C(t) + C^*(t) Q_h(t) + Q_h(t) C(t)] \, dt, \quad t \in [0,T], \\ P(T) = P_T, \end{cases} \tag{5.4}$$

where A_h are the Yosida approximants of A. The definition of mild solution for the above equation is obtained from the one corresponding to (5.1) just by replacing A by A_h. Since A_h is bounded, mild solutions are classical solutions, i.e., they satisfy $\mathbb{P}$-a.s. for all $t \in [0,T]$:

$$\begin{aligned} P_h(t) = P_T &+ \int_t^T \left(A_h^* P_h(s) + P_h(s) A_h + A_\sharp^*(s) P_h(s) + P_h(s) A_\sharp(s) + L(s) \right) ds \\ &+ \int_t^T \mathrm{Tr}[C^*(s) P_h(s) C(s) + C^*(s) Q_h(s) + Q_h(s) C(s)] \, ds + \int_t^T Q_h(s) \, dW(s). \end{aligned}$$

THEOREM 5.4. *Assume hypotheses* (A1)–(A3). *Moreover, assume that* $P_T \in L^2_{\mathcal{P}}(\Omega, \mathcal{F}_T; \Sigma_2(H))$ *and* $L \in L^2_{\mathcal{P}}(\Omega \times [0,T]; \Sigma_2(H))$.

Then problem (5.1) *has a unique mild solution* $(P,Q) \in L^2_{\mathcal{P}}(\Omega, C([0,T]; \Sigma_2(H))) \times L^2_{\mathcal{P}}(\Omega \times [0,T]; L_2(\Xi; \Sigma_2(H)))$. *Moreover, for all* $h \in \mathbb{N}$, *problem* (5.4) *has a unique classical solution* $(P_h, Q_h) \in L^2_{\mathcal{P}}(\Omega, C([0,T]; \Sigma_2(H))) \times L^2_{\mathcal{P}}(\Omega \times [0,T]; L_2(\Xi; \Sigma_2(H)))$.

Finally the following stability result holds:

(5.5)
$$\lim_{h\to\infty}\mathbb{E}\Big(\sup_{t\in[0,T]}|P_h(t)-P(t)|^2_{\Sigma_2(H)}\Big)=0,\quad \lim_{h\to\infty}\mathbb{E}\int_0^T|Q_h(s)-Q(s)|^2_{L_2(\Xi;\Sigma_2(H))}ds=0.$$

Proof. The claim is a special case of Theorem 4.4, letting $K=\Sigma_2(H)$, $G=\mathcal{A}$, $G_h=\mathcal{A}_h$, $\eta=P_T$, and defining, for all $P\in\Sigma_2(H)$, $Q\in L_2(\Xi,\Sigma_2(H))$,

$$\psi(s,P,Q)=\mathrm{Tr}[C^*(s)PC(s)+C^*(s)Q+QC(s)]+L(s)+A^*_\sharp(s)P+PA_\sharp(s).$$

We have just to check that in this specific situation Hypothesis 4.1 holds, but this is a direct consequence of hypotheses (A1)–(A3) and of the fact that $L\in L^2_{\mathcal{P}}(\Omega\times[0,T];\Sigma_2(H))$. □

Remark 5.5. Remark 4.5 gives, in the present case, the following estimate for the solution (P,Q) of (5.1):

$$|(P,Q)|^2_{\mathcal{K}}\le\hat{C}\left[e^{2\beta T}\mathbb{E}|P_T|^2_{\Sigma_2(H)}+\frac{1}{\beta}\int_0^T e^{2\beta s}\mathbb{E}|L(s)|^2_{\Sigma_2(H)}ds\right]\tag{5.6}$$

holding for β large enough, depending on T, M_A, $M_{A_\sharp}$, M_C, and for a suitable constant $\hat{C}$, depending on T, M_A, $M_{A_\sharp}$.

An identical estimate holds (with the same constant) for the solution (P_h,Q_h) of the approximating equation (5.4).

The following result is a key step towards the *fundamental relation* (see Proposition 5.11). Moreover, it gives useful estimates on the solution to (5.1).

THEOREM 5.6. *Besides the hypotheses of Theorem* 5.4*, assume that P_T belongs to $L^\infty_S(\Omega,\mathcal{F}_T;L(H))$ and L belongs to $L^1_{\mathcal{P},S}([0,T];L^\infty(\Omega;L(H)))$. Let (P,Q) be the unique mild solution to* (5.1) *and let $y^{t,x,u}$ be the mild solution to* (2.1)*. Then for all $t\in[0,T]$, $x\in H$, $u\in L^2_{\mathcal{P}}(\Omega\times[0,T];U)$ it holds, $\mathbb{P}$-a.s., that*

$$\begin{aligned}(P(t)x,x)=\mathbb{E}^{\mathcal{F}_t}(P_Ty^{t,x,u}(T),y^{t,x,u}(T))+\mathbb{E}^{\mathcal{F}_t}\int_t^T[(L(s)y^{t,x,u}(s),y^{t,x,u}(s))\\ -2(P(s)B(s)u(s),y^{t,x,u}(s))]ds.\end{aligned}\tag{5.7}$$

Moreover, for all $t\in[0,T]$,

$$|P(t)|_{L(H)}\le C_2\left[|P_T|_{L^\infty_S(\Omega;L(H))}+\int_t^T|L(s)|_{L^\infty_S(\Omega;L(H))}\,ds\right]\qquad\mathbb{P}\text{-a.s.},\tag{5.8}$$

where C_2 is the positive constant depending only on $T,M_B,M_C,M_A,M_{A_\sharp}$ defined in (3.1)*.*

Similarly if, for all $h\in\mathbb{N}$, (P_h,Q_h) is the unique solution of problem (5.4) *and for all $t\in[0,T]$,*

$$|P_h(t)|_{L(H)}\le C_2\left[|P_T|_{L^\infty_S(\Omega;L(H))}+\int_t^T|L(s)|_{L^\infty_S(\Omega;L(H))}\,ds\right]\qquad\mathbb{P}\text{-a.s.}$$

Proof. The proof will be concluded in three steps. In the first we will prove (5.7) for $u\in L^6_{\mathcal{P}}(\Omega\times[0,T];U)$, then we will prove estimate (5.8), and finally we will extend (5.7) to all the admissible controls.

First step. The following argument is simple but has some delicate points; thus we expose it here in all details. Let $y_h = y_h^{t,x,u}$ be the classical solution to (3.3). By Theorem 3.3 we know that $y_h \in L^6_{\mathcal{P}}(\Omega, C([t,T];H))$ and $y_h \to y^{t,x,u}$ in $L^6_{\mathcal{P}}(\Omega, C([t,T];H))$ as $h \to \infty$.

Let $\Psi \in C^2(H)$ with $\Psi(y) = 1$ for $|y| \le 1$, $\Psi(y) = 0$ for $|y| \ge 2$ and $\Psi(y) \in [0,1] \forall y \in H$. Differentiating by the Itô rule we obtain (we consider $\Psi' \in H$, $\Psi'' \in L(H)$)

(5.9)
$$\begin{aligned} d_s[\Psi\left(y_h(s)/N\right)(P_h(s)y_h(s), y_h(s))] &= N^{-1}F_N(s)ds + G_N(s)dW_s \\ &\quad -\Psi(y_h(s)/N)[(L(s)y_h(s), y_h(s))_H - 2(P_h(s)B(s)u(s), y_h(s))_H]ds, \end{aligned}$$

where

$$\begin{aligned} F_N(s) &= \big(\Psi'(N^{-1}y_h(s)), [A_h y_h(s) + A_\sharp(s)y_h(s) \\ &\quad + B(s)u(s)]\big)_H (P_h(s)y_h(s), y_h(s))_H \\ &\quad + 2\sum_{i=1}^{\infty} \big(\Psi'(N^{-1}y_h(s)), C_i(s)y_h(s)\big)_H (P_h(s)C_i(s)y_h(s), y_h(s))_H \\ &\quad + \frac{1}{2N}\sum_{i=1}^{\infty} \big(\Psi''(N^{-1}y_h(s))C_i(s)y_h(s), C_i(s)y_h(s)\big)_H (P_h y_h(s), y_h(s))_H \\ &\quad + \sum_{i=1}^{\infty} \big(\Psi'(N^{-1}y_h(s)), C_i(s)y_h(s)\big)_H (Q_h^i y_h(s), y_h(s))_H, \end{aligned}$$

$$\begin{aligned} G_N(s)f_i &= 2\Psi(N^{-1}y_h(s))(P_h(s)C_i(s)y_h(s), y_h(s))_H \\ &\quad - \Psi(N^{-1}y_h(s))(Q_h^i(s)y_h(s), y_h(s))_H \\ &\quad + \frac{1}{N}(P_h(s)y_h(s), y_h(s))_H (\Psi'(N^{-1}y_h(s)), C_i(s)y_h(s))_H, \end{aligned}$$

where $Q_h^i = Q_h f_i$, with $\{f_i\}_{i\in\mathbb{N}}$ an orthonormal basis of Ξ.

As it can be easily verified $\mathbb{E}\int_t^T |F_N(s)|ds \le$ const. for all $N \in \mathbb{N}$. Moreover, since $\Psi(N^{-1}y) = 0$ and $\Psi'(N^{-1}y) = 0$ if $|y| > 2N$ we have, for all $N \in \mathbb{N}$,

$$\begin{aligned} \sum_{i=1}^{\infty} \mathbb{E}\int_t^T |G_N(s)f_i|_H^2 ds \le c_2 N^4 \Big\{ M_C^2 \mathbb{E}\int_t^T \sup_{s\in[t,T]} \|P_h(s)\|^2_{L(H)}\, ds \\ + \mathbb{E}\int_t^T \|Q_h(s)\|^2_{\Sigma_2(H)}\, ds \Big\} < +\infty, \end{aligned}$$

where c_2 is a positive universal constant. Finally $(Ly_h, y_h)_H$ and $(P_h Bu, y_h)$ belong to $L^1_{\mathcal{P}}(\Omega \times [t,T], \mathbb{R})$ and $\Psi(y_h(s)/N)$ converges to 1 $\mathbb{P}$-a.s. for all s.

Thus, first integrating in $[t,T]$ and then computing conditional expectation with respect to $\mathcal{F}_t$ ($\mathbb{E}^{\mathcal{F}_t}$) and finally letting $N \to 0$, relation (5.9) becomes

$$\begin{aligned} (P_h(t)x, x)_H = \mathbb{E}^{\mathcal{F}_t}(P_T y_h(T), y_h(T))_H + \mathbb{E}^{\mathcal{F}_t}\int_t^T \big[(L(s)y_h(s), y_h(s))_H \\ -2(P_h(s)B(s)u(s), y_h(s))_H\big]ds \end{aligned}$$

and the claim follows, letting $h \to +\infty$ thanks to (3.3) and (5.4).

Second step. The following L^∞ bound for the $L(H)$ norm of the mild solution to (5.1) will be important in the approach to the (nonlinear) Riccati equation. In the finite dimensional case a similar result is proved by a slightly different argument in [19]. Here is our proof. From the *first step* we know that for all $x \in H$, $\mathbb{P}$-a.s.

$$(P(t)x,x) = \mathbb{E}^{\mathcal{F}_t}(P_T y^{t,x,0}(T), y^{t,x,0}(T)) - \mathbb{E}^{\mathcal{F}_t}\int_t^T (L(s)y^{t,x,0}(s), y^{t,x,0}(s))ds; \tag{5.10}$$

consequently

$$|(P(t)x,x)| \le |P_T|_{L^\infty_{\mathcal{S}}(\Omega;L(H))}\mathbb{E}^{\mathcal{F}_t}|y^{t,x,0}(T)|^2 + \int_t^T |L(s)|_{L^\infty_{\mathcal{S}}(\Omega;L(H))}\mathbb{E}^{\mathcal{F}_t}|y^{t,x,0}(s)|^2 ds,$$

and by estimate (3.1) for $u \equiv 0$, with $\mathbb{E}$ replaced by $\mathbb{E}^{\mathcal{F}_t}$,

$$|(P(t)x,x)| \le C_2|P_T|_{L^\infty_{\mathcal{S}}(\Omega;L(H))} + C_2\int_t^T |L(s)|_{L^\infty_{\mathcal{S}}(\Omega;L(H))}ds \quad \forall x \in H,\ |x| \le 1,\ \mathbb{P}\text{-a.s.},$$

and the claim holds, H being separable. The same estimate holds true also for every $|(P_h(t)x,x)|$, since the constant C_2 does not depend on h.

Third step. For a general $u \in L^2_{\mathcal{P}}(\Omega\times[0,T];U)$ we choose a sequence $u_m \to u$ such that u_m is bounded and $u_m \to u$ in $L^2_{\mathcal{P}}(\Omega\times[0,T];U)$. By Theorem 3.3, $y^{t,x,u_m} \to y^{t,x,u}$ in $C_{\mathcal{P}}([t,T];L^2(\Omega;H))$ and, by the *second step*, $P \in L^\infty_{\mathcal{P},\mathcal{S}}(\Omega\times[0,T];L(H))$. Moreover,

$$\begin{aligned}
&\left|\mathbb{E}^{\mathcal{F}_t}\int_t^T (L(s)y^{t,x,u_m}(s), y^{t,x,u_m}(s))_H - (L(s)y^{t,x,u}(s), y^{t,x,u}(s))_H\, ds\right| \\
&\quad \le \Big[\Big(\sup_{s\in[t,T]}\mathbb{E}^{\mathcal{F}_t}|y^{t,x,u_m}(s)|^2\Big)^{1/2} + \Big(\sup_{s\in[t,T]}\mathbb{E}^{\mathcal{F}_t}|y^{t,x,u}(s)|^2\Big)^{1/2}\Big] \\
&\quad \times \Big(\sup_{s\in[t,T]}\mathbb{E}^{\mathcal{F}_t}|y^{t,x,u_m}(s) - y^{t,x,u}(s)|^2\Big)^{1/2}\int_t^T |L(s)|_{L^\infty_{\mathcal{S}}(\Omega,H)}\, ds.
\end{aligned}$$

Thus we can pass relation (5.7) to the limit as $m \to \infty$ obtaining the claim. ∎

5.2. Existence of a unique solution for the Riccati equation and the synthesis of the optimal control. In this section we prove the existence of a unique mild solution for the Riccati equation

$$\begin{cases}
-dP(t) = (A^*P(t) + P(t)A + \mathrm{Tr}[C^*(t)P(t)C(t) + C^*(t)Q(t) + Q(t)C(t)])\, dt \\
\quad -(P(t)B(t)B^*(t)P(t) - A^*_\sharp(t)P(t) - P(t)A_\sharp(t) - S(t))\, dt + Q(t)\, dW(t), \qquad t \in [0,T], \\
P(T) = P_T
\end{cases} \tag{5.11}$$

under assumptions (A1)–(A5).

The presence of a quadratic nonlinear term imposes the following approach (classical when dealing with the Riccati equation; see [22]) in solving the problem: first we will find a local solution and then we will prove some a priori estimate for the solution to guarantee the existence of a global solution. The method we use to prove the a priori bound is based on the so called *fundamental relation* (see Proposition 5.11) and uses, in an essential way, the control-theoretic interpretation of the Riccati equation.

We start extending the notion of *mild* solution given in section 5.1.

DEFINITION 5.7. *Fix $T_0 \in [0,T]$. A mild solution for problem* (5.11)*, considered in $[T_0,T]$ is a pair (P,Q) with*

$$P \in L^2_{\mathcal{P}}(\Omega, C([T_0,T];\Sigma_2(H))) \cap L^\infty_{\mathcal{P},S}(\Omega; C([T_0,T];\Sigma^+(H))),$$
$$Q \in L^2_{\mathcal{P}}(\Omega \times (T_0,T); L_2(\Xi;\Sigma_2(H)))$$

such that for all $t \in [T_0,T]$

(5.12)

$$\begin{aligned} P(t) &= \int_t^T e^{(s-t)A^*}\mathrm{Tr}[C^*(s)P(s)C(s) + C^*(s)Q(s) + Q(s)C(s)]e^{(s-t)A}\,ds \\ &+e^{(T-t)A^*}P_T e^{(T-t)A} + \int_t^T e^{(s-t)A^*}[S(s) + A^*_\sharp(s)P(s) + P(s)A_\sharp(s)]e^{(s-t)A}\,ds \\ &+\int_t^T e^{(s-t)A^*}Q(s)e^{(s-t)A}dW(s) - \int_t^T e^{(s-t)A^*}P(s)B(s)B^*(s)P(s)e^{(s-t)A}\,ds \quad \mathbb{P}\text{-a.s.} \end{aligned}$$

PROPOSITION 5.8 (local existence). *Under Hypotheses* (A1)–(A5) *there exists a $\delta \in]0,T]$ such that problem* (5.11) *has a unique mild solution in the interval $[T-\delta,T]$.*

Proof. To simplify the notation we will set

$$|P_T|_{L^\infty_S(\Omega;L(H))} = M_P, \qquad |S|_{L^1_{\mathcal{P},S}([0,T];L^\infty(\Omega;L(H)))} = M_S.$$

We choose $r > C_2(M_P + M_S)$ and δ such that $C_2[M_P + r^2\delta M_B^2 + M_S] \le r$.

We define

$$B(r) = \Big\{P \in L^2(\Omega; C([T-\delta,T];\Sigma_2(H))) : \sup_{t\in[T-\delta,T]} |P(t,\omega)|_{L(H)} \le r \quad \mathbb{P}\text{-a.e.}\Big\}$$

endowed with the norm

$$|P|^2_\beta = \mathbb{E}\sup_{t\in[T-\delta,T]} e^{2\beta t}|P(t)|^2_{\Sigma_2(H)}.$$

On $B(r)$ we construct the map $\Lambda : B(r) \to B(r)$, letting $\Lambda(K) = P$, where (P,Q) is the unique solution to (5.1) (in $[T-\delta,T]$) with $L = -KBB^*K$; that is,

$$\begin{aligned} P(t) &= \int_t^T e^{(s-t)A^*}\mathrm{Tr}[C^*(s)P(s)C(s) + C^*(s)Q(s) + Q(s)C(s)]e^{(s-t)A}\,ds \\ &+ \int_t^T e^{(s-t)A^*}S(s)e^{(s-t)A}\,ds + e^{(T-t)A^*}P_T e^{(T-t)A} \\ &+ \int_t^T e^{(s-t)A^*}Q(s)e^{(s-t)A}\,dW(s) \\ &- \int_t^T e^{(s-t)A^*}\left[A^*_\sharp(s)P(s) + P(s)A_\sharp(s) - K(s)B(s)B^*(s)K(s)\right]e^{(s-t)A}\,ds. \end{aligned}$$

We claim that the map Λ is a contraction in $B(r)$.

First of all we check that it maps $B(r)$ into itself. By Theorem 5.4 (applied in $[T-\delta,T]$) we know that $\Lambda(K)\in L^2_{\mathcal{P}}(\Omega\times[T-\delta,T];\Sigma_2(H))$. So it is enough to show that for all $t\in[T-\delta,T]$ it holds $|\Lambda(K)(t)|_{L(H)}\le r$ $\mathbb{P}$-a.s. Thanks to (5.8) we have that $\mathbb{P}$-a.s.

$$|\Lambda(K)(t)|_{L(H)}\le C_2\bigg[|P_T|_{L^\infty_S(\Omega,L(H))}+\int_{T-\delta}^T(|K(s)B(s)B^*(s)K(s)|_{L^\infty_S(\Omega,L(H))}$$
$$+|S(s)|_{L^\infty_S(\Omega,L(H))})\,ds\bigg]\le C_2[M_P+r^2\delta M_B^2+M_S]\le r.$$

Moreover, by (4.19) for all K_1 and K_2 in $B(r)$ (since (4.19) is stated in the whole $[0,T]$ we should, to be precise, extend $K_1(s)=K_2(s)=0$ for $s<T-\delta$)

$$|\Lambda(K_2)-\Lambda(K_1)|^2_\beta\le\frac{\hat C}{\beta}\int_{T-\delta}^T e^{2\beta s}\mathbb{E}|K_2BB^*K_2-K_1BB^*K_1|^2_{\Sigma_2(H)}ds.$$

Since $|K_i|_{\Sigma_2(H)}\le r$, $i=1,2$, $\mathbb{P}$-a.s. for all $t\in[T-\delta,T]$ the above relation gives

$$|\Lambda(K_2)-\Lambda(K_1)|^2_\beta\le\frac{\hat C}{\beta}r^2M_B^4\int_{T-\delta}^T e^{2\beta s}\mathbb{E}|K_2-K_1|^2_{\Sigma_2(H)}ds.$$

Therefore, if β is large enough, Λ is a contraction in $B(r)$. If P is its unique fixed point, the mild solution (P,Q) of (5.1) with $L=-PBB^*P$ is the unique mild solution of (5.11). □

Clearly local uniqueness of the solution immediately implies global uniqueness.

COROLLARY 5.9 (global uniqueness). *Let (P_i,Q_i), $i=1,2$, be two mild solutions of the Riccati equation* (5.11) *in the interval $[T_0,T]$ for some $T_0\in[0,T)$. Then $P_1(t)=P_2(t)$, $\mathbb{P}$-a.s. for all $t\in[T_0,T]$ and $Q_1(t)=Q_2(t)$, $\mathbb{P}$-a.s. for almost all $t\in[T_0,T]$.*

Remark 5.10. The length δ of the interval on which the mild solution of the Riccati equation exists depends only on T, M_A, $M_{A_\sharp}$, M_B, M_C, $|S|_{L^1_{\mathcal{P},S}([0,T];L^\infty(\Omega;L(H))}$ and $|P_T|_{L^\infty_S(\Omega;L(H))}$. Thus to extend the solution to the whole $[0,T]$ it is sufficient to establish an a priori bound for the $L^\infty(\Omega;L(H))$ norm of the P part of any local solution, independently on the length of the interval in which it is defined.

This will be done using the following consequence of Theorem 5.6. The next relation also has an obvious control-theoretic interpretation and will be essential in performing the synthesis of the optimal control.

PROPOSITION 5.11 (fundamental relation). *Assume* (A1)–(A5) *and let (P,Q) be the mild solution of* (5.11) *in an interval $[T_0,T]$. Then, for all $t\ge T_0$, $x\in H$ $u\in L^2_{\mathcal{P}}(\Omega\times[t,T];U)$ it holds*

$$(P(t)x,x)_H=J(t,x,u)-\mathbb{E}^{\mathcal{F}_t}\int_t^T|u(s)+B^*P(s)y^{t,x,u}(s)|^2_H\,ds. \tag{5.13}$$

Proof. We start by noticing that, by definition, (P,Q) is a mild solution of the Lyapunov equation (5.1) with $L=-PBB^*P+S$. Thus by (5.7)

$$(P(t)x,x)=\mathbb{E}^{\mathcal{F}_t}(P_Ty^{t,x,u}(T),y^{t,x,u}(T))+\mathbb{E}^{\mathcal{F}_t}\int_t^T|\sqrt{S}(s)y^{t,x,u}(s)|^2_Hds$$
$$-\mathbb{E}^{\mathcal{F}_t}\int_t^T\big[(y^{t,x,u}(s),[P(s)B(s)B^*(s)P(s)+S(s)]y^{t,x,u}(s))_H$$
$$+2(P(s)B(s)u(s),y^{t,x,u}(s))_H\big]ds.$$

Then the claim follows just adding and subtracting $\mathbb{E}^{\mathcal{F}_t}\int_t^T |u(s)|_U^2 ds$. □

PROPOSITION 5.12 (positivity and a priori estimate). *Let (P,Q) be the mild solution to (5.11) in $[T_0,T]$; then*

1. *for every $t\in[T_0,T]$ and $x\in H$, $(P(t)x,x)_H\geq 0$ $\mathbb{P}$-a.s.;*
2. *for every $t\in[T_0,T]$, $|P(t)|_{L(H)}\leq C_2[M_P+M_S]$ $\mathbb{P}$-a.s,*

where C_2 is the constant defined in Theorem 3.2.

Proof. If we apply (5.13) to $u\equiv 0$, we obtain for all $x\in H$ with $|x|_H\leq 1$ and for all $t\in[T_0,T]$

$$\begin{aligned}(P(t)x,x)_H &= \mathbb{E}^{\mathcal{F}_t}(P_T y^{t,x,0}(T), y^{t,x,0}(T))_H + \mathbb{E}^{\mathcal{F}_t}\int_t^T |\sqrt{S}(s)y^{t,x,0}(s)|_H^2\,ds\\ &\leq |P_T|_{L_S^\infty(\Omega,L(H))}\mathbb{E}^{\mathcal{F}_t}|y^{t,x,0}(T)|_H^2 + \int_t^T |S(s)|_{L_S^\infty(\Omega,L(H))}\mathbb{E}^{\mathcal{F}_t}|y^{t,x,0}(s)|_H^2\,ds\end{aligned}$$

and by (3.1)

(5.14)
$$(P(t)x,x)_H \leq C_2[|P_T|_{L_S^\infty(\Omega;L(H))} + |S|_{L^1_{\mathcal{P},S}([0,T];L^\infty(\Omega;L(H)))}] \quad \mathbb{P}\text{-a.s.} \quad \forall x: |x|_H\leq 1.$$

Then consider the following *closed loop* equation, starting at a certain instant $t\geq T_0$ with an arbitrary initial data $x\in H$:

$$\begin{cases} d\overline{y}(r) = [A\overline{y}(r) + A_\sharp(r)\overline{y}(r) - B(r)B^*(r)P(r)\overline{y}(r)]\,dr + C(r)\overline{y}(r)\,dW(r),\\ \overline{y}(t) = x.\end{cases} \tag{5.15}$$

Notice that if we replace $A_\sharp$ by $A_\sharp - BB^*P$, then assumptions of Theorem 3.2 still hold. Thus there exists a unique solution $\overline{y}\in L^p_{\mathcal{P}}(\Omega, C([t,T];H))$ for every $p\geq 2$. Applying then the fundamental relation (5.13) to $\overline{u} = -B^*P\overline{y}$ and consequently to $y^{t,x,\overline{u}} = \overline{y}$ we get

(5.16)
$$(P(t)x,x)_H = \mathbb{E}^{\mathcal{F}_t}(P_T\overline{y}(T),\overline{y}(T))_H + \mathbb{E}^{\mathcal{F}_t}\int_t^T [|\sqrt{S}(r)\overline{y}(r)|_H^2 + |B^*(r)P(r)\overline{y}(r)|_H^2]\,dr;$$

thus $(P(t)x,x)_H\geq 0$, $\mathbb{P}$-a.s. for all $x\in H$, and this together with (5.14) gives the claim. □

We summarize the content of the section in the following result.

THEOREM 5.13. *Assume* (A1)–(A5). *Problem* (5.11) *has a unique mild solution (P,Q) with the following regularity: $P\in L^2_{\mathcal{P}}(\Omega;C([0,T];\Sigma_2^+(H)))\cap L^\infty_{\mathcal{P},S}(\Omega;C([0,T];\Sigma^+(H)))$ and $Q\in L^2_{\mathcal{P}}(\Omega\times[0,T];L_2(\Xi;\Sigma_2(H)))$.*

Proof. The a priori estimate in Proposition 5.12 allows us to apply the local existence result in Proposition 5.8 recursively in time intervals of fixed length (see also Remark 5.10) to obtain a global solution of (5.11). Indeed, let $\tilde{M}_P = C_2(M_P+M_S)$; then it is enough to choose $\tilde{r}$ such that $\tilde{r} > C_2(\tilde{M}_P + M_S)$ and $\tilde{\delta}$ such that

$$C_2[\tilde{M}_P + \tilde{r}^2\tilde{\delta}M_B^2 + M_S]\leq \tilde{r}.$$

Then we can iterate the procedure in $[T-n\tilde{\delta}, T-(n-1)\tilde{\delta}]$ for a finite number of $n\geq 1$ until we cover the whole interval $[0,T]$. □

Now we are ready to solve the finite horizon problem in a standard way.

THEOREM 5.14. *Fix $T>0$ and $x\in H$. Then we have the following:*

1. *There exists a unique* optimal control. *That is a unique control* $\bar{u} \in L^2_{\mathcal{P}}(\Omega \times [0,T]; U)$ *such that*

$$J(0,x,\bar{u}) = \inf_{u \in L^2_{\mathcal{P}}(\Omega\times[0,T];U)} J(0,x,u).$$

2. *If* $\bar{y}$ *is the mild solution of the state equation corresponding to* $\bar{u}$ *(that is the* optimal state*), then* $\bar{y}$ *satisfies the* closed loop *equation*

(5.17)
$$\begin{cases} d_s\bar{y}(s) = [A\bar{y}(s) + A_\sharp(s)\bar{y}(s) - B(s)B(s)^*P(s)\bar{y}(s)]\,ds + C(s)\bar{y}(s)\,dW(s), \\ \bar{y}(0) = x. \end{cases}$$

3. *The following feedback law holds* $\mathbb{P}$*-a.s. for almost every* s*:*

$$\bar{u}(s) = -B^*(s)P(s)\bar{y}(s). \tag{5.18}$$

4. *The optimal cost is given by* $\mathbb{E}J(0,x,\bar{u}) = \mathbb{E}(P(0)x,x)_H$ *for all* $x \in H$.

Proof. Let (P,Q) be the unique mild solution to Riccati equation (5.11). Relation (5.13) becomes

$$J(0,x,u) = (P(0)x,x)_H + \mathbb{E}\int_0^T |u(s) + B^*P(s)y^{t,x,u}(s)|^2\,ds.$$

Thus $J(0,x,u) \geq (P(0)x,x)_H$ for all $u \in L^2_{\mathcal{P}}(\Omega \times [0,T]; U)$ and the equality holds if and only if (5.18) holds, that is, if and only if y solves (5.17) and $u = \bar{u}$. □

6. The general case. In order to get rid of assumption (A5) we introduce the following new notion of solution.

DEFINITION 6.1. *A process* $P \in L^\infty_{\mathcal{P},S}(\Omega\times[0,T];\Sigma^+(H))$ *is a* generalized solution *if there exists a sequence* (S^N,P^N,Q^N) *where*

(i) $S^N \in L^1_{\mathcal{P},S}([0,T];L^\infty(\Omega;\Sigma^+(H))) \cap L^2_{\mathcal{P}}(\Omega\times[0,T];\Sigma_2(H))$ *and there exists a positive function* $c \in L^1([0,T])$ *such that* $|S^N(t)|_{L(H)} \leq c(t)$*, for all* $N \in \mathbb{N}$*,* $\mathbb{P}$*-a.s. for a.e.* $t \in [0,T]$;

(ii) *the pair* (P^N,Q^N) *is a mild solution to the Riccati equation* (5.11) *in the space of Hilbert–Schmidt operators, with forcing term* S^N *and final data* $P^N_T = P^N(T)$*. Namely* (P^N,Q^N) *is the unique mild solution of*

$$\begin{cases} -dP^N(t) = (A^*P^N(t) + P^N(t)A + \mathrm{Tr}[C^*(t)P^N(t)C(t) + C^*(t)Q^N(t) \\ \qquad\qquad +Q^N(t)C(t)])\,dt + (A^*_\sharp(t)P^N(t) + P^N(t)A_\sharp(t) \\ \qquad\qquad - P^N(t)B(t)B^*(t)P^N(t) + S^N(t))\,dt + Q^N(t)\,dW(t), \quad t \in [0,T], \\ P^N(T) = P^N_T \end{cases}$$

such that

(iii) *for all* $x \in H$*,*

$$S^N(t,\omega)x \to S(t,\omega)x \text{ in } H \quad \mathbb{P} \text{ a.s. for a.e. } t \in [0,T];$$

(iv) *for every* $t \in [0,T]$ *and for all* $x \in H$*,*

$$P^N(t,\omega)x \to P(t,\omega)x \quad \text{in } H \qquad \mathbb{P} \text{ a.s.}$$

Remark 6.2. Although in the definition the value of $P(t)$ seem determined for a.e. t, point (iv) in the definition implies that there exists a version such that for all $t \in [0,T]$ and for all $x \in H$ the value of $P(t)x$ is determined $\mathbb{P}$-a.s. Actually we will show extra regularity property for the generalized solution if evaluated at a vector $x \in H$.

Remark 6.3. In the previous definition only the process P in the Riccati equation is characterized. On one hand this is natural for control theory since Q is not involved in the expression for the optimal cost or in the expression for the optimal feedback law (see Theorem 6.6). On the other hand it is a general feature of backward stochastic differential equations that the martingale representation term is only an auxiliary variable that can be determined computing the joint quadratic variation between the other unknown process and the noise; see [18]. We start by showing some regularity properties of the generalized solutions.

LEMMA 6.4. *Every generalized solution fulfills the fundamental relation, i.e., for all* $t \in [0,T]$, *for all* $x \in H$, *and for all* $u \in L^2_{\mathcal{P}}(\Omega \times [0,T]; U)$,

$$(P(t)x,x)_H = J(t,x,u) - \mathbb{E}^{\mathcal{F}_t} \int_t^T |u(s) + B^*(s)P(s)y(s)|^2 \, ds \qquad \mathbb{P}\text{- a.s.} \tag{6.1}$$

Proof. At each fixed N the pair (P^N, Q^N) is the mild solution of the Riccati equation; therefore, by Proposition 5.11, we have that for all $t \in [0,T]$ and $x \in H$:

$$(P^N(t)x,x)_H = \mathbb{E}^{\mathcal{F}_t}(P^N_T y(T), y(T))_H + \mathbb{E}^{\mathcal{F}_t} \int_t^T [|\sqrt{S^N}(s)y(s)|^2 + |u(s)|^2] \, ds$$
$$- \mathbb{E}^{\mathcal{F}_t} \int_t^T |u(s) + B^*(s)P^N(s)y(s)|^2 \, ds \qquad \mathbb{P}\text{-a.s.}$$

Now, we have to pass to the limit as $N \to \infty$ in the identity. We notice that if we show that the right-hand side converges in mean to

$$\mathbb{E}^{\mathcal{F}_t}(P_T y(T), y(T))_H + \mathbb{E}^{\mathcal{F}_t} \int_t^T [|\sqrt{S}(s)y(s)|^2 + |u(s)|^2] \, ds - \mathbb{E}^{\mathcal{F}_t} \int_t^T |u(s) + B^*(s)P(s)y(s)|^2 \, ds,$$

then the proof is completed just by choosing a subsequence on which convergence occurs $\mathbb{P}$ a.s.

Coming now to the proof of convergence, considering for instance the second term, by Jensen inequality it is enough to show that

$$\lim_{N\to\infty} \mathbb{E}\left| \int_t^T ((S - S^N)(s)y(s), y(s))_H ds \right| = 0.$$

Applying a first time dominated convergence theorem, we get $\mathbb{E}|((S - S^N)(t)y(t), y(t))_H| \to 0$ for all fixed $s \in [t,T]$. Then we notice that

$$\mathbb{E}|((S - S^N)(s)y(s), y(s))_H| \le 2c(s)\mathbb{E}|y(s)|^2,$$

where the map c is in $L^1([0,T])$ and the map $s \to \mathbb{E}|y(s)|^2$ is in $C([0,T];\mathbb{R})$. Thus we can apply a second time dominated convergence theorem to obtain the claim. Since the other terms can be treated in an identical way, the proof is completed. □

LEMMA 6.5. *Let $P(t)$ be any generalized solution. Then $P(t)x \in C_{\mathcal{P}}([0,T];L^p(\Omega;H))$ for all $x \in H$ and for all $p \geq 2$.*

Proof. We have to prove that, for all $x \in H$, $\lim_{r\to t} \mathbb{E}|[P(r)-P(t)]x|_H^p = 0$ or, equivalently,

$$\lim_{r\to t} \mathbb{E}|(P(r)x,x)_H - (P(t)x,x)_H|^p = 0 \quad \forall x \in H.$$

Let us consider the state equation corresponding to $u = 0$. Then for all $x \in H$ the following holds $\mathbb{P}$-a.s.:

$$(P(t)x,x)_H = \mathbb{E}^{\mathcal{F}_t}(P_T y(T), y(T))_H + \mathbb{E}^{\mathcal{F}_t}\int_t^T |\sqrt{S}(s)y(s)|^2\,ds$$
$$-\mathbb{E}^{\mathcal{F}_t}\int_t^T |B^*(s)P(s)y(s)|^2\,ds.$$

We set $y^{t,x} = y^{t,x,0}$ and we recall that $y^{t,x} \in L^p_{\mathcal{P}}(\Omega;C([0,T];H))$ for all $p \geq 2$ (see [7, Proposition 3.2]). Moreover, the map $(t,x) \to y^{t,x}(\cdot)$ is continuous from $[0,T]\times H$ to $L^p_{\mathcal{P}}(\Omega;C([0,T];H))$; again see [7, Proposition 3.3].

Taking all these facts into account, we have that, for all $0 \leq t \leq r \leq T$,

(6.2)

$$\begin{aligned}\mathbb{E}|(P(r)x,x)_H - (P(t)x,x)_H|^p &\leq c(p)\mathbb{E}|\mathbb{E}^{\mathcal{F}_r}(P_T y(T),y(T))_H - \mathbb{E}^{\mathcal{F}_t}(P_T y(T),y(T))_H|^p\\
&+ \mathbb{E}|\mathbb{E}^{\mathcal{F}_r}\int_r^T |\sqrt{S(s)}y^{r,x}(s)|_H^2\,ds - \mathbb{E}^{\mathcal{F}_t}\int_t^T |\sqrt{S(s)}y^{t,x}(s)|_H^2\,ds|^p\\
&+ \mathbb{E}|\mathbb{E}^{\mathcal{F}_r}\int_r^T |B^*(s)P(s)y^{r,x}(s)|_H^2\,ds - \mathbb{E}^{\mathcal{F}_t}\int_t^T |B^*(s)P(s)^{t,x}y(s)|_H^2\,ds|^p.\end{aligned}$$

Since $P_T \in L^\infty_{\mathcal{S}}(\Omega,\mathcal{F}_T,\mathbb{P},L(H))$ and consequently $(P_T y(T),y(T))_H \in L^p(\Omega,\mathcal{F}_T,\mathbb{P},\mathbb{R})$, for all $p \in [2,\infty[$ by the Kunita–Watanabe martingale representation theorem there exists a process Z in $L^p_{\mathcal{P}}(\Omega;L^2([0,T];\Xi^*))$ such that

$$\left|\mathbb{E}^{\mathcal{F}_r}(P_T y(T),y(T))_H - \mathbb{E}^{\mathcal{F}_t}(P_T y(T),y(T))_H\right| = \left|\int_r^t Z(s)\,dW(s)\right|.$$

Now fix $\tau \in [0,T]$; by Burkholder–Davies–Gundy inequalities and the dominated convergence theorem we get that

$$\begin{aligned}(6.3)\qquad &\lim_{r\uparrow\tau,\,t\downarrow\tau} \mathbb{E}|\mathbb{E}^{\mathcal{F}_r}(P_T y(T),y(T))_H - \mathbb{E}^{\mathcal{F}_t}(P_T y(T),y(T))_H|^p\\
&\quad = \lim_{r\downarrow\tau,\,t\uparrow\tau} \mathbb{E}\left|\int_t^r Z(s)\,dW(s)\right|^p \leq \lim_{r\downarrow\tau,\,t\uparrow\tau} \mathbb{E}\Big(\int_t^r |Z(s)|_{\Xi^*}^2\,ds\Big)^{p/2} = 0.\end{aligned}$$

Let us consider the second term on the right-hand side in (6.2):

$$\begin{aligned}(6.4)\qquad &\lim_{r\downarrow\tau,\,t\uparrow\tau} \mathbb{E}|\mathbb{E}^{\mathcal{F}_r}\int_r^T |\sqrt{S(s)}y^{r,x}(s)|_H^2\,ds - \mathbb{E}^{\mathcal{F}_t}\int_t^T |\sqrt{S(s)}y^{t,x}(s)|_H^2\,ds|^p\\
&\leq \lim_{r\downarrow\tau,\,t\uparrow\tau} \mathbb{E}|\mathbb{E}^{\mathcal{F}_r}\int_r^T |\sqrt{S(s)}y^{r,x}(s)|_H^2\,ds - \mathbb{E}^{\mathcal{F}_r}\int_t^T |\sqrt{S(s)}y^{t,x}(s)|_H^2\,ds|^p\\
&+ \lim_{r\downarrow\tau,\,t\uparrow\tau} \mathbb{E}|[\mathbb{E}^{\mathcal{F}_r} - \mathbb{E}^{\mathcal{F}_t}]\int_t^T |\sqrt{S(s)}y^{t,x}(s)|_H^2\,ds|^p.\end{aligned}$$

Setting $y^{r,x}(s) = x$ for $s \in [t, r]$, first splitting the above expression, and then applying the Jensen inequality we get

$$\lim_{r\downarrow\tau,\, t\uparrow\tau} \mathbb{E}|\mathbb{E}^{\mathcal{F}_r} \int_r^T |\sqrt{S(s)}y^{r,x}(s)|_H^2\, ds - \mathbb{E}^{\mathcal{F}_r} \int_r^T |\sqrt{S(s)}y^{t,x}(s)|_H^2\, ds$$
$$+\mathbb{E}^{\mathcal{F}_r} \int_t^r |\sqrt{S(s)}y^{t,x}(s)|_H^2\, ds|^p \le c(p)[\lim_{r\downarrow\tau,\, t\uparrow\tau} \mathbb{E}| \int_r^T (|\sqrt{S(s)}[y^{r,x}(s) - y^{t,x}(s)]|_H^2)\, ds|^p$$
$$+ \lim_{r\downarrow\tau,\, t\uparrow\tau} \mathbb{E}|\mathbb{E}^{\mathcal{F}_r} \int_t^r |\sqrt{S(s)}y^{t,x}(s)|_H^2\, ds|^p] \le C(T, p, M_S)$$
$$\times \lim_{r\downarrow\tau,\, t\uparrow\tau} \int_t^T \mathbb{E} \sup_{s\in[t,T]} (|[y^{r,x}(s) - y^{t,x}(s)]|_H^{2p})\, ds + C(T, p)$$
$$\times \lim_{r\downarrow\tau,\, t\uparrow\tau} \mathbb{E} \int_t^r |\sqrt{S(s)}y^{t,x}(s)|_H^{2p}\, ds].$$

Therefore, by the dominated convergence theorem we get that

$$\lim_{r\downarrow\tau,\, t\uparrow\tau} \mathbb{E}|\mathbb{E}^{\mathcal{F}_r} \int_r^T |\sqrt{S(s)}y^{r,x}(s)|_H^2\, ds - \mathbb{E}^{\mathcal{F}_r} \int_t^T |\sqrt{S(s)}y^{t,x}(s)|_H^2\, ds|^p = 0.$$

The second term on the right-hand side of (6.4) can be treated like the term $\mathbb{E}^{\mathcal{F}_r}(P_T y(T), y(T))_H$ in (6.3). The third term in (6.2) follows identically as does the second term in (6.2).

This concludes the proof of the lemma. □

Now we can state the main result of the paper.

THEOREM 6.6. *Assume that hypotheses* (A1)–(A4) *hold true. Then there exists a unique generalized solution of problem* (5.11).

Moreover, we have the following characterization of the optimal control: fix $T > 0$ and $x \in H$. Then

1. *there exists a unique control $\overline{u} \in L^2_{\mathcal{P}}(\Omega \times [0, T]; U)$ such that*
$$J(0, x, \overline{u}) = \inf_{u\in L^2_{\mathcal{P}}(\Omega\times[0,T];U)} J(0, x, u);$$
2. *if $\overline{y}$ is the mild solution of the state equation corresponding to $\overline{u}$ (that is, the optimal state), then $\overline{y}$ is the unique mild solution to the* closed loop *equation*
(6.5)
$$\begin{cases} d\overline{y}(r) = [A\overline{y}(r) + A_\sharp(r)\overline{y}(r) - B(r)B^*(r)P(r)\overline{y}(r)]\, dr + C\overline{y}(r)\, dW(r), \\ \overline{y}(0) = x; \end{cases}$$
3. *the following feedback law holds $\mathbb{P}$-a.s. for almost every s:*
$$\overline{u}(s) = -B^*(s)P(s)\overline{y}(s); \tag{6.6}$$
4. *the optimal cost is given by $J(0, x, \overline{u}) = (P(0)x, x)_H$.*

Proof. We divide the proof into three steps.

First step: *Existence of a generalized solution.* We fix a complete orthonormal basis $\{e_i : i \in \mathbb{N}\}$ in H and introduce, for each $N \in \mathbb{N}$, the finite dimensional projections $\Pi_N : H \to H : v \to \sum_{i=1}^N (v, e_i)_H e_i$. For each $N \in \mathbb{N}$ we define for (t, ω) fixed:

(6.7)
$$\Pi_N P_T(\omega)\Pi_N = P_T^N(\omega) \quad \text{and} \quad S^N(t,\omega) = \begin{cases} \Pi_N S(t,\omega)\Pi_N, & |S(t,\omega)|_{L(H)} \le N, \\ 0, & |S(t,\omega)|_{\Sigma(H)} > N. \end{cases}$$

First of all we notice that, from this definition, $P_T^N \in \Sigma_2^+(H)$, $\mathbb{P}$-a.s., for all $N \in \mathbb{N}$ and that for all $x \in H$, $(P_T^N x, x)_H \nearrow (P_T x, x)_H$, $\mathbb{P}$-a.s. Again from this definition it follows that $S^N(t,\omega) \in \Sigma_2^+(H)$, for all $N \in \mathbb{N}$ and that for all $x \in H$, $(S^N(t)x, x)_H \nearrow (S(t)x, x)_H$, $\mathbb{P}$-a.s. for a.e. $t \in [0,T]$. Moreover, $|S^N(t)|_{\Sigma(H)} \leq |S(t)|_{\Sigma(H)} \mathbb{P}$-a.s. for a.e. $t \in [0,T]$, so in particular (i) and (iii) in Definition 6.1 are verified. The pair (P_T^N, S^N) will become the data of the following approximating problems:

(6.8)
$$\begin{cases} -dP^N(t) = (A^*P^N(t) + P^N(t)A + A_\sharp^*(t)P^N(t) + P^N(t)A_\sharp(t) \\ \qquad\qquad - P^N(t)B(t)B^*(t)P^N(t))\,dt + S^N(t)dt + \mathrm{Tr}[C^*(t)P^N(t)C(t) \\ \qquad\qquad + C^*(t)Q^N(t) + Q^N(t)C(t)]\,dt + Q^N(t)\,dW(t), \quad t \in [0,T], \\ \\ P^N(T) = P_T^N. \end{cases}$$

We notice that the above equation satisfies the assumptions of Theorem 5.13. Thus for each fixed $N \in \mathbb{N}$ there exists a mild solution (P^N, Q^N) with $P^N \in L^2_{\mathcal{P}}(\Omega; C([0,T]; \Sigma_2^+(H))) \cap L^\infty_{\mathcal{P},S}(\Omega; C([0,T]; \Sigma^+(H)))$ and $Q^N \in L^2_{\mathcal{P}}(\Omega \times [0,T]; L_2(\Xi; \Sigma_2(H)))$. Notice that P^N has strongly continuous trajectories, so the final condition $P^N(T) = P_T^N$, $\mathbb{P}$-a.s. is attained.

Points (i)–(iii) in Definition 6.1 are satisfied by construction. We only have to show that (iv) holds true.

We fix $t \in [0,T]$. By Theorem 5.14

$$(P^N(t)x, x)_H = \inf\{J^N(t,x,u) : u \in L^2_{\mathcal{P}}(\Omega \times [0,T]; U)\},$$

where

$$J^N(t,x,u) = \mathbb{E}^{\mathcal{F}_t} \int_t^T \left[\left| \sqrt{S^N(s)}\, y^{t,x,u}(s) \right|_H^2 + |u(s)|^2 \right] ds + \left| \sqrt{P_T^N}\, y^{t,x,u}(T) \right|^2.$$

Clearly at each $u \in L^2_{\mathcal{P}}(\Omega \times [0,T]; U)$ fixed, we have that, for every $t \in [0,T]$ and $x \in H$, the sequence $\{J^N(t,x,u) : N \in \mathbb{N}\}$ is $\mathbb{P}$-a.s. nondecreasing. Moreover, it is bounded by $J(t,x,u) < +\infty$ $\mathbb{P}$-a.s. Thus the sequence of random variables $(P^N(t)x, x)_H = \inf_{u \in L^2_{\mathcal{P}}(\Omega \times [0,T]; U)} J^N(t,x,u)$ is nondecreasing as well. Since it is $\mathbb{P}$-a.s. bounded, it has a limit. It remains for us to show that this limit is actually of the form $(P(t)x, x)_H$ with $P \in L^\infty_{\mathcal{P},S}(\Omega \times [0,T]; \Sigma^+(H))$.

Let $\mathcal{D} \doteq \{x_i\}_{i \in \mathbb{N}}$ be a dense subset of H; then we can find a subset $\Omega_0 \subset \Omega$, with $P(\Omega_0) = 1$, such that for every $x_i \in \mathcal{D}$, $\exists \lim_{N \to +\infty} (P^N(t)x_i, x_i)_H$, for every $\omega \in \Omega_0$. Thus we define the limit $\phi(t, x_i, x_i)$ as follows:

$$\phi(t, x_i, x_i) = \begin{cases} \lim_{N \to +\infty} (P^N(t)x_i, x_i)_H & \forall x_i \in \mathcal{D} \quad \text{if } \omega \in \Omega_0, \\ 0 & \forall x_i \in \mathcal{D} \quad \text{if } \omega \notin \Omega_0. \end{cases}$$

For every $\omega \in \Omega_0$ the quadratic functional $\phi(t, x_i, x_i)$ defines a continuous, positive semidefinite, quadratic form on a dense subset. Indeed thanks to (5.8) one has the following uniform bound, modifying Ω_0 if necessary:

$$|\phi(t, x_i, x_i)| \leq \sup_{N \in \mathbb{N}} |P^N(t)|_{\Sigma(H)} |x_i|_H^2 \leq C_2(M_p + M_s)|x_i|_H^2 \quad \forall \omega \in \Omega_0.$$

Therefore, $\phi(t, x_i, x_i)$ can be extended to the whole $H \times H$ by density. Moreover, by the Riesz theorem we can associate with this quadratic form a linear, bounded symmetric, and positive semidefinite operator $P(t)$ such that for every $t \in [0,T]$

$$\phi(t, x, y) = (P(t)x, y)_H \quad \forall \omega \in \Omega_0 \quad \forall x, y \in H.$$

The following uniform bound is valid for all $t \in [0,T]$:

$$|P(t)|_{L(H)} \le C_2[M_P + M_S] \quad \forall \omega \in \Omega_0. \tag{6.9}$$

The process P is by construction predictable and strongly measurable. Finally, because the operators are positive and symmetric, the weak convergence also implies the strong convergence for all $t \in [0,T]$ we have that

$$|P^N(t)x - P(t)x|_H \to 0 \quad \forall \omega \in \Omega_0, \quad \forall x \in H. \tag{6.10}$$

This concludes the proof of the *first step* since (6.10) implies (iv) in the Definition 6.1.

Second step: *Characterization of the optimal control.* By Lemma 6.4 we know that any generalized solution verifies (6.1). In particular for $t = 0$ we have that

$$J(0,x,u) = (P(0)x,x)_H + \mathbb{E}\int_0^T |u(s) + B^*(s)P(s)y^{t,x,u}(s)|^2\, ds.$$

Thus $J(0,x,u) \ge (P(0)x,x)_H$ for all $u \in L^2_{\mathcal{P}}(\Omega \times [0,T];U)$, and the equality holds if and only if (6.6) holds, that is, if and only if y solves (6.5) and $u = \overline{u}$. This completely characterizes the optimal control. We notice that existence and uniqueness of a solution to the closed loop equation (6.5) are guaranteed since the assumptions of Theorem 3.2 are satisfied if $A_\sharp$ is replaced by $A_\sharp - BB^*P$.

Third step: *Uniqueness of the generalized solution.* Let P_1 and P_2 be two generalized solutions. We choose $\bar{u} = -B^*P_1\bar{y}$, where $\bar{y}$ solves (6.5) with P replaced by P_1. By the fundamental relation (6.1) and the characterization of the optimal control proved above we immediately have

$$(P_1(t)x,x)_H = J(t,x,\bar{u}) = (P_2(t)x,x)_H + \mathbb{E}^{\mathcal{F}_t}\int_t^T |\bar{u}(s) + B^*(s)P_2(s)\bar{y}(s)|^2\, ds \quad \mathbb{P}\text{-a.s.}$$

Thus $(P_1(t)x,x)_H \ge (P_2(t)x,x)_H$ $\mathbb{P}$-a.s. The claim follows by repeating the argument, choosing P_2 instead of P_1 and repeating the argument. □

Remark 6.7. The idea of regularizing the data and then defining a generalized notion of solution is rather classical in the PDE context; see, for instance, the definition of "strong solution" in [16], [1], and the references therein, although it seems to be the first time that is used in the context of the Riccati equation.

Remark 6.8. If assumptions (A1)–(A5) in Hypothesis 2.1 hold, then comparing relation (6.1) and relation (5.13) we immediately deduce that generalized solutions of (5.11) are mild solutions of (5.11) (see section 5 for definition of mild solutions) and vice versa. On the contrary, when (A5) fails, generalized solutions still exist and are unique while mild solutions cannot be defined.

7. Generalized solutions and variation of constant formula. The aim of this section is to give a further characterization of the generalized solution just defined. To this purpose we notice that the state equation defines an evolution operator in a suitable sense, and we recover a variation of the constant formula for the value function. The main ingredient is the fundamental relation (6.1) that on one hand is verified by the generalized solution and on the other hand will turn out to be essential to define the evolution operator.

DEFINITION 7.1. *Assume* (A1)–(A3) *and consider the state equation starting from* x *at time* $t \in [0,T]$ *and with control* $u = 0$*, namely*

$$\begin{cases} dy(s) = (Ay(s) + A_\sharp(s)y(s))\,ds + C(s)y(s)\,dW(s), & s \in [t,T], \\ y(t) = x. \end{cases}$$

We denote by $y^{t,x}$ *its mild solution and define the family of maps* $L_{t,\sigma} : L^\infty_S(\Omega, \mathcal{F}_\sigma; \Sigma(H)) \to L^\infty_S(\Omega, \mathcal{F}_t; \Sigma(H))$ *for* $0 \le t \le \sigma \le T$ *in the following way. For every* $V \in L^\infty_S(\Omega, \mathcal{F}_\sigma; \Sigma(H))$ *we define*

$$(L_{t,\sigma}Vx, x)_H = \mathbb{E}^{\mathcal{F}_t}(Vy^{t,x}(\sigma), y^{t,x}(\sigma))_H. \tag{7.1}$$

We collect some properties for the evolution operator $L_{t,\sigma}$ that can be easily deduced from its definition.

LEMMA 7.2. *The family of operators* $\{L_{t,\sigma} : 0 \le t \le \sigma \le T\}$ *has the following properties:*

1. *for every* $0 \le t \le \sigma \le T$ $L_{t,\sigma}$ *is a linear and bounded operator*
$$L_{t,\sigma} : L^\infty_S(\Omega, \mathcal{F}_\sigma; \Sigma(H)) \to L^\infty_S(\Omega, \mathcal{F}_t; \Sigma(H));$$
2. *for every* $0 \le t \le r \le \sigma \le T$ *one has that*
$$L_{t,\sigma} = L_{t,r} \circ L_{r,\sigma} \quad \mathbb{P}\text{-}a.s.;$$
3. *with fixed* $V \in L^\infty_S(\Omega, \mathcal{F}_\sigma; \Sigma(H))$, $x \in H$, *and* σ *in* $[0,T]$ *the map* $t \to (L_{t,\sigma}Vx, x)_H$ *belongs to* $C_{\mathcal{P}}([0,T]; L^p(\Omega, \mathbb{R}))$ *for all* $p \ge 2$.

Proof.

1. We have that
$$\begin{aligned} \sup_{x\in H, |x|_H \le 1} |(L_{t,\sigma}Vx, x)_H &= \sup_{x\in H, |x|_H \le 1} |\mathbb{E}^{\mathcal{F}_t}(Vy^{t,x}(\sigma), y^{t,x}(\sigma))_H| \\ &\le |V|_{L^\infty_S(\Omega,\mathcal{F}_\sigma;L(H))} \sup_{x\in H, |x|_H \le 1} \mathbb{E}|y^{t,x}(\sigma)|^2_H \le C_2|V|_{L^\infty_S(\Omega,\mathcal{F}_\sigma;L(H))}. \end{aligned}$$
2. The proof follows from the semigroup property of the solution $y^{t,x}(\sigma)$ and the property of conditional expectations with respect to the filtration $\mathcal{F}_t$.
3. The proof is identical to that of Lemma 6.5. □

We notice that the fundamental relation (6.1), evaluated at $u = 0$, can be rewritten in terms of the evolution operator and reads as follows, for all $t \in [0,T]$ and all $x \in H$:

$$\begin{aligned} (P(t)x, x)_H &= \mathbb{E}^{\mathcal{F}_t}(P_T y(T), y(T))_H + \mathbb{E}^{\mathcal{F}_t} \int_t^T |\sqrt{S}(s)y(s)|^2\,ds \\ &\quad - \mathbb{E}^{\mathcal{F}_t} \int_t^T |B^*(s)P(s)y(s)|^2\,ds \\ &= (L_{t,T}P_T x, x)_H + \int_t^T (L_{t,s}S(s)x, x)_H\,ds \\ &\quad - \int_t^T (L_{t,s}P(s)B(s)B^*(s)P(s)x, x)_H\,ds \quad \mathbb{P}\text{-a.s.} \end{aligned} \tag{7.2}$$

This relation suggests a new characterization for a solution of (5.11).

DEFINITION 7.3. *A variation of constants solution to problem* (5.11) *is a map* $P \in L^{\infty}_{\mathcal{P},S}((0,T)\times\Omega;\Sigma^{+}(H))$ *such that for all* $x \in H$, $\forall p \geq 1$, $Px \in C_{\mathcal{P}}([0,T];L^{p}(\Omega;H))$ *and the following variation of constant formula is verified,* $\mathbb{P}$*-a.s.,*

$$(P(t)x,x)_H = (L_{t,T}P_Tx,x)_H + \int_t^T (L_{t,s}S(s)x,x)_H\,ds - \int_t^T (L_{t,s}P(s)B(s)B^*(s)P(s)x,x)_H\,ds. \tag{7.3}$$

We can prove existence and uniqueness of such solutions.

THEOREM 7.4. *Assume* (A1)–(A4)*; then there exists a unique solution of problem* (5.11) *in the sense of Definition* 7.3.

Proof. We already know that the generalized solution defined in the previous section verifies (7.3) and it is regular enough to be a solution in the sense of Definition 7.3. It remains to prove the uniqueness of the solution in this class. Let P_1 and P_2 be two solutions in the sense of Definition 7.3 and denote by $\bar{P}$ their difference $\bar{P}(t) = P_1(t) - P_2(t)$. Then the following holds, for all $t \in [0,T]$ and $x \in H$, $\mathbb{P}$-a.s.:

$$(\bar{P}(t)x,x)_H = \int_t^T (L_{t,s}P_2(s)B(s)B^*(s)\bar{P}(s)x,x)_H + \int_t^T (L_{t,s}\overline{P}(s)B(s)B^*(s)P_1(s)x,x)_H\,ds.$$

Therefore, for every $t \in [0,T]$ we have that

$$\begin{aligned}|\overline{P}(t)|_{L(H)} &= \sup_{x\in H,|x|_H\leq 1}(\overline{P}(t)x,x)_H\\ &\leq \sup_{x\in H,|x|_H\leq 1}\int_t^T (L_{t,s}P_2(s)B(s)B^*(s)\overline{P}(s)x,x)_H\\ &\quad + \int_t^T (L_{t,s}\overline{P}(s)B(s)B^*(s)P_1(s)x,x)_H\,ds\\ &\leq C\int_t^T |\overline{P}(s)|_{L(H)}\,ds \quad \mathbb{P}\text{-a.s.},\end{aligned}$$

where C depends on C_2, M_B, $|P_1|_{L^{\infty}_{\mathcal{P},S}(\Omega\times[0,T];L(H))}$, and $|P_2|_{L^{\infty}_{\mathcal{P},S}(\Omega\times[0,T];L(H))}$.

Applying the Gronwall lemma to $s \to |\overline{P}(s)|_{L^{\infty}(\Omega,\mathcal{F}_s,L(H))}$ we get that $P_1(t) = P_2(t)$, $\mathbb{P}$-a.s. for all $t \in [0,T]$. □

Remark 7.5. Since the solution of Definition 7.3 is also the unique generalized solution, it is obvious that it allows the synthesis of optimal controls as in Theorem 6.6. □

8. Nonhomogeneous problem. As in [13] we consider a simple generalization of our original control problem that enlarges the set of applicability of our abstract results.

We fix $\eta \in L^2(\Omega,\mathcal{F}_T,\mathbb{P},H)$ and $\nu \in L^2_{\mathcal{P}}(\Omega\times[0,T],H)$ and instead of $J(0,x,u)$ we minimize

$$\begin{aligned}\widehat{J}(0,x,u) = \mathbb{E}\int_0^T \big((S(s)(y^{0,x,u}(s)-\nu(s)),(y^{0,x,u}(s)-\nu(s)))_H + |u(s)|_U^2\big)\,ds\\ +\mathbb{E}(P_T(y^{0,x,u}(T)-\eta),(y^{0,x,u}(T)-\eta))_H.\end{aligned}$$

We assume that (A1)–(A4) in Hypothesis 2.1 hold and let P be the unique generalized solution of the Riccati equation (5.11). Moreover, (p,q) with p in $L^2_{\mathcal{P}}(\Omega, C([0,T];H))$ and q in $L^2_{\mathcal{P}}(\Omega, L^2([0,T];L_2(\Xi,H)))$ is the unique mild solution of the backward equation

$$\begin{cases} dp(s)=\big(-A^*p(s) - A^*_\sharp(s)p(s) + P(s)B(s)B^*(s)p(s) - \mathrm{Tr}[C^*(s)q(s)]\big)\,ds \\ \qquad\quad -S(s)\nu(s)ds + q(s)dW(s), \quad s\in[0,T], \\ p(T)=P_T\eta, \end{cases} \tag{8.1}$$

where, defining notation with respect to the usual basis $\{f_i : i\in\mathbb{N}\}$ in Ξ,

$$\mathrm{Tr}[C^*(s)q(s)] = \sum_{i=1}^{\infty} C_i^*(s)(q(s),f_i).$$

Moreover, existence and uniqueness of a mild solution to (8.1) are guaranteed by Theorem 4.4, whose assumptions are easily verified.

The following is the analogue of Theorem 6.6.

THEOREM 8.1. *Assume that hypotheses* (A1)–(A4) *hold true. Then*

1. *there exists a unique control* $\overline{u}\in L^2_{\mathcal{P}}(\Omega\times[0,T];U)$ *such that*

$$\widehat{J}(0,x,\overline{u}) = \inf_{u\in L^2_{\mathcal{P}}(\Omega\times[0,T];U)} \widehat{J}(0,x,u);$$

2. *if* $\overline{y}$ *is the mild solution of the state equation corresponding to* $\overline{u}$ *(that is the optimal state), then* $\overline{y}$ *is the unique mild solution to the* closed loop *equation*

$$\begin{cases} d\overline{y}(r) = [A\overline{y}(r) + A_\sharp(r)\overline{y}(r) - B(r)B^*(r)P(r)\overline{y}(r) + B(r)B^*(r)p(r)]\,dr \\ \qquad\quad +C\overline{y}(r)\,dW(r), \\ \overline{y}(0) = x; \end{cases} \tag{8.2}$$

3. *the following feedback law holds* $\mathbb{P}$*-a.s. for almost every* s*:*

$$\overline{u}(s) = -B^*(s)P(s)\overline{y}(s) + B^*(s)p(s); \tag{8.3}$$

4. *the optimal cost is given by*

$$\begin{aligned}\widehat{J}(0,x,\overline{u}) = (P(0)x,x)_H - 2(p(0),x)_H + \mathbb{E}(P_T\eta,\eta)_H \\ +\mathbb{E}\int_0^T \big((S(s)\nu(s),\nu(s))_H - |B^*(s)p(s)|^2_H\big)ds.\end{aligned}$$

Proof. Let $(p_h,q_h)\in L^2_{\mathcal{P}}(\Omega;C([0,T];H))\times L^2_{\mathcal{P}}(\Omega;L^2([0,T];L_2(\Xi,H)))$, $h=1,2,\dots$, be the unique classical solution of the backward equation

$$\begin{cases} dp(s)=\big(-A^*_hp_h(s) - A^*_\sharp(s)p_h(s) + P(s)B(s)B^*(s)p_h(s) - \mathrm{Tr}[C^*(s)q_h(s)]\big)\,ds \\ \qquad\quad -S(s)\nu(s)ds + q_h(s)dW(s), \quad s\in[0,T], \\ p(T)=P_T\eta. \end{cases} \tag{8.4}$$

We proceed as in the proof of Theorem 5.6. Namely we choose $\Psi\in C^2(H)$ with $\Psi(y)=1$ for $|y|\le 1$, $\Psi(y)=0$ for $|y|\ge 2$, and $\Psi(y)\in[0,1]\forall y\in H$. Then we

differentiate $d\Psi(N^{-1}y_h(s))(p_h(s),y_h(s))_H$ by the Itô rule. We integrate in $[0,T]$ and compute the mean value. Finally we let $N\to+\infty$ to obtain

$$\begin{aligned}\mathbb{E}(P_T\eta,y_h(T))_H &= (p(0),x)_H+\mathbb{E}\int_0^T(P(s)B^*(s)B(s)p_h(s),y_h(s))_H ds\\ &\quad+\mathbb{E}\int_0^T[(u(s),B^*(s)p_h(s))_H-(S(s)\nu(s),y_h(s))_H]ds.\end{aligned}$$

Letting $h\to\infty$ we get by Theorems 3.3 and 4.4

$$\begin{aligned}\mathbb{E}(P_T\eta,y(T))_H &= (p(0),x)_H+\mathbb{E}\int_0^T(P(s)B^*(s)B(s)p(s),y(s))_H ds\\ &\quad+\mathbb{E}\int_0^T[(u(s),B^*(s)p(s))_H-(S(s)\nu(s),y(s))_H]ds.\end{aligned}$$

Thus by easy computations

$$\begin{aligned}\widehat{J}(0,x,u) &= \mathbb{E}\int_0^T|u(s)+B^*(s)P(s)y(s)-B^*(s)p(s)|^2\,ds+(P(0)x,x)_H\\ &\quad-2(p(0),x)_H+\mathbb{E}(P_T\eta,\eta)_H\\ &\quad+\mathbb{E}\int_0^T\Big[(S(s)\nu(s),\nu(s))_H-|B^*(s)p(s)|_H\Big]ds.\end{aligned}$$

The above relation completes the proof (notice that existence and uniqueness of the mild solution of (8.2) can be proved exactly as are existence and uniqueness of the mild solution of (2.1)). □

9. Example: Minimal variance problem for a stochastic equation with delay and random volatility. We consider the controlled stochastic differential equation with memory effects:

$$\text{(9.1)}\quad\begin{cases}d\xi(t)=\left[\displaystyle\int_{-1}^0\xi(t+\theta)\,a(d\theta)+r(t)u(t)\right]dt+\sum_{i=1}^d\sigma_i(t)\xi(t)d\beta_t^i,\quad t\in[0,T],\\ \\ \xi(0)=\mu_0,\quad \xi(\theta)=\nu_0(\theta),\qquad\text{for a.e. }\theta\in(-1,0),\end{cases}$$

where $\mu_0\in\mathbb{R}^n$, $\nu_0\in L^2((-1,0);\mathbb{R}^n)$, $(\Omega,\mathcal{E},\mathcal{P})$ is a complete probability space, $\{\beta_t^i:t\geq0,\ i=1,\dots,d\}$ are independent standard Brownian motions defined in Ω. Moreover, $\mathcal{F}_t$ denotes the σ-algebra generated by $\{\beta_\sigma^i,\ \sigma\in[0,t],\ i=1,\dots,d\}$ and augmented with the sets of $\mathcal{F}$ with $\mathbb{P}$-measure zero (see Remark 2.4 to see how this requirement can be relaxed, and notice that for some $i=1,\dots,d$, σ_i can be null).

We assume that a is a $L(\mathbb{R}^n,\mathbb{R}^n)$-valued finite measure on $[-1,0]$, $r:[0,T]\times\Omega\to L(\mathbb{R}^d,\mathbb{R}^n)$ is bounded and predictable stochastic process, and $\sigma_i:[0,T]\times\Omega\to L(\mathbb{R}^n,\mathbb{R}^n)$ are bounded and predictable stochastic processes, $i=1,\dots,d$.

We also consider the following cost functional of *minimal variance* type:

$$J(0,\mu_0,\nu_0,u)=\mathbb{E}\int_0^T|u(\tau)|^2_{\mathbb{R}^d}d\tau+\mathbb{E}(k(\xi(T)-\zeta),(\xi(T)-\zeta))_{\mathbb{R}^n},$$

where $k\in L^\infty(\Omega,\mathcal{F}_T,\mathbb{P};\Sigma^+(\mathbb{R}^n))$ and $\zeta\in L^2(\Omega,\mathcal{F}_T,\mathbb{P};\mathbb{R}^n)$.

Our purpose is to minimize $J(0,\mu_0,\nu_0,u)$ over all predictable controls $u:[0,T]\times\Omega\to\mathbb{R}^d$.

Following [4] and [6] we set $H=\mathbb{R}^n\times L^2((-1,0);\mathbb{R}^n)$,

$$\mathcal{D}(A)=\left\{\begin{pmatrix}\mu\\ \nu\end{pmatrix}\in H:\nu\in W^{1,2}((-1,0);\mathbb{R}^n)\text{ and }\nu(0)=\mu\right\},$$
$$A\begin{pmatrix}\mu\\ \nu\end{pmatrix}=\begin{pmatrix}\int_{-1}^0\nu(\theta)a(d\theta)\\ \frac{d\nu}{d\theta}\end{pmatrix}.$$

It is proved in [9], among other places, that A generates a strongly continuous semigroup in H (see also [6]). Moreover, if we set $U=\mathbb{R}^d$ and for $t\in[0,T]$, $\mu\in\mathbb{R}^n$, $\nu\in L^2((-1,0);\mathbb{R}^n)$, $u\in\mathbb{R}^d$,

$$x=\begin{pmatrix}\mu_0\\ \nu_0\end{pmatrix},\quad B(t)u=\begin{pmatrix}r(t)u\\ 0\end{pmatrix},\quad C_i(t)\begin{pmatrix}\mu\\ \nu\end{pmatrix}=\begin{pmatrix}\sigma_i(t)\mu\\ 0\end{pmatrix},$$
$$P_T\begin{pmatrix}\mu\\ \nu\end{pmatrix}=\begin{pmatrix}k(t)\mu\\ 0\end{pmatrix},\quad \eta=\begin{pmatrix}\zeta\\ 0\end{pmatrix},\quad y(\tau)=\begin{pmatrix}\xi(\tau)\\ \xi_\tau(\cdot)\end{pmatrix},$$

where $x_\tau(\theta)=x(\theta+\tau)$, $\tau\ge0$, $\theta\in[-1,0]$, then (9.1) is equivalent (see [4] and [6, Chapter 10]) to

$$\begin{cases}dy(\tau)=(Ay(\tau)+Bu_\tau)\ d\tau+\sum_{i=1}^d C_i(\tau)y(\tau)d\beta_\tau^i,\quad \tau\in[0,T],\\ y(0)=y_0.\end{cases}$$

Moreover, the cost functional becomes

$$\widehat{J}(0,x,u)=\mathbb{E}\int_0^T|u(\tau)|^2_{\mathbb{R}^d}ds+\mathbb{E}\left|\sqrt{P_T}(y(T)-\eta)\right|^2_H$$

Moreover, it is easy to verify that (A1)–(A5) of Hypothesis 2.1 hold. Thus Theorem 8.1 can be applied to obtain the synthesis of the optimal control. We notice that in this case the Riccati equation has a unique mild solution in the sense clarified by Definition 5.3.

Remark 9.1. We believe that the present example is interesting on its own because of its simplicity. Notice that the model is finite dimensional, but the presence of a simple delay term and of the stochastic coefficient σ immediately requires us to use backward stochastic Riccati equations in infinite dimensional spaces.

In addition it can be regarded as a first step towards realistic financial applications of the theory. Namely in [14] (see also [25]) the authors showed that the *mean variance hedging* problem for a (incomplete) *Black and Scholes market with stochastic volatility* can be treated as a singular linear quadratic control problem with stochastic coefficients. The solution of such problem requires one to prove existence and uniqueness of the solution of a backward stochastic Riccati equation in finite dimensions. On the other hand, in [8] it was pointed out that *memory effects* can be introduced in the market model describing the evolution of the share prices by a delay equation. Thus the present example can be seen as a contribution towards the solution of the mean variance hedging problem for a market with stochastic volatility and memory effects. To deal with the realistic formulation of the problem it would be necessary to allow

control dependent noise and singular costs. This complicates the form of the Riccati equation and requires careful mixing of the techniques developed in this paper to deal with infinite dimensional stochastic Riccati equations and of the ones developed in [14] and [23] to deal with singular control problems and control dependent noise. This will be the topis of a future work.

10. Example: Optimal control for a wave equation in random media with stochastic damping. In order to show that our general results can be applied to concrete controlled stochastic PDEs arising in applications we consider a stochastic wave equation with diffused control. We assume that the system is evolving in a random media, and this influences its evolution in two ways: through a stochastic force of elastic type (the term $\sum_{i=1}^{\infty} c_i(t,\zeta)\xi(t,\zeta)d\beta_i(t)$ below) and through a stochastic damping (the term $\mu(t,\zeta)\partial_t\xi(t,\zeta)dt$ below). Notice that in this model it is natural to introduce the stochastic coefficient μ; moreover, although only one coefficient is stochastic, the use of backward stochastic Riccati equations is necessary to solve the optimal control problem.

We consider the state equation

(10.1)
$$\begin{cases} d_t\partial_t\xi(t,\zeta) = \Delta_\zeta\xi(t,\zeta)dt + b(t,\zeta)u(t,\zeta)dt + \mu(t,\zeta)\partial_t\xi(t,\zeta)dt \\ \qquad\qquad + \sum_{i=1}^{\infty} c_i(t,\zeta)\xi(t,\zeta)d\beta_i(t), \quad \zeta\in\mathcal{D}, \quad t\in[0,T], \\ \xi(t,\zeta) = 0, \quad \zeta\in\partial\mathcal{D},\ t\in[0,T], \\ \\ \xi(0,\zeta) = x_0(\zeta),\ \partial_t\xi(0,\zeta) = v_0(\zeta), \quad \zeta\in\mathcal{D}, \end{cases}$$

and the cost functional

(10.2)
$$J(0,x,u) = \mathbb{E}\int_0^T\int_{\mathcal{D}}\left[\kappa_1(t,\zeta)\xi^2(t,\zeta) + \kappa_2(t,\zeta)\left(\frac{\partial\xi}{\partial t}(t,\zeta)\right)^2\right]d\zeta\,dt$$
$$+\mathbb{E}\int_0^T\int_{\mathcal{D}} u^2(t,\zeta)d\zeta\,dt\mathbb{E}\int_{\mathcal{D}}\left[\pi_1(\zeta)\xi^2(T,\zeta) + \pi_2(\zeta)\left(\frac{\partial\xi}{\partial t}(T,\zeta)\right)^2\right]d\zeta.$$

In the above formulae $\mathcal{D}\subset\mathbb{R}^d$ is a bounded domain with regular boundary. By $\mathcal{B}(\mathcal{D})$ we denote the Borel σ-field in $\mathcal{D}$.

Moreover, $\{\beta_i : i = 1,2,\dots\}$ are independent standard (real valued) Brownian motions defined on $(\Omega,\mathcal{F},\mathbb{P})$. We set $\mathcal{F}_t = \sigma\{\beta_i(s) : s\in[0,t],\ i = 1,2,\dots\}$ and denote by $\mathcal{P}$ the predictable σ-field in $\Omega\times[0,T]$.

On the coefficients we assume the following.

1. μ is a bounded measurable process defined on $([0,T]\times\Omega)\times\mathcal{D}$ endowed with the σ-field $\mathcal{P}\otimes\mathcal{B}(\mathcal{D})$ with values in $\mathbb{R}^+$ (with Borel σ-field).
2. b, κ_1, κ_2, and c_i, $i = 1,2,\dots$, are bounded measurable maps $[0,T]\times\mathcal{D}\to\mathbb{R}$. We assume that κ_1 and κ_2 have values in $\mathbb{R}^+$.
3. There exists a constant $M > 0$ such that $\sum_{i=1}^{\infty}|c_i(t,\zeta)|^2 \le M$ for a.e. $t\in[0,T]$ and a.e. $\zeta\in\mathcal{D}$.
4. π_1 and π_2 are bounded measurable maps $\mathcal{D}\to\mathbb{R}^+$.

Following, for instance, [1] we set

1. $H = H_0^1(\mathcal{D})\times L^2(\mathcal{D})$, $U = L^2(\mathcal{D})$;
2. $W(t) = \sum_{i=1}^{\infty} f_i\beta_i(t)$, where $\{f_i : i = 1,2,\dots\}$ is an orthonormal basis in an arbitrary separable real Hilbert space Ξ;

3. $\mathcal{D}(A) = \left[H^2(\mathcal{D}) \cap H_0^1(\mathcal{D})\right] \times H_0^1(\mathcal{D})$ and

$$\left(A\begin{pmatrix}\xi\\ v\end{pmatrix}\right)(\zeta) = \begin{pmatrix}v(\zeta)\\ \Delta_\zeta \xi(\zeta)\end{pmatrix}, \quad \begin{pmatrix}\xi\\ v\end{pmatrix} \in \mathcal{D}(A),$$
$$\left(A_\sharp(t)\begin{pmatrix}\xi\\ v\end{pmatrix}\right)(\zeta) = \begin{pmatrix}0\\ \mu(t,\zeta)\xi(\zeta)\end{pmatrix}, \quad \begin{pmatrix}\xi\\ v\end{pmatrix} \in H;$$

4. $(B(t)u)(\zeta) = \begin{pmatrix}0\\ b(t,\zeta)u(\zeta)\end{pmatrix}, \quad \left(C_i(t)\begin{pmatrix}\xi\\ v\end{pmatrix}\right)(\zeta) = \begin{pmatrix}0\\ c_i(t,\zeta)\xi(\zeta)\end{pmatrix};$
5. $\left(S(t)\begin{pmatrix}\xi\\ v\end{pmatrix}\right)(\zeta) = \begin{pmatrix}\kappa_1(t,\zeta)\xi(\zeta)\\ \kappa_2(t,\zeta)v(\zeta)\end{pmatrix}, \left(P_T(t)\begin{pmatrix}\xi\\ v\end{pmatrix}\right)(\zeta) = \begin{pmatrix}\pi_1(\zeta)\xi(\zeta)\\ \pi_2(\zeta)v(\zeta)\end{pmatrix},$
$x = \begin{pmatrix}x_0\\ v_0\end{pmatrix}.$

With this setting the state equation (10.1) is equivalent to (2.1) and the cost (10.2) is equivalent to (2.2). Moreover, it is easy to verify that assumptions (A1)–(A4) in Hypothesis 2.1 are verified. So in this case we can apply the results in Theorem 6.6 to obtain existence of a unique solution of the Riccati equation both in the "generalized" sense of Definition 6.1 and the "variation of constants" sense of Definition 7.3. Moreover, such a solution allows to perform the synthesis of the optimal control as it is stated in Theorem 6.6.

Remark 10.1. Notice that assumption (A5) is in general not satisfied; take, for instance, $\kappa_2 \equiv 1$ or $\pi_2 \equiv 1$.

Acknowledgments. We wish to thank Marco Fuhrman; we are indebted to him for his help and encouragement. We also thank the referees for their useful remarks.

REFERENCES

[1] A. Bensoussan, G. da Prato, M.C. Delfour, and S.K. Mitter, *Representation and Control of Infinite Dimensional Systems*, Vol. 1, Syst. Control Found. Appl., Birkhuser, Boston, MA, 1992.

[2] J.-M. Bismut, *Linear quadratic optimal stochastic control with random coefficients,* SIAM J. Control Optim., 14 (1976), pp. 419–444.

[3] J.-M. Bismut, *Contrôle des systèmes linéaires quadratiques: applications de l'intégrale stochastique,* in Séminaire de Probabilités, XII, Lecture Notes in Math. 649, Springer-Varlag, Berlin, 1978.

[4] A. Chojnowska-Michalik, *Representation theorem for general stochastic delay equations,* Bull. Pol. Acad. Sci. Ser. Sci. Math., 26 (1978), pp. 634–641.

[5] G. da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions,* Cambridge University Press, Cambridge, UK, 1992.

[6] G. da Prato and J. Zabczyk, *Ergodicity for Infinite-Dimensional Systems,* London Math. Soc. Lecture Note Ser. 229, Cambridge University Press, Cambridge, UK, 1996.

[7] M. Fuhrman and G. Tessitore, *Nonlinear Kolmogorov equations in infinite dimesional spaces: The backward stochastic differential equations approach and applications to optimal control,* Ann. Probab., 30 (2002), pp. 1397–1465.

[8] M. Fuhrman and G. Tessitore, *Generalized directional gradients, backward stochastic differential equations and mild solutions of semilinear parabolic equations,* Appl. Math. Optim., to appear.

[9] J. Hale, *Theory of Functional Differential Equations,* Appl. Math. Sci. 3, Springer-Verlag, New York, 1977.

[10] Y. Hu and S. Peng, *Adapted solution of a backward semilinear stochastic evolution equation,* Stochastic Anal. Appl., 9 (1991), pp. 445–459.

[11] M. Kobylanski, *Backward stochastic differential equations and partial differential equations with quadratic growth,* Ann. Probab., 28 (2000), pp. 558–602.

[12] M. Kohlmann and S. Tang, *New developments in backward stochastic Riccati equations and their applications,* in Mathematical Finance (Konstanz, 2000), Trends Math., Birkhuser, Basel, 2001.

[13] M. Kohlmann and S. Tang, *Global adapted solution of one-dimensional backward stochastic Riccati equations, with application to the mean-variance hedging,* Stochastic Process. Appl., 97 (2002), pp. 1255–1288.

[14] M. Kohlmann and S. Tang, *Multidimensional backward btochastic riccati equations and applications,* SIAM J. Control Optim., 41 (2003), pp. 1696–1721.

[15] M. Kohlmann and X.Y. Zhou, *Relationship between backward stochastic differential equations and stochastic controls: A linear-quadratic approach,* SIAM J. Control Optim., 38 (2000), pp. 1392–1407.

[16] A. Lunardi, *Analytic Semigroups and Optimal Regularity in Parabolic Problems,* Progr. Nonlinear Differential Equations Appl. 16, Birkhäuser Verlag, Basel, 1995.

[17] J.-P. Lepeltier and J. San Martín, *Existence for BSDE with superlinear-quadratic coefficient,* Stoch. Stoch. Rep., 63 (1998), pp. 227–240.

[18] E. Pardoux and S. Peng, *Adapted solution of a backward stochastic differential equation,* Systems Control Lett., 14 (1990), pp. 55–61.

[19] S. Peng, *Stochastic Hamilton–Jacobi–Bellman equations,* SIAM J. Control Optim., 30 (1992), pp. 284–304.

[20] S. Peng, *Open problems on backward stochastic differential equations,* in Control of Distributed Parameter and Stochastic Systems (Hangzhou, 1998), Kluwer, Boston, 1999.

[21] S. Tang, *General Linear Quadratic Optimal Control Problems with Random Coefficients: Linear Stochastic Hamilton Systems and Backward Stochastic Riccati Equations,* SIAM J. Control Optim., 42 (2003), pp. 53–75.

[22] G. Tessitore, *Some remarks on the Riccati equation arising in an optimal control problem with state- and control-dependent noise,* SIAM J. Control Optim., 30 (1992), pp. 717–744.

[23] J. Yong and X. Zhou, *Stochastic Controls: Hamiltonian Systems and HJB Equations,* Springer-Verlag, Berlin, 1999.

[24] J. Zabczyk, *Parabolic equations on Hilbert spaces,* in Stochastic PDE's and Kolmogorov Equations in Infinite Dimensions, G. da Prato, ed., Lecture Notes in Math. 1715, Springer-Verlag, Berlin, 1999, pp. 117–213.

[25] X. Zhou and D. Li, *Continuous time mean-variance portfolio selection: A stochastic LQ framework,* Appl. Math. Optim., 42 (2000), pp. 19–33.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 1, pp. 195–209

RELAXATION OF AN OPTIMAL DESIGN PROBLEM WITH AN INTEGRAL-TYPE CONSTRAINT*

ERNESTO ARANDA† AND JOSÉ C. BELLIDO‡

Abstract. We study a new relaxation for a two-dimensional optimal design problem in conductivity consisting of determining how to mix two given conducting materials in order to minimize the amount of one of them, subject to a constraint on the efficiency of the conducting properties of the mixture. Our approach here is different from that obtained in [R. V. Kohn and G. Strang, *Comm. Pure Appl. Math.*, 39 (1986), pp. 113–137, 139–182, 353–377], and is based on a local reformulation of the optimal design problem by means of the introduction of new potentials. The concept of constrained quasiconvexification is used in an important way.

Key words. optimal design, relaxation, constrained quasiconvexification

AMS subject classifications. 49J45, 74P10

DOI. 10.1137/S0363012902418662

1. Introduction. We would like to analyze the following optimal design problem. We have at our disposal two given conducting materials; one of them is a bad and cheap conductor and the other is better but also more expensive, and we want to fill out the domain Ω (a regular, open and simply connected set of $\mathbb{R}^2$) mixing those materials in order to minimize the amount of the most expensive conductor. The respective conductivities are α and β with $0 < \alpha < \beta$. If $\chi(x)$ is the characteristic function of the set where we put the material with conductivity α, the conductivity function in Ω is

$$a(x) = \chi(x)\alpha + (1 - \chi(x))\beta.$$

The electric potential of the body is given by the solution of the diffusion equation

$$\begin{cases} -\operatorname{div}(a(x)\nabla u(x)) = 0 & \text{in } \Omega, \\ a(x)\nabla u \cdot n = f & \text{on } \partial\Omega, \end{cases} \tag{1}$$

where $f \in H^{-\frac{1}{2}}(\partial\Omega)$ stands for the current flux on $\partial\Omega$ and n is the outer normal vector to $\partial\Omega$. We assume the compatibility condition

$$\int_{\partial\Omega} f\, ds = 0,$$

in order to guarantee the existence of the solution of (1). Recall that under that condition, (1) has a unique solution up to an additive constant. The optimal design

*Received by the editors November 26, 2002; accepted for publication (in revised form) October 13, 2004; published electronically July 18, 2005. This work was supported by MCyT (Spain) through grant BMF2001-0738, by Junta de Comunidades de Castilla-La Mancha through grant GC-02-001, and by Universidad de Castilla-La Mancha.

http://www.siam.org/journals/sicon/44-1/41866.html

†E.T.S.I. Industriales, Universidad de Castilla-La Mancha, Campus Universitario s/n, 13071-Ciudad Real, Spain (Ernesto.Aranda@uclm.es).

‡Mathematical Institute, University of Oxford, OX1 3LB, Oxford, UK (JoseCarlos.Bellido@uclm.es). On leave from Universidad de Castilla-La Mancha. The work of this author was supported by a CEC-Marie Curie Individual Fellowship, contract HPMF-CT-2002-02177, and by EU TMR network HMS2000-"Homogenization and multiple scales."

problem we want to address consists of finding a layout of material, i.e., a characteristic function χ, minimizing the functional

$$I(\chi) = \int_\Omega (1 - \chi(x))\, dx$$

under the integral-type constraint

$$J(\chi) = \int_\Omega \frac{1}{a(x)} |\nabla u(x)|^2\, dx \leq \gamma, \tag{2}$$

where $\gamma > 0$ is given. The integral $J(\chi)$ represents the rate at which energy is dissipated to heat in the composite material given by the design χ (the bigger this integral is, the more dissipation of energy and the less efficient the design is) so that (2) constrains the efficiency with which the conductivity a conducts the current load f through Ω.

A typical feature in optimal design problems like the one considered here is the lack of optimal solutions (see, for instance, [14]), so that relaxation is needed in order to understand the behavior of minimizing sequences. For our optimal design problem, relaxation has been analyzed in [10, 11, 12] using a suitable reformulation of the problem as a min-max variational problem amenable to relaxation. In that paper the authors also consider the case of multiple constraints of type (2) for several boundary data of f_i. In our work, we only deal with the case of a single integral constraint. Recent papers about this subject are [1, 7, 13].

In this paper, we propose a different approach to analyze relaxation of this optimal design problem. We reformulate this one, in an equivalent way, as a genuine vector variational problem subject to an integral-type constraint, and then study relaxation for this new problem. As usual, relaxation for variational problems is carried out on two different levels: convexified problems in which we change the energy density by a suitable convex envelope of it, and generalized problems in which we enlarge the set of admissible functions to the set of Young measures generated by sequences of admissible functions for the original problem. Very recently this approach has been successfully used in other optimal design problems in conductivity (see [4, 19] for the two-dimensional case and [5, 6] for the three-dimensional situation).

Let us see how we reformulate the optimal design problem as a vector variational problem subject to an integral constraint. The state equation (1) is equivalent to the existence of a stream function $v \in H^1(\Omega)$ such that

$$a(x)\nabla u(x) + T\nabla v(x) = 0, \quad \text{a.e. } x \in \Omega, \tag{3}$$

where T is the counterclockwise rotation of angle $\frac{\pi}{2}$ (see [9]). Due to the fact that u verifies the boundary condition

$$a\nabla u \cdot n = f,$$

(3) implies that the tangential derivative of v is equal to the negative normal component of $a\nabla u$. Hence, up to an arbitrary constant, the boundary values of v are determined by indefinite integration along the boundary

$$v = v_0 = -\int f\, ds \quad \text{on } \partial\Omega,$$

where $v_0 \in H^{\frac{1}{2}}(\partial\Omega)$.

Now we put the functions u and v together in a single field $U = (u, v)$ and consider the functions

$$W, V : \mathbb{M}^{2\times 2} \to \mathbb{R}^* = \mathbb{R} \cup \{+\infty\}$$

defined by

$$W(A) = \begin{cases} 0 & \text{if } A \in \Lambda_\alpha, \\ 1 & \text{if } A \in \Lambda_\beta \backslash \Lambda_\alpha, \\ +\infty & \text{otherwise,} \end{cases}$$

and

$$V(A) = \begin{cases} \frac{1}{\alpha} \left|A^{(1)}\right|^2 & \text{if } A \in \Lambda_\alpha, \\ \frac{1}{\beta} \left|A^{(1)}\right|^2 & \text{if } A \in \Lambda_\beta, \\ +\infty & \text{otherwise,} \end{cases}$$

where for $\delta > 0$

$$\Lambda_\delta = \left\{ A \in \mathbb{M}^{2\times 2} \ : \ \delta A^{(1)} + T A^{(2)} = 0 \right\}.$$

Here, $A^{(i)}$ stands for the i-row of the matrix A, $i = 1, 2$. Note that

$$\Lambda_\alpha \cap \Lambda_\beta = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

so there is no ambiguity in the definition of V. We must take into account that W and V are not Carathéodory functions since they take on the value $+\infty$ in a noncontinuous way; however, V is continuous where it is finite and W too, except at the origin.[1] This fact will be important in the study of relaxation in the next section.

It is easy to realize that the original optimal design problem is equivalent to the following variational problem:

$$\text{minimize} \int_\Omega W(\nabla U(x))\, dx \tag{4}$$

over the class of admissible functions

$$\mathcal{U} = \left\{ U \in H^1(\Omega; \mathbb{R}^2),\ U^{(2)} = v_0 \text{ on } \partial\Omega \right\}$$

subject to

$$\int_\Omega V(\nabla U(x))\, dx \le \gamma.$$

Due to this equivalence, the new problem does not have a solution, and therefore we are interested in characterization of minimizing sequences for it. To this end, we will analyze relaxation of this variational problem, proving, at first step, a

[1] This fact is not a difficulty due to 0-1 law proved in [3, Theorem 1.3]. See section 3 for more details.

subrelaxation result in terms of the appropriate convex envelope and Young measures generated by sequences of admissible gradients. Indeed, that convex envelope is defined in the following fashion. For a fixed $x \in \Omega$, $\rho > 0$, set

$$W^\sharp(A,\rho) = \inf\left\{ \int_{\mathbb{M}^{2\times 2}} W(F)\, d\nu(F) \ : \ \nu \in \mathcal{A}(A,\rho) \right\}, \tag{5}$$

where $\mathcal{A}(A,\rho)$ is the set of homogeneous H^1 Young measures ν such that

$$\int_{\mathbb{M}^{2\times 2}} F\, d\nu(F) = A$$

and

$$\int_{\mathbb{M}^{2\times 2}} V(F)\, d\nu(F) = \rho.$$

This will be carried out in section 2.

As a second step, being the most important part of this paper and where the greatest emphasis is placed, we focus on the explicit computation of the envelope $W^\sharp$ and the optimal microstructures, that is to say, the minimizers of (5). Section 3 is devoted to that computation, whose conclusion is stated in the following theorem.

THEOREM 1. *Let $\mathcal{B}$ be the set of pairs of matrices and real numbers (A,ρ) defined by the two inequalities*

$$\left(\frac{\beta+\alpha}{\beta-\alpha}\right)\frac{\left|\alpha A^{(1)} + TA^{(2)}\right|^2}{\beta(\det A - \alpha^2\rho)} \le 1 + \left(\frac{\beta+\alpha}{\beta-\alpha}\right)\frac{\left|\beta A^{(1)} + TA^{(2)}\right|^2}{\alpha(\det A - \beta^2\rho)} \tag{6}$$

and

$$\alpha^2 \le \frac{\det A}{\rho} \le \beta^2. \tag{7}$$

The constrained envelope $W^\sharp$ is given by

$$W^\sharp(A,\rho) = \begin{cases} \left(\dfrac{\beta+\alpha}{\beta-\alpha}\right)\dfrac{\left|\alpha A^{(1)} + TA^{(2)}\right|^2}{\beta(\det A - \alpha^2\rho)} & \text{if } (A,\rho)\in\mathcal{B}, \\ +\infty & \text{otherwise.} \end{cases}$$

Moreover, if (6) *happens with equality, there exists a unique first order laminate,*

$$\nu = (1-t_0)\delta_{A_\alpha} + t_0\delta_{A_\beta},$$

which is the optimal microstructure (that is, $W^\sharp(A,\rho) = \int_{\mathbb{M}^{2\times 2}} W(F)\, d\nu(F)$ and ν has barycenter A).

On the contrary, if (6) *holds with strict inequality, the optimal microstructures are second order laminates*

$$\nu_{i,j} = (1-\sigma_{i,j})\delta_{A_{\alpha,j}} + \sigma_{i,j}(\rho_{i,j}\delta_{A_\beta} + (1-\rho_{i,j})\delta_{\bar{A}_{\alpha,i}}), \quad i,j = 1,2,\ i \ne j,$$

where

$$\sigma_{i,j} = \frac{t_0(r_j - r_i)}{t_0(1-r_i) - r_i(1-r_j)}, \qquad \rho_{i,j} = \frac{t_0(1-r_i) - r_i(1-r_j)}{r_j - r_i}.$$

In both cases, $t_0 = \left(\frac{\beta+\alpha}{\beta-\alpha}\right)\frac{\left|\alpha A^{(1)} + TA^{(2)}\right|^2}{\beta(\det A - \alpha^2\rho)}$.

The values r_i and the rest of the matrices will be defined in section 3.

Note also that the function $W^\sharp$ is well-defined for (A,ρ) such that $\det A = \alpha^2\rho$, because in that case, $A \in \Lambda_\alpha$. In this situation, $W(A,\rho) = 0$ and the laminate is $\nu = \delta_A$. In the same way, if $A \in \Lambda_\beta$, $W(A,\rho) = 1$ and the optimal is attained at the same laminate.

Before going into relaxation a word should be said about the envelope $W^\sharp$. Due to the fact that functions W, V are not Carathéodory functions, the infimum in (5) defined over homogeneous gradient Young measures could be different from (less than or equal to) the corresponding infimum defined over gradients, so that being rigorous we should say that $W^\sharp$ is a constrained semiconvex envelope instead of a constrained quasiconvexification. Anyhow, working with $W^\sharp$ is enough to have a relaxation result in our setting, and we are even able to give a full and explicit computation of it. Therefore, with a little abuse of language, we will refer to that function as constrained quasiconvexification.

It is worth saying that the approach presented here is not exclusively two-dimensional as it could seem being based on the fact that, for any solution of (6), there exists a conjugate. In the three-dimensional case the recent paper [6] analyzes classical questions of the calculus of variations for functionals depending on gradients and curls in dimension three, and the conclusions of that paper were implemented in the analysis of the relaxation of a three-dimensional optimal design problem related to the one considered here (cf. [5]). We think that the analysis presented here could be extended to the three-dimensional case following the ideas presented in those papers.

2. Relaxation. A general analysis of relaxation of variational problems under integral constraints has been carried out in [17] where densities W and V are Carathéodory functions. That analysis has been applied in that paper to obtain a subrelaxation (in the sense that the infimum of the relaxed problem is less than or equal to the original one) for general structural design problems in which also the fact of having non-Carathédory densities happen. We could apply those results to directly obtain a subrelaxation; however, we will go into the analysis of relaxation of problem (4) in a slightly different way. We get to improve the subrelaxation result, although we do not get a complete relaxation result; we will deal with this question at the end of the section. We would like to emphasize that the more important and interesting result of this paper is Theorem 1 rather than the results shown in this section, which are mainly of a technical nature (although of considerable theoretical interest). Any reader not particularly interested in the questions analyzed here can pass over this section, keeping in mind the conclusion of Theorem 2, and read directly to section 3.

PROPOSITION 1. *The infimum*

$$m^\sharp = \inf\left\{\int_\Omega W^\sharp(\nabla U(x), t(x))\,dx :\ U \in \mathcal{U},\ \|t\|_{L^\infty(\Omega)} \le M,\ \int_\Omega t(x)\,dx \le \gamma\right\}$$

is attained.

Proof. Let us call

$$I^\sharp(U,t) = \int_\Omega W^\sharp(\nabla U(x), t(x))\,dx.$$

This functional is coercive because of any function taking values on the set where $W^\sharp$ is finite is bounded. This property was proved in [4] in the context of another optimal design problem in conductivity, and the proof uses the essential fact that Sobolev

functions with gradients taking values on the support of function $W^\sharp$ are solutions of linear elliptic PDEs with uniformly bounded coefficients.

To apply the direct method we need to prove that the functional $I^\sharp$ is also weak and lower semicontinuous. To this end, it is enough to show that $W^\sharp$ verifies the following jointly convex property:

$$W^\sharp(A,\theta) \le \frac{1}{|\Omega|}\int_\Omega W^\sharp(A+\nabla V(y),\theta+t(y))\,dy,$$

for all $(A,\theta)\in \mathbb{M}^{2\times 2}\times\mathbb{R}$, $V\in H^1_0(\Omega;\mathbb{R}^2)$ and $t\in L^\infty(\Omega)$ with vanishing mean-value (cf. [8]).

Let $V\in H^1_0(\Omega;\mathbb{R}^2)$ and $t\in L^\infty(\Omega)$ such that $\int_\Omega t(y)\,dy=0$. For a.e. $y\in\Omega$, we can find a minimizing probability measure $\nu^y\in\mathcal{A}(A+\nabla V(y),\theta+t(y))$ such that

$$W^\sharp(A+\nabla V(y),\theta+t(y)) = \int_{\mathbb{M}^{2\times 2}} W(F)\,d\nu^y(F).$$

Proving that such a minimizing measure exists, there is an elementary fact of functional analysis due to the functional to minimize linear over measures, and the set $\mathcal{A}(A+\nabla V(y),\theta+t(y))$ is closed convex and consequently compact in the weak-$\star$ topology on the Radon measures' space.

Let us prove that the family of probability measures $\nu=\{\nu^y\}_{y\in\Omega}$ is an H^1 Young measure. To this end, we use [15, Theorem 8.16], which gives the following sufficient conditions for being a gradient Young measure: in our case, $\nu=\{\nu^y\}_{y\in\Omega}$ is a H^1 Young measure if it verifies

(a) $\nabla U(y)=\int_{\mathbb{M}^{2\times 2}} F\,d\nu^y(F)$ for some $U\in H^1$;

(b) $\int_{\mathbb{M}^{2\times 2}}\varphi(F)\,d\nu^y(F)\ge\varphi(\nabla U(y))$ for any quasiconvex function φ;

(c) $\int_\Omega\int_{\mathbb{M}^{2\times 2}}|F|^2\,d\nu^y(F)\,dy<\infty$.

(a) holds by definition of the set $\mathcal{A}(A,\rho)$, actually $U=V+A\cdot y$ in this case; (b) is true because each ν^y is a Young measure, as any term of a Young measure is a homogeneous Young measure itself, and consequently it verifies Jensen's inequality for any quasiconvex function (see [15]), and (c) holds because of the facts that $\operatorname{supp}(\nu^y)\subset\Delta$, a.e. $y\in\Omega$, and that the quadratic growth conditions on the function V are finite at any time.

Finally, the average measure of ν, $\bar\nu$ is an homogeneous H^1 Young measure belonging to $\mathcal{A}(A,\theta)$ and such that

$$\begin{aligned}\frac{1}{|\Omega|}\int_\Omega W^\sharp(A+\nabla V(y),\theta+t(y))\,dy &= \int_\Omega\int_{\mathbb{M}^{2\times 2}} W(F)\,d\nu^y(F)\,dy\\ &= \int_{\mathbb{M}^{2\times 2}} W(F)\,d\bar\nu(F)\ge W^\sharp(A,\theta).\end{aligned}$$

The averaging measure procedure is a standard technique when dealing with Young measures and can be checked at [15]. □

PROPOSITION 2. *If $\mathcal{A}$ stands for the set of H^1 Young measures such that there exists $U\in H^1(\Omega;\mathbb{R}^2)$ with*

$$\begin{aligned}\nabla U(x) &= \int_{\mathbb{M}^{2\times 2}} F\,d\nu^x(F),\\ U^{(2)} &= v_0 \quad \text{on } \partial\Omega,\end{aligned}$$

then the infimum

$$\tilde{m} = \inf \left\{ \int_\Omega \int_{\mathbb{M}^{2\times 2}} W(F)\, d\nu^x(F)\, dx : \right.$$
$$\left. \nu = \{\nu^x\}_{x\in\Omega} \in \mathcal{A},\ \int_\Omega \int_{\mathbb{M}^{2\times 2}} V(F)\, d\nu^x(F)\, dx \le \gamma \right\}$$

is attained.

The proof is standard and is based on the minimization of a linear functional on a closed convex set; see [15]. The following is a subrelaxation result.

THEOREM 2. *Under above conditions, if we put*

$$m = \inf \left\{ \int_\Omega W(\nabla U(x))\, dx \ :\ U \in \mathcal{U},\ \int_\Omega V(\nabla U(x))\, dx \le \gamma \right\},$$

then

$$\tilde{m} \le m^\sharp \le m.$$

Proof. The inequality

$$m^\sharp \le m$$

is trivial as a consequence of the definition of the constrained semiconvex envelope $W^\sharp$. Let us prove the other inequality. Let $U \in H^1(\Omega;\mathbb{R}^2)$, $t \in L^\infty(\Omega)$ be such that

$$U^{(2)} = v_0 \quad \text{on } \partial\Omega,$$

$$\int_\Omega W^\sharp(\nabla U(x), t(x))\, dx < +\infty,$$

$$\int_\Omega t(x)\, dx \le \gamma,$$

and

$$\|t\|_{L^\infty(\Omega)} \le M.$$

As was shown in the proof of Proposition 1, for a.e. $x \in \Omega$ there exists a homogeneous H^1 Young measure $\nu^x \in \mathcal{A}(\nabla U(x), t(x))$ such that

$$W^\sharp(\nabla U(x), t(x)) = \int_{\mathbb{M}^{2\times 2}} W(F)\, d\nu^x(F)$$

and the family of probability measures $\nu = \{\nu^x\}_{x\in\Omega}$ is an H^1 Young measure. Moreover, ν verifies

$$\int_\Omega \int_{\mathbb{M}^{2\times 2}} W(F)\, d\nu^x(F)\, dx = \int_\Omega W^\sharp(\nabla U(x), t(x))\, dx$$

and

$$\int_\Omega \int_{\mathbb{M}^{2\times 2}} V(F)\, d\nu^x(F)\, dx = \int_\Omega t(x)\, dx \le \gamma.$$

This finishes the proof. □

In order to obtain a full relaxation result it would be enough to prove the equality

$$m = \tilde{m}. \tag{8}$$

This fact would be true if, given any Young measure with support contained in the set where W is finite, Λ, and verifying the integral constrained, there exists a generating sequence of gradients such that it verifies all the admissibility constraints for the problem, and, further, those gradients take values on Λ. Otherwise, it could happen that for an optimal Young measure there is no admissible generating sequence taking values on Λ and therefore the inequality $\tilde{m} \leq m$ would be strict. That is to say, there is a gap between the values of the two infima. From the point of view of applications that is not really important because it just means that small errors in the designs would improve the cost, as is extensively discussed in [16]. In our case the authors have not been able to prove equality (8). Concretely, the difficulty is the following: given a Young measure with support contained on Δ and satisfying the integral constraint, we are able to find a generating sequence for the Young measure taking values on Δ, but it is not clear how to modify a generating sequence in order to make sure that such a constraint is verified.

The following result is an improvement of the subrelaxation result in the sense pointed out above.

THEOREM 3. *For any family of measures $\nu = \{\nu^x\}_{x\in\Omega}$ belonging to $\mathcal{A}$ and such that*

$$\operatorname{supp}(\nu^x) \subset \Lambda, \quad \text{a.e. } x \in \Omega,$$

there exists a sequence of gradients $\{\nabla U_j\}$ generating ν and verifying that, for any j,

(i) $U_j \in H^1(\Omega;\mathbb{R}^2)$, $\quad U_j^{(2)} = v_0$,

(ii) $\{|\nabla U_j|^2\}$ *is equi-integrable,*

(iii) $\nabla U_j(x) \in \Lambda$, $\quad$ *a.e.* $x \in \Omega$.

Proof. Let $\nu = \{\nu^x\}_{x\in\Omega}$, satisfying the hypotheses of the theorem, and let $\{\nabla \bar{U}_j\}$ be a sequence of gradients of functions in $H^1(\Omega;\mathbb{R}^2)$ generating ν such that $\{|\nabla \bar{U}_j|^2\}$ is equi-integrable and

$$\bar{U}_j^{(2)} = v_0 \quad \text{on } \partial\Omega.$$

Let us consider the function

$$\varphi(A) = \min\{|\alpha A^{(1)} + TA^{(2)}|^2, |\beta A^{(1)} + TA^{(2)}|^2\}.$$

It is obvious that the zero set of $\varphi(\cdot)$ is Λ and consequently

$$\int_{\mathbb{M}^{2\times 2}} \varphi(F)\, d\nu^x(F) = 0, \quad \text{a.e. } x \in \Omega.$$

Due to the growth conditions verified by φ and the equi-integrability of the sequence, it is true that

$$\lim_{j\to+\infty} \int_\Omega \varphi\left(\nabla \bar{U}_j(x)\right)\, dx = 0.$$

Moreover,

$$\varphi\left(\nabla \bar{U}_j(x)\right) = |\sigma_j(x)\nabla \bar{U}_j^{(1)}(x) + T\nabla \bar{U}_j^{(2)}(x)|^2,$$

for some $\sigma_j(x)$ taking values on $\{\alpha, \beta\}$. Hence, the above limit reads as

$$\lim_{j\to\infty} \int_\Omega |\sigma_j(x)\nabla \bar{U}_j^{(1)}(x) + T\nabla \bar{U}_j^{(2)}(x)|^2 dx = 0. \tag{9}$$

Solving the Neumann boundary values problem

$$\begin{cases} -\operatorname{div}(\sigma_j(x)\nabla u(x)) = 0 & \text{in } \Omega, \\ \sigma_j \nabla u \cdot n = f & \text{on } \partial\Omega, \end{cases}$$

and denoting by $U_j^{(1)}$ its solution (unique up to an additive constant) and $U_j^{(2)}$ the corresponding stream function, it is clear that this new sequence verifies (i) and (iii). Finally, if we prove that

$$\nabla \bar{U}_j - \nabla U_j \to 0, \quad \text{strong in } L^2, \tag{10}$$

the sequence $\{\nabla U_j\}$ generates ν and will verify (ii).

Let us see (10). By ellipticity,

$$\begin{aligned} \alpha \int_\Omega |\nabla U_j^{(1)}(x) - \nabla \bar{U}_j^{(1)}(x)|^2 \, dx \\ \leq \int_\Omega \langle \sigma_j(x)(\nabla U_j^{(1)}(x) - \nabla \bar{U}_j^{(1)}(x)), \nabla U_j^{(1)}(x) - \nabla \bar{U}_j^{(1)}(x)\rangle \, dx \end{aligned}$$

adding and subtracting $T\nabla U_j^{(2)}(x)$ and $T\nabla \bar{U}_j^{(2)}(x)$ in the first factor of the scalar product,

$$\begin{aligned} = \int_\Omega \langle -(\sigma_j(x)\nabla \bar{U}_j^{(1)}(x) + T\nabla \bar{U}_j^{(2)}(x)), \nabla U_j^{(1)}(x) - \nabla \bar{U}_j^{(1)}(x)\rangle \, dx \\ + \int_\Omega \langle \sigma_j(x)\nabla U_j^{(1)}(x) + T\nabla U_j^{(2)}(x), \nabla U_j^{(1)}(x) - \nabla \bar{U}_j^{(1)}(x)\rangle \, dx \\ + \int_\Omega \langle T\nabla \bar{U}_j^{(2)}(x) - T\nabla U_j^{(2)}(x), \nabla U_j^{(1)}(x) - \nabla \bar{U}_j^{(1)}(x)\rangle \, dx. \end{aligned}$$

The second term vanishes because of $\sigma_j(x)\nabla U_j^{(1)}(x) + T\nabla U_j^{(2)}(x) = 0$ a.e. $x \in \Omega$. The third term is equal, taking into account that the transpose matrix of T is $-T$, so this integral results in

$$\int_\Omega \langle \nabla \bar{U}_j^{(2)}(x) - \nabla U_j^{(2)}(x), T\nabla \bar{U}_j^{(1)}(x) - T\nabla U_j^{(1)}(x)\rangle \, dx$$

and, integrating by parts, it is equal to

$$\begin{aligned} -\int_\Omega \operatorname{div}(T\nabla \bar{U}_j^{(1)}(x) - T\nabla U_j^{(1)}(x))(\bar{U}_j^{(2)}(x) - U_j^{(2)}(x)) \, dx \\ + \int_{\partial\Omega} (\bar{U}_j^{(2)}(x) - U_j^{(2)}(x))((T\nabla \bar{U}_j^{(1)}(x) - T\nabla U_j^{(1)}(x)) \cdot n) \, dS, \end{aligned}$$

but the function $T\nabla\bar{U}_j^{(1)}(x) - T\nabla U_j^{(1)}(x)$ is obviously divergence-free, and the functions $U^{(2)}$ and $\bar{U}^{(2)}$ satisfy the same Dirichlet boundary condition. Therefore the third term also vanishes. Then applying Hölder's inequality,

$$\alpha \int_\Omega |\nabla U_j^{(1)}(x) - \nabla \bar{U}_j^{(1)}(x)|^2\,dx \le \lim_{j\to\infty} \int_\Omega |\sigma_j(x)\nabla \bar{U}_j^{(1)}(x) + T\nabla \bar{U}_j^{(2)}(x)|^2\,dx$$

and

$$\int_\Omega |\nabla U_j^{(1)}(x) - \nabla \bar{U}_j^{(1)}(x)|^2\,dx \to 0$$

by (9). For the second components a similar argument works. □

3. Proof of Theorem 1. This section is devoted to the computation of the constrained envelope defined by

$$W^\sharp(A,\rho) = \inf\left\{ \int_{\mathbb{M}} W(F)\,d\nu(F) : \nu \in \mathcal{A}(A,\rho)\right\},$$

where

$$\mathcal{A}(A,\rho) = \left\{ \nu \text{ homog. } H^1 \text{ Young meas.}, \int_{\mathbb{M}} F\,d\nu(F) = A,\ \int_{\mathbb{M}} V(F)\,d\nu(F) = \rho\right\}.$$

The proof follows along the lines of the computations in [18]. For the sake of clarity, we divide the proof into various steps.

Step 1. Let us consider ν a homogeneous H^1 Young measure with barycenter A. To avoid singular cases we first assume $A \notin \Lambda$.

By definition of W and V, we can restrict our attention to all admissible measures ν with support in Λ. That is,

$$\nu = (1-t)\nu_\alpha + t\nu_\beta, \quad t\in(0,1), \quad \operatorname{supp}(\nu_\alpha)\subset\Lambda_\alpha, \quad \operatorname{supp}(\nu_\beta)\subset\Lambda_\beta.$$

This decomposition would not be well-defined when the null matrix belongs to $\operatorname{supp}(\nu)$. However this situation cannot happen, as was proved in [3, Theorem 1.3]. There, it was shown that if null matrix belongs to $\operatorname{supp}(\nu)$, then $\nu = \delta_0$, and consequently it does not have barycenter $A \neq 0$.[2]

Let us start studying some properties of first and second moments of such admissible measures. Considering the respective first moments of ν_α and ν_β,

$$A_\alpha = \int_{\Lambda_\alpha} F\,d\nu_\alpha(F), \quad A_\beta = \int_{\Lambda_\beta} F\,d\nu_\beta(F).$$

It is clear that $A = (1-t)A_\alpha + tA_\beta$, with $A_\alpha \in \Lambda_\alpha$ and $A_\beta \in \Lambda_\beta$. Then, using the definition of Λ_δ, we can write

$$A_\alpha = \begin{pmatrix} z \\ \alpha T z\end{pmatrix}, \quad A_\beta = \begin{pmatrix} w \\ \beta T w\end{pmatrix}, \tag{11}$$

and therefore, it is an easy computation to obtain

$$z = \frac{1}{(1-t)(\beta-\alpha)}(\beta A^{(1)} + TA^{(2)}), \quad w = \frac{-1}{t(\beta-\alpha)}(\alpha A^{(1)} + TA^{(2)}). \tag{12}$$

[2] In the same way, gradients of admissible functions cannot take zero value.

Now, let us define the second moments

$$x_\alpha = \int_{\Lambda_\alpha} |F^{(1)}|^2 \, d\nu_\alpha(F), \quad x_\beta = \int_{\Lambda_\beta} |F^{(1)}|^2 \, d\nu_\beta(F).$$

From Jensen's inequality follows

$$\left.\begin{aligned} x_\alpha = \int_{\Lambda_\alpha} |F^{(1)}|^2 \, d\nu_\alpha(F) \geq \left| \int_{\Lambda_\alpha} F^{(1)} \, d\nu_\alpha(F) \right|^2 = |A_\alpha^{(1)}|^2 = |z|^2, \\ \text{and similarly} \quad x_\beta \geq |w|^2. \end{aligned}\right\} \tag{13}$$

On the other hand, using the weak continuity of the determinant and the fact that ν can be generated by a sequence satisfying the hypotheses of Theorem 3, it holds that

$$\det A = \int_{\mathbb{M}} \det F \, d\nu(F) = (1-t) \int_{\Lambda_\alpha} \det F \, d\nu_\alpha(F) + t \int_{\Lambda_\beta} \det F \, d\nu_\beta(F). \tag{14}$$

And now, taking into account that if $F \in \Lambda_\alpha$, then $\det F = \alpha |F^{(1)}|^2$, and $F \in \Lambda_\beta$ implies $\det F = \beta |F^{(1)}|^2$, it is obtained that

$$\det A = (1-t)\alpha \int_{\Lambda_\alpha} |F^{(1)}|^2 \, d\nu_\alpha(F) + t\beta \int_{\Lambda_\beta} |F^{(1)}|^2 \, d\nu_\beta(F).$$

So (14) reads as

$$\det A = (1-t)\alpha x_\alpha + t\beta x_\beta. \tag{15}$$

We now consider the integral constraint. By definition of V we have

$$\begin{aligned} \rho &= \int_{\mathbb{M}} V(F) \, d\nu(F) \\ &= (1-t) \int_{\Lambda_\alpha} \frac{1}{\alpha} \left| F^{(1)} \right|^2 \, d\nu_\alpha(F) + t \int_{\Lambda_\beta} \frac{1}{\beta} \left| F^{(1)} \right|^2 \, d\nu_\beta(F) \\ &= \frac{(1-t)}{\alpha} x_\alpha + \frac{t}{\beta} x_\beta. \end{aligned} \tag{16}$$

Therefore, from (15) and (16), the second moments of all admissible measures $\nu \in \mathcal{A}(A, \rho)$ such that $\nu = (1-t)\nu_\alpha + t\nu_\beta$, $t \in (0,1)$, have to verify

$$x_\alpha = \frac{\alpha(\det A - \beta^2 \rho)}{(1-t)(\alpha^2 - \beta^2)}, \qquad x_\beta = \frac{\beta(\det A - \alpha^2 \rho)}{t(\beta^2 - \alpha^2)}. \tag{17}$$

Step 2. Once we have explicit expressions of x_α and x_β, (13) reads as

$$\begin{aligned} \frac{\alpha(\det A - \beta^2 \rho)}{(1-t)(\alpha^2 - \beta^2)} &\geq \frac{1}{(1-t)^2(\beta - \alpha)^2} |\beta A^{(1)} + T A^{(2)}|^2, \\ \frac{\beta(\det A - \alpha^2 \rho)}{t(\beta^2 - \alpha^2)} &\geq \frac{1}{t^2(\beta - \alpha)^2} |\alpha A^{(1)} + T A^{(2)}|^2. \end{aligned} \tag{18}$$

First, note that $\det A - \beta^2\rho \leq 0$ and $\det A - \alpha^2\rho \geq 0$ (otherwise, first and second inequalities have no sense, respectively) and consequently

$$\alpha^2 \leq \frac{\det A}{\rho} \leq \beta^2. \tag{19}$$

On the other hand, if $\det A = \beta^2\rho$ or $\det A = \alpha^2\rho$, then $x_\alpha = 0$ or $x_\beta = 0$, and therefore $A \in \Lambda_\beta$ or $A \in \Lambda_\alpha$, respectively, but we have previously assumed that $A \notin \Lambda$. Then, for $A \notin \Lambda$ we can assume strict inequality in (19).

Therefore, making some easy computations, the first inequality is written as

$$t \leq 1 + \left(\frac{\beta+\alpha}{\beta-\alpha}\right) \frac{\left|\beta A^{(1)} + TA^{(2)}\right|^2}{\alpha(\det A - \beta^2\rho)},$$

and the second one is

$$t \geq \left(\frac{\beta+\alpha}{\beta-\alpha}\right) \frac{\left|\alpha A^{(1)} + TA^{(2)}\right|^2}{\beta(\det A - \alpha^2\rho)},$$

so that

$$\left(\frac{\beta+\alpha}{\beta-\alpha}\right) \frac{\left|\alpha A^{(1)} + TA^{(2)}\right|^2}{\beta(\det A - \alpha^2\rho)} \leq t \leq 1 + \left(\frac{\beta+\alpha}{\beta-\alpha}\right) \frac{\left|\beta A^{(1)} + TA^{(2)}\right|^2}{\alpha(\det A - \beta^2\rho)}, \tag{20}$$

which, in this particular case, implies

$$\left(\frac{\beta+\alpha}{\beta-\alpha}\right) \frac{\left|\alpha A^{(1)} + TA^{(2)}\right|^2}{\beta(\det A - \alpha^2\rho)} \leq 1 + \left(\frac{\beta+\alpha}{\beta-\alpha}\right) \frac{\left|\beta A^{(1)} + TA^{(2)}\right|^2}{\alpha(\det A - \beta^2\rho)}. \tag{21}$$

As a consequence, for all admissible $\nu = (1-t)\nu_\alpha + t\nu_\beta \in \mathcal{A}(A,\rho)$, (A,ρ) has to verify (19) and (21), which are the inequalities given in the statement of Theorem 1.

Step 3. Obtaining the value of $W^\sharp$ is a direct consequence of the definition of W and (20). If (A,ρ) satisfies (19) and (21),

$$\begin{aligned} W^\sharp(A,\rho) &= \inf\left\{\int_{\mathbb{M}} W(F)\,d\nu(F) : \nu \in \mathcal{A}(A,\rho)\right\} \\ &= \inf\left\{(1-t)\int_{\Lambda_\alpha} W(F)\,d\nu_\alpha(F) + t\int_{\Lambda_\beta} W(F)\,d\nu_\beta(F) : \nu \in \mathcal{A}(A,\rho)\right\} \\ &= \inf\{t : \nu \in \mathcal{A}(A,\rho)\} = \left(\frac{\beta+\alpha}{\beta-\alpha}\right) \frac{\left|\alpha A^{(1)} + TA^{(2)}\right|^2}{\beta(\det A - \alpha^2\rho)} \end{aligned}$$

for $A \notin \Lambda$.

Step 4. For the construction of the laminates, we will follow the idea presented in [18, 19]. First, we observe that for

$$t_0 = \left(\frac{\beta+\alpha}{\beta-\alpha}\right) \frac{\left|\alpha A^{(1)} + TA^{(2)}\right|^2}{\beta(\det A - \alpha^2\rho)}, \tag{22}$$

substituting in (12) and (17), we see that

$$x_\beta = \frac{\beta^2(\det A - \alpha^2\rho)^2}{|\alpha A^{(1)} + TA^{(2)}|^2(\beta+\alpha)^2} = |w|^2,$$

so that

$$\int_{\Lambda_\beta} |F^{(1)}|^2 \, d\nu_\beta(F) = |w|^2 = |A_\beta^{(1)}|^2,$$

and due to the strict convexity of the integrand $\nu_\beta = \delta_{A_\beta}$.

On the other hand, if we determine the rank-one directions going through A with extreme points on Λ_α and Λ_β, we must look for $A^\alpha \in \Lambda_\alpha$, $A^\beta \in \Lambda_\beta$, and $r \in [0,1]$ such that

$$A = (1-r)A^\alpha + rA^\beta \quad \text{and} \quad \det(A^\alpha - A^\beta) = 0. \tag{23}$$

After some manipulations, this condition implies that A and r have to satisfy

$$\begin{aligned} \frac{\alpha}{(1-r)^2(\beta-\alpha)^2}\left|\beta A^{(1)} + TA^{(2)}\right|^2 + \frac{\beta}{r^2(\beta-\alpha)^2}\left|\alpha A^{(1)} + TA^{(2)}\right|^2 \\ + \frac{\alpha+\beta}{r(1-r)(\beta-\alpha)^2}(\alpha A^{(1)} + TA^{(2)}) \cdot (\beta A^{(1)} + TA^{(2)}) = 0. \end{aligned} \tag{24}$$

Therefore, there will be such rank-one directions if A is such that the above expression has solutions for $r \in [0,1]$. We assert that if $(A,\rho) \in \mathcal{B}$, then there exists $r \in [0,1]$ satisfying (24).

Let us prove our claim. It is clear that condition $(A,\rho) \in \mathcal{B}$ implies (18). That is to say, $x_\alpha \geq |z|^2$ and $x_\beta \geq |w|^2$. Then, from (15) we deduce that

$$\det A \geq (1-t)\alpha|z|^2 + t\beta|w|^2,$$

which can be written as

$$\begin{aligned} t^2(\beta-\alpha)^2 \det A + t(\alpha|\beta A^{(1)} - TA^{(2)}|^2 - \beta|\alpha A^{(1)} + TA^{(2)}|^2 - (\beta-\alpha)^2 \det A) \\ +\beta|\alpha A^{(1)} + TA^{(2)}|^2 \leq 0. \end{aligned}$$

Now, if $P_A(t)$ stands for the second degree polynomial on the left-hand side, then $(A,\rho) \in \mathcal{B}$ implies $P_A(t)$ has its roots in $[0,1]$ (note that P_A is an upward parabola where $P_A(0)$ and $P_A(1)$ are positive and $t \in (0,1)$). Making some easy computations, it turns out that (24) can be written as $P_A(r) = 0$. That is, $(A,\rho) \in \mathcal{B}$ implies that there exist as many rank-one directions going through A with extreme points in Λ_α and Λ_β as roots in equation $P_A(r) = 0$.

Namely, if (21) happens with equality, then it is easy to realize that $x_\alpha = |z|^2$ and $x_\beta = |w|^2$. Therefore the laminate has to be $\nu = (1-t_0)\delta_{A_\alpha} + t_0\delta_{A_\beta}$ for t_0 given in (22). Note that A_α and A_β are only dependent on t_0 and A, so the laminate is unique. That is, $P_A(r)$ has only one solution.

On the contrary, if (21) is a strict inequality, then there are two solutions of $P_A(r) = 0$, denoted by r_i, $i = 1,2$, and therefore two rank-one directions going through A with extreme points in Λ_α and Λ_β. If we denote by $A_{\alpha,i}$ and $A_{\beta,i}$, $i = 1,2$, the extreme points in Λ_α and Λ_β, respectively, we can construct second order laminates in the following way.

Let us consider $\bar{A}_{\alpha,i} = A_\beta + \zeta_i(A_{\alpha,i} - A_{\beta,i})$ such that $\bar{A}_{\alpha,i} \in \Lambda_\alpha$. We have to adjust the parameter ζ_i conveniently. Making some easy computations,

$$A_\beta + \zeta_i(A_{\alpha,i} - A_{\beta,i}) = \begin{pmatrix} w + \zeta(z_i - w_i) \\ \alpha T\left(\frac{\beta}{\alpha}w + \frac{\zeta_i}{\alpha}(\alpha z_i - \beta w_i)\right) \end{pmatrix},$$

where z_i and w_i are the corresponding vectors for $A_{\alpha,i}$ and $A_{\beta,i}$ in the same way as (12).

This matrix will be in Λ_α if and only if

$$w + \zeta_i(z_i - w_i) = \left(\frac{\beta}{\alpha} w + \frac{\zeta_i}{\alpha}(\alpha z_i - \beta w_i)\right).$$

This implies $w - \zeta_i w_i = 0$, and thus

$$\zeta_i = \frac{r_i}{t_0}.$$

Then,

$$\bar{A}_{\alpha,i} = \begin{pmatrix} \bar{z}_i \\ \alpha T \bar{z}_i \end{pmatrix} \text{ for } \bar{z}_i = \frac{r_i}{t_0(1-r_i)(\beta-\alpha)}(\beta A^{(1)} + T A^{(2)})$$

and t given in (22).

In this situation, the laminates

$$\nu_{i,j} = (1-\sigma_{i,j})\delta_{A_{\alpha,j}} + \sigma_{i,j}(\rho_{i,j}\delta_{A_\beta} + (1-\rho_{i,j})\delta_{\bar{A}_{\alpha,i}}), \quad i,j = 1,2,\ i \neq j,$$

where

$$\sigma_{i,j} = \frac{t_0(r_j - r_i)}{t_0(1-r_i) - r_i(1-r_j)}, \qquad \rho_{i,j} = \frac{t_0(1-r_i) - r_i(1-r_j)}{r_j - r_i},$$

where t_0 is given in (22), are optimal microstructures,[3] and obviously

$$\int_{\mathbb{M}} W(F)\, d\nu_{i,j}(F) = \left(\frac{\beta+\alpha}{\beta-\alpha}\right) \frac{\left|\alpha A^{(1)} + T A^{(2)}\right|^2}{\beta(\det A - \alpha^2\rho)}, \quad i,j = 1,2,\ \ i \neq j.$$

Step 5. Finally, let us see what happens if $A \in \Lambda$. For instance, we assume $A \in \Lambda_\alpha$. Then $A_\alpha = A$, $A_\beta = 0$ and (15), (16) imply $\det A = \alpha^2\rho$, and as a consequence $t_0 = 0$. That is, $W^\sharp(A,\rho) = 0$. In the same way, if $A \in \Lambda_\beta$, $\det A = \beta^2\rho$ and $W^\sharp(A,\rho) = 1$. This finishes the proof.

Acknowledgements. The authors are grateful to Daniel Faraco for suggesting a more elegant proof of Theorem 2. The review work and suggestions of two anonymous referees is also greatly appreciated.

REFERENCES

[1] G. Allaire, *Shape Optimization by the Homogenization Method*, Springer-Verlag, New York, 2002.

[2] E. Aranda and P. Pedregal, *Constrained envelope for a general class of design problems*, Discrete Contin. Dyn. Syst. A, Suppl. (2003), pp. 30–41.

[3] K. Astala and D. Faraco, *Quasiregular mappings and Young measures*, Proc. Roy. Soc. Edinburgh Sect. A, 132 (2002), pp. 1045–1056.

[4] J. C. Bellido and P. Pedregal, *Explicit quasiconvexification for some cost functionals depending on derivatives of the state in optimal design*, Discrete Contin. Dyn. Syst. A, 8 (2002), pp. 967–982.

[3]Note that the optimal microstructure has sense with respect to the original problem. The optimum attains in a laminate with as little mass on Λ_β as possible (that is, with less account of expensive material).

[5] J. C. Bellido, *Explicit computation of the relaxed density coming from a three-dimensional optimal design problem*, Nonlinear Anal., 52 (2003), pp. 1709–1726.
[6] J. C. Bellido and P. Pedregal, *Optimal design via variational principles: The three-dimensional case*, J. Math. Anal. Appl., 287 (2003), pp. 157–176.
[7] A. Cherkaev, *Variational Methods for Structural Optimization*, Springer-Verlag, New York, 2000.
[8] I. Fonseca, D. Kinderlehrer, and P. Pedregal, *Energy functionals depending on elastic strain and chemical composition*, Calc. Var. Partial Differential Equations, 2 (1994), pp. 283–313.
[9] V. Girault and P. A. Raviart, *Finite Elements Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin, 1986.
[10] F. V. Kohn and G. Strang, *Optimal design and relaxation of variational problems,* I, Comm. Pure Appl. Math., 39 (1986), pp. 113–137.
[11] F. V. Kohn and G. Strang, *Optimal design and relaxation of variational problems,* II, Comm. Pure Appl. Math., 39 (1986), pp. 139–182.
[12] F. V. Kohn and G. Strang, *Optimal design and relaxation of variational problems,* III, Comm. Pure Appl. Math., 39 (1986), pp. 353–377.
[13] G. Milton, *Theory of Composites*, Cambridge University Press, Cambridge, UK, 2002.
[14] F. Murat, *Contre-exemples pour divers problèmes ou le controle intervient dans les coefficients*, Ann. Mat. Pura Appl., 112 (1977), pp. 49–68.
[15] P. Pedregal, *Parametrized Measures and Variational Principles*, Progress in Nonlinear Partial Differential Equations, Birkhäuser-Verlag, Basel, Switzerland, 1997.
[16] P. Pedregal, *Optimal design and constrained quasiconvexity*, SIAM J. Math. Anal., 32 (2000), pp. 854–869.
[17] P. Pedregal, *Constrained quasiconvexity and structural optimization*, Arch. Ration. Mech. Anal., 154 (2000), pp. 325–342.
[18] P. Pedregal, *Constrained quasiconvexification of the square of the gradient of the state in optimal design*, Quart. Appl. Math., 62 (2004), pp. 459–470.
[19] P. Pedregal, *Fully explicit quasiconvexification of the mean-square deviation of the gradient of the state in optimal design*, Electron. Res. Announc. Amer. Math. Soc., 7 (2001), pp. 72–78.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 1, pp. 210–239

A CONVERSE LYAPUNOV THEOREM FOR LINEAR PARAMETER-VARYING AND LINEAR SWITCHING SYSTEMS*

FABIAN WIRTH†

Abstract. We study families of linear time-varying systems, where time variations have to satisfy restrictions on the dwell time, that is, on the minimum distance between discontinuities, as well as on the derivative in between discontinuities. Such classes of systems may be formulated as linear flows on vector bundles. The main objective of this paper is to construct parameter-dependent Lyapunov functions, which characterize the exponential growth rate. This is possible in the generic irreducible case. As an application the Gelfand formula is generalized to the class of systems studied here. In other words, the maximal exponential growth rate may be approximated by only considering the periodic systems in the family of time-varying systems. A perspective on the question of continuous dependence of the exponential growth rate on the data is given.

Key words. converse Lyapunov theorem, linear parameter-varying systems, linear switching systems, linear flows on vector bundles, Gelfand formula, periodic systems

AMS subject classifications. 34D08, 37B25, 37B55, 93D09, 93D30

DOI. 10.1137/S0363012903434790

1. Introduction.

In this paper we consider linear time-varying systems of the form

$$\dot{x}(t) = A(t)x(t), \tag{1.1}$$

where $A : \mathbb{R} \to \mathcal{M}$ is a measurable map, and $\mathcal{M}$ is a compact set of real or complex matrices of a given dimension. We are interested in the exponential growth rate of not one individual system but a set of systems described by a subset $\mathcal{A} \subset L^\infty(\mathbb{R}, \mathcal{M})$. The stability and spectral properties of such systems have been actively investigated over the past two decades.

In this paper we present a framework covering many of the systems studied in the areas of linear parameter-varying (LPV) systems with constraints on the derivative and of linear switching systems with dwell times. We introduce a certain class of linear time-varying systems that allows for (i) bounds on the minimal time between discontinuities and (ii) bounds on the derivative of parameter variations between discontinuities.

The main contribution of the present paper lies in the construction of parameterized Lyapunov functions that characterize the exponential growth rate of the system under consideration. The construction is possible in the generic irreducible case, in which the system leaves no nontrivial subspace invariant. For each parameter the corresponding Lyapunov function is a norm. One of the features of the Lyapunov functions is that for any solution the corresponding infinitesimal decay is upper bounded by the maximal growth rate. Also the exponential growth rate can be realized instantaneously from every initial condition of the state and the parameter. Under mild

*Received by the editors September 15, 2003; accepted for publication (in revised form) October 12, 2004; published electronically July 18, 2005.
http://www.siam.org/journals/sicon/44-1/43479.html

†Hamilton Institute, NUI Maynooth, Maynooth, Co. Kildare, Ireland (Fabian.Wirth@nuim.ie). This work was completed while the author was on leave from the Center for Technomathematics, University of Bremen, Germany. Support from Science Foundation Ireland under grant 00/PI.1/C067 is gratefully acknowledged.

assumptions the Lyapunov functions are Lipschitz continuous in both the state and the parameter. As in [25], it would be possible to consider smooth approximations to obtain differentiable Lyapunov functions, which still yield a decay arbitrarily close to the growth rate. This problem is not pursued here, as the method is well described in the literature; see [9, 25, 32].

Using the existence of Lyapunov functions, we give a fairly simple proof of a version of the Gelfand formula. By this result the exponential growth rate can be approximated to arbitrary precision using periodic parameter variations. This result would appear to be new for LPV systems with bounds on the derivative as well as for linear switching systems with dwell time.

The results obtained in this paper are generalizations of [34] on the exponential growth rate of families of time-varying systems with measurable parameter variations. The proofs are often similar, but more preparation has to be undertaken to proceed to the actual results. In [34] it is also shown how the same ideas yield results on the (Lipschitz) continuity of the growth rate as a function of the data. We briefly comment on this problem here; see [36] for further details.

It is interesting to note that the subject of exponential growth of certain sets of linear time-varying systems has been taken up by different communities over time. We will not try to give an overview of the relevant literature, but an effort has been made to at least cite landmarks in each of the areas, and the reader is invited to look for further references in these papers. The literature related to this problem is not readily accessible because the terms *families of linear time-varying systems*, *linear differential inclusions*, *LPV systems*, *linear flows on vector bundles*, and *linear switching systems* are different names for very similar situations. All these names cover at least the case in (1.1), where we consider $\mathcal{A} = L^\infty(\mathbb{R}, \mathcal{M})$.

Probably the oldest exponent of this area is formed by the theory of linear flows on vector bundles, which has been developed in the dynamical systems community at least since the 1970s. For a recent account of the state of the art insofar as it is related to control theory, we refer to [13]. In fact, in this book it is shown that a good deal of work is necessary before system (1.1) with $\mathcal{A} = L^\infty(\mathbb{R}, \mathcal{M})$ can be justifiably interpreted as a linear flow on a vector bundle. Another good general reference in this area is [8]. The problem of exponential growth rates is treated in [18].

Papers concerned with linear differential inclusions and families of time-varying systems often treat the case when $\mathcal{A} = L^\infty(\mathbb{R}, \mathcal{M})$. In this area a detailed description of spectral concepts is available (see [11, 12, 13]) and a good Lyapunov theory has been developed [5, 26, 34]. Furthermore, it is known that the uniform exponential growth rate can be approximated arbitrarily well by periodic systems. This result is sometimes called the Gelfand formula in reminiscence of the characterization of the spectral radius of bounded linear operators as the infimum of norms of its powers; see [7, 11, 16].

The control and robustness analysis of LPV systems have been actively investigated during the last decade. In particular, parameter-dependent quadratic Lyapunov functions for such systems are frequently discussed in the literature, and many sufficient results for the existence of Lyapunov functions have been obtained in the framework of linear matrix inequalities (LMIs); see [2, 3, 4, 6, 17, 21, 30, 31]. In some papers, however, the interesting added feature is that time variations are restricted by requiring certain bounds on the derivative of the parameter variations as well; see, e.g., [2]. Also for this case sufficient conditions for the existence of Lyapunov functions are available in terms of LMIs. It is interesting to note that the parameter variations

in this case may be interpreted as a solution set to a differential inclusion so that the results in [32] can be interpreted in such a manner as to yield a converse Lyapunov theorem also in this case; see Remark 3.3(i). A preliminary version of the present paper treats exclusively the case of parameter variations without discontinuities [35].

To complete the enumeration of different concepts we have to mention the term *linear switching system*, which is to be found most often in engineering literature. For an overview and much of the related literature we refer to [15, 23, 24]. For instance, the paper [1] analyzes conditions for exponential stability and gives a complete solution to the question of which systems stability can be determined based on knowledge of the Lie algebra generated by the systems matrices. While is often assumed in this area that the set of matrices $\mathcal{M}$ is a finite set, this does not really change the overall problem; as for inclusions at least, the exponential growth rate defined by $\mathcal{M}$ and its convex hull is the same.[1]

However, also in the analysis of linear switching systems a certain twist has been added, which consists of a condition on the minimal time that has to elapse between two discontinuities of the switching signal. This minimal time is called the *dwell time*. This approach derives its motivation in part from adaptive control and has been discussed in [27, 28]. Sufficient conditions for the existence of Lyapunov functions in terms of LMIs are available; see, e.g., [20].

We proceed as follows. In section 2 we introduce the exponential growth rate under the (essential) assumption of shift-invariance. This is one of the primary interests in this paper. The precise definition of the class of systems studied in the paper is given in section 3 introducing parameter variations defined by a value set, a set of admissible derivatives, and a dwell time.

One of our initial results will be that each system in this class defines a linear flow on a vector bundle. This concept from dynamical systems theory treats the following situation: Given a compact metric space M and a vector space $\mathbb{K}^n$, we consider a continuous dynamical system

$$\Phi : \mathbb{R} \times M \times \mathbb{K}^n \to M \times \mathbb{K}^n,$$

where each time-t map $\Phi_t : M \times \mathbb{K}^n \to M \times \mathbb{K}^n$ can be represented in the form $\Phi_t = (\Phi_t^1, \Phi_t^2)$ such that $\Phi_t^1 : M \to M$ is continuous and $\Phi_t^2 : M \times \mathbb{K}^n \to \mathbb{K}^n$ is a linear map in the second component. (Here we have described only *trivial* vector bundles, which are all that is needed in this paper. More generally, the described situation is only valid in appropriate local coordinates.)

So in particular, any LPV system and linear switching system with dwell time can be interpreted as such a linear flow. While this result is mostly of interest for classification purposes, it has the advantage nonetheless that the general results on linear flows are available. In particular, the general theory on linear flows provides results on growth rates, fiberwise Lyapunov functions, bifurcation theory, and Hartman–Grobman-type results; see, e.g., [8, 13].

In section 4 a rather tedious analysis of the concatenation structure within the set of admissible parameter variations is undertaken, which turns out to be vital in the

[1] In the literature on switching systems it is often assumed that parameter variations have to be piecewise constant with an arbitrarily small, positive, lower bound on the distance between discontinuities. With respect to the problem treated in this paper, note that there is no difference in the exponential growth rate, whether one considers parameter variations or switching signals in $L^\infty(\mathbb{R}, \mathcal{M})$ or in the subset thereof consisting of piecewise continuous functions with an (arbitrarily small) lower bound on the distance between discontinuities.

subsequent construction of Lyapunov functions. In section 5 irreducibility of a system is introduced and some immediate consequences of this property are shown. The assumption of irreducibility is used in section 6 to construct parameter-dependent Lyapunov norms that characterize the exponential growth rate. We particularly discuss the case of linear switching systems with dwell time, for which an easy interpretation is available. Finally, in section 7 the Gelfand formula is proved, and we comment on the question of continuous dependence on the systems parameters in section 8. The paper concludes with some final comments in section 9.

Finally, we would like to warn the reader that our use of the term *Lyapunov function* is not quite standard. It will be used to denote functions that characterize the exponential growth rate of the system if evaluated along trajectories. Now if the system is stable, then this will give the usual decrease condition. However, if the system is not exponentially stable, then we still speak of a Lyapunov function because of the characterization of the growth rate.

2. Families of linear time-varying systems. Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$ denote the real or the complex field. In this paper we study families of continuous time LPV systems in $\mathbb{K}^n$ that are given in the form of linear systems subject to (time-varying) variations of certain parameters entering the equation. The parameter space Θ is taken to be a compact subset of $\mathbb{K}^m$, and the map $A : \Theta \to \mathbb{K}^{n\times n}$ that associates a matrix to a given parameter is assumed to be continuous. Parameter variations are always taken to be elements of $L^\infty(\mathbb{R}, \Theta)$. Every such parameter variation $\theta(\cdot)$ induces a time-varying linear system of the form

$$\dot{x}(t) = A(\theta(t))x(t), \quad t \in \mathbb{R}. \tag{2.1}$$

The corresponding evolution operator is denoted by $\Phi_\theta(t, s),\ t, s \in \mathbb{R}$.

The main object of this paper is to discuss families of linear time-varying systems defined by a set of admissible parameter variations $\mathcal{U} \subset L^\infty(\mathbb{R}, \Theta)$. An important property of these sets is the following.

Definition 2.1. *A set $\mathcal{U} \subset L^\infty(\mathbb{R}, \Theta)$ is called* shift-invariant *if for all $u \in \mathcal{U}$ and all $t \in \mathbb{R}$ the function $w(\cdot) := u(t + \cdot)$, defined by $w(s) = u(t + s)$, is an element of $\mathcal{U}$.*

We now define the primary interest in this paper, which is the (uniform) exponential growth rate associated with system (2.1). Given the map $A : \Theta \to \mathbb{K}^{n\times n}$ and the set of admissible parameter variations $\mathcal{U} \subset L^\infty(\mathbb{R}, \Theta)$, define for $t \geq 0$ the sets of finite time evolution operators

$$\mathcal{S}_t(A,\mathcal{U}) := \{\, \Phi_u(t, 0) \,|\, u \in \mathcal{U}\}, \quad \mathcal{S}(A,\mathcal{U}) := \bigcup_{t\geq 0} \mathcal{S}_t(A,\mathcal{U}).$$

We now introduce for $t > 0$ finite time growth constants given by

$$\widehat{\rho}_t(A,\mathcal{U}) := \sup\left\{ \frac{1}{t} \log \|S\| \;|\; S \in \mathcal{S}_t(A,\mathcal{U}) \right\}.$$

It is easy to see that under the assumption of shift-invariance of $\mathcal{U}$, the function $t \longmapsto t\widehat{\rho}_t(A,\mathcal{U})$ is subadditive. Using a folklore result (see, e.g., [22, pp. 27–28]), this implies that the following limit exists:

$$\widehat{\rho}(A,\mathcal{U}) := \lim_{t\to\infty} \widehat{\rho}_t(A,\mathcal{U}) = \inf_{t\geq 0} \widehat{\rho}_t(A,\mathcal{U}). \tag{2.2}$$

It is well known that an alternative way to describe $\hat{\rho}$ is given by

(2.3)
$$\widehat{\rho}(A,\mathcal{U}) = \inf\{\beta \in \mathbb{R} \mid \exists\, M \geq 1 \text{ such that } \|\Phi_u(t,0)\| \leq Me^{\beta t} \text{ for all } u \in \mathcal{U}, t \geq 0\}.$$

For this reason the quantity $\hat{\rho}(A,\mathcal{U})$ is called the *uniform exponential growth rate* of the family of linear time-varying systems of the form (2.1) given by $\mathcal{U}$ and A. An alternative way to define exponential growth is to employ a trajectorywise definition. In this case we define the Lyapunov exponent corresponding to an initial condition $x_0 \in \mathbb{K}^n \setminus \{0\}$ and $u \in \mathcal{U}$ by

$$\lambda(x_0,u) := \limsup_{t\to\infty} \frac{1}{t} \log \|\Phi_u(t,0)x_0\| \tag{2.4}$$

and define the exponential growth rate as $\kappa(A,\mathcal{U}) := \sup\{\lambda(x,u) \mid 0 \neq x \in \mathbb{K}^n,\, u \in \mathcal{U}\}$.

If $\mathcal{U}$ is shift-invariant and closed in the weak-* topology induced by $L^\infty(\mathbb{R},\mathbb{K}^m)$, then $\mathcal{U}$ is metrizable (recall that Θ is compact), and by [13, Lem. 4.2.4] the shift is continuous on $\mathcal{U}$ endowed with that topology. If also the map $(t,x,u) \mapsto \Phi_u(t,0)x$ is continuous jointly in all variables (which is affirmative if $u \mapsto \Phi_u(t,0)$ is uniformly continuous on compact time intervals), then the following time-t maps define a continuous dynamical system or a flow on $\mathcal{U} \times \mathbb{K}^n$:

$$(u,x) \mapsto (u(t+\cdot), \Phi_u(t,0)x)\,,$$

and in fact define a linear flow on the vector bundle $\pi : \mathcal{U} \times \mathbb{K}^n \to \mathcal{U}$. Under this assumption it follows using Fenichel's uniformity lemma that $\kappa(A,\mathcal{U}) = \hat{\rho}(A,\mathcal{U})$; see [13, Prop. 5.4.15].

The outlined setup works if we assume that the set $\mathcal{U}$ is convex and that the function A is affine in θ; see [13, Chap. 4]. These assumptions, however, are somewhat restrictive. In the following section, it is shown that LPV systems and linear switching systems with dwell times may be formulated as linear flows on vector bundles.

As it is our aim to construct a certain class of parameter-dependent Lyapunov functions, it should be noted that a general theory of quadratic Lyapunov functions for linear flows on vector bundles exists; see [8, Chap. 3]. However, this theory works with Lyapunov functions defined individually in every fiber; in our case, individually for every $u \in \mathcal{U}$. This is too fine a point of view for the results that we want to obtain. In particular, despite some effort on the part of the author, the fine point of view has not yielded a way of proving the Gelfand formula.

One might now be tempted to take a very coarse point of view and to look for norms that are Lyapunov functions for the whole system and characterize the quantity $\hat{\rho}(A,\mathcal{U})$, as for the case of linear differential inclusions [34]. However, the following lemma shows that this is not a very fruitful enterprise.

LEMMA 2.2. *Let $\mathcal{U} \subset L^\infty(\mathbb{R},\Theta)$ be shift-invariant and assume system* (2.1) *defines a linear flow on the vector bundle $\pi : \mathcal{U} \times \mathbb{K}^n \to \mathcal{U}$. Assume that the constant functions $u \equiv \theta, \theta \in \Theta$ are contained in $\mathcal{U}$. If there is a norm v on $\mathbb{K}^n$, such that for all $x \in \mathbb{K}^n$, $u \in \mathcal{U}$ and the corresponding evolution operator $\Phi_u(t,s)$ it holds that*

$$v(\Phi_u(t,0)x) \leq e^{\hat{\rho}(A,\mathcal{U})t}\, v(x) \quad \forall t \geq 0, \tag{2.5}$$

then $\hat{\rho}(A,\mathcal{U}) = \rho := \max\{\lambda(x,B) \mid 0 \neq x \in \mathbb{K}^n,\, B : \mathbb{R} \to A(\Theta)$ measurable$\}$, where $\lambda(x,B)$ denotes the Lyapunov exponent corresponding to the initial condition x and B defined as in (2.4).

Proof. Clearly, we only have to show that $\hat{\rho}(A,\mathcal{U}) \geq \rho$. Let v^* be the dual norm to v; see [19]. The assumption (2.5) implies that for all $A \in A(\Theta)$, all $x \in \mathbb{K}^n$, and all $l \in \mathbb{K}^n$ with $\langle l, x\rangle = v(x) = v^*(l) = 1$ we have $\langle l, Ax\rangle \leq \hat{\rho}(A,\mathcal{U})$ by [10, Thm. 4.6.3]. This, however, implies that $\rho \leq \hat{\rho}(A,\mathcal{U})$ by [5, Thm. 5]. □

By the previous lemma, a norm satisfying (2.5) can only exist for (2.1) if the parameter-varying system realizes the exponential growth, which is obtained by allowing all measurable functions with values in $A(\Theta)$; in other words, by studying (2.1) with $\mathcal{U} = L^\infty(\mathbb{R}, \Theta)$. For general sets of parameter variations this situation is rarely encountered. For this reason we use a different approach that introduces a family of norms with an extremal property. The idea of using parameter-dependent Lyapunov functions, proposed by several authors (see, e.g., [2, 20, 21]), can be made exact in this way. That is, a family of parameter-dependent Lyapunov norms may be constructed such that the exponential growth rate of system (2.1) is the incremental growth rate with respect to this family. Note that we cannot restrict our attention to quadratic norms to perform such a construction.

Remark 2.3. The main technical problem in this paper is that $\mathcal{S}(A,\mathcal{U})$ does not naturally carry the structure of a semigroup. As an example consider the case when $\mathcal{U}$ consists of all globally Lipschitz continuous functions with values in Θ and fixed Lipschitz constant L. For $u_1, u_2 \in \mathcal{U}$ the concatenation of $u_{1|[0,t]}$ and $u_{2|(t,\infty)}$ is an admissible parameter variation if and only if $u_1(t) = u_2(t)$. This complicates matters compared to the case of linear inclusions of the form

$$\dot{x} \in \{Ax \mid A \in A(\Theta)\},$$

as studied in [5, 11, 15, 16, 34] and references therein.

3. Parameter variations. We denote the space of nonempty, compact subsets of $\mathbb{K}^m$ by $\mathcal{K}(\mathbb{K}^m)$ and the subset of nonempty, convex, compact subsets of $\mathbb{K}^m$ by $\mathrm{Co}(\mathbb{K}^m)$. Both these spaces are complete metric spaces if endowed with the Hausdorff metric defined by

$$d_H(X,Y) := \max\left\{\max_{x\in X} \operatorname{dist}(x,Y), \quad \max_{y\in Y} \operatorname{dist}(y,X)\right\}.$$

All ensuing topological statements on $\mathcal{K}(\mathbb{K}^m), \mathrm{Co}(\mathbb{K}^m)$ should be understood with respect to this metric. The convex hull of a set X is denoted by $\operatorname{conv} X$. We denote by $X - y$ the set $\{x - y \mid x \in X\}$, as usual.

In the remainder of the paper the admissible parameter variations are described by the following data: a space of parameters $\Theta \in \mathcal{K}(\mathbb{K}^m)$ given as a finite union of pairwise disjoint compact convex sets Ω_j, $j = 1, \ldots, l$; a space describing the rate of parameter variation $\Theta_1 \in \mathrm{Co}(\mathbb{K}^m)$; a *dwell time* $h \in (0, \infty]$ that describes the minimal time between discontinuities; and a continuous map $A \in C(\mathbb{K}^m, \mathbb{K}^{n\times n})$. A system is therefore now a quadruple $\Sigma = (h, \Theta, \Theta_1, A) \in (0,\infty] \times \mathcal{K}(\mathbb{K}^m) \times \mathrm{Co}(\mathbb{K}^m) \times C(\mathbb{K}^m, \mathbb{K}^{n\times n})$. We will always assume that the following assumptions are satisfied:

(A1) $h \in (0, \infty]$;
(A2) $\Theta \subset \mathbb{K}^m$ is a finite, disjoint union of sets $\Omega_j \in \mathrm{Co}(\mathbb{K}^m)$, $j \in \{1, \ldots, l\}$. If $h = \infty$, then $l = 1$, i.e., Θ is compact and convex;
(A3) $\Theta_1 \in \mathrm{Co}(\mathbb{K}^m)$;
(A4) $0 \in \Theta_1$;
(A5) $A : \Theta \to \mathbb{K}^{n\times n}$ is a continuous map.

In some cases we will need an additional assumption that allows for additional freedom in the construction of parameter variations. Recall that the relative interior of a

convex set $\mathcal{M}$, denoted by $\operatorname{ri}\mathcal{M}$, is the interior of $\mathcal{M}$ in the relative topology of the affine space generated by $\mathcal{M}$. Or in other words, the interior of $\mathcal{M}$ relative to the smallest affine space containing $\mathcal{M}$. With this we formulate the following condition.

(A6) $0 \in \operatorname{ri}\Theta_1$ and $\operatorname{span}\Theta_1 \supset \operatorname{span}(\Omega_j - \eta_j)$, $j = 1, \dots, l$, for some $\eta_j \in \Omega_j$.

In order to denote the discontinuities of parameter variations, which for the purposes of this paper are discrete sets, we consider (bounded or unbounded) index sets $\mathcal{I} \subset \mathbb{Z}$. In the following it will always be tacitly assumed that these index sets are given as the intersection of a real interval with $\mathbb{Z}$, i.e., of the form $\mathcal{I} := [a, b] \cap \mathbb{Z}$, where $a, b \in \mathbb{R} \cup \{\pm\infty\}$.

DEFINITION 3.1. *Consider a system $\Sigma = (h, \Theta, \Theta_1, A)$ satisfying* (A1)–(A5). *If $h \in (0, \infty)$, a parameter variation $\theta : \mathbb{R} \to \Theta$ is called* admissible (*with respect to Σ*) *if there is an index set $\mathcal{I}_\theta \subset \mathbb{Z}$ and times $t_k, k \in \mathcal{I}_\theta$, such that*

(i) *$h \le t_{k+1} - t_k$ for $k \in \mathcal{I}_\theta, k < \sup \mathcal{I}_\theta$,*

(ii) *for $k \in \mathcal{I}_\theta, k < \sup \mathcal{I}_\theta$ the function θ is absolutely continuous on the interval $[t_k, t_{k+1})$ and satisfies*

$$\dot\theta(t) \in \Theta_1 \quad \text{a.e.} \tag{3.1}$$

(*This condition also applies to $(-\infty, \inf \mathcal{I}_\theta)$, resp., $(\sup \mathcal{I}_\theta, \infty)$, if $\inf \mathcal{I}_\theta$, resp., $\sup \mathcal{I}_\theta$, is finite.*)

If $h = \infty$, the admissible parameter variations are given as the set of absolutely continuous functions $\theta : \mathbb{R} \to \Theta$ satisfying (3.1) *a.e. on $\mathbb{R}$.*

The set of admissible parameter variations is denoted by $\mathcal{U}$ or $\mathcal{U}(h, \Theta, \Theta_1, A)$ if dependence on the data needs to be emphasized. By convention we let $t_0 > 0$, and $t_0(u)$ denotes the smallest positive discontinuity of a parameter variation u. If there is no such discontinuity, then we set $t_0(u) := \infty$.

Remark 3.2. (i) Note that the set $\mathcal{U}$ defined above is clearly shift-invariant, but not convex in general, because convex combinations of the admissible parameter variations would in general have too many switches. Thus [13, Chap. 4] is not directly applicable to our situation. We will be able to show the necessary properties of $\mathcal{U}$ by a different strategy, which also allows us to dispense with the assumption that A is affine.

(ii) In the case $h = \infty$, it is reasonable to assume that Θ itself is convex, as parameter variations cannot leave the sets Ω_j. So with the notation of (A2) we have $\hat\rho(\infty, \Theta, \Theta_1, A) = \max_j \hat\rho(\infty, \Omega_j, \Theta_1, A)$. Hence it is sufficient to assume Θ is convex.

(iii) Assumption (A4) guarantees that the constant trajectories $u \equiv \theta$, $\theta \in \Theta$ are admissible parameter variations. This assumption is not absolutely essential but simplifies several of the ensuing statements. It would, of course, be interesting to consider systems in which only the interplay of the continuous and discontinuous behavior allows for trajectories defined on $\mathbb{R}$. An example of this kind is given by $\Theta = [0, 2], \Theta_1 = [1, 2],\ h = 1$.

(iv) If Assumption (A6) is satisfied, then for a fixed convex component Ω_j of Θ the set of derivatives Θ_1 contains a neighborhood of 0 in the linear subspace $\operatorname{span}(\Omega_j - \eta_j)$ for $\eta_j \in \Omega_j$. Thus there is a constant $c > 0$ such that for any pair $\theta, \eta \in \Omega_j$ we have $c(\theta - \eta) \in \Theta_1$. Hence for all $t > \|\theta - \eta\|/c$ there is a $u \in \mathcal{U}$ with $u(0) = \theta, u(t) = \eta$. In particular, as the Ω_j are compact, there is a constant $\bar c > 0$ such that any pair $\theta, \eta \in \Omega_j$ may be connected by an admissible parameter variation in a time equal to $\bar c$ for all $j = 1, \dots, l$.

(v) We explicitly exclude the case in which the parameter variations $\theta(\cdot)$ are arbitrary (measurable) functions taking values in Θ. This corresponds to taking

$h = 0$ in a way that can be made precise. For this case the results analogous to those obtained in this paper are already available in the literature; see [5, 13, 16, 34, 37].

Remark 3.3. (i) In the literature on LPV systems it is often assumed that the parameter variations $\theta(\cdot)$ are continuously differentiable and that the derivative satisfies certain constraints. However, it can be shown that the exponential growth rates defined by the sets

$$\{\theta : \mathbb{R} \to \Theta \mid \theta \text{ is Lipschitz continuous and } \dot\theta(t) \in \Theta_1 \quad \text{a.e.}\}$$

and

$$\{\theta : \mathbb{R} \to \Theta \mid \theta \text{ is continuously differentiable and } \dot\theta(t) \in \Theta_1 \text{ for all } t \in \mathbb{R}\}$$

are the same [37] so that our setup from the point of view of stability theory encompasses this standard case. We find the set of Lipschitz continuous parameter variations easier to handle.

In fact, LPV systems are a special case, which may be subsumed under the following more general framework; see [37]. Consider systems of the form

$$\begin{aligned} \dot x(t) &= A(\theta(t))x(t), \quad t \in \mathbb{R}, \\ \dot\theta(t) &\in \mathcal{F}(\theta(t)) \quad \text{a.e. } t \in \mathbb{R}, \end{aligned} \tag{3.2}$$

where $A : \Theta \to \mathbb{K}^{n\times n}$ is a given continuous map, $\Theta \subset \mathbb{K}^m$ is a compact, pathwise connected set, and $\mathcal{F} : \Theta \to \mathbb{K}^m$ is an upper semicontinuous set-valued map with compact values that defines a complete dynamical system on Θ. Under controllability assumptions for the parameter variations, a number of the basic results of this paper hold. Let us point out that for systems of the form (3.2) with $\hat\rho < 0$ the natural attractor to consider is $\{0\} \times \Theta$. For this case a Lyapunov function theory exists. Namely, $\hat\rho < 0$ if and only if there exists a smooth Lyapunov function on $\mathbb{K}^n \times \Theta$ for the overall system (3.2); see [9, 32]. This result is therefore also applicable to the LPV systems commonly studied in the literature. With respect to this case, the contribution of the present paper is merely a construction of a particular type of Lyapunov functions (and a proof of the Gelfand formula, of course).

(ii) A further class of families of linear time-varying systems that has attracted widespread interest recently is the so-called *linear switching systems* with dwell times, as discussed in the introduction. These systems are often given by a finite set of matrices $\Theta = \{A_1, \dots, A_k\}$ and a restriction on discontinuities by two numbers $h > 0$ and $N \in \mathbb{N}$. In our terminology a parameter variation (in this context often called a switching function) is a piecewise constant function $\theta : \mathbb{R} \to \Theta$ such that on any compact time interval $[a, b]$ the number of discontinuities is bounded from above by

$$\frac{b-a}{h} + N.$$

The class of systems we have set up encompasses the case in which $N = 1$. The Ω_j are then simply singleton sets, and Θ_1 is irrelevant. There does not seem to be a significant technical obstacle to generalizing the results of this paper to the case $N > 1$. However, the framework used here does become rather tedious for larger N so that we have chosen to restrict the system class for the time being.

4. Concatenation of admissible parameter variations. In this section the basic machinery for describing our problem is set up. We introduce sets of parameter

variations that can be concatenated to a given one, and we analyze the associated sets of evolution operators. To this end some topological properties of the space of parameter variations are needed. These imply, in particular, that we are indeed dealing with certain linear flows on vectors bundles. Then several useful properties of the sets of evolution operators are collected that arise from the concatenation restrictions. As a by-product, it is obtained that the exponential growth rate is at least an upper semicontinuous function of the data.

As we will be dealing with set-valued maps, let us briefly recall that a set-valued map F from $X \subset \mathbb{K}^m$ to $\mathbb{K}^n$ is a map that associates to every point in X a subset of $\mathbb{K}^n$. We will encounter only the easy case, in which the images are compact sets. Such a map is called upper semicontinuous at $x \in X$ if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that $\|x - \tilde{x}\| < \delta$ implies $F(\tilde{x}) \subset F(x) + \varepsilon B$, where B is the open unit ball in $\mathbb{K}^n$. The map F is called upper semicontinuous, if it is so at every $x \in X$, and locally Lipschitz continuous, if for every compact subset $K \subset X$ there is a constant L such that $d_H(F(x), F(y)) < L\|x - y\|$ for all $x, y \in K$.

If F is a set-valued map from $X_1 \times X_2$ to $\mathbb{K}^n$, then we call the above properties in x_1 uniform with respect to x_2 if the δ corresponding to an ε, resp., the L, can be chosen for x_1 uniformly for all $x_2 \in X_2$.

In this section we assume the system $\Sigma = (h, \Theta, \Theta_1, A)$ to be given. For ease of notation we will therefore suppress the dependence of $\hat{\rho}(A,\mathcal{U})$, $\mathcal{S}_t(A,\mathcal{U})$, etc. on these data. As we have noted before, simple concatenation of admissible parameter variations does not in general result in an admissible parameter variation. In contrast, for every admissible parameter variation $u \in \mathcal{U}$ and $t \geq 0$ there is a certain subset of $\mathcal{U}$ of admissible parameter variations w for which the following concatenation is also admissible:

$$(u \diamond_t w)(s) := \begin{cases} u(s), & s < t, \\ w(s-t), & t \leq s. \end{cases} \tag{4.1}$$

It is easy to see that this subset depends on the continuous extension of u at t from the left and, in the case $h \in (0, \infty)$, on the difference between the time instance t and the largest discontinuity of u smaller than t. To denote these quantities we define

$$u(t^-) := \lim_{s \nearrow t} u(s) \tag{4.2}$$

and

$$\tau^-(u,t) := \min\{h, t - \max\{t_k \mid t_k < t, \quad \text{where } t_k \text{ is a discontinuity of } u\}\}. \tag{4.3}$$

We first treat the case $h \in (0, \infty)$ and define for $(\theta, \tau) =: \omega \in \Theta \times [0, h)$ the set of concatenable parameter variations by

$$\mathcal{U}(\omega) := \mathcal{U}(\theta, \tau) := \{u \in \mathcal{U} \mid u(0) = \theta \quad \text{and} \quad h \leq t_0(u) + \tau\};$$

here τ represents the time elapsed since the last discontinuity. For $\tau = h$ and $\omega = (\theta, h)$,

$$\mathcal{U}(\omega) := \mathcal{U}(\theta, h) := \{u \in \mathcal{U} \mid u(0) = \theta \quad \text{or} \quad h \leq t_0(u)\}.$$

Note that with this definition we clearly have $\mathcal{U} = \cup_{\omega \in \Theta \times [0,h]} \mathcal{U}(\omega)$, as every admissible parameter variation is continuous on some interval of the form $[0, \tau]$.

The interpretation of the set $\mathcal{U}(\theta,\tau)$ is the following. Consider a parameter variation u defined on the interval $(-\infty,t)$ and the concatenation (4.1). If a discontinuity of u occurs in the interval $(t-h,t)$, then admissible concatenations in t have to result in a continuous function in t. This requires $u(t)=w(0)$. Additionally, w has to wait for a time span of length at least $h-\tau^-(u,t)$ until it is allowed to have a discontinuity, so $t_0(w)\geq h-\tau^-(u,t)$ is also necessary. If there is no discontinuity of u in $(t-h,t)$, equivalently if $\tau^-(u,t)=h$, then we can either introduce a discontinuity at t, in which case $t_0(w)\geq h$ is necessary, or we can continue continuously with $u(t)=w(0)$, in which case there is no restriction on the first discontinuity of w. In all, for $w\in\mathcal{U}$ the concatenation $u\diamond_t w$ defines an admissible parameter variation if and only if

$$w\in\mathcal{U}(u(t^-),\tau^-(u,t)).$$

Note that for $0\leq\tau_1<\tau_2\leq h$ we have

$$\mathcal{U}(\theta,\tau_1)\subset\mathcal{U}(\theta,\tau_2).$$

This implies that for $0\leq\tau\leq\tau^-(u,t)$ we have at least the property that if $w\in\mathcal{U}(u(t^-),\tau)$, then $u\diamond_t w$ from (4.1) defines an admissible parameter variation. Furthermore, it should be noted that the sets $\mathcal{U}(\theta,0)$ are not really needed for concatenation purposes but are included for continuity reasons.

In the case $h=\infty$ there is no need to account for discontinuities. We thus define for $\theta\in\Theta$ the set

$$\mathcal{U}(\theta):=\{u\in\mathcal{U}\mid u(0)=\theta\}.$$

For the sake of a unified notation, we define

$$\Pi(\Theta,h):=\begin{cases}\Theta\times[0,h] & \text{if } h\in(0,\infty),\\ \Theta & \text{if } h=\infty.\end{cases}$$

In the following we denote the restriction of a parameter variation u to an interval (a,b) by $u_{|(a,b)}$. Given the sets $\mathcal{U}(\omega),\omega\in\Pi(\Theta,h)$, we now define parameter variations that may be an "initial piece" for all parameter variations $w\in\mathcal{U}(\omega)$ by

$$\begin{aligned}\mathcal{B}(\theta,\tau)&:=\left\{u_{|(-\infty,t)}\mid u\in\mathcal{U},\ u(t^-)=\theta,\ \tau\leq\tau^-(u,t)\right\}\quad\text{if } h\in(0,\infty),\\ \mathcal{B}(\theta)&:=\left\{u_{|(-\infty,t)}\mid u\in\mathcal{U},\ u(t^-)=\theta\right\}\quad\text{else.}\end{aligned}$$

Note that any parameter variation defined on a finite interval (s,t) can be extended to an admissible parameter variation on $\mathbb{R}$ if the conditions of Definition 3.1 are respected on (s,t). We will therefore also use the notation $u_{|(s,t)}\in\mathcal{B}_t(\omega)$. The interpretation of this is that a suitable extension of $u_{|(s,t)}$ to $(-\infty,t)$ lies in $\mathcal{B}_t(\omega)$ for $\omega\in\Pi(\Theta,h)$.

In all we have introduced notation just to be able to make the following statement, which is now obvious.

LEMMA 4.1. *Consider a system $\Sigma=(h,\Theta,\Theta_1,A)$ satisfying* (A1)–(A5) *and let $u,w\in\mathcal{U}$. The concatenation* (4.1) *yields an admissible parameter variation $u\diamond_t w$ if and only if there exists $\omega\in\Pi(\Theta,h)$ such that*

$$u_{|(-\infty,t)}\in\mathcal{B}(\omega)\quad\textit{and}\quad w\in\mathcal{U}(\omega).$$

For each $\omega\in\Pi(\Theta,h)$ and $t\geq 0$ we define the set of evolution operators "starting in ω" by

$$\mathcal{S}_t(\omega):=\{\,\Phi_u(t,0)\mid u\in\mathcal{U}(\omega)\}. \tag{4.4}$$

Similarly, we define for $\omega, \zeta \in \Pi(\Theta, h)$ and for $t \geq 0$ the sets of evolution operators "starting in ω and ending at ζ" by

$$\begin{aligned}\mathcal{R}_t(\omega,\zeta) := \{\Phi_u(t,0) \,|\, & u \in \mathcal{U}(\omega), u_{|(-\infty,t)} \in \mathcal{B}(\zeta), \\ & \text{and for all } w \in \mathcal{U}(\zeta) \text{ it holds that } u \diamond_t w \in \mathcal{U}(\omega)\}.\end{aligned} \tag{4.5}$$

Thus by definition if $R \in \mathcal{R}_s(\omega,\zeta)$ and $S \in \mathcal{S}_t(\zeta)$, then $SR \in \mathcal{S}_{t+s}(\omega)$.

Remark 4.2. The definition of $\mathcal{R}_t(\omega,\zeta)$ might seem peculiar at first glance. In fact, in the case $h = \infty$ the third condition in (4.5) is superfluous. It is sufficient that $u(0) = \theta, u(t) = \eta$ in order for $u \diamond_t w \in \mathcal{U}(\theta)$ for all $w \in \mathcal{U}(\eta)$. However, if $h \in (0,\infty)$, then although the condition $u_{|(-\infty,t)} \in \mathcal{B}(\zeta)$ implies that $u \diamond_t w$ defines an admissible parameter variation if $w \in \mathcal{U}(\zeta)$, it does not automatically imply that this concatenation lies in $\mathcal{U}(\omega)$. For this, further restrictions regarding the discontinuities have to be observed. Namely, if $\omega = (\theta,\tau)$ and $\zeta = (\eta,\sigma)$, a short calculation shows that it is necessary that $t \geq \sigma - \tau$ to guarantee $u \diamond_t w \in \mathcal{U}(\omega)$ for all $w \in \mathcal{U}(\zeta)$. In particular, if $t \geq h$, then again the third condition in (4.5) is superfluous.

We now define

$$\mathcal{S}_{\leq T}(\omega) := \bigcup_{0\leq t\leq T} \mathcal{S}_t(\omega) \quad \text{and} \quad \mathcal{S}(\omega) := \bigcup_{t\geq 0} \mathcal{S}_t(\omega), \quad \text{resp.},$$
$$\mathcal{R}_{\leq T}(\omega,\zeta) := \bigcup_{0\leq t\leq T} \mathcal{R}_t(\omega,\zeta) \quad \text{and} \quad \mathcal{R}(\omega,\zeta) := \bigcup_{t\geq 0} \mathcal{R}_t(\omega,\zeta).$$

Note that for every $\omega \in \Pi(\Theta,h)$ the set $\mathcal{R}(\omega,\omega)$ is a semigroup.

Remark 4.3. It is useful to keep in mind the following remark on parameter variations connecting two points $\omega, \zeta \in \Pi(\Theta,h)$. If $h \in (0,\infty)$, then for all $\omega, \zeta \in \Theta \times [0,h]$ the set $\mathcal{R}_{2h}(\omega,\zeta)$ is not empty. For if $\omega = (\theta,\tau), \zeta = (\eta,\sigma)$, then it suffices to define $u(s) = \theta, 0 \leq s < h$ and $u(s) = \eta, h \leq s \leq 2h$. Similarly, if $h = \infty$ and (A6) holds, then it follows from Remark 3.2(iv) and the constant $\bar{c}$ used in that remark that $\mathcal{R}_{\bar{c}}(\theta,\eta) \neq \emptyset$ for all $\theta, \eta \in \Theta$.

In a first step let us clarify the continuity properties of the sets just defined. To this end we note the following consequence of the Arzela–Ascoli theorem.

LEMMA 4.4. *Let $\Theta \in \mathcal{K}(\mathbb{K}^m)$, $\Theta_1 \in \mathrm{Co}(\mathbb{K}^m), h \in (0,\infty]$ satisfy* (A1)–(A4) *and consider the space $\mathcal{U}$ of admissible parameter variations in the sense of Definition* 3.1.

Given $T > 0$ and sequences $\omega_k, \zeta_k \in \Pi(\Theta,h)$, $u_k \in \mathcal{U}(\omega_k)$ with $\Phi_{u_k}(T,0) \in \mathcal{R}(\omega_k,\zeta_k)$, there exist subsequences such that

(i) *the limits $\lim_{\mu\to\infty} \omega_{k_\mu} =: \omega$ and $\lim_{\mu\to\infty} \zeta_{k_\mu} =: \zeta$ exist;*
(ii) *$\{u_{k_\mu}\}_{\mu\in\mathbb{N}}$ converges in the weak-* topology on $[0,T]$ to an admissible parameter variation $u \in \mathcal{U}(\omega)$ with $\Phi_u(T,0) \in \mathcal{R}(\omega,\zeta)$.*

Furthermore,

$$\Phi_{u_{k_\mu}}(t,0) \to \Phi_u(t,0) \text{ uniformly on } [0,T].$$

Proof. Fix $T > 0$. By compactness we may assume that $\omega_k \to \omega$ and $\zeta_k \to \zeta$. For the case $h = \infty$ the claims are immediate from the Arzela–Ascoli theorem.

We now treat the case $h \in (0,\infty)$ and let $\omega_k =: (\theta_k,\tau_k) \to (\theta,\tau)$ and $\zeta_k =: (\eta_k,\sigma_k) \to (\eta,\sigma)$. For each k the function u_k has finitely many discontinuities on $[0,T]$, the number of which is bounded by $T/h + 1$. By choosing an appropriate subsequence we may therefore assume that the number of discontinuities of u_k is equal to a certain number $0 \leq l \leq T/h + 1$ independent of k. Furthermore, without

loss of generality the discontinuities $0 < s_{1k} < \cdots < s_{lk} \leq T$ of u_k converge to points $s_1, \ldots, s_l$ as $k \to \infty$. Clearly, $s_{j+1} - s_j \geq h, j = 1, \ldots, l-1$, as the same is true for the points $s_{1k}, \ldots, s_{lk}$ for all k.

As Θ and Θ_1 are bounded, the conditions of the Arzela–Ascoli theorem are satisfied by the u_k on $[s_j + \varepsilon, s_{j+1} - \varepsilon]$ for all $\varepsilon > 0$ small enough. By applying a diagonal sequence argument, we may assume that u_k converges to a function u uniformly on any interval of the form $[s_j + \varepsilon, s_{j+1} - \varepsilon]$ for $\varepsilon > 0$ small enough. If $s_1 > 0$, the same argument applies to the interval $[0, s_1 - \varepsilon]$. Similarly, if $s_l < T$, we can treat the interval $[s_l + \varepsilon, T]$ in this way. It follows that u is well defined on $[0, T] \setminus \{s_1, \ldots, s_l\}$. By continuous extension from the right in the points $s_1, \ldots, s_l$ we obtain that u is Lipschitz continuous on each of the intervals $[s_j, s_{j+1})$. By construction, $u(t) \in \Theta$ for all $t \in [0, T]$. Furthermore, $\dot{u}(\cdot)$ is the weak-* limit of an appropriate subsequence of the $\dot{u}_k(\cdot)$ (as Θ_1 is compact). By the convexity of Θ_1 it follows that $\dot{u}(t) \in \Theta_1$ for almost all $t \in [0, T]$. Hence u is admissible.

We now show that $u \in \mathcal{U}(\theta, \tau)$. If $\tau \in [0, h)$, then $s_1 > 0$ because $s_{1k} + \tau_k \geq h$ by definition, and hence $s_1 \geq h - \tau > 0$. Thus $u_k(0) = \theta_k \to \theta = u(0)$ by uniform convergence on $[0, s_1 - \varepsilon]$ for some $\varepsilon > 0$ small enough. This shows that $u \in \mathcal{U}(\theta, \tau)$. If $\tau = h$ and $s_1 > 0$, the same argument is applicable so that it remains to treat the case when $\tau = h$ and $s_1 = 0$. In this case we have defined $u(0)$ as the continuous extension of $u_{|(0,s_2)}$, so that $u(0) \neq \theta$ is possible. However, we also have $s_2 \geq h$, and so the first discontinuity of u occurs after time h. Thus $u \in \mathcal{U}(\theta, h)$ according to Definition 3.1. The arguments showing that $u_{|[0,T)} \in \mathcal{B}(\eta, \sigma)$ are completely analogous. To show that $\Phi_u(T, 0) \in \mathcal{R}_T(\omega, \zeta)$ we finally have to check that $T \geq \sigma - \tau$ by Remark 4.2. This follows by the assumption $T \geq \sigma_k - \tau_k$ for all k.

The final statement is now immediate from the uniform convergence of the u_k on $[0, T] \setminus \cup_{j=1}^{l} (s_j - \varepsilon, s_j + \varepsilon)$ for all small $\varepsilon > 0$. □

We note an immediate consequence, which is of independent interest, and which will turn out to be useful in section 7.

COROLLARY 4.5. *Given a system $\Sigma = (h, \Theta, \Theta_1, A)$ satisfying* (A1)–(A5)*, the set $\mathcal{U}$ is a metrizable compact space, and the map*

$$(t, u, x) \mapsto (u(t + \cdot), \Phi_u(t, 0)x) \tag{4.6}$$

defines a linear flow on the vector bundle $\pi : \mathcal{U} \times \mathbb{K}^n \to \mathcal{U}$.

Proof. It is a standard result that $L^\infty(\mathbb{R}, \operatorname{conv}\Theta)$ endowed with the weak-* topology is compact and metrizable. The shift $u(\cdot) \mapsto u(t + \cdot)$ is continuous on that space by [13, Lem. 4.2.4]. Lemma 4.4 shows that $\mathcal{U}$ is a compact subset of that space, and so in particular is also metrizable. Furthermore, by the same lemma it follows that (4.6) is continuous as a function of t, u, x. Linearity in the x component is clear by construction. □

We are now ready to prove an essential though fairly basic lemma concerning the dependence of the parameterized sets of transition operators on time and the parameters. To this end we introduce the set

$$W := \{(t, \omega, \zeta) \in \mathbb{R}_+ \times \Pi(\Theta, h)^2 \mid \mathcal{R}_t(\omega, \zeta) \neq \emptyset\}.$$

LEMMA 4.6. *Consider system* (2.1) *given by Σ satisfying* (A1)–(A5)*. Then*

(i) *for all $(t, \omega, \zeta) \in [0, \infty) \times \Pi(\Theta, h)^2$ the sets $\mathcal{S}_t(\omega)$ and $\mathcal{R}_t(\omega, \zeta)$ are compact.*

(ii) *the maps $\mathcal{S} : \mathbb{R}_+ \times \Pi(\theta, h) \to \mathcal{K}(\mathbb{K}^n)$, $\mathcal{R} : W \to \mathcal{K}(\mathbb{K}^n)$ given by*

$$(t, \omega) \mapsto \mathcal{S}_t(\omega), \quad (t, \omega, \zeta) \mapsto \mathcal{R}_t(\omega, \zeta) \tag{4.7}$$

are upper semicontinuous.

(iii) *assume* $h \in (0,\infty)$ *and denote* $\omega = (\theta,\tau), \zeta = (\eta,\sigma)$. *Then for fixed* $\theta \in \Theta$ *the maps in* (4.7) *are locally Lipschitz continuous in* t,τ *(resp., in* t,τ,σ *for fixed* θ,η*). For* $h = \infty$ *and* $\theta \in \Theta$ *(resp.,* $\theta,\eta \in \Theta$*) fixed, the maps are locally Lipschitz continuous in* t.

(iv) *if additionally* (A6) *holds, the maps from* (4.7) *are locally Lipschitz continuous on* $\mathbb{R}_+ \times \Pi(\Theta,h)$ *(resp.,* W*).*

(v) *if* (A6) *holds and* $\mathcal{S}(A,\mathcal{U})$ *is bounded, the Lipschitz constants with respect to* $\omega \in \Pi(\Theta,h)$ *(resp.,* $(\omega,\zeta) \in W$*) may be chosen uniformly in* t.

(vi) *if* $h \in (0,\infty)$ *and* $\mathcal{S}(A,\mathcal{U})$ *is bounded, the maps from* (4.7) *are upper semicontinuous in* (θ,τ) *(resp.,* $(\theta,\tau,\eta,\sigma)$*) uniformly in* t.

Proof. It is clear that each of the sets $\mathcal{S}_t(\theta,\tau), \mathcal{R}_t(\theta,\tau,\eta,\sigma)$ is bounded by the boundedness of $A(\Theta)$. From Lemma 4.4 it is now immediate that they are also closed, so that the proof of (i) is complete. Assertion (ii) is another immediate consequence of Lemma 4.4.

For the remaining statements we restrict our attention to the case $h \in (0,\infty)$ and the sets $\mathcal{S}_t(\theta,\tau)$, as the arguments for $h = \infty$, resp., $\mathcal{R}_t(\theta,\tau,\eta,\sigma)$, are of a very similar nature.

In order to show (iii), let θ be fixed and consider a compact time interval $[0,T]$. Let $t_1,t_2 \in [0,T]$ and $\tau_1,\tau_2 \in [0,h]$. We may assume without loss of generality that $\tau_1 \le \tau_2$. Note that in this case we have $\mathcal{S}_t(\theta,\tau_1) \subset \mathcal{S}_t(\theta,\tau_2)$ for all $t \ge 0$. Let $S = \Phi_u(t_2,0) \in \mathcal{S}_{t_2}(\theta,\tau_2)$ for some $u \in \mathcal{U}$. As $0 \in \Theta_1$, this implies that

$$\tilde{S} := \begin{cases} e^{A(\theta)t_1} & \text{if } t_1 \le \tau_2 - \tau_1, \\ \Phi_u(t_1 - (\tau_2 - \tau_1),0)e^{A(\theta)(\tau_2-\tau_1)} & \text{else} \end{cases}$$

is an element of $\mathcal{S}_{t_1}(\theta,\tau_1)$. We obtain for the second case that

$$\begin{aligned} \|S - \tilde{S}\| &\le \|S\|\|I - e^{A(\theta)(\tau_2-\tau_1)}\| + \|\Phi_u(t_2,t_1-(\tau_2-\tau_1)) - I\|\|\tilde{S}\| \\ &\le L|\tau_2 - \tau_1| + L(|t_2 - t_1| + |\tau_2 - \tau_1|) \end{aligned} \tag{4.8}$$

for a suitable constant L independent of S and θ (which exists as, by the compactness of Θ, the set of evolution operators of length t generated by the system is uniformly bounded for $t \in [0,T]$). It is now easy to check that the same estimates apply to the first case if we use $t_1 \le \tau_2 - \tau_1$ along the way.

Conversely, let $S = \Phi_u(t_1,0) \in \mathcal{S}_{t_1}(\theta,\tau_1)$ for some $u \in \mathcal{U}$. Then $S \in \mathcal{S}_{t_1}(\theta,\tau_2)$ by definition. If $t_2 \le t_1$, then $\tilde{S} := \Phi_u(t_2,0) \in \mathcal{S}_{t_2}(\theta,\tau_2)$. Otherwise, letting $\eta := u(t_1^-)$ we have $\tilde{S} := e^{A(\eta)(t_2-t_1)}S \in \mathcal{S}_{t_2}(\theta,\tau_2)$. Using this, the required Lipschitz estimate in $|t_1 - t_2|$ can be obtained easily.

Thus we have obtained the desired local Lipschitz estimate in (t,τ).

In order to show (iv) note that we have shown local Lipschitz continuity in t,τ uniformly in θ. Thus, if we prove Lipschitz continuity with respect to θ locally uniformly in t,τ, then we have overall local Lipschitz continuity. To this end it is sufficient to restrict our attention to one of the convex components Ω_j of Θ, which we now assume to be fixed. Fix $\theta_1,\theta_2 \in \Omega_j$. As (A6) holds, we may use Remark 3.2(iv) to obtain that the map $s \mapsto \theta_1 + sc(\theta_2 - \theta_1)/\|\theta_2 - \theta_1\|$, $s \in [0,\|\theta_2 - \theta_1\|/c]$ is the initial part of an admissible parameter variation connecting θ_1 and θ_2. Here $c > 0$ is a suitable constant depending only on Θ,Θ_1. Denote by $R \in \mathcal{S}_{\|\theta_2-\theta_1\|/c}(\theta_1,\tau)$ the corresponding evolution operator. For any $S = \Phi_u(t,0) \in \mathcal{S}_t(\theta_2,\tau)$ with $t \ge \|\theta_2 - \theta_1\|/c$, it follows that $\tilde{S} := \Phi_u(t - \|\theta_2 - \theta_1\|/c,0)R \in \mathcal{S}_t(\theta_1,\tau)$. Then again

$$\|S - \tilde{S}\| \le \|S\|\|I - R\| + \|\Phi_u(t,t - \|\theta_2 - \theta_1\|/c) - I\|\|\tilde{S}\|, \tag{4.9}$$

which allows for a Lipschitz estimate in $\|\theta_1 - \theta_2\|$ independently of $t \in [\|\theta_2 - \theta_1\|/c, T], \tau \in [0, h]$ as in (4.8), and using symmetry, the proof is complete. The case that $t < \|\theta_2 - \theta_1\|/c$ is an easy exercise.

(v) If the set of evolution operators of the system is bounded, then the expressions in (4.8) and (4.9) can be bounded independently of $S, \tilde{S}$ so that L does not depend on t, as desired.

(vi) On the bounded interval $[0, 3h]$ the assertion is clear from Lemma 4.4 so that we restrict our attention to $t \geq 3h$.

Fix $(\theta_0, \tau_0) \in \Theta \times [0, h]$. According to (i) the map $(\theta, \tau) \mapsto \mathcal{S}_{3h}(\theta, \tau)$ is upper semicontinuous at (θ_0, τ_0) so that for every $\varepsilon > 0$ there exists a $\delta > 0$ such that $\|\theta - \theta_0\| + |\tau - \tau_0| < \delta$ implies $\mathcal{S}_{3h}(\theta, \tau) \subset \mathcal{S}_{3h}(\theta_0, \tau_0) + \varepsilon B$. Let $t \geq 3h$, $\Phi_u(t, 3h)\Phi_u(3h, 0) \in \mathcal{S}_t(\theta, \tau)$ be arbitrary, and let $w \in \mathcal{U}(\theta_0, \tau_0)$ be such that $\|\Phi_u(s, 0) - \Phi_w(s, 0)\| < \varepsilon$ for all $s \in [0, 3h]$. The proof of Lemma 4.4 shows that we may assume that the discontinuities of u and w are no more than ε apart.

Let $s_u, s_w \in [0, 3h]$ be two discontinuities of u, resp., w with $|s_u - s_w| < \varepsilon$ (assuming they exist; if not, set $s_u := s_w := 3h/2$) and define

$$\tilde{u}(t) := \begin{cases} w(t), & t < s_w, \\ u(t - s_w + s_u), & t \geq s_w. \end{cases}$$

Then $\Phi_{\tilde{u}}(t, 0) \in \mathcal{S}_t(\theta_0, \tau_0)$ and we obtain that

$$\begin{aligned} \|\Phi_u(t, 0) - \Phi_{\tilde{u}}(t, 0)\| &\leq \|\Phi_u(t, s_u)\| \|\Phi_u(s_u, 0) - \Phi_w(s_w, 0)\| \\ &\quad + \|\Phi_u(t, t - s_w + s_u) - I\| \|\Phi_{\tilde{u}}(t, 0)\| \\ &\leq M(\|\Phi_u(s_u, 0) - \Phi_w(s_w, 0)\| + \|\Phi_u(t, t - s_w + s_u) - I\|), \end{aligned}$$

where M is some bound on the norm of $\Phi_u(t, 0), u \in \mathcal{U}, t \geq 0$. Using that $\|\Phi_u(s, 0) - \Phi_w(s, 0)\| < \varepsilon$ for all $s \in [0, 3h]$ and that $|s_u - s_w| < \varepsilon$, we see that the last bound may be made arbitrarily small by choosing δ small enough. As the bound is independent of t, this shows the assertion. □

Remark 4.7. It should be noted that without assumption (A6) the maps studied in the previous lemma need not be continuous in θ. As an example consider the convex subset of $\mathbb{R}^3$ given by

$$\Theta := \operatorname{conv}\{[0 \quad 0 \quad 1]'\} \cup \{[x \quad x^2 \quad 0]' \mid x \in [0, 1]\},$$

and let $\Theta_1 = \{0\} \times \{0\} \times [-1, 1]$, $A(z_1, z_2, z_3) = z_3 \in \mathbb{R}$, $h \in (0, \infty]$. For fixed $0 < t \leq 1$ and the initial value $\theta(0) = [0, 0, 0]$, the function

$$u(s) = [0, 0, s]', s \in [0, t]$$

defines an admissible parameter variation which yields the evolution operator $\Phi_u(t, 0) = \exp(t^2/2) \in \mathcal{S}_t(\theta(0), \tau), \tau \in [0, h]$. On the other hand, for arbitrary $1 \geq \varepsilon > 0$ and the parameter value $\theta(\varepsilon) = [\varepsilon, \varepsilon^2, 0]$ the only admissible parameter variation is the function $u_\varepsilon \equiv \theta(\varepsilon)$, as no point of the form $[\varepsilon, \varepsilon^2, z_3], z_3 \neq 0$, is contained in Θ. Hence

$$\mathcal{S}_t(\theta(\varepsilon), \tau) = \{1\}$$

as long as $t + \tau < h$. In particular, for all $t > 0$ small enough the map $\varepsilon \mapsto \mathcal{S}_t(\theta(\varepsilon), \tau)$ is discontinuous in $\varepsilon = 0$.

With arguments very similar to those employed in the proof of Lemma 4.4, a semicontinuity property of $\hat{\rho}$ may be shown. We denote the space of systems

$$\mathcal{L} := \{\Sigma := (h, \Theta, \Theta_1, A) \mid \Sigma \text{ satisfies (A1)–(A5)}\}$$

and endow it with the product topology inherited from $(0,\infty]\times\mathcal{K}(\mathbb{R}^{n\times n})\times \mathrm{Co}(\mathbb{R}^{n\times n})\times \mathcal{C}(\mathbb{R}^m, \mathbb{R}^{n\times n})$, where we consider the topology of locally uniform convergence on $\mathcal{C}(\mathbb{R}^m, \mathbb{R}^{n\times n})$.

PROPOSITION 4.8. *The map*

$$\hat{\rho} : \mathcal{L} \to \mathbb{R}, \qquad (h, \Theta, \Theta_1, A) \mapsto \hat{\rho}(h, \Theta, \Theta_1, A)$$

is upper semicontinuous.

Proof. It is sufficient to show that the maps $(h, \Theta, \Theta_1, A) \mapsto \hat{\rho}_t(h, \Theta, \Theta_1, A)$ are upper semicontinuous, as by (2.2) we have $\hat{\rho} = \inf_{t>0} \hat{\rho}_t$ and the infimum of upper semicontinuous maps is upper semicontinuous. So fix $t \geq 0$ and a sequence $\Sigma_k = (h_k, \Theta_k, \Theta_{1,k}, A_k) \to \Sigma = (h, \Theta, \Theta_1, A) \in \mathcal{L}$. We first consider the case $h \in (0,\infty)$. Let $u_k \in \mathcal{U}(\Sigma_k)$ be such that $\|\Phi_{u_k}(t,0)\| = \hat{\rho}_t(\Sigma_k)$. We may assume that $\lim_{k\to\infty} \Phi_{u_k}(t,0) =: S$ exists, and we now have to show that $S \in \mathcal{S}_t(\Sigma)$ because in this case $\hat{\rho}_t(\Sigma) \geq \limsup_{k\to\infty} \hat{\rho}_t(\Sigma_k)$.

Now, as in the proof of Lemma 4.4 we may choose a subsequence of the u_k such that the discontinuities of u_k on $[0,t]$ converge to finitely many points $s_1, \dots, s_l$. These are at least distance h apart. On the intervals of the form $[s_j + \varepsilon, s_{j+1} - \varepsilon]$, $j = 1, \dots, l$, we may (after going to a further subsequence) assume that the u_k converge uniformly and that their derivatives converge in the weak-* sense. Then it follows again that $u \in \mathcal{U}(\Sigma)$ and that $S = \Phi_u(t,0)$, as desired.

If $h = \infty$ and $h_k = \infty$, the same argument is applicable. We finally have to treat the case $h_k \in (0,\infty), h_k \to \infty$. In this case the number of discontinuities of u_k on $[0,t]$ is bounded by $t/h_k + 1$. Thus it may happen that for a given choice of t and $u_k \in \mathcal{U}(\Sigma_k)$ the discontinuities of u_k in $[0,t]$ converge to one point $s_1 \in [0,t]$. In this case the limit function u is not an element of $\mathcal{U}(h, \Theta, \Theta_1, A)$. However, we have $\Phi_u(t, s_1) \in \mathcal{S}(h, \Theta, \Theta_1, A)$ as well as $\Phi_u(s_1, 0) \in \mathcal{S}(h, \Theta, \Theta_1, A)$. Thus using (2.3), for every $\varepsilon > 0$ there is a constant M_ε such that

$$\|\Phi_u(t,0)\| \leq \|\Phi_u(t,s_1)\|\|\Phi_u(s_1,0)\| \leq M_\varepsilon^2 e^{(\hat{\rho}(h,\Theta,\Theta_1,A)+\varepsilon)t}.$$

As t is arbitrary, the last inequality implies that also in this case $\hat{\rho}(h, \Theta, \Theta_1, A) \geq \limsup_{k\to\infty} \hat{\rho}(h_k, \Theta_k, \Theta_{1k}, A_k)$, as desired. ∎

If we describe the exponential growth rate within the subsets of evolution operators with given initial and end conditions, this leads to the definitions

$$\widehat{\rho}_t(\omega) := \max\left\{\frac{1}{t}\log\|S\| \mid S \in \mathcal{S}_t(\omega)\right\}, \quad \widehat{\rho}_t(\omega,\zeta) := \max\left\{\frac{1}{t}\log\|S\| \mid S \in \mathcal{R}_t(\omega,\zeta)\right\}.$$

With this, the problem arises in which the functions $t \mapsto t\widehat{\rho}_t(\omega)$ and $t \mapsto t\widehat{\rho}_t(\omega,\zeta)$ are no longer subadditive, so that it does not follow automatically to what value they are converging, if at all. It is therefore useful to point out the following.

LEMMA 4.9. *Consider the system* (2.1) *with* (A1)–(A5) *and let one of the following assumptions be satisfied:*

(a) $h \in (0,\infty)$.

(b) $h = \infty$ *and* (A6) *is satisfied.*

Then there is a constant $C \in \mathbb{R}$ such that for all $\omega, \zeta \in \Pi(\Theta, h)$ we have

$$t\widehat{\rho}_t(\omega,\zeta) \geq t\hat{\rho} - C \quad \textit{for all } t > 0. \tag{4.10}$$

In particular, it follows for all $\omega, \zeta \in \Pi(\Theta, h)$ that

$$\widehat{\rho} = \lim_{t\to\infty} \widehat{\rho}_t(\omega,\zeta) = \lim_{t\to\infty} \widehat{\rho}_t(\omega).$$

Proof. Fix $\omega, \eta \in \Pi(\Theta, h)$. Clearly, for all $t \geq 0$ we have $\widehat{\rho}_t(\omega,\zeta) \leq \widehat{\rho}_t(\omega) \leq \widehat{\rho}_t$ so that in order to show the second assertion it is sufficient to show that $\widehat{\rho} \leq \liminf_{t\to\infty} \widehat{\rho}_t(\omega,\zeta)$. This, however, is an immediate consequence of (4.10).

In order to show (4.10), note that by (2.2) we can for each $t > 0$ choose a matrix $S_t \in \mathcal{S}_t$ with $\log \|S_t\| = t\widehat{\rho}_t \geq t\widehat{\rho}$. Then $S_t \in \mathcal{R}(\omega_1, \zeta_1)$ for suitable ω_1, ζ_1 (depending on t). If (a) holds, then we may by Remark 4.3 for each such S_t choose an $R_1 \in \mathcal{R}_{2h}(\omega, \omega_1)$ and an $R_2 \in \mathcal{R}_{2h}(\zeta_1, \zeta)$. With this choice we obtain that

$$R_2 S_t R_1 \in \mathcal{R}_{t+4h}(\omega,\zeta),$$

and so

$$\begin{aligned}(t+4h)\hat{\rho}_{t+4h}(\omega,\zeta) &\geq \log\|R_2 S_t R_1\| \geq \log\|S_t\|\|R_2^{-1}\|^{-1}\|R_1^{-1}\|^{-1} \\ &\geq t\widehat{\rho}_t - 2\log\max\{\|S^{-1}\| \mid S \in \mathcal{S}_{2h}\} \\ &\geq t\hat{\rho} - 2\log\max\{\|S^{-1}\| \mid S \in \mathcal{S}_{2h}\},\end{aligned}$$

which shows the assertion under assumption (a). To prove the assertion when (b) holds, we can use Remarks 4.3 and 3.2(iv), by which all pairs $\theta, \eta \in \Theta$ can be connected in time $\bar{c}$ independently of θ, η. The remaining arguments are then exactly the same as before. ∎

5. Irreducibility. We aim to construct parameter-dependent Lyapunov functions that exactly reflect the exponential growth rate of the system $\Sigma = (h, \Theta, \Theta_1, A)$. To this end it is crucial to assume the irreducibility of $A(\Theta)$. Recall that a set of matrices $\mathcal{M} \subset \mathbb{K}^{n\times n}$ is called irreducible, if only the trivial subspaces $\{0\}$ and $\mathbb{K}^n$ are invariant under all $A \in \mathcal{M}$, and called reducible otherwise.

Remark 5.1. (i) Note that the set of systems Σ for which $A(\Theta)$ is irreducible is open and dense in the set $\mathcal{L}$ of all systems satisfying (A1)–(A5), with the topology introduced just before Proposition 4.8.

(ii) If $A(\Theta)$ is reducible, we can find a similarity transformation T such that for all $\theta \in \Theta$ the transformed matrix $TA(\theta)T^{-1}$ is of the form

$$\begin{bmatrix} A_{11}(\theta) & A_{12}(\theta) & \dots & A_{1d}(\theta) \\ 0 & A_{22}(\theta) & \dots & A_{2d}(\theta) \\ & \ddots & \ddots & \vdots \\ 0 & & 0 & A_{dd}(\theta) \end{bmatrix}, \tag{5.1}$$

where the sets $A_{ii}(\Theta) \subset \mathbb{K}^{n_i\times n_i}$ are irreducible or $\{0\}$, $i = 1, \dots, d$. It is an easy exercise to show that in this case $\hat{\rho}(A, \mathcal{U}) = \max_{i=1,\dots,d} \hat{\rho}(A_i, \mathcal{U})$, where $A_i : \Theta \to \mathbb{K}^{n_i\times n_i}$ is the map $\theta \mapsto A_{ii}(\theta)$. Having said this, it is clear that for the analysis of $\hat{\rho}$ with respect to one system we can assume irreducibility without loss of generality.

The next simple lemma is crucial in the following construction.

LEMMA 5.2. *Let $\mathbb{K} = \mathbb{R}, \mathbb{C}$ and let $\mathcal{S} \subset \mathbb{K}^{n\times n}$ be an irreducible semigroup. For any family of sets $\mathcal{S}_t$, $t \in \mathbb{R}_+$, with*

$$\mathcal{S} = \bigcup_{t\geq 0} \mathcal{S}_t,$$

there are $\varepsilon > 0$ and $T \in \mathbb{R}_+$ such that for all $z \in \mathbb{K}^n, A \in \mathbb{K}^{n\times n}$ there is an $S \in \bigcup_{1\leq t\leq T} \mathcal{S}_t$ with

$$\|ASz\| \geq \varepsilon \|A\| \|z\|.$$

Proof. This is a minute generalization of [34, Lem. 3.1]. □

We now begin to study the consequences of irreducibility. The following properties are essential in our construction of Lyapunov functions.

PROPOSITION 5.3. *Consider system* (2.1) *with Assumptions* (A1)–(A5). *Assume that $A(\Theta)$ is irreducible and let one of the following assumptions be satisfied:*

(a) *$h \in (0, \infty)$.*

(b) *$h = \infty$ and* (A6) *is satisfied.*

Then for all $\omega, \zeta \in \Pi(\Theta, h)$,

(i) *the set $\mathcal{R}(\omega, \zeta)$ is irreducible.*

(ii) *the set $\mathcal{S}(\omega)$ is irreducible.*

Proof. (i) We first show the claim assuming (a). Fix an arbitrary nontrivial subspace X and let $\Phi_u(t, 0) \in \mathcal{R}(\omega, \zeta)$ with $t \geq 2h$ be such that $\Phi_u(t, 0)X = X$. (If no such Φ exists, we are done.) Let $t^* \in (0, t)$ be a discontinuity of u, or if such a discontinuity does not exist, let $t^* = t/2$. Denote $Y := \Phi_u(t^*, 0)X$. As $A(\Theta)$ is irreducible, there exists a $\theta^* \in \Theta$ such that $\exp(A(\theta^*)s)Y \not\subset Y$ for some $s \geq h$. Hence $\Phi_u(t, t^*) \exp(A(\theta^*)s)\Phi_u(t^*, 0)X \not\subset X$. On the other hand, $\Phi_u(t, t^*) \exp(A(\theta^*)s)\Phi_u(t^*, 0) \in \mathcal{R}(\omega, \zeta)$ because we may at time t^* jump to θ^*, remain there for the time s, and jump back to $u(t^*)$. This defines an admissible parameter variation, and the assertion follows.

Now assume that (b) holds and let X be a nontrivial invariant subspace for all $\Phi_u(t, 0) \in \mathcal{R}(\theta, \eta)$. Fix one of the corresponding parameter variations u. As $0 \in \Theta_1$, we also have for arbitrary $0 \leq s \leq t$ and all $r \geq 0$ that $\Phi_u(t, s) \exp(A(u(s))r)\Phi_u(s, 0) \in \mathcal{R}(\theta, \eta)$. Denoting $Y_s := \Phi_u(s, 0)X$ we obtain that $\exp(A(u(s))r)Y_s = Y_s$ for all $r \geq 0$, $s \in [0, t]$, so that $A(u(s))Y_s \subset Y_s$ for all $s \in [0, t]$.

Assume that $\dim Y_s = m$ for some $1 \leq m < n$ and denote the Grassmannian of m-dimensional subspaces of $\mathbb{K}^n$ by $G(n, m)$. Consider the induced differential equation on $G(n, m)$ given by

$$\dot{X}(s) = A(u(s))X(s). \tag{5.2}$$

Then the function $s \mapsto Y_s, s \in [0, t]$ is a solution of (5.2), as we have by the previous construction for all $s \in [0, t]$ that $\Phi_u(s, 0)X = Y_s$. On the other hand, we have

$$\frac{d}{ds} Y_s = \frac{d}{ds} \Phi_u(s, 0)X = A(u(s))\Phi_u(s, 0)X = A(u(s))Y_s \subset Y_s,$$

or in other words $\frac{d}{ds} Y_s = 0$ for all $s \in [0, t]$ in the Grassmannian. This shows that $Y_s \equiv X$ so that X is a common invariant subspace for all $A(u(s)), s \in [0, t]$. Under condition (A6), however, we may for arbitrary $\theta_1 \in \Theta$ choose an admissible parameter variation u such that for suitable times $0 \leq s \leq t$ we have $u(0) = \theta, u(s) = \theta_1, u(t) = \eta$.

By the previous argument this implies that X is an invariant subspace of $A(\theta_1)$ so that X is a common invariant subspace for all $A \in A(\Theta)$, which contradicts irreducibility of $A(\Theta)$. This completes the proof.

(ii) This is immediate from (i) as $\mathcal{S}(\omega) = \cup_{\zeta\in\Pi(\Theta,h)}\mathcal{R}(\omega,\zeta)$. □

6. Parameterized Lyapunov functions. In this section the main result of the paper is derived. In Theorem 6.4 we obtain the existence of parameterized Lyapunov functions that characterize the exponential growth rate. Also some results of the Lipschitz continuous dependence of the Lyapunov function on the parameter are presented.

The main step of the proof relies on the following construction. By Lemma 4.9 the exponential growth in $\mathcal{S}$ and in the subsets $\mathcal{S}(\omega)$, $\mathcal{R}(\omega,\eta)$ is essentially the same. It therefore makes sense to define limit sets as follows:

$$\mathcal{S}_\infty(\omega) := \{S \in \mathbb{K}^{n\times n} \mid \exists\, t_k \to \infty, S_k \in \mathcal{S}_{t_k}(\omega) : e^{-\hat\rho t_k} S_k \to S\}, \tag{6.1}$$

$$\mathcal{R}_\infty(\omega,\zeta) := \{S \in \mathbb{K}^{n\times n} \mid \exists\, t_k \to \infty, S_k \in \mathcal{R}_{t_k}(\omega,\zeta) : e^{-\hat\rho t_k} S_k \to S\}. \tag{6.2}$$

We note the following properties of $\mathcal{S}_\infty(\omega)$ and $\mathcal{R}_\infty(\omega,\zeta)$.

LEMMA 6.1. *Consider the system* (2.1) *with* (A1)–(A5). *Assume that* $A(\Theta)$ *is irreducible and let one of the following assumptions be satisfied:*

(a) $h \in (0,\infty)$.

(b) $h = \infty$ *and* (A6) *is satisfied.*

Then

(i) *the set* $\cup_{\omega\in\Pi(\Theta,h)}\mathcal{S}_\infty(\omega)$ *is bounded,*

and for all $\omega,\zeta \in \Pi(\Theta,h)$ *it holds that*

(ii) $\mathcal{R}_\infty(\omega,\zeta)$ *is a compact, nonempty set not equal to* $\{0\}$.

(iii) $\mathcal{S}_\infty(\omega)$ *is a compact, nonempty set not equal to* $\{0\}$.

(iv) *for every* $t \geq 0$ *we have that if* $R \in \mathcal{R}_t(\omega,\zeta)$ *and* $S \in \mathcal{S}_\infty(\zeta)$, *or if* $R \in \mathcal{R}_\infty(\omega,\zeta)$ *and* $S \in \mathcal{S}_t(\zeta)$, *then* $e^{-\hat\rho t}SR \in \mathcal{S}_\infty(\omega)$.

(v) *for every* $S \in \mathcal{S}_\infty(\omega)$ *and every* $t \in \mathbb{R}_+$ *there exist* $\zeta \in \Pi(\Theta,h)$, $R \in \mathcal{R}_t(\omega,\zeta)$, *and* $T \in \mathcal{S}_\infty(\zeta)$ *such that* $S = e^{-\hat\rho t}TR$.

(vi) $\mathcal{R}_\infty(\omega,\omega)$, $\mathcal{S}_\infty(\omega)$ *are irreducible.*

Proof. Without loss of generality, we may assume in this proof that $\hat\rho = 0$ by considering the map $\tilde A(\theta) := A(\theta) - \hat\rho I$.

(i) For ease of notation define

$$\delta := \min\{\|R^{-1}\|^{-1} \mid R \in \mathcal{S}_{\leq\gamma}\} > 0,$$

where $\gamma = 2h$ in the case (a) or $\gamma = \bar c$ in the case (b) is the constant described in Remark 4.3.

If the assertion is false, then there are $t_k \to \infty$, $S_k \in \mathcal{S}_{t_k}(\omega_k)$ with $\|S_k\| \to \infty$. Without loss of generality, we may assume that $S_k \in \mathcal{R}_{t_k}(\omega_k,\omega_k)$. To see this, note that by Remark 4.3 we can always ensure that $R_kS_k \in \mathcal{R}_{t_k}(\omega_k,\omega_k)$ for some $R_k \in \mathcal{S}_\gamma$. It is easy to see that $\|R_kS_k\| \geq \|S_k\|\|R_k^{-1}\|^{-1} \geq \|S_k\|\delta \to \infty$ as $k \to \infty$.

Fix some $\omega \in \Pi(\Theta,h)$. The set $\mathcal{R}(\omega,\omega)$ is a semigroup and irreducible by Proposition 5.3. We may therefore use Lemma 5.2 to find constants $1 \geq \varepsilon_1 > 0$ and $T > 0$ such that for all $x \in \mathbb{K}^n$ and all $B \in \mathbb{K}^{n\times n}$ there is an $R \in \mathcal{R}_{\leq T}(\omega,\omega)$ with $\|BRx\| \geq \varepsilon_1\|B\|\|x\|$.

Now define $\varepsilon := \min\{1,\varepsilon_1\delta^2\}$ and choose k large enough such that

$$\|S_k\| > 4/\varepsilon.$$

Fix $U \in \mathcal{R}_{\leq\gamma}(\omega_k, \omega)$ and $V \in \mathcal{R}_{\leq\gamma}(\omega, \omega_k)$ and pick an arbitrary $x_0 \in \mathbb{K}^n, \|x_0\| = 1$, such that $\|S_k x_0\| \geq \|S_k\| \varepsilon/2$. Then we can choose $R_1 \in \mathcal{R}_{\leq T}(\omega, \omega)$ such that

$$\begin{aligned}\|S_k V R_1 U S_k x_0\| &\geq \varepsilon_1 \|S_k V\| \|U S_k x_0\| \geq \varepsilon_1 \|S_k\| \|V^{-1}\|^{-1} \|U^{-1}\|^{-1} \|S_k x_0\| \\ &\geq \left(\|S_k\| \frac{\varepsilon}{2} \right)^2.\end{aligned}$$

Note that by construction $S_k V R_1 U S_k \in \mathcal{R}_{\leq 2t_k+T+2\gamma}(\omega_k, \omega_k)$. Applying the same arguments again, we can choose $R_2 \in \mathcal{R}_{\leq T}(\omega, \omega)$ such that

$$\|S_k V R_2 U S_k V R_1 U S_k x_0\| \geq \left(\|S_k\| \frac{\varepsilon}{2} \right)^3.$$

Arguing inductively we construct times τ_l with $l t_k \leq \tau_l \leq l(t_k + T + 2\gamma)$ and matrices $T_l \in \mathcal{R}_{\tau_l}(\omega_k, \omega_k)$ with

$$\frac{1}{\tau_l} \log \|T_l\| \geq \frac{l}{\tau_l} \log \left(\|S_k\| \frac{\varepsilon}{2} \right) \geq \frac{l}{\tau_l} \log 2 \geq \frac{1}{t_k + T + 2\gamma} \log 2 > 0.$$

This contradicts the assumption that $\limsup_{l\to\infty} \frac{1}{\tau_l} \log \|T_l\| \leq 0$, which follows from $\hat{\rho} = 0$.

(ii) A standard argument shows that $\mathcal{R}_\infty(\omega, \zeta)$ is closed and by part (i) it is bounded. Thus we have to show that there are nonzero elements. Now Lemma 4.9 shows that there exists a constant $C > 0$ and sequences $t_k \to \infty, S_k \in \mathcal{R}_{t_k}(\omega, \zeta)$ with $\|S_k\| \geq C$ for all $k \in \mathbb{N}$. By (i) the sequence is bounded so that it has a convergent subsequence with nonzero limit. By definition this limit is contained in $\mathcal{R}_\infty(\omega, \zeta)$.

(iii) As $\mathcal{R}_\infty(\omega, \zeta) \subset \mathcal{S}_\infty(\omega)$, it is clear from (i) that $\mathcal{S}_\infty(\omega)$ is nonempty and not equal to $\{0\}$. Closedness is immediate from the definition and so compactness follows from (i).

(iv) This is an easy exercise.

(v) Let $t_k \to \infty$, $u_k \in \mathcal{U}(\omega)$ be sequences such that $\Phi_{u_k}(t_k, 0) \to S \in \mathcal{S}_\infty(\omega)$. Fix $t \geq 0$. Applying Lemma 4.4 we may assume that there exists a $u \in \mathcal{U}(\omega)$ such that $\Phi_{u_k}(s, 0) \to \Phi_u(s, 0)$ uniformly for $s \in [0, t + 3h]$. For some $\zeta \in \Pi(\Theta, h)$, we have that $\Phi_u(t, 0) \in \mathcal{R}(\omega, \zeta)$.

We now treat the case $h \in (0, \infty)$. If the limit function u has no discontinuity in $(t, t+3h)$, then for all k large enough the parameter variations u_k have no discontinuity in $(t + h/2, t + 5h/2)$. This implies that we may introduce a discontinuity at $s = t + 3h/2$, and the functions

$$v_k(\sigma) := \begin{cases} u(\sigma) & \text{if } \sigma < t + 3h/2, \\ u_k(\sigma) & \text{if } \sigma \geq t + 3h/2 \end{cases}$$

are admissible parameter variations. Furthermore,

$$\Phi_{v_k}(t_k, 0) = \Phi_{u_k}(t_k, t + 3h/2) \Phi_u(t + 3h/2, 0)$$

and so

$$\begin{aligned}\|\Phi_{v_k}(t_k, 0) - \Phi_{u_k}(t_k, 0)\| \leq \|\Phi_{u_k}(t_k, t + 3h/2)\| \\ \times \|\Phi_{u_k}(t + 3h/2, 0) - \Phi_u(t + 3h/2, 0)\|,\end{aligned} \tag{6.3}$$

which converges to 0 for $k \to \infty$. (Here we are using (i) to bound the first factor on the right independently of t_k.) Now the construction implies that

$$\Phi_{v_k}(t_k, t) \in \mathcal{S}(\zeta).$$

If we extract a convergent subsequence of $\Phi_{v_k}(t_k, t)$ with limit T, then we have $T \in \mathcal{S}_\infty(\zeta)$. Also by (6.3) we have $S = T\Phi_u(t, 0)$. This shows the assertion.

If u has a discontinuity $s \in (t, t + 3h)$, then there exists a sequence $s_k \to s$, where each s_k is a discontinuity of u_k. This implies that the following function is an admissible parameter variation:

$$v_k(\sigma) := \begin{cases} u(\sigma), & 0 \le \sigma < s, \\ u_k(\sigma - s + s_k), & s \le \sigma \le t_k + s - s_k. \end{cases}$$

Again we see

$$\|\Phi_{v_k}(t_k + s - s_k, 0) - \Phi_{u_k}(t_k, 0)\| \le \|\Phi_{u_k}(t_k, s_k)\| \|\Phi_{u_k}(s_k, 0) - \Phi_u(s, 0)\|,$$

which converges to 0 by the uniform convergence of the u_k and as $s - s_k \to 0$. As before we may extract a convergent subsequence of the sequence $\Phi_{v_k}(t_k + s - s_k, t) \in \mathcal{S}(\zeta)$, and for the limit we have that $S = T\Phi_u(t, 0)$.

If $h = \infty$ and (A6) holds, then by Remark 3.2(iv) there are nonnegative times $s_k \to 0$ and $S_k \in \mathcal{R}_{s_k}(\zeta, u_k(t))$. Then we have

$$\Phi_{u_k}(t_k, t) S_k \Phi_u(t, 0) \in \mathcal{S}_{t_k + s_k}(\omega).$$

Defining $T_k := \Phi_k(t_k, t) S_k \in \mathcal{S}(\zeta)$ we may assume, without loss of generality, that $T_k \to T \in \mathcal{S}_\infty(\zeta)$, and it follows that $T\Phi_u(t, 0) = S$. This shows the assertion.

(vi) Fix $\omega \in \Pi(\Theta, h)$. As we have noted, the set $\mathcal{R}(\omega, \omega)$ is a semigroup, which is irreducible by Proposition 5.3. By (iv) it is easy to see that if $S \in \mathcal{R}(\omega, \omega) \cup \mathcal{R}_\infty(\omega, \omega)$ and $T \in \mathcal{R}_\infty(\omega, \omega)$, then $ST, TS \in \mathcal{R}_\infty(\omega, \omega)$ (where we have used the assumption $\hat{\rho} = 0$; otherwise some further factors appear according to (iv)). Using (ii) this shows that $\mathcal{R}_\infty(\omega, \omega)$ is a nonzero semigroup ideal of the irreducible semigroup

$$\mathcal{R}_\infty(\omega, \omega) \cup \mathcal{R}(\omega, \omega).$$

By [29, Lem. 1] this shows irreducibility of $\mathcal{R}_\infty(\omega, \omega)$. The second assertion follows from $\mathcal{R}_\infty(\omega, \omega) \subset \mathcal{S}_\infty(\omega)$. □

The following interesting observation is obtained through the previous proof.

COROLLARY 6.2. *Under the assumption of Lemma* 6.1 *the set* $\mathcal{S}(A, \mathcal{U})$ *is bounded if* $\hat{\rho} = 0$.

Proof. If the assertion is false, then there exists a sequence $\|S_k\| \to \infty$. This is brought to a contradiction in the proof of Lemma 6.1(i) of the previous theorem. □

We note the following corollary with respect to the maps $\omega \mapsto \mathcal{S}_\infty(\omega)$, $(\omega, \zeta) \mapsto \mathcal{R}_\infty(\omega, \zeta)$.

COROLLARY 6.3. *Consider system* (2.1) *with* (A1)–(A5). *Assume that* $A(\Theta)$ *is irreducible and let* (A6) *hold. Then the set-valued maps*

$$\omega \mapsto \mathcal{S}_\infty(\omega), \tag{6.4}$$

$$(\omega, \zeta) \to \mathcal{R}_\infty(\omega, \zeta) \tag{6.5}$$

are Lipschitz continuous on $\Pi(\Theta,h)$*, resp.,* $(\Pi(\Theta,h))^2$*, with respect to the Hausdorff topology.*

Proof. Without loss of generality, we may assume that $\hat\rho = 0$ so that in particular the set of evolution operators $\mathcal{S}(A,\mathcal{U})$ is bounded by Corollary 6.2. This and the assertions imply that Lemma 4.6(v) is applicable and the map $(\omega,t) \mapsto S_t(\omega)$ is Lipschitz continuous in ω uniformly in t. Thus if $S_k \to S$ for $S_k \in \mathcal{S}_{t_k}(\omega_1), t_k \to \infty$, then for $\omega_2 \in \Pi(\Theta,h)$ there exist evolution operators $R_k \in \mathcal{S}_{t_k}(\omega_2)$ with $\|S_k - R_k\| \le L\|\omega_1 - \omega_2\|$. We extract a convergent subsequence from the sequence $\{R_k\}_{k\in\mathbb{N}}$ with limit R. Then $\|S - R\| \le L\|\omega_1 - \omega_2\|$. By symmetry this implies the assertion. The proof for (6.5) is, of course, exactly the same. ∎

We now define for $\omega \in \Pi(\Theta,h)$ the function $v_\omega : \mathbb{K}^n \to \mathbb{R}_+$ by setting

$$v_\omega(x) := \max\{\|Sx\| \mid S \in \mathcal{S}_\infty(\omega)\}. \tag{6.6}$$

Using Lemma 6.1(iii) and (vi) it is easy to see that for every $\omega \in \Pi(\Theta,h)$ the function defined in (6.6) is a norm on $\mathbb{K}^n$. The following result shows that in this manner we have defined a family of parameterized Lyapunov functions for our system.

THEOREM 6.4. *Consider system* (2.1) *with* (A1)–(A5)*. Assume that* $A(\Theta)$ *is irreducible and let* $\omega \in \Pi(\Theta,h)$ *be arbitrary. Then*

(i) *for all* $u \in \mathcal{U}(\omega), t \ge 0$ *and all* $x \in \mathbb{K}^n$ *it holds that*

$$v_\zeta(\Phi_u(t,0)x) \le e^{\hat\rho t}\, v_\omega(x) \tag{6.7}$$

whenever $\Phi_u(t,0) \in \mathcal{R}_t(\omega,\zeta)$ *for* $\zeta \in \Pi(\Theta,h)$*. In particular, for all* $t \ge s \ge 0$ *it holds that*

$$v_{u(t^-),\tau^-(u,t)}(\Phi_u(t,0)x) \le e^{\hat\rho(t-s)}\, v_{u(s^-),\tau^-(u,s)}(\Phi(s,0)x).$$

(ii) *for every* $x \in \mathbb{K}^n$*,* $\omega \in \Pi(\Theta,h)$*, and* $t \ge 0$ *there exist* $u \in \mathcal{U}(\omega)$ *and a piecewise continuous map* $\zeta : [0,t] \to \Pi(\Theta,h)$*, with* $\zeta(0) = \omega$*, and such that for all* $s \in [0,t]$ *we have*

$$v_{\zeta(s)}(\Phi_u(s,0)x) = e^{\hat\rho s}\, v_\omega(x).$$

If $h = \infty$*, then* ζ *may be chosen to be continuous. If* $h < \infty$ *and* $\omega = (\theta,\tau) \in \Theta \times [0,h)$*, the function* ζ *may be chosen so that its discontinuities on* $[0,t)$ *coincide with those of* u*. Otherwise,* ζ *may have one further discontinuity at* 0*.*

Proof. Without loss of generality, we may assume that $\hat\rho = 0$.

(i) Fix $u \in \mathcal{U}(\omega), t \ge 0$ and assume that $\Phi_u(t,0) \in \mathcal{R}_t(\omega,\zeta)$ for a suitable $\zeta \in \Pi(\Theta,h)$. Assume furthermore that $v_\zeta(\Phi_u(t,0)x) > v_\omega(x)$. Then by definition $\|T\Phi_u(t,0)x\| > v_\omega(x)$ for some $T \in \mathcal{S}_\infty(\zeta)$. Now Lemma 6.1(iv) shows that $T\Phi_u(t,0) \in \mathcal{S}_\infty(\omega)$. Therefore $v_\omega(x) \ge \|T\Phi_u(t,0)x\|$, which is a contradiction.

The second assertion is simply a special case of the first statement.

(ii) Fix $x \in \mathbb{K}^n$, $\omega \in \Pi(\Theta,h)$, and $t \ge 0$ and let $S \in \mathcal{S}_\infty(\omega)$ be such that $\|Sx\| = v_\omega(x)$. By Lemma 6.1(iv) there exist $\hat\zeta \in \Pi(\Theta,h)$ and $\Phi_u(t,0) \in \mathcal{R}_t(\omega,\hat\zeta)$, $T \in \mathcal{S}_\infty(\hat\zeta)$ such that $S = T\Phi_u(t,0)$.

If $h = \infty$, we set $\zeta(s) = u(s)$, $s \in [0,t]$. To treat the case $h \in (0,\infty)$ let $\omega = (\theta,\tau)$. If $0 \le \tau < h$ and $t_0 \le t$ is the smallest positive discontinuity of u, then define $\zeta(s) = (u(s), \min\{\tau + s, h\})$ for $s \in [0,t_0]$ and $\zeta(s) = (u(t^-),\tau^-(u,t))$ for $s \in (t_0,t)$ and $\zeta(t) = \hat\zeta$. This is clearly a piecewise continuous map, whose discontinuities coincide with those

of u on $[0,t)$ and which satisfies $\zeta(0) = (\theta,\tau)$ as by assumption $u(0) = \theta$. This construction also works if $\omega = (\theta,h)$ and $u(0) = \theta$. Otherwise, if $u(0) \neq \omega$, we define $\zeta(0) = \omega$ and $\zeta(s) = (u(t^-), \tau^-(u,t))$ for $s \in (0,t)$ and $\zeta(t) = \hat{\zeta}$.

In all, ζ is defined in such a manner that for all $s \in (0,t]$ we have $\Phi_u(s,0) \in \mathcal{R}(\omega,\zeta(s))$, and for $s \in [0,t)$ it holds that $u(s+\cdot) \in \mathcal{U}(\zeta(s))$. Then it follows from Lemma 6.1(iv) that $T\Phi_u(t,s) \in \mathcal{S}_\infty(\zeta(s))$ for $s \in [0,t]$ and we have by part (i) for $s \in [0,t]$ that

$$v_\omega(x) = \|Sx\| = \|T\Phi_u(t,s)\Phi_u(s,0)x\| \leq v_{\zeta(s)}(\Phi_u(s,0)x) \leq v_\omega(x).$$

This concludes the proof. □

The previous result has a particularly easy interpretation in the case of linear switching systems, which we briefly discuss. Let $A(\Theta) = \{A_1,\dots,A_m\}$ be a finite, irreducible set and assume we are given a dwell time $h \in (0,\infty)$. As the system has no other possibility than to stay in a certain A_i for a time period of at least length h after a discontinuity, we see that for $\tau \in [0,h)$ we have $\mathcal{S}_\infty(i,\tau) = \mathcal{S}_\infty(i,h)e^{-\hat{\rho}(h-\tau)}e^{A(i)(h-\tau)}$. Thus the norms $v_{i,\tau}$ are related through the equality

$$v_{i,\tau}(x) = e^{-\hat{\rho}(h-\tau)}v_{i,h}(e^{A(i)(h-\tau)}x), \quad \tau \in [0,h].$$

It is therefore sufficient to consider the norms $v_i := v_{i,h}$. If we investigate (6.7) with this in mind, we see that after discontinuities this equation contains no information. To be precise, if u has a discontinuity at 0 and $u(t) = i, t \in [0,h)$, then for $t \in [0,h)$ (6.7) is equivalent to the tautology $v_i(e^{A(i)h}x) = v_i(e^{A(i)h}x)$. So after switching, a transient phase is allowed. The interesting information is contained in the other times and the result yields a finite number of norms, which are of interest. We summarize this in the following statement.

COROLLARY 6.5. *Let $\{A_1,\dots,A_m\} \subset \mathbb{K}^{n\times n}$ be a finite irreducible set and let $h \in (0,\infty)$. Then the following two statements are equivalent:*

(i) $\hat{\rho}(A_1,\dots,A_m,h) \leq \rho$.

(ii) *There are norms $v_1,\dots,v_m$ on $\mathbb{K}^n$ with the following properties:*

$$v_i(e^{A_i t}x) \leq e^{\rho t}v_i(x) \quad \text{for all } t \geq 0, x \in \mathbb{K}^n, i = 1,\dots,m, \tag{6.8}$$

$$v_j(e^{A_j t}x) \leq e^{\rho t}v_i(x) \quad \text{for all } t \geq h, x \in \mathbb{K}^n, i,j = 1,\dots,m. \tag{6.9}$$

Proof. (i)⇒(ii): By assumption we may apply the results of Theorem 6.4 to the system $\Sigma = (\Theta,\Theta_1,h,A)$ given by $\Theta = \{1,\dots,m\}, \Theta_1 = \{0\}, A(i) = A_i$. Define the norms v_i by $v_i := v_{i,h}$, where $v_{i,h}$ is defined according to (6.6). Now consider the admissible parameter variation $u \equiv i$. For this we have $u \in \mathcal{U}(i,h)$ and $\Phi_u(t,0) = e^{A_i t} \in \mathcal{R}_t((i,h),(i,h))$ for all $t \geq 0$ so that (6.7) implies (6.8). If we consider

$$u(t) := \begin{cases} i & \text{for } t < 0, \\ j & \text{for } t \geq 0, \end{cases}$$

then $u \in \mathcal{U}(i,h)$ and $\Phi_u(t,0) = e^{A_j t} \in \mathcal{R}_t((i,h),(j,h))$ for all $t \geq h$. In this case, (6.7) implies (6.9).

(ii)⇒(i): By the discussion in section 2 and by Corollary 4.5, it is sufficient to show that all Lyapunov exponents $\lambda(x,u)$ are upper bounded by ρ. So fix $0 \neq x \in \mathbb{K}^n$ and an admissible parameter variation u. If u has no discontinuities on an interval of the form (a,∞), where $a \geq 0$, the assertion is obvious from (6.8). Otherwise let $t_0,t_1,\dots$ denote the switching times of u and let $i(k)$ be such that $u(t) = i(k)$

for $t \in [t_k, t_{k+1})$. Without loss of generality let $t_0 = 0$, which we may assume as $\lambda(x,u) = \lambda(x, u(\cdot - t_0))$. Then we have by (6.8) that

$$v_{i(0)}(\exp(A_{i(0)}t)x) \leq e^{\rho t} v_{i(0)}(x) \quad \text{for } t \in [t_0 + h, t_1],$$

and so for $t \in [t_1 + h, t_2]$ it follows, again using (6.9), that

$$\begin{aligned} v_{i(1)}(\Phi_u(t,0)x) &= v_{i(1)}(\exp(A_{i(1)}(t-t_1))\exp(A_{i(0)}t_1)x) \\ &\leq e^{\rho(t-t_1)} v_{i(0)}(\exp(A_{i(0)}t_1)x) \leq e^{\rho t} v_{i(0)}(x). \end{aligned}$$

By induction we obtain for $t \in [t_k + h, t_{k+1}]$ that

$$\frac{1}{t}\log\left(v_{i(k)}(\Phi_u(t,0)x)\right) \leq \rho + \frac{1}{t}\log(v_{i(0)}(x)).$$

As the growth in the intervals $[t_k, t_k + h]$ is bounded, and as $v_{i(0)} \leq C v_i, i = 1, \ldots, m$, for a suitable constant C, this implies that $\lambda(x,u) \leq \rho$, as desired. ∎

We are now aiming at a continuity result for the norms v_ω. To this end we need a notion of distance between norms. We therefore introduce the space of continuous, positively homogeneous functions on $\mathbb{K}^n$ defined by

$$\text{Hom}\,(\mathbb{K}^n, \mathbb{R}) := \{f : \mathbb{K}^n \to \mathbb{R} \,|\, \forall \alpha \geq 0 : f(\alpha x) = \alpha f(x) \text{ and } f \text{ is continuous on } \mathbb{K}^n\}.$$

Clearly, all norms on $\mathbb{K}^n$ are elements of Hom $(\mathbb{K}^n, \mathbb{R})$. This space becomes a Banach space if equipped with the norm

$$\|f\|_{\infty,\text{hom}} := \max\{|f(x)| \mid \|x\|_2 = 1\}.$$

PROPOSITION 6.6. *Consider system* (2.1) *with* (A1)–(A5). *Assume that* $A(\Theta)$ *is irreducible and let* (A6) *hold. Then the map*

$$\omega \mapsto v_\omega \tag{6.10}$$

is Lipschitz continuous from $\Pi(\Theta, h)$ *to* Hom $(\mathbb{K}^n, \mathbb{R})$.

Proof. Fix $\omega, \zeta \in \Pi(\Theta, h)$. By definition we have

$$\|v_\omega - v_\zeta\|_{\infty,\text{hom}} = \max_{\|x\|_2 = 1} |v_\omega(x) - v_\zeta(x)|.$$

Fix $x \in \mathbb{K}^n$ and let $v_\omega(x) = \|\tilde{S}x\|$ for a suitable $\tilde{S} \in \mathcal{S}_\infty(\omega)$. Then there is a $T \in \mathcal{S}_\infty(\zeta)$ such that $\|\tilde{S} - T\| \leq d_H(\mathcal{S}_\infty(\omega), \mathcal{S}_\infty(\zeta))$ and we obtain

$$v_\omega(x) - v_\zeta(x) \leq \|\tilde{S}x\| - \|Tx\| \leq \|\tilde{S} - T\|\|x\| \leq C d_H(\mathcal{S}_\infty(\omega), \mathcal{S}_\infty(\zeta))\|x\|_2,$$

where C is a constant such that $\|x\| \leq C\|x\|_2$. This shows that

$$\|v_\omega - v_\zeta\|_{\infty,\text{hom}} \leq C d_H(\mathcal{S}_\infty(\omega), \mathcal{S}_\infty(\zeta)).$$

Now the assertion follows from Corollary 6.3. ∎

COROLLARY 6.7. *Consider system* (2.1) *with* (A1)–(A5). *Assume that* $A(\Theta)$ *is irreducible and let one of the following assumptions be satisfied:*

(a) $h \in (0, \infty)$.

(b) $h = \infty$ *and* (A6) *is satisfied.*

Then there exists a constant $1 \leq C \in \mathbb{R}$ *such that for all* $\omega, \zeta \in \Pi(\Theta, h)$ *and all* $x \in \mathbb{K}^n$ *we have*

$$C^{-1} v_\omega(x) \leq v_\zeta(x) \leq C v_\omega(x). \tag{6.11}$$

Proof. We may assume that $\hat{\rho} = 0$.

It is clearly sufficient to prove the inequality on the left-hand side, as the other follows by symmetry. Let $\omega, \zeta \in \Pi(\Theta, h)$ be arbitrary. Fix $x \in \mathbb{K}^n$ and let $S \in \mathcal{S}_\infty(\omega)$ be such that $v_\omega(x) = \|Sx\|$. Fix an arbitrary $\omega_0 \in \Pi(\Theta, h)$. Using Remark 4.3 we have that $\mathcal{R}_{\leq \max\{2h, \bar{c}\}}(\omega, \omega_0), \mathcal{R}_{\leq \max\{2h, \bar{c}\}}(\omega_0, \zeta) \neq \emptyset$. By Lemma 5.2 there exists $\varepsilon > 0$ such that for all $x \in \mathbb{K}^n, B \in \mathbb{K}^{n \times n}$ there is an $R \in \mathcal{R}_\infty(\omega_0, \omega_0)$ with

$$\|BRx\| \geq \varepsilon \|B\| \|x\|.$$

Choose $T_1 \in \mathcal{R}_{s_1}(\zeta, \omega_0), T_2 \in \mathcal{R}_{s_2}(\omega_0, \omega)$ for $s_1, s_2 \leq \max\{2h, \bar{c}\}$. Then we may choose $R \in \mathcal{R}_\infty(\omega_0, \omega_0)$ such that $ST_2RT_1 \in \mathcal{S}_\infty(\zeta)$ (by Lemma 6.1(iv)) and so that

$$\begin{aligned} v_\zeta(x) &\geq \|ST_2RT_1x\| \geq \varepsilon \|ST_2\| \|T_1x\| \\ &\geq \varepsilon (\min\{\|\Phi_u(s,0)^{-1}\|^{-1} \mid u \in \mathcal{U}, s \in [0, \max\{2h, \bar{c}\}]\})^2 \|Sx\| \geq C^{-1} v_\omega(x) \end{aligned}$$

for a constant $C \geq 1$ and independent of ω, ζ, and x. This shows the assertion. ∎

Remark 6.8. Note that the construction of parameterized Lyapunov functions for reducible systems is now an easy exercise by using the upper block triangular structure (5.1). In general, however, only a decay of $\hat{\rho} + \varepsilon$, where $\varepsilon > 0$ is arbitrary, will be achievable. See [16, 34] for related results in the case of linear inclusions.

7. The Gelfand formula. In this section we give an application of the existence of the parameterized Lyapunov functions we have described so far. One of the classical results in the analysis of families of linear time-varying systems states that under certain conditions the exponential growth rate can be approximated by just considering the subset of periodic systems within the family. Results to this effect can be found in [7, 11, 13, 16]. We now show that the same statement is true for our class of systems. In our case periodicity of the underlying parameter variation is the natural assumption, which is analyzed in what follows.

For $t \in \mathbb{R}_+$ we define the set of evolution operators corresponding to periodic $u \in \mathcal{U}$ by

$$\mathcal{P}_t := \bigcup_{\omega \in \Pi(\Theta, h)} \mathcal{R}_t(\omega, \omega).$$

Then we may define the normalized supremum over the spectral radii by

$$\bar{\rho}_t := \sup \left\{ \frac{1}{t} \log r(S) \mid S \in \mathcal{P}_t \right\},$$

and the supremum of the exponential growth rates obtainable by periodic parameter variations is defined by

$$\bar{\rho} := \limsup_{t \to \infty} \bar{\rho}_t.$$

As it is clear that $\bar{\rho}_t \leq \hat{\rho}_t$ for all $t \geq 0$, we obtain immediately that $\bar{\rho} \leq \hat{\rho}$. We intend to show that these quantities are equal. To this end we need the following lemma.

LEMMA 7.1. *Consider system* (2.1) *with* (A1)–(A5). *Assume that* $A(\Theta)$ *is irreducible and let one of the following assumptions be satisfied:*

(a) $h \in (0, \infty)$.
(b) $h = \infty$ *and* (A6) *is satisfied.*
Then there exist $\omega \in \Pi(\Theta, h)$, $x \in \mathbb{K}^n, v_\omega(x) = 1$, *and a sequence* $S_k \in \mathcal{R}_{t_k}(\omega, \omega)$, $t_k \geq 1$, *with*

$$e^{-\hat{\rho} t_k} S_k x \to x.$$

Proof. We may assume that $\hat{\rho} = 0$. Pick an arbitrary $\omega_0 \in \Pi(\Theta, h)$ and $z \in \mathbb{K}^n$ such that $v_{\omega_0}(z) = 1$. By Theorem 6.4(ii) there exist an ω_1 and $S_1 \in \mathcal{R}_1(\omega_0, \omega_1)$ such that $v_{\omega_1}(S_1 z) = v_{\omega_0}(z) = 1$. Applying this argument again, we see that there exist ω_2 and $S_2 \in \mathcal{R}_1(\omega_1, \omega_2)$ such that $v_{\omega_2}(S_2 S_1 z) = 1$. Repeating this argument inductively we obtain sequences $\{\omega_k\}_{k\in\mathbb{N}}$ and $\{S_k\}_{k\in\mathbb{N}}$ with

$$v_{\omega_k}(S_k S_{k-1}, \ldots, S_1 z) = 1 \quad \text{for all } k \in \mathbb{N}.$$

As $\Pi(\Theta, h)$ is compact, there exists a convergent subsequence $\omega_{k_l} \to \omega \in \Pi(\Theta, h)$. Applying Corollary 6.7 we may assume, without loss of generality, that $z_{k_l} := S_{k_l} S_{k_l - 1}, \ldots, S_1 z \to x$. We denote $T_{k_l} := S_{k_l} S_{k_l - 1}, \ldots, S_{k_{l-1}} \in \mathcal{R}(\omega_{k_{l-1}}, \omega_{k_l})$. After relabeling we return to the index k.

Now by Lemma 4.6(vi) and using assumption (a) or (b), the map $(\omega, \zeta) \to \mathcal{R}_t(\omega, \zeta)$ is upper semicontinuous uniformly in t (which is crucial, as we have no control over the length of the intervals needed to define the sequence $\{T_k\}$). Thus by convergence of $\omega_k \to \omega$ and for every $\varepsilon > 0$ there exists a k_0 such that for every $k \geq k_0$ there exists an $R_k \in \mathcal{R}(\omega, \omega)$ with $\|T_k - R_k\| < \varepsilon$ and so that $v_\omega(z_k - x) \leq \varepsilon$. Then we obtain that

$$\begin{aligned} v_\omega(R_k x - x) &\leq v_\omega(R_k - T_k) v_\omega(x) + v_\omega(T_k x - T_k z_k) + v_\omega(z_{k+1} - x) \\ &\leq \varepsilon \left(v_\omega(x) + v_\omega(T_k) + 1 \right). \end{aligned}$$

This implies that there exists a sequence $\{R_k\} \subset \mathcal{R}(\omega, \omega)$ with $R_k x - x \to 0$, as desired. □

Before we can state the main result of this section, we need a further observation for the case $h = \infty$.

PROPOSITION 7.2. *Let* $\Theta, \Theta_1 \in \mathrm{Co}\,(\mathbb{K}^m)$, $A \in C(\mathbb{K}^m, \mathbb{K}^{n\times n})$, *and* $h = \infty$ *satisfying* (A1)–(A5) *be given. Let* Θ_2 *be the largest convex set contained in* Θ_1 *such that* $0 \in \mathrm{ri}\,\Theta_2$. *Then*

$$\hat{\rho}(\infty, \Theta, \Theta_1, A) = \hat{\rho}(\infty, \Theta, \Theta_2, A).$$

Proof. It is clear that $\hat{\rho}(\infty, \Theta, \Theta_1, A) \geq \hat{\rho}(\infty, \Theta, \Theta_2, A)$ so that we only have to show the converse direction.

If $0 \in \mathrm{ri}\,\Theta_1$, there is nothing to show. Otherwise denote by X_2 the linear subspace generated by Θ_2 and denote by $X_2^\perp$ its orthogonal complement. Recall the definition (2.4) and choose $\theta(\cdot) \in \mathcal{U}$ such that for some $x_0 \neq 0$ we have

$$\hat{\rho}(\infty, \Theta, \Theta_1, A) = \lambda(x_0, \theta(\cdot)).$$

As mentioned before, this choice is possible using Corollary 4.5 and [13, Prop. 5.4.15].

Now θ may be decomposed as $\theta = \theta_1 + \theta_2$ such that $\dot{\theta}_1 : \mathbb{R}_+ \to X_2^\perp$ and $\dot{\theta}_2 : \mathbb{R}_+ \to \Theta_2$. Furthermore, as 0 is contained in the boundary of Θ_1, there exists a supporting hyperplane X in 0, which has to contain X_2. Hence there is a vector $d \neq 0$ such that

$\langle d, \dot\theta_1(t)\rangle \geq 0$ and $\langle d, \dot\theta_2(t)\rangle \equiv 0$ for all $t \geq 0$. Now Θ is compact and so $\langle d, \theta\rangle$ is bounded over $\theta \in \Theta$. This implies that the expression

$$c := \langle d, \theta(0)\rangle + \int_0^\infty \langle d, \dot\theta_1(t)\rangle dt = \lim_{t\to\infty} \langle d, \theta(t)\rangle$$

is well defined. If we introduce the set $\Theta_c := \{\eta \in \Theta \mid \langle d, \eta\rangle = c\}$, we see that

$$\operatorname{dist}(\theta(t), \Theta_c) \to 0 \text{ for } t \to \infty.$$

Thus for the set $\Theta_{c,\varepsilon} := \{\eta \in \Theta \mid \operatorname{dist}(\eta, \Theta_c) \leq \varepsilon\}$ we obtain $\theta(t) \in \Theta_\varepsilon$ for all t large enough. This implies that for all $\varepsilon > 0$ and for t large enough we have

$$\hat\rho(\infty, \Theta, \Theta_1, A) \geq \hat\rho(\infty, \Theta_{c,\varepsilon}, \Theta_1, A) \geq \lambda(\Phi_\theta(t,0)x_0, \theta(t+\cdot)) = \lambda(x_0, \theta(\cdot))$$

so that equality holds throughout. Now by Proposition 4.8 it follows that

$$\hat\rho(\infty, \Theta_c, \Theta_1, A) \geq \lim_{\varepsilon\to 0} \hat\rho(\infty, \Theta_{c,\varepsilon}, \Theta_1, A) = \hat\rho(\infty, \Theta, \Theta_1, A),$$

and the converse inequality holds because $\Theta_c \subset \Theta$. Furthermore, any admissible parameter variation with derivative in Θ_1, that remains in Θ_c, has to satisfy $\langle d, \theta(t)\rangle \equiv 0$. This implies $\langle d, \dot\theta(t)\rangle = 0$ a.e., from which it follows that $\dot\theta(t) \in \Theta_2$ a.e. Hence we have

$$\hat\rho(\infty, \Theta_c, \Theta_1, A) = \hat\rho(\infty, \Theta_c, \Theta_2, A).$$

This completes the proof. □

We are now ready to prove the main result of this section: the exponential growth rate $\hat\rho$ coincides with the maximum of the growth rates corresponding to periodic parameter variations $\bar\rho$.

THEOREM 7.3. *Consider a system* $\Sigma = (h, \Theta, \Theta_1, A)$ *satisfying* (A1)–(A5); *then*

$$\bar\rho(h, \Theta, \Theta_1, A) = \hat\rho(h, \Theta, \Theta_1, A). \tag{7.1}$$

Proof. Without loss of generality we may assume that $\hat\rho = 0$.

If $h = \infty$ and (A6) does not hold, then we may first assume that $0 \in \operatorname{ri}\Theta_1$ using Proposition 7.2. Let $X = \operatorname{span}\Theta_1$. Then with the notation $\Theta_z := \Theta \cap (z + X)$ we may write

$$\Theta = \bigcup_{z\in X^\perp} \Theta_z.$$

As each (nonempty) Θ_z is invariant under parameter variations with derivative in Θ_1, we see that

$$\hat\rho(\infty, \Theta, \Theta_1, A) = \sup_{z, \Theta_z \neq \emptyset} \hat\rho(\infty, \Theta_z, \Theta_1, A).$$

Thus if we can show the assertion for each of the terms on the right-hand side, it follows also for $(\infty, \Theta, \Theta_1, A)$. Note that (A6) is satisfied for $(\infty, \Theta_z, \Theta_1, A)$ so that from now on we may assume that $h \in (0, \infty)$ or (A6) is satisfied.

Furthermore, if $A(\Theta)$ is reducible, then there exists a regular $T \in \mathbb{K}^{n\times n}$ such that all matrices $A_0 \in A(\Theta)$ can be transformed to upper block triangular form as in (5.1). For this form it is easy to see that

$$\hat\rho(A,\mathcal{U}) = \max_{i=1,\dots,d} \hat\rho(A_i,\mathcal{U}) \quad \text{and} \quad \bar\rho(A,\mathcal{U}) = \max_{i=1,\dots,d} \bar\rho(A_i,\mathcal{U}). \tag{7.2}$$

Hence, if we show (7.1) for each of the irreducible blocks, then it follows for the overall system.

So assume now that $A(\Theta)$ is irreducible and that $h \in (0,\infty)$ or (A6) holds. By Lemma 7.1 there exist $\omega \in \Pi(\Theta,h), x \in \mathbb{K}^n, v_\omega(x) = 1$, and a sequence $S_k \in \mathcal{R}(\omega,\omega)$ such that $S_k x - x \to 0$. Then we have by [16, Lem. 2] for the eigenvalues $\lambda_i(k)$ of S_k that

$$0 \le \min_{1\le i\le n} 1 - |\lambda_i(k)| \le \min_{1\le i\le n} |1-\lambda_i(k)| \le C\|S_k x - x\|^{1/n},$$

where C is a constant depending only on the upper bound of $\|S_k\|$. Denoting by $\tilde\lambda_k$ an eigenvalue of S_k for which the minimum on the left is attained, we see that $|\tilde\lambda_k| \to 1$ as $k \to \infty$. As we have $|\tilde\lambda_k| \le 1$ and $t_k \ge 1$, we obtain $\bar\rho \ge 1/t_k \log|\tilde\lambda_k| \ge \log|\tilde\lambda_k|$, and it follows that $\bar\rho \ge 0$. This completes the proof. □

Remark 7.4. Note that the proof of the previous result shows for the particular case $\hat\rho = 0$ that

$$\limsup_{t\to\infty} \max\{r(S) \mid S \in \mathcal{P}_t\} = 0$$

holds. This statement is slightly stronger than that of Theorem 7.3.

8. Continuity of the exponential growth rate. One of the basic questions in stability theory is whether stability is a robust property in the space of systems. A first step toward answering this question is obtained by showing that the exponential growth rate is an upper semicontinuous function, because then the set of exponentially stable systems given by $\{\hat\rho < 0\}$ is open. It is, however, even more desirable to have continuous dependence of the growth rate on the data. We will first show that the Gelfand formula, which we just proved in Theorem 7.3, allows for an easy criterion of continuity. Unfortunately, we then have to present an example showing that, in the setup we have studied so far, $\hat\rho$ is not a continuous function of the data.

COROLLARY 8.1. *Let $\mathcal{N}$ be a subset of $\mathcal{L}$ such that the maps*

$$(h,\Theta,\Theta_1,A) \mapsto \mathcal{S}_t(h,\Theta,\Theta_1,A)$$

are continuous on $\mathcal{N}$ for all t large enough. Then the map

$$(h,\Theta,\Theta_1,A) \mapsto \hat\rho(h,\Theta,\Theta_1,A)$$

is continuous on $\mathcal{N}$.

Proof. We already know that $\hat\rho$ is upper semicontinuous on $\mathcal{N}$ by Proposition 4.8. The assumption implies that the maps $\bar\rho_t : \mathcal{N} \to \mathbb{R}$ are continuous for all t large enough by continuity of the spectral radius. Now $\bar\rho = \sup_{t>0} \bar\rho_t$ is lower semicontinuous as the supremum of continuous functions. Using Theorem 7.3, the function $\hat\rho = \bar\rho$ is both upper semicontinuous and lower semicontinuous, and thus continuous on $\mathcal{N}$. □

Example 8.2. Let $h = \infty$, $\Theta_1 := [-1,1] \times \{0\} \subset \mathbb{R}^2$, define $\Theta(0) = [0,2\pi] \times \{0\}$, and define the sets $\Theta(\phi) = B_\phi \Theta(0)$, where B_ϕ is the rotation matrix

$$B_\phi := \begin{bmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{bmatrix}, \quad \phi \in (-\pi,\pi).$$

Define furthermore

$$A(\theta_1,\theta_2) = \begin{bmatrix} -1+3/2\cos^2\theta_1 & 1-3/2\sin\theta_1\cos\theta_1 \\ -1-3/2\sin\theta_1\cos\theta_1 & -1+3/2\sin^2\theta_1 \end{bmatrix}.$$

(The reader will most likely recognize the famous example of a periodic system of Hurwitz stable matrices that is unstable; see, e.g., [14, 33]. We recall the well-known fact that the characteristic polynomial of $A(\theta_1, \theta_2)$ is equal to $p(z) = z^2 + 1/2z + 1/2$ with zeros $-1/4 \pm i\sqrt{7}/4$ independent of θ.)

We will show that the exponential growth rate as a function of $\Theta(\phi)$ with all the other data left fixed has a discontinuity at 0. Clearly, the map $\phi \mapsto \Theta(\phi)$ is Lipschitz continuous.

For $0 \neq \phi \in (-\pi, \pi)$ only the constant functions are admissible parameter variations because Θ_1 only allows for variations in the first component. Hence for $\phi \neq 0$ we have $\hat{\rho}(\phi) = \max\{\operatorname{Re}\lambda \mid \lambda \in \sigma(B); B \in A(\Theta(\phi))\} = -1/4$.

On the other hand, time-varying systems are possible for $\phi = 0$ because $\Theta(0)$ is collinear to the admissible derivatives in Θ_1. In particular, we cannot expect the assumption of Corollary 8.1 to be satisfied, as with time-varying parameter variations we expect to be able to construct a much richer set of transition operators. In particular, if we define the admissible parameter variation

$$\theta(t) = \begin{cases} t, & t \in [0, 2\pi], \\ 4\pi - t, & t \in [2\pi, 4\pi], \end{cases}$$

and continue this function periodically, then we have the classical example on the interval $[0, 2\pi]$, where it is well known that

$$\Phi_\theta(2\pi, 0) = \begin{bmatrix} e^{\pi} & 0 \\ 0 & e^{-2\pi} \end{bmatrix}.$$

For the calculation of $\Phi_\theta(4\pi, 2\pi)$, numerical evaluation yields

$$\Phi_\theta(4\pi, 2\pi) = \begin{bmatrix} 0.0597 & -0.178 \\ 0.178 & 0.1932 \end{bmatrix}.$$

By calculating the spectral radius $r(\Phi_\theta(4\pi, 0)) = r(\Phi_\theta(4\pi, 2\pi)\Phi_\theta(2\pi, 0)) \approx 1.3799$, we see that the exponential growth rate corresponding to $\Theta(0)$ is positive.

The previous example is a bit unfair because the constraint on the derivative that can be effectively used is simply $\Theta_1 = \{0\}$ for $\phi \neq 0$. Another way of saying this is that there is a discontinuity hidden in the data in the previous example: at $\phi = 0$ the derivative constraint set changes discontinuously from $\{0\}$ to Θ_1. This shows that, so far, we were too lenient in our description of the system data.

With reasonable extra assumptions, however, it is possible to obtain (Lipschitz) continuity results in the spirit of [34], which for reasons of space appears in [36]; see also [37].

9. Conclusions. In this paper we have studied certain classes of families of LPV systems that are basically described by constraints on the distance between discontinuities and on the derivative in the time between discontinuities. Both the LPV system class and linear switching system class are special cases of the presented setup. For these classes parameter-dependent Lyapunov functions, which are norms for each fixed parameter, have been constructed in such a way that the resulting Lyapunov function characterizes the exponential growth rate in an infinitesimal manner. This result complements constructions of Lyapunov functions for linear inclusions in [5, 26, 34]. It was shown how the existence of such norms can be used to obtain a fairly simple proof of the Gelfand formula in this case. Conditions for continuous dependence of the growth rate on the data can be derived using the tools developed here. This is discussed in [36].

Acknowledgment. The hospitality of the members of the Hamilton Institute has been very much appreciated.

REFERENCES

[1] A. A. Agrachev and D. Liberzon, *Lie-algebraic stability criteria for switched systems*, SIAM J. Control Optim., 40 (2001), pp. 253–269.
[2] F. Amato, M. Corless, M. Mattei, and R. Setola, *A multivariable stability margin in the presence of time-varying bounded rate gains*, Internat. J. Robust Nonlinear Control, 7 (1997), pp. 127–143.
[3] P. Apkarian and R. J. Adams, *Advanced gain-scheduling techniques for uncertain systems*, in Advances in Linear Matrix Inequality Methods in Control, L. El Ghaoui and S.-I. Niculescu, eds., SIAM, Philadelphia, 2000, pp. 209–228.
[4] P. Apkarian and H. D. Tuan, *Parameterized LMIs in control theory*, SIAM J. Control Optim., 38 (2000), pp. 1241–1264.
[5] N. E. Barabanov, *Absolute characteristic exponent of a class of linear nonstationary systems of differential equations*, Siberian Math. J., 29 (1988), pp. 521–530.
[6] G. Becker and A. Packard, *Robust performance of linear parametrically varying systems using parametrically-dependent linear feedback*, Systems Control Lett., 23 (1994), pp. 205–215.
[7] M. A. Berger and Y. Wang, *Bounded semigroups of matrices*, Linear Algebra Appl., 166 (1992), pp. 21–27.
[8] I. U. Bronstein and A. Ya. Kopanskii, *Smooth Invariant Manifolds and Normal Forms*, World Scientific, Singapore, 1994.
[9] F. H. Clarke, Y. S. Ledyaev, and R. J. Stern, *Asymptotic stability and smooth Lyapunov functions*, J. Differential Equations, 149 (1998), pp. 69–114.
[10] F. H. Clarke, Y. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth Analysis and Control Theory*, Graduate Texts in Math. 178, Springer-Verlag, New York, 1998.
[11] F. Colonius and W. Kliemann, *Maximal and minimal Lyapunov exponents of bilinear control systems*, J. Differential Equations, 101 (1993), pp. 232–275.
[12] F. Colonius and W. Kliemann, *The Lyapunov spectrum of families of time varying matrices*, Amer. Math. Soc. Transl., 348 (1996), pp. 4389–4408.
[13] F. Colonius and W. Kliemann, *The Dynamics of Control*, Birkhäuser Boston, Cambridge, MA, 2000.
[14] W. A. Coppel, *Dichotomies in Stability Theory*, Lecture Notes in Math. 629, Springer-Verlag, Berlin, New York, 1978.
[15] W. P. Dayawansa and C. F. Martin, *A converse Lyapunov theorem for a class of dynamical systems which undergo switching*, IEEE Trans. Automat. Control, 44 (1999), pp. 751–760.
[16] L. Elsner, *The generalized spectral-radius theorem: An analytic-geometric proof*, Linear Algebra Appl., 220 (1995), pp. 151–159.
[17] P. Gahinet, P. Apkarian, and M. Chilali, *Affine parameter-dependent Lyapunov functions and real parametric uncertainty*, IEEE Trans. Automat. Control, 41 (1996), pp. 436–442.
[18] L. Grüne, *A uniform exponential spectrum for linear flows on vector bundles*, J. Dynam. Differential Equations, 12 (2000), pp. 435–448.
[19] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1990.
[20] M. Johansson and A. Rantzer, *Computation of piecewise quadratic Lyapunov functions for hybrid systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 555–559.
[21] U. Jönsson and A. Rantzer, *Systems with uncertain parameters—Time-variations with bounded derivative*, Internat. J. Robust Nonlinear Control, 6 (1996), pp. 969–983.
[22] T. Kato, *Perturbation Theory for Linear Operators*, 2nd ed., Springer-Verlag, Berlin, Heidelberg, New York, 1980.
[23] D. Liberzon, *Switching in Systems and Control*, Systems Control Found. Appl., Birkhäuser Boston, Boston, MA, 2003.
[24] D. Liberzon and A. S. Morse, *Basic problems in stability design and design of switched systems*, IEEE Control Systems Magazine, 19 (1999), pp. 59–70.
[25] Y. Lin, E. D. Sontag, and Y. Wang, *A smooth converse Lyapunov theorem for robust stability*, SIAM J. Control Optim., 34 (1996), pp. 124–160.
[26] A. P. Molchanov and E. S. Pyatnitskij, *Criteria of asymptotic stability of differential and difference inclusions encountered in control theory*, Systems Control Lett., 13 (1989), pp. 59–64.

[27] A. S. Morse and A. Stephen, *Supervisory control of families of linear set-point controllers. I. Exact matching*, IEEE Trans. Automat. Control, 41 (1996), pp. 1413–1431.

[28] A. S. Morse and A. Stephen, *Supervisory control of families of linear set-point controllers. II. Robustness*, IEEE Trans. Automat. Control, 42 (1997), pp. 1500–1515.

[29] H. Radjavi, *On irreducibility of semigroups of compact operators*, Indiana Univ. Math. J., 39 (1990), pp. 499–515.

[30] J. S. Shamma and M. Athans, *Guaranteed properties of gain scheduled control for linear parameter-varying plants*, Automatica J. IFAC, 27 (1991), pp. 559–564.

[31] J. S. Shamma and D. Xiong, *Set-valued methods for linear parameter varying systems*, Automatica, 35 (1999), pp. 1081–1089.

[32] A. R. Teel and L. Praly, *A smooth Lyapunov function from a class-$\mathcal{KL}$ estimate involving two positive semidefinite functions*, ESAIM Control Optim. Calc. Var., 5 (2000), pp. 313–367.

[33] M. Vidyasagar, *Nonlinear Systems Analysis*, 2nd ed., Prentice–Hall, Englewood Cliffs, NJ, 1993.

[34] F. Wirth, *The generalized spectral radius and extremal norms*, Linear Algebra Appl., 342 (2002), pp. 17–40.

[35] F. Wirth, *Parameter dependent extremal norms for linear parameter varying systems*, in Proceedings of the 15th International Symposium on Mathematical Theory of Networks and Systems (MTNS 2002), Notre Dame, IN, 2002, paper 9024 (CD-Rom). Also available online at http://www.nd.edu/∼mtns/papers/9024.pdf

[36] F. Wirth, *On Lipschitz continuity of the top Lyapunov exponent of linear parameter varying and linear switching systems*, Stoch. Dyn., 4 (2004), pp. 461–481.

[37] F. Wirth, *Stability Theory for Perturbed Systems: Joint Spectral Radii and Stability Radii*, Lecture Notes in Math., Springer-Verlag, New York, to appear.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 1, pp. 240–267

COMPUTATION OF THE (J, J')-LOSSLESS FACTORIZATION FOR GENERAL RATIONAL MATRICES*

DELIN CHU† AND DANIEL W. C. HO‡

Abstract. (J, J')-lossless factorization plays a central role in H_∞-control because it gives a simple and unified framework of H_∞-control from the viewpoint of classical network theory, and it includes the well-known inner–outer factorization of rational matrices, Wiener–Hopf factorization, and spectral factorization of positive rational matrices as special cases. However, up to now, there is still a lack of numerically reliable methods for this important factorization problem in a general setting. In this paper, we present necessary and sufficient solvability conditions and develop a numerically reliable algorithm based on a generalized eigenvalue approach for the (J, J')-lossless factorization of general rational matrices.

Key words. (J, J')-lossless factorization, rational matrix, eigenfactorization orthogonal transformation

AMS subject classifications. 93B05, 93B40, 93B52, 65F35

DOI. 10.1137/040609136

1. Introduction.

Throughout this paper the following notation will be used:

- $J \in \mathbf{R}^{p\times p}$ *and* $J' \in \mathbf{R}^{m\times m}$ *are two given symmetric matrices.*
- $M \geq 0$ *means that* M *is symmetric and positive semidefinite.*
- *For any* $M, N \in \mathbf{R}^{n\times n}$ *with* M *nonsingular,* $\rho(M, N)$ *denotes the spectral radius of the pencil* $-sM + N$*, and* $\rho(N) := \rho(I, N)$*.*
- $\mathbf{C}_0, \mathbf{C}_+$ *denote the imaginary axis and open right half complex plane, respectively.*
- $\mathcal{R}^{p\times m}(s), \mathcal{RL}_\infty^{p\times m}(s)$ *denote the set of* $p \times m$ *real rational matrices and set of* $p \times m$ *proper real rational matrices having no poles on* $\mathbf{C}_0$*, respectively.*
- $G(s) = \left[\begin{array}{c|c} -sE + A & B \\ \hline C & D \end{array}\right]$ *means that* $G(s)$ *has a realization* $G(s) = D + C(sE - A)^{-1}B$*.*

The H_∞-control problem has been studied extensively based on several different approaches; see, e.g., [10, 14, 15, 17, 21, 22, 33, 35, 40] and the references therein. Among these approaches, the linear matrix inequality (LMI) approach [21, 22] has become very popular because it converts the H_∞-control problem into LMIs [19], where one works with larger size matrices to keep matrix equations linear rather than having to solve nonlinear matrix equations of Riccati type. This approach employs methods of semidefinite programming to compute the desired optimal H_∞-controllers. This is very attractive, because easy-to-use methods for semidefinite programming are readily available; see, e.g., [16, 20]. However, the computational complexity of this approach for a control plant with dimension n is up to $O(n^6)$ [44], which is rather high. Another remarkable approach is that of (J, J')-lossless factorization, which provides a simple and unified framework of an H_∞-control problem from the point of view of

*Received by the editors May 27, 2004; accepted for publication (in revised form) October 1, 2004; published electronically July 18, 2005.

http://www.siam.org/journals/sicon/44-1/60913.html

†Department of Mathematics, National University of Singapore, 2 Science Drive 2, Singapore 117543 (matchudl@math.nus.edu.sg).

‡Department of Mathematics, City University of Hong Kong, Hong Kong (madaniel@cityu.edu.hk). The work of this author was supported by a grant from RGC of the Hong Kong Special Administrative Region (CityU 101103).

classical network theory [2, 6, 9, 10, 13, 27, 30, 32, 33, 36, 38, 39]. This approach reduces the H_∞-control problem to a (J, J')-lossless factorization of a chain-scattering representation of the system. Generally, the existence and solution of the (J, J')-lossless factorization can be characterized by Riccati equations. Consequently, by the (J, J')-lossless factorization approach, the optimal controllers for the H_∞-control problem are obtained by solving the associated Riccati equations. Numerical methods for solving such Riccati equations have been developed in [4, 28, 34, 44], but these Riccati equations may become very ill-conditioned when the computed optimal H_∞-norm approaches the exact optimal H_∞-norm, which leads to these Riccati equations being very difficult to solve. Therefore, although every existing approach, including the LMI and (J, J')-lossless factorization approaches, has a method of solving the H_∞-control problem, many numerical problems associated with it remain to be studied.

In this paper, motivated by the importance of the (J, J')-lossless factorization for H_∞-control, we study the (J, J')-lossless factorization problem for general rational matrices.

DEFINITION 1 (see [36, 41]). (i) *A matrix* $\Theta(s) \in \mathcal{RL}_\infty^{p\times m}(s)$ *is* (J, J')*-unitary if*

$$\Theta^T(-s)J\Theta(s) = J' \quad \forall s \in \mathbf{C}.$$

(ii) *A matrix* $\Theta(s) \in \mathcal{RL}_\infty^{p\times m}(s)$ *is* (J, J')*-lossless if it is* (J, J')*-unitary and*

$$\Theta^T(\bar{s})J\Theta(s) \le J' \quad \forall s \in \mathbf{C}_0 \cup \mathbf{C}_+,$$

where $\bar{s}$ *is the complex conjugate of* s.

DEFINITION 2 (see [36, 41]). $G(s) \in \mathcal{R}^{p\times m}(s)$ *has a* (J, J')*-lossless factorization if it can be represented as a product*

$$G(s) = \Theta(s)\Xi(s),$$

where $\Theta(s) \in \mathcal{RL}_\infty^{p\times m}(s)$ *is* (J, J')*-lossless, and* $\Xi(s) \in \mathcal{R}^{m\times m}(s)$ *has neither zeros nor poles in* $\mathbf{C}_+$.

The notion of (J, J')-lossless factorization was first introduced in [41] in a geometrical context. It is a generalization of the well-known inner–outer factorization of rational matrices. It also includes spectral factorization and Wiener–Hopf factorization for positive rational matrices as special cases. Some connections between the (J, J')-lossless factorization and the chain-scattering formulation of H_∞-control were discussed in [10, 31]. The (J, J')-lossless factorization of proper rational matrices without zeros on $\mathbf{C}_0$ was studied in [29] based on the theory of conjugation developed in [36]. Later, the (J, J')-lossless factorization of general proper rational matrices was considered in [23], and the solvability conditions and state-space realizations of the factors $\Pi(s)$ and $\Theta(s)$ were derived there, based on a generalized eigenvalue approach. The (J, J')-lossless factorization problems in the setting of discrete-time systems with/without zeros on the unit circle were investigated in [7, 8]. The main existing result for the (J, J')-lossless factorization of general proper rational matrices can be summarized in Theorem 3 below, which is a slight extension of Theorem 1 in [23] and a continuous-time version of Theorem 7 in [7].

Assume that $G(s) = \left[\begin{array}{c|c} -sI + A & B \\ \hline C & D \end{array}\right] \in \mathcal{RL}_\infty^{p\times m}(s)$ is left invertible, i.e.,

$$\max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} -sI + A & B \\ C & D \end{bmatrix} = n + m, \tag{1}$$

where $A \in \mathbf{R}^{n\times n}$, $B \in \mathbf{R}^{n\times m}$, $C \in \mathbf{R}^{p\times n}$, and $D \in \mathbf{R}^{p\times m}$. Then there exist an orthogonal matrix S and a nonsingular matrix T [25, 42] such that

$$S\begin{bmatrix} -sI+A & B \\ C & D \end{bmatrix}T = \begin{array}{c} \begin{matrix} n-n_{0\infty} & n_{0\infty} & m \end{matrix} \\ \begin{bmatrix} -sE_{nf}+A_{nf} & 0 & 0 \\ \star & -sE_{11}+A_{11} & A_{12} \\ \star & A_{21} & A_{22} \end{bmatrix} \end{array} \begin{matrix} \\ \}n_{0\infty} \\ \}m \end{matrix},$$

where E_{nf} is of full column rank, E_{11} is nonsingular, and

$$\begin{aligned} &\text{rank}\,(-sE_{nf}+A_{nf}) = n-n_{0\infty} \quad \forall s \in \mathbf{C}_0, \\ &\text{rank}\begin{bmatrix} -sE_{11}+A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = n_{0\infty}+m \quad \forall s \in \mathbf{C}\backslash\mathbf{C}_0. \end{aligned}$$

Partition S and T into

$$S = \begin{array}{c} \begin{matrix} n & p \end{matrix} \\ \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \\ S_{31} & S_{32} \end{bmatrix} \end{array} \begin{matrix} \\ \}n_{0\infty} \\ \}m \end{matrix}, \quad T = \begin{array}{c} \begin{matrix} n-n_{0\infty} & n_{0\infty} & m \end{matrix} \\ \begin{bmatrix} T_{11} & T_{12} & T_{13} \\ T_{21} & T_{22} & T_{23} \end{bmatrix} \end{array} \begin{matrix} \}n \\ \}m \end{matrix}.$$

Furthermore, let the columns of the full column rank matrix

$$\begin{array}{c} \begin{matrix} r \end{matrix} \\ \begin{bmatrix} L_1 \\ L_2 \\ L_3 \end{bmatrix} \end{array} \begin{matrix} \}n \\ \}n \\ \}m \end{matrix}$$

span the stable eigenspace of the matrix pencil

$$\begin{bmatrix} -sI+A & 0 & B \\ -C^TJC & -sI-A^T & -C^TJD \\ D^TJC & B^T & D^TJD \end{bmatrix},$$

and let there exist a stable matrix $\Lambda \in \mathbf{R}^{r\times r}$ such that

$$\begin{bmatrix} A & 0 & B \\ -C^TJC & -A^T & -C^TJD \\ D^TJC & B^T & D^TJD \end{bmatrix}\begin{bmatrix} L_1 \\ L_2 \\ L_3 \end{bmatrix} = \begin{bmatrix} I_n & 0 & 0 \\ 0 & I_n & 0 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} L_1 \\ L_2 \\ L_3 \end{bmatrix}\Lambda.$$

THEOREM 3 (cf. [7, 23]). *Given a $G(s) \in \mathcal{RL}_\infty^{p\times m}(s)$, let be its stabilizable and detectable realization, i.e.,*

$$\text{rank}\,[-sI+A \;\; B] = \text{rank}\begin{bmatrix} -sI+A \\ C \end{bmatrix} = n \qquad \forall s \in \mathbf{C}_0 \cup \mathbf{C}_+,$$

where $A \in \mathbf{R}^{n\times n}$, $B \in \mathbf{R}^{n\times m}$, $C \in \mathbf{R}^{p\times n}$, and $D \in \mathbf{R}^{p\times m}$. Then $G(s)$ has a (J, J')-lossless factorization if and only if the following conditions hold:

(i) *$G(s)$ is left invertible, i.e., (1) holds true.*

(ii) *There exists a nonsingular matrix $D_0 \in \mathbf{R}^{m\times m}$ such that $D_0^T S_{32} J S_{32}^T D_0 = J'$.*

(iii) *$r + n_{0\infty} = n$, matrix $[L_1\ T_{12}]$ is nonsingular, $X := [L_2\ 0][L_1\ T_{12}]^{-1} \geq 0$, and the algebraic Riccati equation*

$$YA^T + AY + YC^TJCY = 0 \tag{2}$$

has a solution $Y \geq 0$ such that $A + YC^TJC$ is stable.

(iv) *$\rho(XY) < 1$.*

Moreover, if the above conditions hold, then a (J, J')-lossless factorization is given by the factors $\Theta(s)$ and $\Xi(s)$:

$$\Theta(s) = \left[\begin{array}{cc|c} -sI+\Lambda & 0 & Z_1 \\ 0 & -sI+A & Z_2 \\ \hline CL_1+DL_3 & C & -S_{32}^T \end{array}\right] D_0, \tag{3}$$

$$\Xi(s) = -(J')^{-1} D_0^T \left[\begin{array}{c|c} -sI+A+YC^TJC & B+YC^TJD \\ \hline (S_{31}X+S_{32}JC)(I-YX)^{-1} & S_{32}JD \end{array}\right], \tag{4}$$

where

$$Z_1 = -\begin{bmatrix} I_r & 0 \end{bmatrix} \begin{bmatrix} L_1 - YL_2 & T_{12} \end{bmatrix}^{-1} (S_{31}^T + YC^TJS_{32}^T),$$
$$Z_2 = (I-YX)^{-1}Y(XS_{31}^T + C^TJS_{32}^T).$$

It is clear that Theorem 3 excludes all improper rational matrices. It is known that the state-space representation can only be used to describe proper rational matrices, while the descriptor-form representation can be used to describe any rational matrices. In [24], the (J, J')-lossless factorization problem for general rational matrices has been considered using the descriptor-form representation approach based on the concept of J-lossless conjugation [36]. The elegant results in [24] are based on (i) a realization of $G(s) = \left[\begin{array}{c|c} -sE+A & B \\ \hline C & D \end{array}\right]$, which is in standard form (i.e., there do not exist nonsingular matrices M and N and an integer $r > 0$ such that $M(-sE+A)N = \left[\begin{array}{cc} -s\hat{E}+\hat{A} & 0 \\ 0 & I_r \end{array}\right]$) and satisfies $E^2 = E$; (ii) the generalized Lyapunov equation

$$\begin{cases} AYE^T + EYA^T + EYC^TJCYE^T = 0, \quad E \text{ is singular,} \\ A^T + C^TJCYE^T - sE^T \text{ is nonsingular } \forall s \in \mathbf{C}_+, \quad EYE^T \geq 0, \\ \text{the null space of } YE^T \text{ contains the eigenspace of} \\ \quad -sE^T + A^T \text{ corresponding to the eigenvalues on } \mathbf{C}_0 \cup \{\infty\}; \end{cases} \tag{5}$$

and (iii) the existence of matrices $D_\pi \in \mathbf{R}^{m\times m}$, $K \in \mathbf{R}^{m\times n}$ and a (J, J')-lossless matrix D_c satisfying

$$\begin{bmatrix} \tilde{C} & D \end{bmatrix} = D_c \begin{bmatrix} K & D_\pi \end{bmatrix}, \ \tilde{C} \text{ is known.} \tag{6}$$

Theoretically, for any realization of $G(s)$, we can always find a new realization, which is in the standard form and satisfies $E^2 = E$. However, the computation of a realization of $G(s) \in \mathcal{R}^{p\times m}(s)$, which is in standard form and satisfies $E^2 = E$, is very ill-conditioned [37, 42] and cannot be obtained in a numerically reliable manner. This issue is easy to understand; for instance, let us consider a very simple example. Let $G(s)$ be of the form

$$G(s) = \left[\begin{array}{ccc|c} -sE_{11}+A_{11} & A_{12} & A_{13} & B_1 \\ A_{21} & 0 & 0 & B_2 \\ 0 & 0 & A_{33} & B_3 \\ \hline C_1 & C_2 & C_3 & D \end{array}\right], \quad E_{11} \text{ and } A_{33} \text{ are nonsingular.}$$

Then, in order to get a realization of $G(s)$, which is in standard form, we have to compute A_{33}^{-1}. However, it is well known that the computation of A_{33}^{-1} is numerically unstable and will contain a large error if A_{33} is ill-conditioned. This implies that the

computation of a realization of a given rational matrix $G(s)$, which is in standard form, is a difficult task in general and should be avoided if possible. Furthermore, the generalized Lyapunov equation (5) is very difficult to solve, and it is not clear under what conditions there exist D_π, K, and a (J, J')-lossless matrix D_c satisfying (6) because of the requirement that D_c be (J, J')-lossless. Thus, the computation of matrices $D_\pi \in \mathbf{R}^{m\times m}$, $K \in \mathbf{R}^{m\times n}$, and a (J, J')-lossless matrix D_c has to be studied further.

Although the (J, J')-lossless factorization has been studied by many researchers, up to now there is still a lack of numerically reliable methods for solving it with general rational matrices. In this paper we will develop a numerically reliable method to verify the solvability and construct a solution for the (J, J')-lossless factorization problem of general rational matrices. Our idea can be outlined as follows:

- For any rational matrix $G(s) \in \mathcal{R}^{p\times m}(s)$, $G(s)$ has a minimal realization of the descriptor-form $G(s) = [\begin{smallmatrix} -sE+A & B \\ \hline C & 0 \end{smallmatrix}]$ with $E, A \in \mathbf{R}^{n\times n}$, $B \in \mathbf{R}^{n\times m}$, and $C \in \mathbf{R}^{p\times n}$, i.e.,
$$\operatorname{rank}[\alpha E + \beta A \quad B] = \begin{bmatrix} \alpha E + \beta A \\ C \end{bmatrix} = n \quad \forall (\alpha, \beta) \in \mathbf{C}^2 \backslash \{(0, 0)\}.$$
Such a minimal realization can always be obtained from any given realization of $G(s)$ by using the well-known controllability–observability staircase form algorithm [25, 26, 42], which is numerically backward stable.
- $G(s)$ can be factored as
$$G(s) = \left[\begin{array}{c|c} -sE + A + BF & B \\ \hline C & 0 \end{array}\right] \left[\begin{array}{c|c} -sE + A & B \\ \hline -F & I \end{array}\right] \quad \forall F \in \mathbf{R}^{m\times n}.$$
We can choose F so that $G_1(s) := [\begin{smallmatrix} -sE+A+BF & B \\ \hline C & 0 \end{smallmatrix}]$ is proper, $G_2(s) := [\begin{smallmatrix} -sE+A & B \\ \hline -F & I \end{smallmatrix}]$, and $G_2^{-1}(s)$ has neither zeros nor poles in $\mathbf{C}_+$;
- By applying Theorem 3 to $G_1(s)$ we can get necessary and sufficient solvability conditions and a desired solution for the (J, J')-lossless factorization of $G(s)$. We will show that *the solvability conditions and the constructed solution in this way are independent of the parameter matrix F.*

2. Main results. In this section we will present numerically verifiable necessary and sufficient solvability conditions and establish a numerically reliable algorithm for solving the (J, J')-lossless factorization problem of general rational matrices.

The main result in this section is based on the following factorization.

THEOREM 4. *Given a rational matrix $G(s) \in \mathcal{R}^{p\times m}(s)$, let*

$$G(s) = \left[\begin{array}{c|c} -sE + A & B \\ \hline C & 0 \end{array}\right], \; E, A \in \mathbf{R}^{n\times n}, \; B \in \mathbf{R}^{n\times m}, \; C \in \mathbf{R}^{p\times n}, \tag{7}$$

be its minimal realization. There exist nonnegative integers n_1, n_2, and n_3 with $n_1 + n_2 + n_3 = n$, and orthogonal matrices $P, Q, U \in \mathbf{R}^{n\times n}$, $W \in \mathbf{R}^{m\times m}$, and $V \in \mathbf{R}^{(n_1+n_3)\times(n_1+n_3)}$, with U and V being partitioned as

$$U = \begin{array}{c} \begin{matrix} n_1+n_2 & n_3 \end{matrix} \\ \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix} \begin{matrix} \}n_1+n_2 \\ \}n_3 \end{matrix} \end{array}, \quad V = \begin{array}{c} \begin{matrix} n_3 & n_1 \end{matrix} \\ \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \begin{matrix} \}n_3 \\ \}n_1 \end{matrix} \end{array},$$
$$\operatorname{rank}(U_{11}) = n_1 + n_2, \quad \operatorname{rank}(V_{11}) = n_3$$

such that

$$\begin{bmatrix} U_{11} & 0 & U_{12} \\ 0 & I & 0 \\ U_{21} & 0 & U_{22} \end{bmatrix} \begin{bmatrix} P & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & P \end{bmatrix} \begin{bmatrix} -sE+A & B & -sE+A \\ C & 0 & C \\ -sE+A & B & 0 \end{bmatrix}$$
$$\times \begin{bmatrix} Q & 0 & 0 \\ 0 & W & 0 \\ 0 & 0 & Q \end{bmatrix} \begin{bmatrix} I_{n_1+n_2} & 0 & 0 & 0 & 0 \\ 0 & V_{11} & 0 & V_{12} & 0 \\ 0 & 0 & I_m & 0 & 0 \\ 0 & V_{21} & 0 & V_{22} & 0 \\ 0 & 0 & 0 & 0 & I \end{bmatrix}$$

$$= \begin{array}{c} \begin{matrix} n_1 & n_2 & n_3 & n_3 & m-n_3 & n \end{matrix} \\ \begin{bmatrix} -sE_{11}+A_{11} & -sE_{12}+A_{12} & A_{13} & 0 & B_{12} & \star \\ 0 & -sE_{22}+A_{22} & A_{23} & 0 & B_{22} & \star \\ 0 & A_{32} & A_{33} & B_{31} & 0 & \star \\ C_1 & C_2 & C_3 & 0 & 0 & \star \\ \star & \star & \star & \star & \star & \star \end{bmatrix} \end{array} \begin{matrix} \}n_1 \\ \}n_2 \\ \}n_3\,, \\ \}p \\ \}n \end{matrix} \tag{8}$$

where $\star$ *denotes the subblock, which we are not interested in, and*

$$\operatorname{rank}(E_{11}) = n_1,\ \operatorname{rank}(E_{22}) = n_2,\ \operatorname{rank}(B_{31}) = n_3, \tag{9}$$

$$\operatorname{rank} \begin{bmatrix} -sE_{22}+A_{22} & A_{23} \\ A_{32} & A_{33} \end{bmatrix} = n_2 + n_3 \quad \forall s \in \mathbf{C}. \tag{10}$$

Proof. The factorization (8) with properties (9) and (10) is constructed in the appendix. □

The numerical procedure in the appendix for computing the factorization (8) needs only $O(n^3 + m^3)$ flops. Furthermore,

$$\begin{bmatrix} U_{11} & 0 & U_{12} \\ 0 & I & 0 \\ U_{21} & 0 & U_{22} \end{bmatrix} \begin{bmatrix} P & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & P \end{bmatrix} \text{ and } \begin{bmatrix} Q & 0 & 0 \\ 0 & W & 0 \\ 0 & 0 & Q \end{bmatrix} \begin{bmatrix} I_{n_1+n_2} & 0 & 0 & 0 & 0 \\ 0 & V_{11} & 0 & V_{12} & 0 \\ 0 & 0 & I_m & 0 & 0 \\ 0 & V_{21} & 0 & V_{22} & 0 \\ 0 & 0 & 0 & 0 & I \end{bmatrix}$$

are orthogonal, and thus the computation of the factorization (8) is numerically backward stable [12].

COROLLARY 5. *With respect to factorization* (8),

$$G(s) = \left[\begin{array}{ccc|cc} -sE_{11}+A_{11} & -sE_{12}+A_{12} & A_{13} & 0 & B_{12} \\ 0 & -sE_{22}+A_{22} & A_{23} & 0 & B_{22} \\ 0 & A_{32} & A_{33} & B_{31} & 0 \\ \hline C_1 & C_2 & C_3 & 0 & 0 \end{array}\right] W^T. \tag{11}$$

Proof. First, a direct calculation yields that

$$\begin{bmatrix} P & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & P \end{bmatrix} \begin{bmatrix} -sE+A & B & -sE+A \\ C & 0 & C \\ -sE+A & B & 0 \end{bmatrix} \begin{bmatrix} Q & 0 & 0 \\ 0 & W & 0 \\ 0 & 0 & Q \end{bmatrix}$$
$$= \begin{bmatrix} P(-sE+A)Q & PBW & P(-sE+A)Q \\ CQ & 0 & CQ \\ P(-sE+A)Q & PBW & 0 \end{bmatrix}. \tag{12}$$

Next, since U and V are orthogonal, we have from (8) and using (12) that

$$\begin{bmatrix} P(-sE+A)Q & PBW & P(-sE+A)Q \\ CQ & 0 & CQ \\ P(-sE+A)Q & PBW & 0 \end{bmatrix} \begin{bmatrix} I_{n_1} \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} U_{11}^T & 0 & U_{21}^T \\ 0 & I & 0 \\ U_{12}^T & 0 & U_{22}^T \end{bmatrix} \begin{bmatrix} -sE_{11}+A_{11} \\ 0 \\ 0 \\ C_1 \\ \star \end{bmatrix}.$$

Note that $n = n_1 + n_2 + n_3$, and thus,

$$\begin{aligned} &[0 \;\; I_{n_3}]\, P(-sE+A)Q \begin{bmatrix} I_{n_1} \\ 0 \end{bmatrix} \\ &\quad = [0_{n_3\times(n_1+n_2)} \;\; I_{n_3} \;\; 0] \begin{bmatrix} P(-sE+A)Q & PBW & P(-sE+A)Q \\ CQ & 0 & CQ \\ P(-sE+A)Q & PBW & 0 \end{bmatrix} \begin{bmatrix} I_{n_1} \\ 0 \end{bmatrix} \\ (13) \quad &\quad = [0_{n_3\times(n_1+n_2)} \;\; I_{n_3} \;\; 0] \begin{bmatrix} U_{11}^T & 0 & U_{21}^T \\ 0 & I & 0 \\ U_{12}^T & 0 & U_{22}^T \end{bmatrix} \begin{bmatrix} -sE_{11}+A_{11} \\ 0 \\ 0 \\ C_1 \\ \star \end{bmatrix} = 0. \end{aligned}$$

Then we have from factorization (8), using (12), and equality $n = n_1 + n_2 + n_3$ that

$$\begin{bmatrix} U_{11} & 0 & 0 & 0 & U_{12} \\ 0 & I_{n_3} & 0 & 0 & 0 \\ 0 & 0 & I_p & 0 & 0 \end{bmatrix} \begin{bmatrix} P(-sE+A)Q & PBW & P(-sE+A)Q \\ CQ & 0 & CQ \\ P(-sE+A)Q & PBW & 0 \end{bmatrix}$$

$$\times \begin{bmatrix} I_{n_1+n_2} & 0 & 0 \\ 0 & V_{11} & 0 \\ 0 & 0 & I_m \\ 0 & V_{21} & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} -sE_{11}+A_{11} & -sE_{12}+A_{12} & A_{13} & 0 & B_{12} \\ 0 & -sE_{22}+A_{22} & A_{23} & 0 & B_{22} \\ 0 & A_{32} & A_{33} & B_{31} & 0 \\ C_1 & C_2 & C_3 & 0 & 0 \end{bmatrix},$$

which give, along with the facts

$$\begin{bmatrix} [U_{11} \;\; 0]P(-sE+A)Q + [0 \;\; U_{12}]P(-sE+A)Q \\ [0 \;\; I_{n_3}]P(-sE+A)Q \end{bmatrix} = \begin{bmatrix} U_{11} & U_{12} \\ 0 & I \end{bmatrix} P(-sE+A)Q,$$

$$\begin{bmatrix} [U_{11} \;\; 0]PBW + [0 \;\; U_{12}]PBW \\ [0 \;\; I_{n_3}]PBW \end{bmatrix} = \begin{bmatrix} U_{11} & U_{12} \\ 0 & I \end{bmatrix} PBW,$$

$$CQ \begin{bmatrix} I_{n_1+n_2} & 0 \\ 0 & V_{11} \end{bmatrix} + CQ \begin{bmatrix} 0 & V_{21} \\ 0 & 0 \end{bmatrix} = CQ \begin{bmatrix} I & 0 & V_{21} \\ 0 & I & 0 \\ 0 & 0 & V_{11} \end{bmatrix} \quad (\text{since } V_{21} \in \mathbf{R}^{n_1\times n_3}),$$

that

$$\begin{bmatrix} U_{11} & U_{12} \\ 0 & I \end{bmatrix} P(-sE+A)Q \begin{bmatrix} I_{n_1+n_2} & 0 \\ 0 & V_{11} \end{bmatrix} + \begin{bmatrix} U_{11} & 0 \\ 0 & I \end{bmatrix} P(-sE+A)Q \begin{bmatrix} 0 & V_{21} \\ 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} -sE_{11}+A_{11} & -sE_{12}+A_{12} & A_{13} \\ 0 & -sE_{22}+A_{22} & A_{23} \\ 0 & A_{32} & A_{33} \end{bmatrix} \tag{14}$$

and

$$\begin{bmatrix} U_{11} & U_{12} \\ 0 & I \end{bmatrix} PBW = \begin{bmatrix} 0 & B_{12} \\ 0 & B_{22} \\ B_{31} & 0 \end{bmatrix}, \quad CQ \begin{bmatrix} I & 0 & V_{21} \\ 0 & I & 0 \\ 0 & 0 & V_{11} \end{bmatrix} = [C_1 \;\; C_2 \;\; C_3]. \tag{15}$$

Because (13) means that

$$[0 \;\; U_{12}]\, P(-sE+A)Q \begin{bmatrix} V_{21} \\ 0 \end{bmatrix} = U_{12}\, [0 \;\; I_{n_3}]\, P(-sE+A)Q \begin{bmatrix} I_{n_1} \\ 0 \end{bmatrix} V_{21} = 0,$$

thus,

$$\begin{bmatrix} U_{11} & 0 \\ 0 & I \end{bmatrix} P(-sE+A)Q \begin{bmatrix} 0 & V_{21} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} U_{11} & U_{12} \\ 0 & I \end{bmatrix} P(-sE+A)Q \begin{bmatrix} 0 & V_{21} \\ 0 & 0 \end{bmatrix},$$

and hence we obtain, using (14) and the equality

$$\begin{bmatrix} I_{n_1+n_2} & 0 \\ 0 & V_{11} \end{bmatrix} + \begin{bmatrix} 0 & V_{21} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} I & 0 & V_{21} \\ 0 & I & 0 \\ 0 & 0 & V_{11} \end{bmatrix} \text{ (since } V_{21} \in \mathbf{R}^{n_1 \times n_3}),$$

that

$$\begin{bmatrix} U_{11} & U_{12} \\ 0 & I \end{bmatrix} P(-sE+A)Q \begin{bmatrix} I & 0 & V_{21} \\ 0 & I & 0 \\ 0 & 0 & V_{11} \end{bmatrix} = \begin{bmatrix} -sE_{11}+A_{11} & -sE_{12}+A_{12} & A_{13} \\ 0 & -sE_{22}+A_{22} & A_{23} \\ 0 & A_{32} & A_{33} \end{bmatrix}. \tag{16}$$

Obviously, (15) and (16) can be combined into the following compact form:

$$\begin{cases} \begin{bmatrix} U_{11} & U_{12} \\ 0 & I \end{bmatrix} P(-sE+A)Q \begin{bmatrix} I & 0 & V_{21} \\ 0 & I & 0 \\ 0 & 0 & V_{11} \end{bmatrix} = \begin{bmatrix} -sE_{11}+A_{11} & -sE_{12}+A_{12} & A_{13} \\ 0 & -sE_{22}+A_{22} & A_{23} \\ 0 & A_{32} & A_{33} \end{bmatrix}, \\ \begin{bmatrix} U_{11} & U_{12} \\ 0 & I \end{bmatrix} PBW = \begin{bmatrix} 0 & B_{12} \\ 0 & B_{22} \\ B_{31} & 0 \end{bmatrix}, \quad CQ \begin{bmatrix} I & 0 & V_{21} \\ 0 & I & 0 \\ 0 & 0 & V_{11} \end{bmatrix} = [C_1 \;\; C_2 \;\; C_3]. \end{cases} \tag{17}$$

Hence, Corollary 5 follows. □

Realization (11) is computed by a numerically backward stable manner in the sense that the factorization (8) is numerically backward stable. In the following, we show that the computation of realization (11) is numerically forward stable.

Let us denote the computed X using finite precision arithmetic by $\bar{X}$, as opposed to exact arithmetic, and denote the machine precision by ϵ. Let

$$
\begin{aligned}
P(-sE+A)Q &= -s\mathcal{E}^{(1)}+\mathcal{A}^{(1)},\\
P(-sE+A)Q\begin{bmatrix} V_{22} & 0 & V_{21}\\ 0 & I & 0\\ V_{12} & 0 & V_{11}\end{bmatrix} &= -s\mathcal{E}^{(2)}+\mathcal{A}^{(2)},\\
\begin{bmatrix} U_{11} & U_{12}\\ U_{21} & U_{22}\end{bmatrix} P(-sE+A)Q &= -s\mathcal{E}^{(3)}+\mathcal{A}^{(3)},\\
\begin{bmatrix} U_{11} & U_{12}\\ U_{21} & U_{22}\end{bmatrix} P(-sE+A)Q\begin{bmatrix} V_{22} & 0 & V_{21}\\ 0 & I & 0\\ V_{12} & 0 & V_{11}\end{bmatrix} &= -s\mathcal{E}^{(4)}+\mathcal{A}^{(4)},\\
\begin{bmatrix} U_{11} & U_{12}\\ 0 & I\end{bmatrix} P(-sE+A)Q\begin{bmatrix} I & 0 & V_{21}\\ 0 & I & 0\\ 0 & 0 & V_{11}\end{bmatrix} &= -s\mathcal{E}+\mathcal{A}
\end{aligned}
$$

and

$$
\begin{aligned}
\bar{P}(-sE+A)\bar{Q} &= -s\bar{\mathcal{E}}^{(1)}+\bar{\mathcal{A}}^{(1)} =: -s(\mathcal{E}^{(1)}+\Delta\mathcal{E}^{(1)})+(\mathcal{A}^{(1)}+\Delta\mathcal{A}^{(1)}),\\
\bar{P}(-sE+A)\bar{Q}\begin{bmatrix} \bar{V}_{22} & 0 & \bar{V}_{21}\\ 0 & I & 0\\ \bar{V}_{12} & 0 & \bar{V}_{11}\end{bmatrix} &= -s\bar{\mathcal{E}}^{(2)}+\bar{\mathcal{A}}^{(2)} =: -s(\mathcal{E}^{(2)}+\Delta\mathcal{E}^{(2)})+(\mathcal{A}^{(2)}+\Delta\mathcal{A}^{(2)}),\\
\begin{bmatrix} \bar{U}_{11} & \bar{U}_{12}\\ \bar{U}_{21} & \bar{U}_{22}\end{bmatrix}\bar{P}(-sE+A)\bar{Q} &= -s\bar{\mathcal{E}}^{(3)}+\bar{\mathcal{A}}^{(3)} =: -s(\mathcal{E}^{(3)}+\Delta\mathcal{E}^{(3)})+(\mathcal{A}^{(3)}+\Delta\mathcal{A}^{(3)}),\\
\begin{bmatrix} \bar{U}_{11} & \bar{U}_{12}\\ \bar{U}_{21} & \bar{U}_{22}\end{bmatrix}\bar{P}(-sE+A)\bar{Q}\begin{bmatrix} \bar{V}_{22} & 0 & \bar{V}_{21}\\ 0 & I & 0\\ \bar{V}_{12} & 0 & \bar{V}_{11}\end{bmatrix} &= -s\bar{\mathcal{E}}^{(4)}+\bar{\mathcal{A}}^{(4)}\\
&=: -s(\mathcal{E}^{(4)}+\Delta\mathcal{E}^{(4)})+(\mathcal{A}^{(4)}+\Delta\mathcal{A}^{(4)}),\\
\begin{bmatrix} \bar{U}_{11} & \bar{U}_{12}\\ 0 & I\end{bmatrix}\bar{P}(-sE+A)\bar{Q}\begin{bmatrix} I & 0 & \bar{V}_{21}\\ 0 & I & 0\\ 0 & 0 & \bar{V}_{11}\end{bmatrix} &= -s\bar{\mathcal{E}}+\bar{\mathcal{A}}\\
&=: -s(\mathcal{E}+\Delta\mathcal{E})+(\mathcal{A}+\Delta\mathcal{A}).
\end{aligned}
$$

Then we have

$$
-s\mathcal{E}+\mathcal{A} = \begin{bmatrix} [I_{n_1+n_2}\ \ 0]\,(-s\mathcal{E}^{(3)}+\mathcal{A}^{(3)})\begin{bmatrix} I_{n_1+n_2}\\ 0\end{bmatrix} & [I_{n_1+n_2}\ \ 0]\,(-s\mathcal{E}^{(4)}+\mathcal{A}^{(4)})\begin{bmatrix} 0\\ I_{n_3}\end{bmatrix}\\ [0\ \ I_{n_3}]\,(-s\mathcal{E}^{(1)}+\mathcal{A}^{(1)})\begin{bmatrix} I_{n_1+n_2}\\ 0\end{bmatrix} & [0\ \ I_{n_3}]\,(-s\mathcal{E}^{(2)}+\mathcal{A}^{(2)})\begin{bmatrix} I_{n_3}\\ 0\end{bmatrix}\end{bmatrix}
$$

and

$$
-s\Delta\mathcal{E}+\Delta\mathcal{A} = \begin{bmatrix} [I_{n_1+n_2}\ \ 0]\,(-s\Delta\mathcal{E}^{(3)}+\Delta\mathcal{A}^{(3)})\begin{bmatrix} I_{n_1+n_2}\\ 0\end{bmatrix} & [I_{n_1+n_2}\ \ 0]\,(-s\Delta\mathcal{E}^{(4)}+\Delta\mathcal{A}^{(4)})\begin{bmatrix} 0\\ I_{n_3}\end{bmatrix}\\ [0\ \ I_{n_3}]\,(-s\Delta\mathcal{E}^{(1)}+\Delta\mathcal{A}^{(1)})\begin{bmatrix} I_{n_1+n_2}\\ 0\end{bmatrix} & [0\ \ I_{n_3}]\,(-s\Delta\mathcal{E}^{(2)}+\Delta\mathcal{A}^{(2)})\begin{bmatrix} I_{n_3}\\ 0\end{bmatrix}\end{bmatrix}.
$$

Since we have from [12] that

$$\|\Delta\mathcal{E}^{(i)}\|_2 \approx \epsilon\|E\|_2, \quad \|\Delta\mathcal{A}^{(i)}\|_2 \approx \epsilon\|A\|_2, \quad i = 1, 2, 3, 4,$$

we thus have

$$\|\Delta\mathcal{E}\|_2 \approx \epsilon\|E\|_2, \quad \|\Delta\mathcal{A}\|_2 \approx \epsilon\|A\|_2. \tag{18}$$

Similarly, if we denote

$$\begin{bmatrix} U_{11} & U_{12} \\ 0 & I \end{bmatrix} PBW = \mathcal{B}, \quad CQ \begin{bmatrix} I & 0 & V_{21} \\ 0 & I & 0 \\ 0 & 0 & V_{11} \end{bmatrix} = \mathcal{C}$$

and

$$\begin{bmatrix} \bar{U}_{11} & \bar{U}_{12} \\ 0 & I \end{bmatrix} \bar{P}B\bar{W} = \bar{\mathcal{B}} =: \mathcal{B} + \Delta\mathcal{B}, \quad C\bar{Q} \begin{bmatrix} I & 0 & \bar{V}_{21} \\ 0 & I & 0 \\ 0 & 0 & \bar{V}_{11} \end{bmatrix} = \bar{\mathcal{C}} =: \mathcal{C} + \Delta\mathcal{C},$$

then we also have

$$\|\Delta\mathcal{B}\|_2 \approx \epsilon\|B\|_2, \quad \|\Delta\mathcal{C}\|_2 \approx \epsilon\|C\|_2. \tag{19}$$

Therefore, we have from (17), (18), and (19) that the computation of realization (11) is numerically forward stable [12].

We can now conclude that the importance of the factorization (8) is that it provides a numerically reliable way, with complexity $O(n^3 + m^3)$, to compute realization (11) of $G(s)$.

The following considerations are necessary preliminaries for Theorem 6 below.

Assume that $G(s) \in \mathcal{R}^{p\times m}(s)$, with a minimal realization (7), is left invertible, or equivalently,

$$\max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} -sE + A & B \\ C & 0 \end{bmatrix} = n + m. \tag{20}$$

Then we have, using (17) and the nonsingularity of B_{31}, that

$$\max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} -sE_{11} + A_{11} & -sE_{12} + A_{12} & A_{13} & B_{12} \\ 0 & -sE_{22} + A_{22} & A_{23} & B_{22} \\ C_1 & C_2 & C_3 & 0 \end{bmatrix} = \max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} -sE + A & B \\ C & 0 \end{bmatrix} - n_3$$

$$= (n + m) - n_3 = n_1 + n_2 + n_3 + (m - n_3). \tag{21}$$

Therefore, the generalized lower triangular form [25, 42] of the pencil

$$\begin{bmatrix} -sE_{11} + A_{11} & -sE_{12} + A_{12} & A_{13} & B_{12} \\ 0 & -sE_{22} + A_{22} & A_{23} & B_{22} \\ C_1 & C_2 & C_3 & 0 \end{bmatrix}$$

is of the form

$$\mathcal{S}\begin{bmatrix} -sE_{11}+A_{11} & -sE_{12}+A_{12} & A_{13} & B_{12} \\ 0 & -sE_{22}+A_{22} & A_{23} & B_{22} \\ C_1 & C_2 & C_3 & 0 \end{bmatrix}\mathcal{T}$$

$$= \begin{array}{c} \begin{matrix} n_1+n_2-n_{0\infty} & n_{0\infty} & n_3 & m-n_3 \end{matrix} \\ \begin{bmatrix} -s\mathcal{E}_{nf}+\mathcal{A}_{nf} & 0 & 0 & 0 \\ -s\mathcal{E}_{10}+\mathcal{A}_{10} & -s\mathcal{E}_{11}+\mathcal{A}_{11} & \mathcal{A}_{12} & \mathcal{A}_{13} \\ -s\mathcal{E}_{20}+\mathcal{A}_{20} & \mathcal{A}_{21} & \mathcal{A}_{22} & \mathcal{A}_{23} \\ -s\mathcal{E}_{30}+\mathcal{A}_{30} & \mathcal{A}_{31} & \mathcal{A}_{32} & \mathcal{A}_{33} \end{bmatrix} \begin{matrix} \}n_{0\infty} \\ \}n_3 \\ \}m-n_3 \end{matrix} \end{array}, \tag{22}$$

where $\mathcal{S}$ and $\mathcal{T}$ are orthogonal, $\mathcal{E}_{nf}$ is of full column rank, $\mathcal{E}_{11}$ is nonsingular, and

$$\operatorname{rank}(-s\mathcal{E}_{nf}+\mathcal{A}_{nf}) = n_1+n_2-n_{0\infty} \quad \forall s \in \mathbf{C}_0,$$

$$\operatorname{rank}\begin{bmatrix} -s\mathcal{E}_{11}+\mathcal{A}_{11} & \mathcal{A}_{12} & \mathcal{A}_{13} \\ \mathcal{A}_{21} & \mathcal{A}_{22} & \mathcal{A}_{23} \\ \mathcal{A}_{31} & \mathcal{A}_{32} & \mathcal{A}_{33} \end{bmatrix} = n_{0\infty}+m \quad \forall s \in \mathbf{C}\backslash\mathbf{C}_0.$$

We partition $\mathcal{S}$ and $\mathcal{T}$ into

$$\mathcal{S} = \begin{array}{c} \begin{matrix} n_1 & n_2 & p \end{matrix} \\ \begin{bmatrix} \mathcal{S}_{11} & \mathcal{S}_{12} & \mathcal{S}_{13} \\ \mathcal{S}_{21} & \mathcal{S}_{22} & \mathcal{S}_{23} \\ \mathcal{S}_{31} & \mathcal{S}_{32} & \mathcal{S}_{33} \\ \mathcal{S}_{41} & \mathcal{S}_{42} & \mathcal{S}_{43} \end{bmatrix} \begin{matrix} \}n_{0\infty} \\ \}n_3 \\ \}m-n_3 \end{matrix} \end{array},$$

$$\mathcal{T} = \begin{array}{c} \begin{matrix} n_1+n_2-n_{0\infty} & n_{0\infty} & n_3 & m-n_3 \end{matrix} \\ \begin{bmatrix} \mathcal{T}_{11} & \mathcal{T}_{12} & \mathcal{T}_{13} & \mathcal{T}_{14} \\ \mathcal{T}_{21} & \mathcal{T}_{22} & \mathcal{T}_{23} & \mathcal{T}_{24} \\ \mathcal{T}_{31} & \mathcal{T}_{32} & \mathcal{T}_{33} & \mathcal{T}_{34} \\ \mathcal{T}_{41} & \mathcal{T}_{42} & \mathcal{T}_{43} & \mathcal{T}_{44} \end{bmatrix} \begin{matrix} \}n_1 \\ \}n_2 \\ \}n_3 \\ \}m-n_3 \end{matrix} \end{array}. \tag{23}$$

Obviously, factorization (22) has isolated the zeros of $G(s)$ on $\mathbf{C}_0$ and at infinity to

$$\begin{bmatrix} -s\mathcal{E}_{11}+\mathcal{A}_{11} & \mathcal{A}_{12} & \mathcal{A}_{13} \\ \mathcal{A}_{21} & \mathcal{A}_{22} & \mathcal{A}_{23} \\ \mathcal{A}_{31} & \mathcal{A}_{32} & \mathcal{A}_{33} \end{bmatrix}.$$

Let the columns of full column rank matrix $[L_1^T \;\; L_2^T \;\; L_3^T \;\; L_4^T \;\; L_5^T \;\; L_6^T]^T$ with $L_1, L_3 \in \mathbf{R}^{n_1\times r}$, $L_2, L_4 \in \mathbf{R}^{n_2\times r}$, $L_5 \in \mathbf{R}^{n_3\times r}$, and $L_6 \in \mathbf{R}^{(m-n_3)\times r}$ span the stable eigenspace of the pencil

$$\begin{bmatrix} -sE_{11}+A_{11} & -sE_{12}+A_{12} & 0 & 0 & A_{13} & B_{12} \\ 0 & -sE_{22}+A_{22} & 0 & 0 & A_{23} & B_{22} \\ -C_1^TJC_1 & -C_1^TJC_2 & -(sE_{11}+A_{11})^T & 0 & -C_1^TJC_3 & 0 \\ -C_2^TJC_1 & -C_2^TJC_2 & -(sE_{12}+A_{12})^T & -(sE_{22}+A_{22})^T & -C_2^TJC_3 & 0 \\ C_3^TJC_1 & C_3^TJC_2 & A_{13}^T & A_{23}^T & C_3^TJC_3 & 0 \\ 0 & 0 & B_{12}^T & B_{22}^T & 0 & 0 \end{bmatrix},$$

which gives

$$\begin{bmatrix} A_{11} & A_{12} & 0 & 0 & A_{13} & B_{12} \\ 0 & A_{22} & 0 & 0 & A_{23} & B_{22} \\ -C_1^T JC_1 & -C_1^T JC_2 & -A_{11}^T & 0 & -C_1^T JC_3 & 0 \\ -C_2^T JC_1 & -C_2^T JC_2 & -A_{12}^T & -A_{22}^T & -C_2^T JC_3 & 0 \\ C_3^T JC_1 & C_3^T JC_2 & A_{13}^T & A_{23}^T & C_3^T JC_3 & 0 \\ 0 & 0 & B_{12}^T & B_{22}^T & 0 & 0 \end{bmatrix} \begin{bmatrix} L_1 \\ L_2 \\ L_3 \\ L_4 \\ L_5 \\ L_6 \end{bmatrix}$$

$$= \begin{bmatrix} E_{11} & E_{12} & 0 & 0 & 0 & 0 \\ 0 & E_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & E_{11}^T & 0 & 0 & 0 \\ 0 & 0 & E_{12}^T & E_{22}^T & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} L_1 \\ L_2 \\ L_3 \\ L_4 \\ L_5 \\ L_6 \end{bmatrix} \Delta, \tag{24}$$

where $\Delta \in \mathbf{R}^{r\times r}$ is stable.

Now we are ready to present our main result.

THEOREM 6. *Given $G(s) \in \mathcal{R}^{p\times m}(s)$ with a minimal realization* (7)*, assume that factorization* (8) *and eigenfactorizations* (22) *and* (24) *have been determined. Then $G(s)$ has a (J, J')-lossless factorization if and only if the following conditions hold:*

(a) *$G(s)$ is left invertible; i.e., property* (20) *holds.*

(b) *There exists a nonsingular matrix $\mathcal{D}_0 \in \mathbf{R}^{m\times m}$ such that*

$$\mathcal{D}_0^T \begin{bmatrix} \mathcal{S}_{33} \\ \mathcal{S}_{43} \end{bmatrix} J \begin{bmatrix} \mathcal{S}_{33}^T & \mathcal{S}_{43}^T \end{bmatrix} \mathcal{D}_0 = J'. \tag{25}$$

(c) *We have*

$$r + n_{0\infty} = n_1 + n_2, \quad \begin{bmatrix} L_1 & \mathcal{T}_{12} \\ L_2 & \mathcal{T}_{22} \end{bmatrix} \text{ is nonsingular}, \tag{26}$$

$$\begin{bmatrix} E_{11}L_1 + E_{12}L_2 & E_{11}\mathcal{T}_{12} + E_{12}\mathcal{T}_{22} \\ E_{22}L_2 & E_{22}\mathcal{T}_{22} \end{bmatrix}^T \begin{bmatrix} L_3 & 0 \\ L_4 & 0 \end{bmatrix} \geq 0, \tag{27}$$

and the algebraic Riccati equation

$$E_{11}\mathcal{Y}_{11}A_{11}^T + A_{11}\mathcal{Y}_{11}E_{11}^T + E_{11}\mathcal{Y}_{11}C_1^T JC_1\mathcal{Y}_{11}E_{11}^T = 0 \tag{28}$$

has a solution $\mathcal{Y}_{11} \geq 0$ such that the pencil $-sE_{11} + A_{11} + E_{11}\mathcal{Y}_{11}C_1^T JC_1$ is stable.

(d) *We have*

$$\rho\left(\begin{bmatrix} L_1 & \mathcal{T}_{12} \\ L_2 & \mathcal{T}_{22} \end{bmatrix}, \begin{bmatrix} \mathcal{Y}_{11}E_{11}^T L_3 & 0 \\ 0 & 0 \end{bmatrix} \right) < 1. \tag{29}$$

Furthermore, in the case when conditions (a), (b), (c), *and* (d) *hold, if we define the following two QR factorizations:*

$$\hat{\mathcal{W}} \begin{bmatrix} S_{21}^T \\ S_{22}^T \end{bmatrix} = \begin{bmatrix} 0 \\ \mathcal{R}_{\hat{\mathcal{W}}} \end{bmatrix}, \quad \hat{\mathcal{W}}\hat{\mathcal{W}}^T = I, \quad \operatorname{rank}(\mathcal{R}_{\hat{\mathcal{W}}}) = n_{0\infty}, \tag{30}$$

$$\tilde{\mathcal{W}}\left(\begin{bmatrix} I_{n_1} & 0 & 0 \\ 0 & I_r & 0 \end{bmatrix}\begin{bmatrix} E_{11}\mathcal{Y}_{11}E_{11}^T L_3 \\ E_{11}L_1 + E_{12}L_2 \\ E_{22}L_2 \end{bmatrix}\right) = \begin{bmatrix} 0 \\ \mathcal{R}_{\tilde{\mathcal{W}}} \end{bmatrix},$$

$$\tilde{\mathcal{W}}\tilde{\mathcal{W}}^T = I, \quad \operatorname{rank}(\mathcal{R}_{\tilde{\mathcal{W}}}) = r, \tag{31}$$

and partition

$$\begin{bmatrix} \tilde{\mathcal{W}} & 0 \\ 0 & I \end{bmatrix}\begin{bmatrix} I & 0 \\ 0 & \hat{\mathcal{W}} \end{bmatrix}\begin{bmatrix} I & 0 & 0 \\ I & I & 0 \\ 0 & 0 & I_{n_2} \end{bmatrix} =: \begin{matrix} \\ \\ \end{matrix}\begin{matrix} n_1 & n_1+n_2 \\ \end{matrix}\begin{bmatrix} \mathcal{W}_{11} & \mathcal{W}_{12} \\ \mathcal{W}_{21} & \mathcal{W}_{22} \end{bmatrix}\begin{matrix} \}n_1 \\ \}n_1+n_2 \end{matrix}, \tag{32}$$

then a (J, J')-lossless factorization of $G(s)$ is given by the factors $\Theta(s)$ and $\Xi(s)$,

$$\Theta(s) = \left[\begin{array}{cc|c} -sE_\Theta + A_\Theta & 0 & \mathcal{Z}_1 \\ 0 & \mathcal{W}_{11}(-sE_{11} + A_{11}) & \mathcal{Z}_2 \\ \hline C_1L_1 + C_2L_2 + C_3L_5 & C_1 & -\begin{bmatrix} \mathcal{S}_{33} \\ \mathcal{S}_{43} \end{bmatrix}^T \end{array}\right]\mathcal{D}_0, \tag{33}$$

$$\Xi(s) = -(J')^{-1}\mathcal{D}_0^T\left[\begin{array}{c|c} sE_\Xi + A_\Xi & B_\Xi \\ \hline C_\Xi & 0 \end{array}\right]W^T, \tag{34}$$

where

$$-sE_\Theta + A_\Theta = [I_r \;\; 0]\,\hat{\mathcal{W}}\begin{bmatrix} E_{11} & E_{12} \\ 0 & E_{22} \end{bmatrix}\begin{bmatrix} L_1 - \mathcal{Y}_{11}E_{11}^T L_3 \\ L_2 \end{bmatrix}(-sI + \Delta),$$

$$\mathcal{Z}_1 = -[I_r \;\; 0]\,\hat{\mathcal{W}}\left(\begin{bmatrix} \mathcal{S}_{31} & \mathcal{S}_{32} \\ \mathcal{S}_{41} & \mathcal{S}_{42} \end{bmatrix}^T + \begin{bmatrix} E_{11}\mathcal{Y}_{11}C_1^T \\ 0 \end{bmatrix} J \begin{bmatrix} \mathcal{S}_{33} \\ \mathcal{S}_{43} \end{bmatrix}^T\right),$$

$$\mathcal{Z}_2 = -\mathcal{W}_{12}\begin{bmatrix} \mathcal{S}_{31} & \mathcal{S}_{32} \\ \mathcal{S}_{41} & \mathcal{S}_{42} \end{bmatrix}^T + \left(\mathcal{W}_{11} - \mathcal{W}_{12}\begin{bmatrix} I_{n_1} \\ 0 \end{bmatrix}\right)E_{11}\mathcal{Y}_{11}C_1^T J\begin{bmatrix} \mathcal{S}_{33} \\ \mathcal{S}_{43} \end{bmatrix}^T,$$

$$-sE_\Xi + A_\Xi = \left(\begin{bmatrix} -sE_{11} + A_{11} & -sE_{12} + A_{12} & A_{13} \\ 0 & -sE_{22} + A_{22} & A_{23} \\ 0 & A_{32} & A_{33} \end{bmatrix} + \begin{bmatrix} E_{11}\mathcal{Y}_{11}C_1^T J \\ 0 \\ 0 \end{bmatrix}[C_1 \;\; C_2 \;\; C_3]\right)\begin{bmatrix} L_1 - \mathcal{Y}_{11}E_{11}^T L_3 & \mathcal{T}_{12} & 0 \\ L_2 & \mathcal{T}_{22} & 0 \\ 0 & 0 & I \end{bmatrix},$$

$$B_\Xi = \begin{bmatrix} 0 & B_{12} \\ 0 & B_{21} \\ B_{31} & 0 \end{bmatrix},$$

$$C_\Xi = \begin{bmatrix} \mathcal{S}_{31} & \mathcal{S}_{32} & \mathcal{S}_{33}J \\ \mathcal{S}_{41} & \mathcal{S}_{42} & \mathcal{S}_{43}J \end{bmatrix}\begin{bmatrix} L_3 & 0 & 0 \\ L_4 & 0 & 0 \\ C_1L_1 + C_2L_2 & C_1\mathcal{T}_{12} + C_2\mathcal{T}_{22} & C_3 \end{bmatrix}.$$

Proof. The proof is given in section 3. ∎

Theorem 6 leads to the following algorithm for solving the (J, J')-lossless factorization problem.

ALGORITHM 1.

Input: $G(s) \in \mathcal{R}^{p\times m}(s)$ *with a minimal realization* (7).

Output: *A (J, J')-lossless factorization $G(s) = \Theta(s)\Xi(s)$ of $G(s)$ if possible.*

Step 1. *Compute the* $\max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} -sE+A & B \\ C & 0 \end{bmatrix}$ *using the generalized lower triangular form* [25, 42] *of the pencil* $\begin{bmatrix} -sE+A & B \\ C & 0 \end{bmatrix}$. *If it equals $n + m$, continue the process. Otherwise, stop.*

Step 2. *Compute factorization* (8) *and the eigenfactorizations* (22) *and* (24).

Step 3. *Solve the algebraic Riccati equation* (28).

Step 4. *Verify conditions* (25), (26), (27), *and* (29). *If these conditions hold, continue. Otherwise, stop.*

Step 5. *Compute QR factorizations* (30) *and* (31) *and then do the partitioning* (32).

Step 6. *Compute the factors $\Theta(s)$ and $\Xi(s)$ by* (33) *and* (34). *Output $\Theta(s)$ and $\Xi(s)$ and then stop.*

We comment on Algorithm 1 as follows:

- The basis of Algorithm 1 is factorization (8), whose computation is numerically backward stable.
- Steps 1, 2, 4, and 5 are all implemented by only orthogonal transformations, which are numerically backward stable [12].
- The algebraic Riccati equation (28) in Step 3 can be solved by MATLAB code *care.m*, which is known to be numerically reliable.
- J' is symmetric and its inverse in Step 6 can be computed by SVDs or QR factorizations [12], which are numerically reliable. Moreover, its computation has no effect on Steps 1–5. Here we emphasize that it is almost impossible to avoid the computation of $(J')^{-1}$ in the (J, J')-lossless factorization problem.

Therefore, Algorithm 1 can be implemented in a numerically reliable manner.

In the following we give an example, produced by MATLAB, to illustrate Algorithm 1.

Example 1. Consider $G(s) \in \mathcal{R}^{3\times 2}(s)$ of the form

$$G(s) =: \left[\begin{array}{c|c} -sE+A & B \\ \hline C & 0 \end{array}\right],$$

where

$$E = \begin{bmatrix} -0.79648662531102 & -0.34408628966218 & -0.92919020476514 & -0.86759340177264 \\ -1.59637666711651 & -0.94742772442409 & -1.88152958422816 & -1.64655996556517 \\ -1.15895570567892 & -0.61226050785142 & -1.39417360448597 & -1.19527660654597 \\ -0.73210108350959 & -0.49401721140691 & -0.89220026767141 & -0.71378640906087 \end{bmatrix},$$

$$A = \begin{bmatrix} -0.54614732946469 & -0.33335809799803 & -0.58359526918812 & -0.51254534118206 \\ -0.41339349292044 & -0.33309306657012 & -0.67775758611413 & -0.19305827569214 \\ -0.17411088650387 & -0.13421301361811 & -0.36658641418862 & 0.02806619555373 \\ 0.02762027911712 & -0.02883001314100 & -0.17294923202402 & 0.19154703875993 \end{bmatrix},$$

$$B = \begin{bmatrix} 0.09377433103503 & 0.57046350715079 \\ 0.21622307574552 & 0.25020142641665 \\ 0.21193192933125 & 0.26835035557552 \\ 0.09482924580770 & -0.22040923528986 \end{bmatrix},$$

$$C = \begin{bmatrix} -1.42431328897985 & -2.68997314918496 & -0.51966015469511 & -2.34686842451761 \\ 0.36818972326874 & -1.72243610987038 & 2.02945944128574 & -1.02577475830025 \\ -3.36933791418090 & -2.94407864412045 & -3.90940468898360 & -3.24308590192695 \end{bmatrix}.$$

$G(s)$ is not proper and not stable (it has poles at 0.99992821324864, 1.0000025, 0740078, and ∞), and $G(s)$ has zeros at 0 and ∞. Let

$$J = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}, \quad J' = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}.$$

We aim at solving the (J, J')-lossless factorization problem using Algorithm 1.

First, we have that

$$\max_{s \in \mathbf{C}} \begin{bmatrix} -sE + A & B \\ C & 0 \end{bmatrix} = 6 = n + m.$$

Then we compute factorization (8) and eigenfactorizations (22) and (24) to get

$$n_1 = 2, \quad n_2 = n_3 = 1, \quad n_{0\infty} = 2, \quad r = 1.$$

Next, we solve the algebraic Riccati equation (28) to get $\mathcal{Y}_{11}$. We have verified that conditions (25), (26), (27), and (29) hold. Hence, the (J, J')-lossless factorization problem is solvable. Finally, we obtain the factor $\Theta(s)$ and $\Xi(s)$ as follows:

$$\Theta(s) = \left[\begin{array}{c|c} -sE_1 + A_1 & B_1 \\ \hline C_1 & D_1 \end{array}\right], \quad \Xi(s) = \left[\begin{array}{c|c} -sE_2 + A_2 & B_2 \\ \hline C_2 & 0 \end{array}\right],$$

where

$$E_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.01774809157525 & -0.06358913345474 \\ 0 & -0.00797313544766 & -0.10278316529096 \end{bmatrix},$$

$$A_1 = \begin{bmatrix} -1.00024697137514 & 0 & 0 \\ 0 & 0.01774668569757 & -0.06358649788333 \\ 0 & -0.00797309297453 & -0.10278354338532 \end{bmatrix},$$

$$B_1 = \begin{bmatrix} 0.00057439346062 & 0.00329848142888 \\ 0.07316225222884 & 0.03930299095489 \\ 0.14590579936467 & 0.28566107869004 \end{bmatrix},$$

$$D_1 = \begin{bmatrix} -0.00043278110916 & 1.73218301798980 \\ -0.57776923753157 & 1.63292687247408 \\ 0.81620027901348 & 1.15499257009050 \end{bmatrix},$$

$$C_1 = \begin{bmatrix} 0.00001056932855 & -0.72504644099477 & 0.40237228536648 \\ 0.00000747363261 & -1.09338239540910 & 1.45479511894656 \\ 0.00000528467535 & 0.09618079240102 & -1.25264068782599 \end{bmatrix}$$

and

$$E_2 = \begin{bmatrix} 0.01774809157525 & -0.06358913345474 & 0.47848254444975 & 0 \\ -0.00797313544766 & -0.10278316529096 & 0.10775627253774 & 0 \\ 0 & 0 & -0.26584884230443 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$A_2 = \begin{bmatrix} -0.01774720011926 & 0.06358729062381 & -0.14556383589018 & 0.00308271113730 \\ 0.00797077340375 & 0.10278851884865 & -1.90110351010103 & 0.02210167323039 \\ 0 & 0 & -0.39065207241599 & 0.00592368556175 \\ 0 & 0 & 0.22044279058618 & 0 \end{bmatrix},$$

$$B_2 = \begin{bmatrix} 0 & 0.04378720103054 \\ 0 & -0.52343364403428 \\ 0 & -0.17552423949623 \\ -0.03645315939670 & 0 \end{bmatrix},$$

$$C_2 = \begin{bmatrix} 0.70991171521262 & -1.86276740788912 & 5.43175549553726 & -0.16061664495765 \\ -0.41841225671249 & 0.23180092585230 & 1.89204730143784 & 0.11350319804946 \end{bmatrix}.$$

3. Proof of Theorem 6. In this section we always assume that *factorization* (8) *and eigenfactorizations* (22) *and* (24) *have been determined.* Our purpose here is to prove Theorem 6. For this, we need two supporting lemmas.

LEMMA 7. (i) *Given factorization* (8)*, there exist matrices* $K \in \mathbf{R}^{n_3 \times n_2}$ *and*

$$\begin{matrix} n_2 & n_3 \\ \end{matrix}$$
$$\begin{bmatrix} F_{12} & F_{13} \\ F_{22} & 0 \end{bmatrix} \begin{matrix} \}n_3 \\ \}m - n_3 \end{matrix}$$

such that

$$\text{(35)} \qquad \begin{cases} A_{32} + B_{31}F_{12} = B_{31}K, \ A_{33} + B_{31}F_{13} = -B_{31}, \\ \textit{the pencil } -sE_{22} + A_{22} + A_{23}K + B_{22}F_{22} \textit{ is stable.} \end{cases}$$

(ii) *The following equality holds:*

$$G(s) = G_1(s)G_2(s),$$

where

$$\text{(36)} \begin{cases} G_1(s) = \left[\begin{array}{cc|cc} -sE_{11} + A_{11} & -sE_{12} + A_{12} + A_{13}K + B_{12}F_{22} & A_{13} & B_{12} \\ 0 & -sE_{22} + A_{22} + A_{23}K + B_{22}F_{22} & A_{23} & B_{22} \\ \hline C_1 & C_2 + C_3K & C_3 & 0 \end{array}\right] \\ \qquad\quad =: \left[\begin{array}{c|c} -s\Theta + \Phi & \Psi \\ \hline \Pi & \mathcal{D} \end{array}\right], \\ G_2(s) = \left[\begin{array}{cc|cc} -sE_{22} + A_{22} & A_{23} & 0 & B_{22} \\ A_{32} & A_{33} & B_{31} & 0 \\ \hline -F_{12} & -F_{13} & I & 0 \\ -F_{22} & 0 & 0 & I \end{array}\right] W^T, \end{cases}$$

and $G_2(s)$ *and* $G_2^{-1}(s)$ *have neither zeros nor poles in* $\mathbf{C}_+$.

(iii) $G(s)$ *has a* (J, J')*-lossless factorization if and only if* $G_1(s)$ *has a* (J, J')*-lossless factorization. Moreover,* $G(s) = \Theta(s)\Xi(s)$ *is a* (J, J')*-lossless factorization*

of $G(s)$ if and only if $G_1(s) = \Theta(s)\hat{\Xi}(s)$ with $\hat{\Xi}(s) = \Xi(s)G_2^{-1}(s)$ is a (J, J')-lossless factorization of $G_1(s)$.

Proof. The minimality of realization (7) implies that

$$\operatorname{rank}[-sE + A \quad B] = n \quad \forall s \in \mathbf{C},$$

which, along with (17) gives

$$\operatorname{rank}[-sE_{22} + A_{22} \quad A_{23} \quad B_{22}] = n_2 \quad \forall s \in \mathbf{C},$$

and thus there exist matrices K and F_{22} such that the pencil $-sE_{22} + A_{22} + A_{23}K + B_{22}F_{22}$ is stable [42]. Since B_{31} is nonsingular, the existence of matrices F_{12} and F_{13} is obvious. Parts (ii) and (iii) follow directly from a simple verification. □

LEMMA 8. *Let Θ, Φ, Ψ, and Π be the same as those defined in* (36). *Then $\mathcal{Y} \geq 0$ satisfies*

$$\mathcal{Y}(\Phi\Theta^{-1})^T + (\Phi\Theta^{-1})\mathcal{Y} + \mathcal{Y}(\Pi\Theta^{-1})^T J(\Pi\Theta^{-1})\mathcal{Y} = 0, \tag{37}$$

and the matrix $\Phi\Theta^{-1} + \mathcal{Y}(\Pi\Theta^{-1})^T J(\Pi\Theta^{-1})$ is stable if and only if

$$\mathcal{Y} = \begin{bmatrix} E_{11}\mathcal{Y}_{11}E_{11}^T & 0 \\ 0 & 0 \end{bmatrix},$$

where $\mathcal{Y}_{11} \geq 0$ satisfies (28) *and the pencil $-sE_{11} + A_{11} + E_{11}\mathcal{Y}_{11}C_1^T JC_1$ is stable.*

Proof. It is easy to see that $\Phi\Theta^{-1}$ and $(\Pi\Theta^{-1})^T J(\Pi\Theta^{-1})$ are of the forms

$$\Phi\Theta^{-1} = \begin{bmatrix} H_{11} & H_{12} \\ 0 & H_{22} \end{bmatrix}, \quad (\Pi\Theta^{-1})^T J(\Pi\Theta^{-1}) = \begin{bmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{12}^T & \Pi_{22} \end{bmatrix},$$

where

$$\begin{aligned}
H_{11} &= A_{11}E_{11}^{-1}, \quad H_{22} = (A_{22} + A_{23}K + B_{22}F_{22})E_{22}^{-1}, \\
H_{12} &= (A_{12} + A_{13}K + B_{12}F_{22})E_{22}^{-1} - A_{11}E_{11}^{-1}E_{12}E_{22}^{-1}, \\
\Pi_{11} &= (C_1E_{11}^{-1})^T J(C_1E_{11}^{-1}), \quad \Pi_{12} = (C_1E_{11}^{-1})^T\{(C_2 + C_3K)E_{22}^{-1} - C_1E_{11}^{-1}E_{12}E_{22}^{-1}\}, \\
\Pi_{22} &= \{(C_2 + C_3K)E_{22}^{-1} - C_1E_{11}^{-1}E_{12}E_{22}^{-1}\}^T J(C_1E_{11}^{-1})^T \\
&\quad \{(C_2 + C_3K)E_{22}^{-1} - C_1E_{11}^{-1}E_{12}E_{22}^{-1}\}.
\end{aligned}$$

Since

$$\begin{bmatrix} (\Phi\Theta^{-1})^T & (\Pi\Theta^{-1})^T J(\Pi\Theta^{-1}) \\ 0 & -(\Phi\Theta^{-1}) \end{bmatrix} = \begin{bmatrix} H_{11}^T & 0 & \Pi_{11} & \Pi_{12} \\ H_{12}^T & H_{22}^T & \Pi_{12}^T & \Pi_{22} \\ 0 & 0 & -H_{11} & -H_{12} \\ 0 & 0 & 0 & -H_{22} \end{bmatrix},$$

and the stability of the pencil $-sE_{22} + A_{22} + A_{23}K + B_{22}F_{22}$ implies that H_{22}^T is stable and $-H_{22}$ is antistable, thus the stable eigendecomposition of the matrix

$$\begin{bmatrix} (\Phi\Theta^{-1})^T & (\Pi\Theta^{-1})^T J(\Pi\Theta^{-1}) \\ 0 & -(\Phi\Theta^{-1}) \end{bmatrix}$$

is of the form

$$(38)\quad \begin{bmatrix} (\Phi\Theta^{-1})^T & (\Pi\Theta^{-1})^T J(\Pi\Theta^{-1}) \\ 0 & -(\Phi\Theta^{-1}) \end{bmatrix} \begin{bmatrix} \mathcal{X}_{11} & 0 \\ \mathcal{X}_{21} & I \\ \mathcal{X}_{31} & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \mathcal{X}_{11} & 0 \\ \mathcal{X}_{21} & I \\ \mathcal{X}_{31} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \Omega_{11} & 0 \\ \Omega_{21} & H_{22}^T \end{bmatrix},$$

where the numbers of rows of $\mathcal{X}_{11}, \mathcal{X}_{21}$, and $\mathcal{X}_{31}$ are n_1, n_2, and n_1, respectively, and Ω_{11} is stable. It is well known [18] that

$$\begin{aligned} &\mathcal{Y} \ge 0 \text{ is the solution of the Lyapunov equation (37) such that} \\ &\qquad \Phi\Theta^{-1} + \mathcal{Y}(\Pi\Theta^{-1})^T J(\Pi\Theta^{-1}) \text{ is stable} \\ \iff &\mathcal{Y} = \begin{bmatrix} \mathcal{X}_{31} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathcal{X}_{11} & 0 \\ \mathcal{X}_{21} & I \end{bmatrix}^{-1} = \begin{bmatrix} \hat{\mathcal{Y}}_{11} & 0 \\ 0 & 0 \end{bmatrix} \ge 0, \quad H_{11} + \hat{\mathcal{Y}}_{11}\Pi_{11} \text{ is stable,} \\ (39)\quad \iff &\hat{\mathcal{Y}}_{11} \ge 0, \quad H_{11} + \hat{\mathcal{Y}}_{11}\Pi_{11} \text{ is stable,} \end{aligned}$$

where

$$\hat{\mathcal{Y}}_{11} := \mathcal{X}_{31}\mathcal{X}_{11}^{-1}.$$

Note that

$$H_{11} + \hat{\mathcal{Y}}_{11}\Pi_{11} = A_{11}E_{11}^{-1} + \hat{\mathcal{Y}}_{11}(C_1E_{11}^{-1})^T J(C_1E_{11}^{-1})$$

and

$$\begin{bmatrix} (A_{11}E_{11}^{-1})^T & (C_1E_{11}^{-1})^T J(C_1E_{11}^{-1}) \\ 0 & -(A_{11}E_{11}^{-1}) \end{bmatrix} \begin{bmatrix} \mathcal{X}_{11} \\ \mathcal{X}_{31} \end{bmatrix} = \begin{bmatrix} H_{11}^T & \Pi_{11} \\ 0 & -H_{11} \end{bmatrix} \begin{bmatrix} \mathcal{X}_{11} \\ \mathcal{X}_{31} \end{bmatrix} = \begin{bmatrix} \mathcal{X}_{11} \\ \mathcal{X}_{31} \end{bmatrix} \Omega_{11},$$

and thus we have from [18] that $\hat{\mathcal{Y}}_{11} \ge 0$ and $H_{11} + \hat{\mathcal{Y}}_{11}\Pi_{11}$ is stable if and only if

$$(40)\quad \begin{cases} \hat{\mathcal{Y}}_{11}(A_{11}E_{11}^{-1})^T + (A_{11}E_{11}^{-1})\hat{\mathcal{Y}}_{11} + \hat{\mathcal{Y}}_{11}(C_1E_{11}^{-1})^T J(C_1E_{11}^{-1})\hat{\mathcal{Y}}_{11} = 0, \\ \hat{\mathcal{Y}}_{11} \ge 0, \quad A_{11}E_{11}^{-1} + \hat{\mathcal{Y}}_{11}(C_1E_{11}^{-1})^T J(C_1E_{11}^{-1}) \text{ is stable.} \end{cases}$$

A simple calculation yields that $\hat{\mathcal{Y}}_{11}$ satisfies (40) if and only if $\hat{\mathcal{Y}}_{11} = E_{11}\mathcal{Y}_{11}E_{11}^T$, $\mathcal{Y}_{11} \ge 0$, (28) is satisfied, and the pencil $-sE_{11} + A_{11} + E_{11}\mathcal{Y}_{11}C_1^T JC_1$ is stable. Hence, Lemma 8 follows directly from (39). ∎

We are now ready to prove Theorem 6.

Proof. We prove Theorem 6 by the following four arguments.

Argument 1. Since $\mathcal{S}$ and $\mathcal{T}$ are orthogonal, we have from (22) that

$$\begin{bmatrix} E_{11} & E_{12} & 0 & 0 \\ 0 & E_{22} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathcal{T}_{11} & \mathcal{T}_{12} & \mathcal{T}_{13} & \mathcal{T}_{14} \\ \mathcal{T}_{21} & \mathcal{T}_{22} & \mathcal{T}_{23} & \mathcal{T}_{24} \\ \mathcal{T}_{31} & \mathcal{T}_{32} & \mathcal{T}_{33} & \mathcal{T}_{34} \\ \mathcal{T}_{41} & \mathcal{T}_{42} & \mathcal{T}_{43} & \mathcal{T}_{44} \end{bmatrix}$$

$$= \begin{bmatrix} \mathcal{S}_{11}^T & \mathcal{S}_{21}^T & \mathcal{S}_{31}^T & \mathcal{S}_{41}^T \\ \mathcal{S}_{12}^T & \mathcal{S}_{22}^T & \mathcal{S}_{32}^T & \mathcal{S}_{42}^T \\ \mathcal{S}_{13}^T & \mathcal{S}_{23}^T & \mathcal{S}_{33}^T & \mathcal{S}_{43}^T \end{bmatrix} \begin{bmatrix} \mathcal{E}_{nf} & 0 & 0 & 0 \\ \mathcal{E}_{10} & \mathcal{E}_{11} & 0 & 0 \\ \mathcal{E}_{20} & 0 & 0 & 0 \\ \mathcal{E}_{30} & 0 & 0 & 0 \end{bmatrix},$$

and thus

$$\begin{bmatrix} E_{11} & E_{12} \\ 0 & E_{22} \end{bmatrix} \begin{bmatrix} \mathcal{T}_{13} & \mathcal{T}_{14} \\ \mathcal{T}_{23} & \mathcal{T}_{24} \end{bmatrix} = 0, \quad \mathcal{S}_{23}^T \mathcal{E}_{11} = 0.$$

Because $\left[\begin{smallmatrix} E_{11} & E_{12} \\ 0 & E_{22} \end{smallmatrix}\right]$ and $\mathcal{E}_{11}$ are nonsingular, thus

$$\begin{bmatrix} \mathcal{T}_{13} & \mathcal{T}_{14} \\ \mathcal{T}_{23} & \mathcal{T}_{24} \end{bmatrix} = 0, \quad \mathcal{S}_{23} = 0. \tag{41}$$

Note that $\left[\begin{smallmatrix} \mathcal{T}_{11} & \mathcal{T}_{12} \\ \mathcal{T}_{21} & \mathcal{T}_{22} \end{smallmatrix}\right]$ is square and

$$\begin{bmatrix} \mathcal{T}_{11} & \mathcal{T}_{12} & \mathcal{T}_{13} & \mathcal{T}_{14} \\ \mathcal{T}_{21} & \mathcal{T}_{22} & \mathcal{T}_{23} & \mathcal{T}_{24} \\ \mathcal{T}_{31} & \mathcal{T}_{32} & \mathcal{T}_{33} & \mathcal{T}_{34} \\ \mathcal{T}_{41} & \mathcal{T}_{42} & \mathcal{T}_{43} & \mathcal{T}_{44} \end{bmatrix}$$

is orthogonal. Thus (41) yields that

$$\begin{bmatrix} \mathcal{T}_{31} & \mathcal{T}_{32} \\ \mathcal{T}_{41} & \mathcal{T}_{42} \end{bmatrix} = 0. \tag{42}$$

Therefore, the orthogonal matrices $\mathcal{S}$ and $\mathcal{T}$ in (22) are of the forms

$$\mathcal{S} = \begin{array}{c} \begin{matrix} n_1 & n_2 & p \end{matrix} \\ \begin{bmatrix} \mathcal{S}_{11} & \mathcal{S}_{12} & \mathcal{S}_{13} \\ \mathcal{S}_{21} & \mathcal{S}_{22} & 0 \\ \mathcal{S}_{31} & \mathcal{S}_{32} & \mathcal{S}_{33} \\ \mathcal{S}_{41} & \mathcal{S}_{42} & \mathcal{S}_{43} \end{bmatrix} \begin{matrix} \}n_{0\infty} \\ \}n_3 \\ \}m-n_3 \end{matrix} \end{array},$$

$$\mathcal{T} = \begin{array}{c} \begin{matrix} n_1+n_2-n_{0\infty} & n_{0\infty} & n_3 & m-n_3 \end{matrix} \\ \begin{bmatrix} \mathcal{T}_{11} & \mathcal{T}_{12} & 0 & 0 \\ \mathcal{T}_{21} & \mathcal{T}_{22} & 0 & 0 \\ 0 & 0 & \mathcal{T}_{33} & \mathcal{T}_{34} \\ 0 & 0 & \mathcal{T}_{43} & \mathcal{T}_{44} \end{bmatrix} \begin{matrix} \}n_1 \\ \}n_2 \\ \}n_3 \\ \}m-n_3 \end{matrix} \end{array}. \tag{43}$$

Moreover, we have from (22), (23), and (24) that

$$\begin{bmatrix} -s\Theta+\Phi & \Psi \\ \Pi & \mathcal{D} \end{bmatrix} \begin{bmatrix} I_{n_1} & 0 & 0 & 0 \\ 0 & I_{n_2} & 0 & 0 \\ 0 & -K & I & 0 \\ 0 & -F_{22} & 0 & I \end{bmatrix} = \begin{bmatrix} -sE_{11}+A_{11} & -sE_{12}+A_{12} & A_{13} & B_{12} \\ 0 & -sE_{22}+A_{22} & A_{23} & B_{22} \\ C_1 & C_2 & C_3 & 0 \end{bmatrix},$$

$$M \begin{bmatrix} -s\Theta+\Phi & 0 & \Psi \\ -\Pi^T J\Pi & -s\Theta^T-\Phi^T & -\Pi^T J\mathcal{D} \\ \mathcal{D}^T J\Pi & \Psi^T & \mathcal{D}^T J\mathcal{D} \end{bmatrix} N$$

$$= \begin{bmatrix} -sE_{11}+A_{11} & -sE_{12}+A_{12} & 0 & 0 & A_{13} & B_{12} \\ 0 & -sE_{22}+A_{22} & 0 & 0 & A_{23} & B_{22} \\ -C_1^T JC_1 & -C_1^T JC_2 & -(sE_{11}+A_{11})^T & 0 & -C_1^T JC_3 & 0 \\ -C_2^T JC_1 & -C_2^T JC_2 & -(sE_{12}+A_{12})^T & -(sE_{22}+A_{22})^T & -C_2^T JC_3 & 0 \\ C_3^T JC_1 & C_3^T JC_2 & A_{13}^T & A_{23}^T & C_3^T JC_3 & 0 \\ 0 & 0 & B_{12}^T & B_{22}^T & 0 & 0 \end{bmatrix},$$

where

$$M = \begin{bmatrix} I_{n_1} & 0 & 0 & 0 & 0 & 0 \\ 0 & I_{n_2} & 0 & 0 & 0 & 0 \\ 0 & 0 & I_{n_1} & 0 & 0 & 0 \\ 0 & 0 & 0 & I_{n_2} & K^T & F_{22}^T \\ 0 & 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & 0 & I \end{bmatrix}, \; N = \begin{bmatrix} I_{n_1} & 0 & 0 & 0 & 0 & 0 \\ 0 & I_{n_2} & 0 & 0 & 0 & 0 \\ 0 & 0 & I_{n_1} & 0 & 0 & 0 \\ 0 & 0 & 0 & I_{n_2} & 0 & 0 \\ 0 & -K & 0 & 0 & I & 0 \\ 0 & -F_{22} & 0 & 0 & 0 & I \end{bmatrix}.$$

Hence,

$$\begin{bmatrix} \mathcal{S}_{11} & \mathcal{S}_{12} & \mathcal{S}_{13} \\ \mathcal{S}_{21} & \mathcal{S}_{22} & 0 \\ \mathcal{S}_{31} & \mathcal{S}_{32} & \mathcal{S}_{33} \\ \mathcal{S}_{41} & \mathcal{S}_{42} & \mathcal{S}_{43} \end{bmatrix} \begin{bmatrix} -sI + \Phi\Theta^{-1} & \Psi \\ \Pi\Theta^{-1} & \mathcal{D} \end{bmatrix}$$

$$\times \begin{bmatrix} E_{11}\mathcal{T}_{11} + E_{12}\mathcal{T}_{21} & E_{11}\mathcal{T}_{12} + E_{12}\mathcal{T}_{22} & 0 & 0 \\ E_{22}\mathcal{T}_{21} & E_{22}\mathcal{T}_{22} & 0 & 0 \\ -K\mathcal{T}_{21} & -K\mathcal{T}_{22} & \mathcal{T}_{33} & \mathcal{T}_{34} \\ -F_{22}\mathcal{T}_{21} & -F_{22}\mathcal{T}_{22} & \mathcal{T}_{43} & \mathcal{T}_{44} \end{bmatrix}$$

$$= \begin{bmatrix} -s\mathcal{E}_{nf} + \mathcal{A}_{nf} & 0 & 0 & 0 \\ -s\mathcal{E}_{10} + \mathcal{A}_{10} & -s\mathcal{E}_{11} + \mathcal{A}_{11} & \mathcal{A}_{12} & \mathcal{A}_{13} \\ -s\mathcal{E}_{20} + \mathcal{A}_{20} & \mathcal{A}_{21} & \mathcal{A}_{22} & \mathcal{A}_{23} \\ -s\mathcal{E}_{30} + \mathcal{A}_{30} & \mathcal{A}_{31} & \mathcal{A}_{32} & \mathcal{A}_{33} \end{bmatrix},$$

the columns of

$$\begin{bmatrix} E_{11}L_1 + E_{12}L_2 \\ E_{22}L_2 \\ L_3 \\ L_4 \\ L_5 - KL_2 \\ L_6 - F_{22}L_2 \end{bmatrix}$$

span the stable eigenspace of the pencil

$$\begin{bmatrix} -sI + \Phi\Theta^{-1} & 0 & \Psi \\ -(\Pi\Theta^{-1})^T J(\Pi\Theta^{-1}) & -sI - (\Phi\Theta^{-1})^T & -(\Pi\Theta^{-1})^T J\mathcal{D} \\ \mathcal{D}^T J(\Pi\Theta^{-1}) & \Psi^T & \mathcal{D}^T J\mathcal{D} \end{bmatrix},$$

and

$$\begin{bmatrix} \Phi\Theta^{-1} & 0 & \Psi \\ -(\Pi\Theta^{-1})^T J(\Pi\Theta^{-1}) & -(\Phi\Theta^{-1})^T & -(\Pi\Theta^{-1})^T J\mathcal{D} \\ \mathcal{D}^T J(\Pi\Theta^{-1}) & \Psi^T & \mathcal{D}^T J\mathcal{D} \end{bmatrix} \begin{bmatrix} E_{11}L_1 + E_{12}L_2 \\ E_{22}L_2 \\ L_3 \\ L_4 \\ L_5 - KL_2 \\ L_6 - F_{22}L_2 \end{bmatrix}$$

$$= \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} E_{11}L_1 + E_{12}L_2 \\ E_{22}L_2 \\ L_3 \\ L_4 \\ L_5 - KL_2 \\ L_6 - F_{22}L_2 \end{bmatrix} \Delta.$$

Argument 2. Note that $G_1(s) = \left[\begin{array}{c|c} -sI + \Phi\Theta^{-1} & \Psi \\ \hline \Pi\Theta^{-1} & \mathcal{D} \end{array}\right]$, so the minimality of realization (7) and the stability of the pencil $-sE_{22} + A_{22} + A_{23}K + B_{22}F_{22}$ give that

$$\operatorname{rank}\begin{bmatrix} -sI + \Phi\Theta^{-1} & \Psi \end{bmatrix} = \operatorname{rank}\begin{bmatrix} -sI + \Phi\Theta^{-1} \\ \Pi\Theta^{-1} \end{bmatrix} = n_1 + n_2 \quad \forall s \in \mathbf{C}_0 \cup \mathbf{C}_+.$$

Thus, we have from Lemma 7(iii) and Theorem 3 that $G(s)$ has a (J, J')-lossless factorization if and only if $G_1(s)$ has a (J, J')-lossless factorization, or equivalently,

- the conditions in (a), (b) hold, $r + n_{0\infty} = n_1 + n_2$;
- $\begin{bmatrix} E_{11}L_1 + E_{12}L_2 & E_{11}\mathcal{T}_{12} + E_{12}\mathcal{T}_{22} \\ E_{22}L_2 & E_{22}\mathcal{T}_{22} \end{bmatrix}$ is nonsingular;
- $\mathcal{X} := \begin{bmatrix} L_3 & 0 \\ L_4 & 0 \end{bmatrix}\begin{bmatrix} E_{11}L_1 + E_{12}L_2 & E_{11}\mathcal{T}_{12} + E_{12}\mathcal{T}_{22} \\ E_{22}L_2 & E_{22}\mathcal{T}_{22} \end{bmatrix}^{-1} \geq 0$;
- the Lyapunov equation (37) has a solution $\mathcal{Y} \geq 0$ such that the matrix $\Phi\Theta^{-1} + \mathcal{Y}(\Pi\Theta^{-1})^T J(\Pi\Theta^{-1})$ is stable, and furthermore, $\rho(\mathcal{X}\mathcal{Y}) < 1$.

Argument 3. Since

$$\begin{bmatrix} E_{11}L_1 + E_{12}L_2 & E_{11}\mathcal{T}_{12} + E_{12}\mathcal{T}_{22} \\ E_{22}L_2 & E_{22}\mathcal{T}_{22} \end{bmatrix} \text{ is nonsingular} \Leftrightarrow \begin{bmatrix} L_1 & \mathcal{T}_{12} \\ L_2 & \mathcal{T}_{22} \end{bmatrix} \text{ is nonsingular},$$

$$\mathcal{X} \geq 0 \Leftrightarrow \begin{bmatrix} E_{11}L_1 + E_{12}L_2 & E_{11}\mathcal{T}_{12} + E_{12}\mathcal{T}_{22} \\ E_{22}L_2 & E_{22}\mathcal{T}_{22} \end{bmatrix}^T \begin{bmatrix} L_3 & 0 \\ L_4 & 0 \end{bmatrix} \geq 0,$$

and Lemma 8 yields that

$$\rho(\mathcal{X}\mathcal{Y}) < 1 \;\Leftrightarrow\; \rho\left(\begin{bmatrix} L_1 & \mathcal{T}_{12} \\ L_2 & \mathcal{T}_{22} \end{bmatrix}, \begin{bmatrix} \mathcal{Y}_{11}E_{11}^T L_3 & 0 \\ 0 & 0 \end{bmatrix}\right) < 1,$$

thus $G(s)$ has a (J, J')-lossless factorization if and only if conditions (a), (b), (c), and (d) hold.

Argument 4. Under the conditions in (a)–(d), according to Lemma 7(iii), Lemma 8, and Theorem 3, factors $\Theta(s)$ and $\Xi(s)$ are given by

$$\Theta(s) = \left[\begin{array}{cc|c} -sI + \Delta & 0 & Z_1 \\ 0 & -sI + \Phi\Theta^{-1} & Z_2 \\ \hline \Pi\Theta^{-1}\begin{bmatrix} E_{11}L_1 + E_{12}L_2 \\ E_{22}L_2 \end{bmatrix} + \mathcal{D}\begin{bmatrix} L_5 - KL_2 \\ L_6 - F_{22}L_2 \end{bmatrix} & \Pi\Theta^{-1} & -\begin{bmatrix} \mathcal{S}_{33} \\ \mathcal{S}_{43} \end{bmatrix}^T \end{array}\right] \mathcal{D}_0$$

$$= \left[\begin{array}{cc|c} -sI + \Delta & 0 & Z_1 \\ 0 & -s\Theta + \Phi & Z_2 \\ \hline \Pi\Theta^{-1}\begin{bmatrix} E_{11}L_1 + E_{12}L_2 \\ E_{22}L_2 \end{bmatrix} + \mathcal{D}\begin{bmatrix} L_5 - KL_2 \\ L_6 - F_{22}L_2 \end{bmatrix} & \Pi & -\begin{bmatrix} \mathcal{S}_{33} \\ \mathcal{S}_{43} \end{bmatrix}^T \end{array}\right] \mathcal{D}_0,$$

$$
\begin{aligned}
&\Xi(s)\\
&= -(J')^{-1}\mathcal{D}_0^T \left[\begin{array}{c|c} -sI + \Phi\Theta^{-1} + \mathcal{Y}(\Pi\Theta^{-1})^T J(\Pi\Theta^{-1}) & \Psi + \mathcal{Y}(\Pi\Theta^{-1})^T J\mathcal{D} \\ \hline \left(\begin{bmatrix} \mathcal{S}_{31} & \mathcal{S}_{32} \\ \mathcal{S}_{41} & \mathcal{S}_{42} \end{bmatrix}\mathcal{X} + \begin{bmatrix} \mathcal{S}_{33} \\ \mathcal{S}_{43} \end{bmatrix} J\Pi\Theta^{-1}\right)(I - \mathcal{Y}\mathcal{X})^{-1} & \begin{bmatrix} \mathcal{S}_{33} \\ \mathcal{S}_{43} \end{bmatrix} J\mathcal{D} \end{array}\right] G_2(s)\\
&= -(J')^{-1}\mathcal{D}_0^T \left[\begin{array}{c|c} -s\Theta + \Phi + \mathcal{Y}(\Pi\Theta^{-1})^T J\Pi & \Psi + \mathcal{Y}(\Pi\Theta^{-1})^T J\mathcal{D} \\ \hline \left(\begin{bmatrix} \mathcal{S}_{31} & \mathcal{S}_{32} \\ \mathcal{S}_{41} & \mathcal{S}_{42} \end{bmatrix}\mathcal{X} + \begin{bmatrix} \mathcal{S}_{33} \\ \mathcal{S}_{43} \end{bmatrix} J\Pi\Theta^{-1}\right)(I - \mathcal{Y}\mathcal{X})^{-1}\Theta & \begin{bmatrix} \mathcal{S}_{33} \\ \mathcal{S}_{43} \end{bmatrix} J\mathcal{D} \end{array}\right] G_2(s),
\end{aligned}
$$

where

$$
\Pi\Theta^{-1}\begin{bmatrix} E_{11}L_1 + E_{12}L_2 \\ E_{22}L_2 \end{bmatrix} + \mathcal{D}\begin{bmatrix} L_5 - KL_2 \\ L_6 - F_{22}L_2 \end{bmatrix} = C_1L_1 + C_2L_2 + C_3L_5,
$$

$$
\begin{aligned}
Z_1 &= -\begin{bmatrix} I_r & 0 \end{bmatrix}\left(\begin{bmatrix} E_{11}L_1 + E_{12}L_2 & E_{11}\mathcal{T}_{12} + E_{12}\mathcal{T}_{22} \\ E_{22}L_2 & E_{22}\mathcal{T}_{22} \end{bmatrix} - \mathcal{Y}\begin{bmatrix} L_3 & 0 \\ L_4 & 0 \end{bmatrix}\right)^{-1}\\
&\quad\times\left(\begin{bmatrix} \mathcal{S}_{31} & \mathcal{S}_{32} \\ \mathcal{S}_{41} & \mathcal{S}_{42} \end{bmatrix}^T + \mathcal{Y}(\Pi\Theta^{-1})^T J\begin{bmatrix} \mathcal{S}_{33} \\ \mathcal{S}_{43} \end{bmatrix}^T\right)\\
&= -\begin{bmatrix} I_r & 0 \end{bmatrix}\left(\begin{bmatrix} E_{11} & E_{12} \\ 0 & E_{22} \end{bmatrix}\begin{bmatrix} L_1 - \mathcal{Y}_{11}E_{11}^T L_3 & \mathcal{T}_{12} \\ L_2 & \mathcal{T}_{22} \end{bmatrix}\right)^{-1}\\
&\quad\times\left(\begin{bmatrix} \mathcal{S}_{31} & \mathcal{S}_{32} \\ \mathcal{S}_{41} & \mathcal{S}_{42} \end{bmatrix}^T + \begin{bmatrix} E_{11}\mathcal{Y}_{11}C_1^T \\ 0 \end{bmatrix} J\begin{bmatrix} \mathcal{S}_{33} \\ \mathcal{S}_{43} \end{bmatrix}^T\right),
\end{aligned}
$$

$$
\begin{aligned}
Z_2 &= (I - \mathcal{Y}\mathcal{X})^{-1}\mathcal{Y}\left(\mathcal{X}\begin{bmatrix} \mathcal{S}_{31} & \mathcal{S}_{32} \\ \mathcal{S}_{41} & \mathcal{S}_{42} \end{bmatrix}^T + (\Pi\Theta^{-1})^T J\begin{bmatrix} \mathcal{S}_{33} \\ \mathcal{S}_{43} \end{bmatrix}^T\right)\\
&= \begin{bmatrix} I \\ 0 \end{bmatrix}\left\{[E_{11}\mathcal{Y}_{11}E_{11}^T L_3 \;\; 0]\begin{bmatrix} E_{11}L_1 + E_{12}L_2 - E_{11}\mathcal{Y}_{11}E_{11}^T L_3 & E_{11}\mathcal{T}_{12} + E_{12}\mathcal{T}_{22} \\ E_{22}L_2 & E_{22}\mathcal{T}_{22} \end{bmatrix}^{-1}\right.\\
&\quad\left.\times\left(\begin{bmatrix} \mathcal{S}_{31} & \mathcal{S}_{32} \\ \mathcal{S}_{41} & \mathcal{S}_{42} \end{bmatrix}^T + \begin{bmatrix} E_{11}\mathcal{Y}_{11}C_1^T \\ 0 \end{bmatrix} J\begin{bmatrix} \mathcal{S}_{33} \\ \mathcal{S}_{43} \end{bmatrix}^T\right) + E_{11}\mathcal{Y}_{11}C_1^T J\begin{bmatrix} \mathcal{S}_{33} \\ \mathcal{S}_{43} \end{bmatrix}^T\right\},
\end{aligned}
$$

$$
\begin{aligned}
&\left(\begin{bmatrix} \mathcal{S}_{31} & \mathcal{S}_{32} \\ \mathcal{S}_{41} & \mathcal{S}_{42} \end{bmatrix}\mathcal{X} + \begin{bmatrix} \mathcal{S}_{33} \\ \mathcal{S}_{43} \end{bmatrix} J\Pi\Theta^{-1}\right)(I - \mathcal{Y}\mathcal{X})^{-1}\Theta\\
&\quad= \begin{bmatrix} \mathcal{S}_{31} & \mathcal{S}_{32} & \mathcal{S}_{33}J \\ \mathcal{S}_{41} & \mathcal{S}_{42} & \mathcal{S}_{43}J \end{bmatrix}\left(\begin{bmatrix} L_3 & 0 \\ L_4 & 0 \\ C_1L_1 + C_2L_2 & C_1\mathcal{T}_{12} + C_2\mathcal{T}_{22} \end{bmatrix}\right.\\
&\qquad\qquad\left.\times\begin{bmatrix} L_1 - \mathcal{Y}_{11}E_{11}^T L_3 & \mathcal{T}_{12} \\ L_2 & \mathcal{T}_{22} \end{bmatrix}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & C_3K \end{bmatrix}\right)\\
&\quad= \begin{bmatrix} \Pi_1 & \Pi_2 + \Pi_3K \end{bmatrix}.
\end{aligned}
$$

In the above, $\begin{bmatrix} \Pi_1 & \Pi_2 & \Pi_3 \end{bmatrix}$ is defined by

$$\begin{bmatrix} \Pi_1 & \Pi_2 & \Pi_3 \end{bmatrix} = C_\Xi \begin{bmatrix} L_1 - \mathcal{Y}_{11} E_{11}^T L_3 & \mathcal{T}_{12} & 0 \\ L_2 & \mathcal{T}_{22} & 0 \\ 0 & 0 & I \end{bmatrix}^{-1}.$$

Note that (22) with (43) together gives that

$$\begin{bmatrix} E_{11} & E_{12} \\ 0 & E_{22} \end{bmatrix} \begin{bmatrix} \mathcal{T}_{12} \\ \mathcal{T}_{22} \end{bmatrix} = \begin{bmatrix} \mathcal{S}_{21}^T \\ \mathcal{S}_{22}^T \end{bmatrix} \mathcal{E}_{11}, \quad \hat{\mathcal{W}} \left(\begin{bmatrix} E_{11} & E_{12} \\ 0 & E_{22} \end{bmatrix} \begin{bmatrix} \mathcal{T}_{12} \\ \mathcal{T}_{22} \end{bmatrix} \right) = \begin{bmatrix} 0 \\ \mathcal{R}_{\hat{\mathcal{W}}} \mathcal{E}_{11} \end{bmatrix},$$

which implies that

$$\left[\begin{array}{c|c} -sI + \Delta & Z_1 \\ \hline C_1 L_1 + C_2 L_2 + C_3 L_3 & 0 \end{array} \right] = \left[\begin{array}{c|c} -sE_\Theta + A_\Theta & \mathcal{Z}_1 \\ \hline C_1 L_1 + C_2 L_2 + C_3 L_3 & 0 \end{array} \right].$$

Furthermore, we have from (30), (31), and (32) that

$$\begin{bmatrix} \mathcal{W}_{11} & \mathcal{W}_{12} \\ \mathcal{W}_{21} & \mathcal{W}_{22} \end{bmatrix} \begin{bmatrix} E_{11} \mathcal{Y}_{11} E_{11}^T L_3 & 0 \\ E_{11} L_1 + E_{12} L_2 - E_{11} \mathcal{Y}_{11} E_{11}^T L_3 & E_{11} \mathcal{T}_{12} + E_{12} \mathcal{T}_{22} \\ E_{22} L_2 & E_{22} \mathcal{T}_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ \Sigma \end{bmatrix}$$

with

$$\Sigma \in \mathbf{R}^{(n_1+n_2)\times(n_1+n_2)}, \quad \operatorname{rank}(\Sigma) = n_1 + n_2,$$

which yields [3, 11, 43] that

$$\begin{bmatrix} E_{11} \mathcal{Y}_{11} E_{11}^T L_3 & 0 \end{bmatrix} \begin{bmatrix} E_{11} L_1 + E_{12} L_2 - E_{11} \mathcal{Y}_{11} E_{11}^T L_3 & E_{11} \mathcal{T}_{12} + E_{12} \mathcal{T}_{22} \\ E_{22} L_2 & E_{22} \mathcal{T}_{22} \end{bmatrix}^{-1} = -\mathcal{W}_{11}^{-1} \mathcal{W}_{12},$$

and consequently, $Z_2 = \begin{bmatrix} \mathcal{W}_{11}^{-1} \mathcal{Z}_2 \\ 0 \end{bmatrix}$. Hence, we have that

$$\begin{aligned}
\Theta(s) &= \left[\begin{array}{cc|c} -sE_\Theta + A_\Theta & 0 & \mathcal{Z}_1 \\ 0 & -s\Theta + \Phi & \begin{bmatrix} \mathcal{W}_{11}^{-1} \mathcal{Z}_2 \\ 0 \end{bmatrix} \\ \hline C_1 L_1 + C_2 L_2 + C_3 L_5 & \Pi & -\begin{bmatrix} \mathcal{S}_{33} \\ \mathcal{S}_{43} \end{bmatrix}^T \end{array} \right] \mathcal{D}_0 \\
&= \left[\begin{array}{ccc|c} -sE_\Theta + A_\Theta & 0 & 0 & \mathcal{Z}_1 \\ 0 & -sE_{11} + A_{11} & -sE_{12} + A_{12} + A_{13}K + B_{12}F_{22} & \mathcal{W}_{11}^{-1} \mathcal{Z}_2 \\ 0 & 0 & -sE_{22} + A_{22} + A_{23}K + B_{22}F_{22} & 0 \\ \hline C_1 L_1 + C_2 L_2 + C_3 L_5 & C_1 & C_2 + C_3 K & -\begin{bmatrix} \mathcal{S}_{33} \\ \mathcal{S}_{43} \end{bmatrix}^T \end{array} \right] \mathcal{D}_0 \\
&= \left[\begin{array}{cc|c} -sE_\Theta + A_\Theta & 0 & \mathcal{Z}_1 \\ 0 & \mathcal{W}_{11}(-sE_{11} + A_{11}) & \mathcal{Z}_2 \\ \hline C_1 L_1 + C_2 L_2 + C_3 L_5 & C_1 & -\begin{bmatrix} \mathcal{S}_{33} \\ \mathcal{S}_{43} \end{bmatrix}^T \end{array} \right] \mathcal{D}_0,
\end{aligned}$$

and

$$\Xi(s) = -(J')^{-1}\mathcal{D}_0^T$$

$$\times \left[\begin{array}{ccc|cc} -sE_{11}+A_{11}+E_{11}\mathcal{Y}_{11}C_1^TJC_1 & -sE_{12}+A_{12}+A_{13}K+B_{12}F_{22}+E_{11}\mathcal{Y}_{11}C_1^TJ(C_2+C_3K) & A_{13}+E_{11}\mathcal{Y}_{11}C_1^TJC_3 & B_{12} \\ 0 & -sE_{22}+A_{22}+A_{23}K+B_{22}F_{22} & A_{23} & B_{22} \\ \hline \Pi_1 & \Pi_2+\Pi_3\mathcal{K} & \Pi_3 & 0 \end{array}\right] G_2(s)$$

$$= -(J')^{-1}\mathcal{D}_0^T$$

$$\times \left[\begin{array}{ccc|cc} -sE_{11}+A_{11}+E_{11}\mathcal{Y}_{11}C_1^TJC_1 & -sE_{12}+A_{12}+B_{12}F_{22}+E_{11}\mathcal{Y}_{11}C_1^TJC_2 & A_{13}+E_{11}\mathcal{Y}_{11}C_1^TJC_3 & 0 & B_{12} \\ 0 & -sE_{22}+A_{22}+B_{22}F_{22} & A_{23} & 0 & B_{22} \\ 0 & A_{32}+B_{31}F_{12} & A_{33}+B_{31}F_{13} & B_{31} & 0 \\ \hline \Pi_1 & \Pi_2 & \Pi_3 & 0 & 0 \end{array}\right] G_2(s)$$

(since $A_{32}+B_{31}F_{12}=B_{31}K,\ A_{33}+B_{31}F_{13}=-B_{31}$)

$$= -(J')^{-1}\mathcal{D}_0^T \left[\begin{array}{ccc|cc} -sE_{11}+A_{11}+E_{11}\mathcal{Y}_{11}C_1^TJC_1 & -sE_{12}+A_{12}+E_{11}\mathcal{Y}_{11}C_1^TJC_2 & A_{13}+E_{11}\mathcal{Y}_{11}C_1^TJC_3 & 0 & B_{12} \\ 0 & -sE_{22}+A_{22} & A_{23} & 0 & B_{22} \\ 0 & A_{32} & A_{33} & B_{31} & 0 \\ \hline \Pi_1 & \Pi_2 & \Pi_3 & 0 & 0 \end{array}\right] W^T$$

$$= -(J')^{-1}\mathcal{D}_0^T \left[\begin{array}{c|c} -sE_\Xi+A_\Xi & B_\Xi \\ \hline C_\Xi & 0 \end{array}\right] W^T. \qquad \square$$

4. Conclusions. We have obtained necessary and sufficient solvability conditions and developed a numerical algorithm based on a generalized eigenvalue approach for the (J, J')-lossless factorization of any general rational matrix $G(s) \in \mathcal{R}^{p\times m}(s)$. Our algorithm consists of factorization (8), eigenfactorizations (22) and (24), and the algebraic Riccati equation (28). Thus, the (J, J')-lossless factorization can be computed in a numerically reliable manner. A numerical example has also been given to illustrate the proposed algorithm.

Appendix. We construct factorization (8) by the following numerical procedure.

Step 1. Compute the generalized upper triangular form [25, 42] of the pencil $-sE + A$ to get orthogonal matrices P and Q such that

$$P(-sE+A)Q =: \begin{matrix} \\ \}n_1 \\ \}n_2 \\ \}n_3 \end{matrix} \begin{matrix} \begin{matrix} n_1 & n_2 & n_3 \end{matrix} \\ \begin{bmatrix} -sE_{11}^{(1)} + A_{11}^{(1)} & -sE_{12}^{(1)} + A_{12}^{(1)} & -sE_{13}^{(1)} + A_{13}^{(1)} \\ 0 & -sE_{22}^{(1)} + A_{22}^{(1)} & A_{23}^{(1)} \\ 0 & A_{32} & A_{33}^{(1)} \end{bmatrix} \end{matrix}, \tag{44}$$

where $E_{11}^{(1)}$ and $E_{22}^{(1)}$ are nonsingular and

$$\operatorname{rank} \begin{bmatrix} -sE_{22}^{(1)} + A_{22}^{(1)} & A_{23}^{(1)} \\ A_{32} & A_{33}^{(1)} \end{bmatrix} = n_2 + n_3 \quad \forall s \in \mathbf{C}.$$

Define

$$PB =: \begin{matrix} \begin{bmatrix} B_1^{(1)} \\ B_2^{(1)} \\ B_3^{(1)} \end{bmatrix} \end{matrix} \begin{matrix} \}n_1 \\ \}n_2 \\ \}n_3 \end{matrix}, \quad CQ =: \begin{matrix} \begin{matrix} n_1 & n_2 & n_3 \end{matrix} \\ \begin{bmatrix} C_1 & C_2 & C_3^{(1)} \end{bmatrix} \end{matrix}.$$

The minimality of realization (7) gives that

$$\operatorname{rank}\,(B_3^{(1)}) = n_3.$$

Step 2. Compute the QR factorization of $(B_3^{(1)})^T$ to get orthogonal matrix W such that

$$B_3^{(1)} W =: \begin{matrix} \begin{matrix} n_3 & m-n_3 \end{matrix} \\ \begin{bmatrix} B_{31} & 0 \end{bmatrix} \end{matrix}, \quad \operatorname{rank}(B_{31}) = n_3. \tag{45}$$

Define

$$\begin{bmatrix} B_1^{(1)} \\ B_2^{(1)} \end{bmatrix} W =: \begin{matrix} \begin{matrix} n_3 & m-n_3 \end{matrix} \\ \begin{bmatrix} B_{11}^{(2)} & B_{12}^{(2)} \\ B_{21}^{(2)} & B_{22}^{(2)} \end{bmatrix} \end{matrix} \begin{matrix} \}n_1 \\ \}n_2 \end{matrix}.$$

Step 3. Compute QR factorizations to get orthogonal matrices $\tilde{U}$, $\hat{U}$, and V with partitioning

$$\tilde{U} = \begin{matrix} \begin{matrix} n_1 & n_3 \end{matrix} \\ \begin{bmatrix} \tilde{U}_{11} & \tilde{U}_{12} \\ \tilde{U}_{21} & \tilde{U}_{22} \end{bmatrix} \end{matrix} \begin{matrix} \}n_1 \\ \}n_3 \end{matrix}, \quad \hat{U} = \begin{matrix} \begin{matrix} n_2 & n_3 \end{matrix} \\ \begin{bmatrix} \hat{U}_{11} & \hat{U}_{12} \\ \hat{U}_{21} & \hat{U}_{22} \end{bmatrix} \end{matrix} \begin{matrix} \}n_2 \\ \}n_3 \end{matrix}, \quad V = \begin{matrix} \begin{matrix} n_3 & n_1 \end{matrix} \\ \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \end{matrix} \begin{matrix} \}n_3 \\ \}n_1 \end{matrix}$$

such that

$$(46)\quad \hat{U}\begin{bmatrix} B_{21}^{(2)} \\ B_{31} \end{bmatrix} =: \begin{bmatrix} 0 \\ \hat{R}_B \end{bmatrix}, \quad \tilde{U}\begin{bmatrix} B_{11}^{(2)} \\ \hat{R}_B \end{bmatrix} =: \begin{bmatrix} 0 \\ \tilde{R}_B \end{bmatrix}, \quad \begin{bmatrix} E_{13}^{(1)} & E_{11}^{(1)} \end{bmatrix} V =: \begin{bmatrix} 0 & R_E \end{bmatrix},$$

where $\tilde{R}_B$, $\hat{R}_B$, and R_E are nonsingular. Since $E_{11}^{(1)}$ and B_{31} are nonsingular, we have that $\tilde{U}_{11}$, $\hat{U}_{11}$, and V_{11} are nonsingular [1, 3, 11, 43].

Step 4. Define

$$\begin{matrix} n_1+n_2 & n_3 \end{matrix}$$
$$\begin{matrix} \\ \\ \end{matrix}\begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix}\begin{matrix} \}n_1+n_2 \\ \}n_3 \end{matrix} := U = \begin{bmatrix} \tilde{U}_{11} & 0 & \tilde{U}_{12} \\ 0 & I & 0 \\ \tilde{U}_{21} & 0 & \tilde{U}_{22} \end{bmatrix}\begin{bmatrix} I & 0 \\ 0 & \hat{U} \end{bmatrix}$$
$$= \begin{bmatrix} \tilde{U}_{11} & \tilde{U}_{12}\hat{U}_{21} & \tilde{U}_{12}\hat{U}_{22} \\ 0 & \hat{U}_{11} & \hat{U}_{12} \\ \tilde{U}_{21} & \tilde{U}_{22}\hat{U}_{21} & \tilde{U}_{22}\hat{U}_{22} \end{bmatrix}.$$

Then orthogonal matrices P, Q, U, V, and W above give factorization (8) with

$$\begin{bmatrix} -sE_{11}+A_{11} & -sE_{12}+A_{12} & -sE_{13}^{(4)}+A_{13}^{(4)} & B_{12} \\ 0 & -sE_{22}+A_{22} & A_{23}^{(4)} & B_{22} \end{bmatrix}$$
$$= \begin{bmatrix} U_{11} & U_{12} \end{bmatrix}\begin{bmatrix} -sE_{11}^{(1)}+A_{11}^{(1)} & -sE_{12}^{(1)}+A_{12}^{(1)} & -sE_{13}+A_{13}^{(1)} & B_{12}^{(2)} \\ 0 & -sE_{22}^{(1)}+A_{22}^{(1)} & A_{23}^{(1)} & B_{22}^{(2)} \\ 0 & A_{32} & A_{33}^{(1)} & 0 \end{bmatrix}$$

and

$$\begin{bmatrix} A_{13} \\ A_{23} \\ A_{33} \\ C_3 \end{bmatrix} = \begin{bmatrix} A_{13}^{(4)} & A_{11} \\ A_{23}^{(4)} & 0 \\ A_{33}^{(1)} & 0 \\ C_3^{(1)} & C_1 \end{bmatrix}\begin{bmatrix} V_{11} \\ V_{21} \end{bmatrix}.$$

A direct calculation yields that properties (9), (10), and (11) hold. □

The above procedures involves only the generalized upper triangular form (44) and four QR factorizations in (45) and (46). Hence, it needs only $O(n^3+m^3)$ flops [5].

REFERENCES

[1] D. Chu, L. De Lathauwer, and B. De Moor, *A QR-type reduction for computing the SVD of a general matrix product/quotient.* Numer. Math., 95 (2003), pp. 101–121.

[2] P. Suchomski, *J-lossless and extended J-lossless factorizations approach for δ-domain H_∞-control,* Internat. J. Control, 76 (2003), pp. 794–809.

[3] P. Benner and R. Byers, *Evaluating products of matrix pencils and collapsing matrix products,* Numer. Linear Algebra Appl., 8 (2001), pp. 357–380.

[4] W.-W. Lin, C.-S. Wang, and Q.-F. Xu, *Numerical computation of the minimal H_∞ norm of the discrete-time output feedback control problem,* SIAM J. Numer. Anal., 38 (2000), pp. 515–547.

[5] A. Varga and P. Van Dooren, *Basic Software Tools for Standard and Generalized State-Space Systems and Transfer Matrix Factorizations,* SLICOT Working Note SLWN1999-17, 1999. Available online at http://www.dlr.de/rm/PortalData/3/Resources//papers/m_t/varga_slwn1999-17.pdf.

[6] P. H. Lee, H. Kimura, and Y. C. Soh, *(J, J')-lossless conjugations, (J, J')-lossless factorization and chain-scattering approach to time-varying H^∞-control–one and two-block cases,* Internat. J. Control, 71 (1998), pp. 195–218.

[7] Y. S. Hung and D. Chu, *On extended (J, J')–lossless factorization,* Linear Algebra Appl. 271 (1998), pp. 117–138.

[8] Y. S. Hung and D. Chu, *(J, J')–lossless factorization for discrete-time systems,* Internat. J. Control, 71 (1998), pp. 517–533.

[9] W. Kongprawechnon and H. Kimura, *J-lossless factorization and H_∞-control for discrete-time systems,* Internat. J. Control, 70 (1998), pp. 423–446.

[10] H. Kimura, *Chain-Scattering Approach to H_∞-Control,* Birkhäuser Boston, Inc., Boston, MA, 1997.

[11] Z. Bai, J. Demmel, and M. Gu, *An inverse free parallel spectral divide and conquer algorithm for nonsymmetric eigenproblems,* Numer. Math., 76 (1997), pp. 279–308.

[12] G. H. Golub and C. F. Van Loan, *Matrix Computations,* 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.

[13] M. Sato and M. Suzuki, *An L-lossless factorization approach to the positive real control problem,* J. Franklin Inst. B, 333 (1996), pp. 225–243.

[14] M. Green and D. J. N. Limebeer, *Robust Linear Control,* 5th ed., Prentice-Hall, Englewood Cliffs, NJ, 1994.

[15] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control,* Prentice-Hall, Upper Saddle River, NJ, 1995.

[16] P. Gahinet, A. Nemirovski, A. Laub, and M. Chilali, *The LMI Control Toolbox,* The MathWorks, Inc., Natick, MA, 1995.

[17] B. R. Copeland and M. G. Safonov, *A zero compensation approach to singular H_2 and H_∞ problems,* Internat. J. Robust Nonlinear Control, 5 (1995), pp. 71–106.

[18] P. Lancaster and L. Rodman, *Algebraic Riccati Equations,* Oxford University Press, Oxford, UK, 1995.

[19] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory,* SIAM, Philadelphia, 1994.

[20] Yu. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming,* SIAM, Philadelphia, 1994.

[21] P. Gahinet and P. Apkarian, *A linear matrix inequality approach to H^∞-control,* Internat. J. Robust and Nonlinear Control, 4 (1994), pp. 421–448.

[22] T. Iwasaki and R. E. Skelton, *All controllers for the general H^∞-control problem: LMI existence conditions and state space formulas,* Automatica, 30 (1994), pp. 1307–1317.

[23] X. Xin and H. Kimura, *Singular (J, J')-lossless factorization for strictly proper functions,* Internat. J. Control, 59 (1994), pp. 1383–1400.

[24] X. Xin and H. Kimura, *(J, J')-lossless factorization for descriptor systems,* Linear Algebra Appl., 205/206 (1994), pp. 1289–1318.

[25] J. W. Demmel and B. Kågström, *The generalized Schur decomposition of an arbitrary pencil $A - \lambda B$: Robust software with error bounds and applications. Part* I: *Theory and algorithms,* ACM Trans. Math. Software, 19 (1993), pp. 160–174.

[26] G. Miminis, *Deflation in eigenvalue assignment of descriptor systems using state feedback,* IEEE Trans. Automat. Control, AC-38 (1993), pp. 1322–1336.

[27] B. R. Copeland and M. G. Safonov, *Zero cancelling compensators for singular control problems and their application to the inner-outer factorization problem,* Internat. J. Robust Nonlinear Control, 2 (1992), pp. 139–164.

[28] B. R. Copeland and M. G. Safonov, *A generalized eigenproblem solution for singular H_2 and H_∞ problems,* in Robust Control System Techniques and Applications, Part I, Control Dynam. Systems Adv. Theory Appl. 50, Academic Press, San Diego, 1992, pp. 331–394.

[29] H. Kimura, *(J, J')-lossless factorization based on conjugation,* Systems Control Lett., 19 (1992), pp. 95–109.

[30] M. Green, *H_∞-controller synthesis by J-lossless coprime factorization,* SIAM J. Control Optim., 30 (1992), pp. 522–547.

[31] J. A. Ball, J. W. Helton, and M. Verma, *A factorization principle for stabilization of linear control systems,* Internat. J. Robust Nonlinear Control, 1 (1991), pp. 229–294.

[32] M. C. Tsai and I. Postlethwait, *On J-lossless coprime factorization approach to H^∞ control,* Internat. J. Robust Nonlinear Control, 1 (1991), pp. 47–68.

[33] M. Green, K. Glover, D. Limebeer, and J. Doyle, *A J-spectral factorization approach to H_∞ control,* SIAM J. Control Optim., 28 (1990), pp. 1350–1371.

[34] C. Scherer, *H_∞-control by state-feedback and fast algorithms for the computation of optimal H_∞-norm,* IEEE Trans. Automat. Control, AC-35 (1990), pp. 1090–1099.

[35] J. C. Doyle, K. Glover, P. P. Khargonekar, and B. A. Francis, *State space solutions to standard H_2 and H_∞ control problems,* IEEE Trans. Automat. Control, AC-34 (1989), pp. 831–847.

[36] H. Kimura, *Conjugation, interpolation and model-matching in H_∞,* Internat. J. Control, 49 (1989), pp. 269–307.

[37] J. W. Demmel and B. Kågström, *Accurate solutions of ill-posed problems in control theory,* SIAM J. Matrix Anal. Appl., 9 (1988), pp. 126–145.

[38] J. A. Ball and A. C. M. Ran, *Optimal Hankel norm model reductions and Wiener–Hopf factorization* I: *The canonical case,* SIAM J. Control Optim., 25 (1987), pp. 362–382.

[39] J. A. Ball and N. Cohen, *The sensitivity minimization in an H_∞ norm: Parameterization of all optimal solutions,* Internat. J. Control, 46 (1987), pp. 785–816.

[40] B. A. Francis, *A Course in H_∞ Control,* Springer, New York, 1987.

[41] J. A. Ball and J. W. Helton, *A Beurling-Lax theorem for the Lie group $U(m, n)$ which contains most classical interpolation theory,* J. Oper. Theory, 9 (1983), pp. 107–142.

[42] P. Van Dooren, *The generalized eigenstructure problem in linear system theory,* IEEE Trans. Automat. Control, AC-26 (1981), pp. 111–129.

[43] P. Benner and R. Byers, *An Arithmetic for Matrix Pencils: Theory and New Algorithms,* Tech. Report, Institute für Mathematik, TU Berlin, Germany, 2004.

[44] P. Benner, R. Byers, V. Mehrmann, and H. Xu, *Robust Numerical Methods for Robust Control,* Tech. Report, Institute für Mathematik, TU Berlin, Germany, 2004.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 1, pp. 268–282

ANALYSIS OF OPTIMAL CONTROL PROBLEMS FOR THE TWO-DIMENSIONAL THERMISTOR SYSTEM*

HYUNG-CHUN LEE† AND TIMOFEY SHILKIN‡

Abstract. An optimal control problem for the thermistor system is considered. First, the precise mathematical problem is established and the proof of existence of the optimal solution is given with appropriate function spaces. Then, Gâteaux differentiability is shown for the thermistor system, with respect to control, and the optimality system is obtained.

Key words. optimal control, thermistor system, Gâteaux derivatives

AMS subject classifications. 49J20, 49K20

DOI. 10.1137/S0363012903434765

1. Introduction. Let Ω be a bounded domain in $\mathbb{R}^2$ with a smooth boundary $\partial\Omega$, and let $Q_T = \Omega \times (0,T)$. In this paper, we consider an optimal control problem for the system

$$\partial_t u - \operatorname{div}(\mu(\theta)\nabla u) = f \quad \text{in } Q_T, \tag{1.1}$$

$$\partial_t \theta - \Delta\theta = \mu(\theta)|\nabla u|^2 \quad \text{in } Q_T. \tag{1.2}$$

Here $u : Q_T \to \mathbb{R}$, $\theta : Q_T \to \mathbb{R}$ are the unknowns and $f : Q_T \to \mathbb{R}$, $\mu : \mathbb{R} \to \mathbb{R}$ are given. The function f will be the control of our optimal control problem. We assume that μ is a function of class C^1 satisfying the following assumption: there are positive constants μ_0, μ_1, and μ_2 such that

$$0 < \mu_0 \le \mu(s) \le \mu_1, \quad |\mu'(s)| \le \mu_2 \quad \forall s \in \mathbb{R}. \tag{1.3}$$

In this paper we study the following initial-boundary value problem for the system (1.1)–(1.2):

$$u|_{\partial\Omega} = 0, \quad u|_{t=0} = u_0, \tag{1.4}$$

$$\theta|_{\partial\Omega} = 0, \quad \theta|_{t=0} = \theta_0. \tag{1.5}$$

The main difficulty of the problem under consideration arises from the right-hand side of (1.2), which has the quadratic growth with respect to the gradient of unknown function u. Formally speaking, this means that our system belongs to *the class of systems having the strong nonlinearity.* Indeed, assuming that μ is smooth, we can introduce a new unknown vector function $U = (u, \theta)$ and get for it the system of type

$$\partial_t U - a(U)\Delta U = b(U, \nabla U), \tag{1.6}$$

*Received by the editors September 16, 2003; accepted for publication (in revised form) August 26, 2004; published electronically July 18, 2005. This work was supported by KRF-2002-041-C00033.
http://www.siam.org/journals/sicon/44-1/43476.html
†Department of Mathematics, Ajou University, Suwon 442-749, South Korea (hclee@ajou.ac.kr).
‡V. A. Steklov Institute of Mathematics, St. Petersburg Department, Fontanka 27, 191011 St. Petersburg, Russia (shilkin@pdmi.ras.ru).

where function b has quadratic growth with respect to ∇U:

$$|b(z,F)| \le C(1+|F|^2). \tag{1.7}$$

For the case of the scalar elliptic and parabolic equations having strong nonlinearities, the existence and regularity theory (based on the Leray–Schauder theory and a priori estimates in the smooth classes of functions) was developed by Ladyzhenskaya and Uraltseva [17] and Ladyzhenskaya, Solonnikov, and Uraltseva [18]. In contrast with the scalar case, there is no general existence and regularity theory for systems of types (1.6) and (1.7) and, moreover, examples show that without additional assumptions (such as special structure, etc.) such theories cannot be established; see, for instance, [10].

Existence of the weak solutions to the system (1.1)–(1.2) was proved in [21], where the mathematical treatment of this system appears, apparently for the first time. Unfortunately, the class of weak solutions considered in [21] does not provide Gâteaux differentiability of the nonlinear operator corresponding to the system (1.1)–(1.2). To study the optimal control problem for this system, we have to work with the class of strong solutions. The regularity theorem requires additional assumptions on f, which can be provided only by the appropriate choice of the cost functional of our problem. On the other hand, a carelessly chosen cost functional leads to the more complex form of the optimality system, which is unpleasant from the point of view of applications. So, to study an optimal control problem for the system (1.1)–(1.2), it is necessary to find the balance between the assumptions of f, which provide the regularity of our optimal solution and the form of the cost functional which, at least in principle, allows us to produce some reasonable numeric calculations. In our work, we suggest a form of the cost functional that satisfies these two requirements.

Regularity of weak solutions to the system (1.1)–(1.2) for the stationary case was studied, for example, in [5]. The complete regularity theory in the "elliptic-parabolic" case, i.e., the case when, in (1.1), the term containing the time derivative of u is absent, was developed in [2]. For the parabolic case, regularity of weak solutions to the system (1.1)–(1.2) was studied by Rodrigues in [21]. He studied the only special case, where the function μ is close to a constant (in the sense that $\mu_1 - \mu_0 \ll 1$). In the present work, we get rid of this restriction and prove the regularity of weak solutions of (1.1)–(1.2), from which Gâteaux differentiability of the nonlinear operator (1.1)–(1.2) on the functional class of strong solutions follows. This allows us to study the optimal control problem and derive the optimality system for our optimal control problem. We believe that our regularity theorem for the parabolic system (1.1)–(1.2) is one of the results of our work which is of independent interest.

The system (1.1)–(1.2) arises in many applications. For example, nonlinearities in (1.1) and (1.2) are typical for the description of *thermistors*, i.e., in studying the heat transfer in the resistor device whose resistance $\mu(\theta)$ depends on the temperature θ, and the volume heating is given by the Joule–Lentz law; see, for instance, [2, 5, 6, 7, 8, 13, 21] and references therein.

In our work, we explore the motivation of the system (1.1)–(1.2), which is borrowed from the paper [21]. Namely, in [21] it was shown that the system (1.1)–(1.2) describes eddy currents induced by a unidirectional external magnetic field of the form

$$H(x_1,x_2,x_3,t) = u(x_1,x_2,t)\vec{e}_3, \tag{1.8}$$

where $\vec{e}_3$ is a unit vector in the x_3-direction; see [21] for more details. Our problem is to find a control f such that the magnetic field of the form (1.8) and the temperature

field θ close to the given fields exist. We remark that in [21] the right-hand side f of (1.1) arises after reduction of the nonhomogeneous boundary conditions for u to the homogeneous conditions. So, in the physical case, f is the only time-dependent function, and hence the optimality system (2.14)–(2.16) obtained in Theorem 2.4 consists of a system of PDEs for the Lagrange multipliers and a second order ODE for f; see Remark 2.4 below. But for the sake of generality we also consider f depending on both time and spatial variables.

Note also that (1.1)–(1.2) can be considered also as a system modelling a unidirectional Poiseuille-type flow of a homogeneous incompressible Newtonian fluid whose viscosity is the temperature-dependent function; see [21] for more information. As we see, in most applications the system (1.1)–(1.2) arises in the two-dimensional case ($\Omega \subset \mathbb{R}^2$).

2. Notation and main results. We use the following notation for functional spaces:

- $L^p(\Omega)$, $L^p(Q_T)$ are the usual Lebesgue spaces with the notation
$$\|f\|_{p,\Omega} \equiv \|f\|_{L^p(\Omega)}, \qquad \|f\|_{p,Q_T} \equiv \|f\|_{L^p(Q_T)};$$
- $W_p^k(\Omega)$ is the usual Sobolev space (or Slobodetskii–Sobolev space in the case of noninteger k); see, for instance, [18, Chap. II, sections 2–3] for the definitions;
- $\mathring{W}_p^k(\Omega)$, $k > 1 - \frac{1}{p}$, is the subspace of functions from $W_p^k(\Omega)$ having zero traces on the boundary;
- $W_p^{-k}(\Omega) = (\mathring{W}_{p'}^k(\Omega))^*$, $\frac{1}{p} + \frac{1}{p'} = 1$, $p \geq 1$;
- $L^{r,s}(Q_T) \equiv L^s(0,T;L^r(\Omega))$, $\quad \|f\|_{L^{r,s}(Q_T)} \equiv (\int_0^T \|f(\cdot,t)\|_{r,\Omega}^s \, dt)^{1/s}$;
- $W_p^{1,0}(Q_T) \equiv L^p(0,T;W_p^1(\Omega)) = \{u \in L^p(Q_T) : \ \nabla u \in L^p(Q_T)\}$, $\|u\|_{W_p^{1,0}(Q_T)} \equiv \|u\|_{p,Q_T} + \|\nabla u\|_{p,Q_T}$;
- $W_p^{2,1}(Q_T) = \{u \in W_p^{1,0}(Q_T) : \ \nabla^2 u, \partial_t u \in L^p(Q_T)\}$, $\|u\|_{W_p^{2,1}(Q_T)} \equiv \|u\|_{W_p^{1,0}(Q_T)} + \|\nabla^2 u\|_{p,Q_T} + \|\partial_t u\|_{p,Q_T}$.

Let the functions $U, \Theta : Q_T \to \mathbb{R}$ and initial data $u_0, \theta_0 : \Omega \to \mathbb{R}$ be given and assume that

$$U, \Theta \in L^2(Q_T). \tag{2.1}$$

Let us introduce the cost functional

$$J(u,\theta,f) := \frac{1}{2}\|u - U\|_{2,Q_T}^2 + \frac{1}{2}\|\theta - \Theta\|_{2,Q_T}^2 + \frac{\beta_1}{2}\|f\|_{2q_0,Q_T}^{2q_0} + \frac{\beta_2}{2}\|\partial_t f\|_{2,Q_T}^2 \tag{2.2}$$

and suppose

$$q_0 > 1, \quad \beta_1 > 0, \quad \beta_2 \geq 0 \tag{2.3}$$

(without loss of generality, we can consider $q_0 \leq \frac{3}{2}$). Dealing with the system (1.1)–(1.2) we shall focus on the strong solutions (i.e., those for which both (1.1) and (1.2) hold a.e. in Q_T). Hence, we must assume some regularity of the initial data as well as compatibility conditions between initial and boundary data. For the sake of simplicity we consider the initial data satisfying the following restrictions:

$$u_0, \theta_0 \in C^2(\bar{\Omega}), \quad u_0|_{\partial\Omega} = \theta_0|_{\partial\Omega} = 0. \tag{2.4}$$

Denote $V := \{u \in W_2^{1,0}(Q_T) : \partial_t u \in L^2(0,T;W_2^{-1}(\Omega))\}$. Our optimal control problem is then the following.

Problem P. Find $(\hat{u},\hat{\theta},\hat{f}) \in V \times W_1^{2,1}(Q_T) \times L^2(Q_T)$, which minimize $J(u,\theta,f)$ among all the functions (u,θ,f) satisfying (1.1) in the sense of distributions, (1.2) a.e. in Q_T, and satisfying also (1.4) and (1.5) in the sense of traces.

Remark 2.1. In the case of $\beta_2 > 0$ we require also $\partial_t f \in L^2(Q_T)$.

Our principal results are the following theorems.

THEOREM 2.1 (existence of the optimal solution). *Assume that the conditions* (1.3), (2.1), (2.3), *and* (2.4) *hold. Then there is a number* $q > 1$ *depending only on* μ_0, μ_1, q_0, *and* Q_T *such that there is at least one optimal solution* $(\hat{u},\hat{\theta},\hat{f})$ *of Problem* P *belonging to the spaces*

$$\hat{u} \in C([0,T];L^2(\Omega)) \cap W_{2q}^{1,0}(Q_T), \quad \partial_t \hat{u} \in L^2(0,T;W_2^{-1}(\Omega)),$$
$$\hat{\theta} \in W_q^{2,1}(Q_T), \quad \hat{f} \in L^{2q_0}(Q_T),$$

and satisfying (1.1) *in the sense of distributions,* (1.2) *a.e. in* Q_T, *and* (1.4)–(1.5) *in the sense of traces. Moreover, if*

$$\beta_2 > 0, \tag{2.5}$$

we also have

$$\partial_t \hat{f} \in L^2(Q_T). \tag{2.6}$$

Remark 2.2. Theorem 2.1 remains true in the multidimensional case $\Omega \subset \mathbb{R}^n$, $n \geq 2$.

THEOREM 2.2 (regularity of the optimal solution). *Assume that the conditions* (1.3) *and* (2.3) *hold and let* $(\hat{u},\hat{\theta},\hat{f})$ *be an optimal solution obtained in Theorem* 2.1. *Then there is* $\alpha > 0$ *such that*

$$\hat{u} \in C^{\alpha,\alpha/2}(\bar{Q}_T), \tag{2.7}$$

the following inclusion holds:

$$\hat{u},\hat{\theta} \in W_4^{1,0}(Q_T), \tag{2.8}$$

and, moreover,

$$\hat{u},\hat{\theta} \in W_2^{2,1}(Q_T). \tag{2.9}$$

Finally, if we assume the condition (2.5) *holds, then*

$$\partial_t \hat{u}, \partial_t \hat{\theta} \in L^\infty(0,T;L^2(\Omega)) \cap W_2^{1,0}(Q_T), \tag{2.10}$$

$$\hat{\theta} \in C^{1/4}(\bar{Q}_T), \tag{2.11}$$

and

$$\hat{\theta} \in W_{2q_0}^{2,1}(Q_T), \quad \hat{u} \in W_{2q_0}^{2,1}(Q_T). \tag{2.12}$$

Now, we consider the nonlinear operator

$$(u,\theta) \mapsto F(u,\theta) = \begin{pmatrix} \partial_t u - \operatorname{div}(\mu(\theta)\nabla u), & \gamma_0 u - u_0 \\ \partial_t \theta - \Delta\theta - \mu(\theta)|\nabla u|^2, & \gamma_0 \theta - \theta_0 \end{pmatrix},$$

where γ_0 is the usual trace operator $\gamma_0 u = u|_{t=0}$. We also introduce the Banach spaces

$$\mathcal{W} = \{u \in W_{2q}^{2,1}(Q_T) : \ u|_{S_T} = 0\},$$
$$\mathcal{V} = \{f \in L^{2q}(Q_T) : \ \partial_t f \in L^2(Q_T)\},$$
$$\mathcal{H} = L^{2q}(Q_T) \times \mathring{W}_{2q}^{2-1/q}(\Omega),$$

where $S_T := \partial\Omega \times (0,T)$.

THEOREM 2.3 (Gâteaux differentiability). *Assume conditions* (1.3) *and* (2.4) *hold and function* μ *is of class* C^2. *Suppose also* $q > 1$. *Then the transformation* F *is Gâteaux differentiable as a map*

$$F : \mathcal{W} \times \mathcal{W} \to \mathcal{H} \times \mathcal{H}$$

and its derivative is

$$\delta F(u,\theta)(w,e) = \begin{pmatrix} \partial_t w - \operatorname{div}(\mu(\theta)\nabla w) - \operatorname{div}(\mu'(\theta)e\nabla u), & \gamma_0 w \\ \partial_t e - \Delta e - \mu'(\theta)e|\nabla u|^2 - 2\mu(\theta)\nabla u \cdot \nabla w, & \gamma_0 e \end{pmatrix}.$$

Moreover, if $(\hat{u},\hat{\theta},\hat{f})$ *is an optimal solution of Problem* P *satisfying* (2.12) *and* $q \in (1, q_0)$, *then*

$$\text{image of } \ \delta F(\hat{u},\hat{\theta}) = \mathcal{H} \times \mathcal{H}. \tag{2.13}$$

THEOREM 2.4 (optimality system). *Assume that all conditions of Theorems* 2.2 *and* 2.3 *hold and let* $(\hat{u},\hat{\theta},\hat{f})$ *be a solution of Problem* P *satisfying* (2.12). *Then there are functions* $(\hat{p},\hat{e}) \in W_2^{2,1}(Q_T) \times W_2^{2,1}(Q_T)$ *satisfying the system*

$$\partial_t \hat{p} + \operatorname{div}(\mu(\hat{\theta})\nabla\hat{p}) - \operatorname{div}(2\mu(\hat{\theta})\hat{e}\nabla\hat{u}) = \hat{u} - U, \quad \hat{p}|_{t=T} = 0, \quad \hat{p}|_{\partial\Omega} = 0, \tag{2.14}$$

$$\partial_t \hat{e} + \Delta\hat{e} - \mu'(\hat{\theta})\nabla\hat{u}\cdot\nabla\hat{p} + \mu'(\hat{\theta})|\nabla\hat{u}|^2\hat{e} = \hat{\theta} - \Theta, \quad \hat{e}|_{t=T} = 0, \quad \hat{e}|_{\partial\Omega} = 0, \tag{2.15}$$

$$-\beta_2 \frac{\partial^2 \hat{f}}{\partial t^2} + 2q_0\beta_1|\hat{f}|^{2q_0-2}\hat{f} = \hat{p}, \qquad \left.\frac{\partial \hat{f}}{\partial t}\right|_{t=0} = \left.\frac{\partial \hat{f}}{\partial t}\right|_{t=T} = 0. \tag{2.16}$$

Remark 2.3. Instead of the Neumann conditions it is possible to get in (2.16) some other boundary conditions. For instance, to get in (2.16) the homogeneous Dirichlet boundary conditions, one should look in Problem P for the minimum of J among all $f \in \mathcal{V}$ such that $f|_{t=0} = f|_{t=T} = 0$.

Remark 2.4. All results of our paper remain true if we consider the problem of minimization of the cost functional (2.2) (with $\beta_2 > 0$) among all functions $f \in W_2^1(0,T)$, which depend only on t and do not depend on the spatial variables. This case corresponds to the physically reasonable boundary conditions for the magnetic field. In this case we must put $\mathcal{V} = W_2^1(0,T)$ and the relation (2.16) must be substituted with the ODE

$$-\beta_2 \frac{d^2 \hat{f}}{dt^2} + 2q_0\beta_1|\hat{f}|^{2q_0-2}\hat{f} = [\hat{p}]_\Omega, \qquad \left.\frac{\partial \hat{f}}{\partial t}\right|_{t=0} = \left.\frac{\partial \hat{f}}{\partial t}\right|_{t=T} = 0, \tag{2.17}$$

where

$$[\hat{p}]_\Omega(t) \equiv \int_\Omega \hat{p}(x,t)\, dx.$$

Moreover, in this case we can put $q_0 = 1$ in (2.2) (and hence obtain the linear equation in (2.17)). All improvements that one should put into the proof of Theorem 2.4 to cover this case are obvious.

3. Proof of Theorem 2.1. Before we prove Theorem 2.1 let us formulate here one lemma on solutions of the linear parabolic systems with measurable coefficients. We are going to use it later.

LEMMA 3.1 (higher integrability of solutions to parabolic equations). *Assume $Q_T = \Omega \times (0,T)$, $\Omega \subset \mathbb{R}^n$, is a bounded domain with a boundary of class C^1, and a matrix $A(x,t) = (A_{ij}(x,t))$ satisfies the conditions*

$$\begin{gathered} \exists\, \nu_0 > 0 \quad \text{such that} \quad A_{ij}(x,t)\xi_i\xi_j \geq \nu_0|\xi|^2 \quad \forall \xi \in \mathbb{R}^n, \\ A_{ij} \in L^\infty(Q_T), \quad A_{ij} = A_{ji}. \end{gathered} \tag{3.1}$$

Assume also $f \in L^{2q_0}(Q_T)$, $u_0 \in W^1_{2q_0}(\Omega)$ for some $q_0 > 1$ and let $u \in C([0,T]; L^2(\Omega)) \cap W_2^{1,0}(Q_T)$ be a weak solution to the equation

$$\begin{gathered} \partial_t u - \operatorname{div}(A(x,t)\nabla u) = f \quad \text{in } Q_T, \\ u|_{\partial\Omega} = 0, \quad u|_{t=0} = u_0. \end{gathered} \tag{3.2}$$

Then there is a constant $q > 1$ depending only on n, q_0, ν_0, $\|A\|_{\infty,Q_T}$, and Q_T such that $u \in W^{1,0}_{2q}(Q_T)$, and the estimate

$$\|\nabla u\|_{2q,Q_T} \leq C(\|f\|_{2q,Q_T} + \|u_0\|_{W^1_{2q}(\Omega)})$$

holds.

Lemma 3.1 is proved in [3, Thm. 2.2, p. 272]. For the case of the homogeneous initial data see also [4, Thm. 3.III].

Now we turn to the proof of Theorem 2.1. Let $(u^m, \theta^m, f^m) \in V \times W_2^{2,1}(Q_T) \times L^{2q_0}(Q_T)$ be a sequence minimizing $J(u,\theta,f)$. Obviously, $\{f^m\}$ is bounded in $L^{2q_0}(Q_T)$. We also have the energy estimate

$$\sup_t \|u^m\|^2_{2,\Omega} + \|\nabla u^m\|^2_{2,Q_T} \leq C\|f^m\|^2_{2,Q_T} \tag{3.3}$$

and the following higher integrability of the first spatial gradient of u (see Lemma 3.1 above): there is $q > 1$ such that

$$\|\nabla u^m\|_{2q,Q_T} \leq C(\|f^m\|_{2q,Q_T} + \|u_0\|_{W^1_{2q}(\Omega)}). \tag{3.4}$$

From the heat equation (1.2) we get

$$\|\theta^m\|_{W^{2,1}_q(Q_T)} \leq \text{const. does not depend on } m. \tag{3.5}$$

Moreover, from (1.1) we obtain that $\{\partial_t u^m\}$ is bounded in $L^2(0,T; W_2^{-1}(\Omega))$. Hence $\{u^m\}$ is compact in $L^2(Q_T)$ and we can extract subsequences $\{u^m\}$, $\{\theta^m\}$, and $\{f^m\}$ such that

$$\begin{aligned} u^m &\rightharpoonup \hat{u} && \text{in } W^{1,0}_{2q}(Q_T), \\ \partial_t u^m &\rightharpoonup \partial_t \hat{u} && \text{in } L^2(0,T; W_2^{-1}(\Omega)), \\ \theta^m &\rightharpoonup \hat{\theta} && \text{in } W^{2,1}_q(Q_T), \\ f^m &\rightharpoonup \hat{f} && \text{in } L^{2q_0}(Q_T), \\ u^m &\to \hat{u} && \text{in } L^2(Q_T), \\ u^m &\to \hat{u} && \text{a.e. in } Q_T, \\ \theta^m &\to \hat{\theta} && \text{a.e. in } Q_T, \end{aligned} \tag{3.6}$$

and $(\hat{u}, \hat{\theta})$ satisfy (1.4) and (1.5). Convergences (3.6) make it possible to pass to the limit in the first equation, (1.1). So we have

$$\int_{Q_T} (-\hat{u}\partial_t w + \mu(\hat{\theta})\nabla\hat{u}\cdot\nabla w)\, dxdt = \int_{Q_T} \hat{f} w\, dxdt \quad \forall w \in C_0^1(Q_T).$$

As we lack strong convergence for ∇u^m, we cannot pass directly to the limit in the second equation, (1.2). Instead, we use the representation

$$\mu(\theta^m)|\nabla u^m|^2 = \operatorname{div}(\mu(\theta^m)u^m\nabla u^m) + f^m u^m - \frac{1}{2}\partial_t|u^m|^2 \quad \text{in } \mathcal{D}'(Q_T), \tag{3.7}$$

which follows from (1.1). Hence

$$\begin{aligned} &c\int_{Q_T} (-\theta^m\partial_t\eta + \nabla\theta^m\cdot\nabla\eta)\, dxdt \\ &= \int_{Q_T} \mu(\theta^m)u^m\nabla u^m\cdot\nabla\eta\, dxdt + \int_{Q_T}\left(f^m u^m\eta + \frac{1}{2}|u^m|^2\partial_t\eta\right) dxdt \quad \forall\eta\in C_0^1(Q_T). \end{aligned}$$

Passing to the limit in this equation, we obtain the same identity for $\hat{u}$, $\hat{\theta}$, and $\hat{f}$. Then using for $\hat{u}$ the relation (3.7) again, we obtain the identity

$$\int_{Q_T} (-\hat{\theta}\partial_t\eta + \nabla\hat{\theta}\cdot\nabla\eta)\, dxdt = \int_{Q_T} \mu(\hat{\theta})|\nabla\hat{u}|^2\eta\, dxdt \quad \forall\eta\in C_0^1(Q_T).$$

Hence $(\hat{u}, \hat{\theta})$ also satisfy (1.2). As J is lower semicontinuous with respect to the weak convergence, we obtain that $(\hat{u}, \hat{\theta}, \hat{f})$ is a solution of Problem P. In the case of $\beta_2 > 0$ we also have the boundedness of $\{\partial_t f^m\}$ in $L^2(Q_T)$, and hence (2.6) holds. Theorem 2.1 is proved.

4. Proofs of Theorems 2.2 and 2.3. In this section we study regularity of weak solutions obtained in Theorem 2.1. Our approach is very close to the method used in [2] for investigation of the thermistor system in the "semistationary" case, i.e., the system (1.1)–(1.2) in the case when the term $\partial_t u$ in the left-hand side of (1.1) is absent. Roughly speaking, our method differs from the approach of [2] only in the use of an additional Step 4 shown below. This step involves so-called "weak coercive estimates" for the spatial gradient of a solution to the heat equation by the appropriate negative norm of the right-hand side. Both the method in [2] and our method follow the general methodology developed in [17, 18], so here we only briefly sketch the proof of Theorem 2.2, assuming the solution is sufficiently smooth and deriving only a priori estimates for it. These estimates can be easily justified (when it is necessary) with the help of appropriate approximations. But to make our presentation more intelligible we skip technical details in our proof and refer readers to [17, 18] for the formal justification.

Step 1. To get (2.7) we use the following lemma from [18, Chap. III, Thms. 7.1 and 10.1].

LEMMA 4.1 (De Giorgi–Nash–Ladyzhenskaya–Uraltseva theorem). *Assume $Q_T = \Omega\times(0,T)$, $\Omega\subset\mathbb{R}^n$, is a bounded domain with a boundary of class C^1, and let $f\in L^{s,r}(Q_T)$, $u_0\in C^{\alpha_0}(\bar{\Omega})$ for some $\alpha_0>0$, $u_0|_{\partial\Omega}=0$, and*

$$\frac{1}{r} + \frac{n}{2s} < 1. \tag{4.1}$$

Assume (3.1) *holds and let* $u \in W_2^{1,0}(Q_T)$ *be a weak solution of* (3.2). *Then there is* $\alpha > 0$ *such that* $u \in C^{\alpha,\alpha/2}(\bar{Q}_T)$ *and*

$$\|u\|_{C^{\alpha,\alpha/2}(\bar{Q}_T)} \le C(\|f\|_{L^{s,r}(Q_T)} + \|u_0\|_{C^{\alpha}(\bar{\Omega})}). \tag{4.2}$$

Boundedness and Hölder continuity of solutions of the scalar equations with measurable coefficients were proved first by De Giorgi for elliptic equations and by Nash for homogeneous parabolic equations; see [9, 20] and [12, 15]. The nonhomogeneous parabolic analogue of the De Giorgi–Nash theorem we use here was proved by Ladyzhenskaya and Uraltseva. We note that from the inclusion $\hat{f} \in L^{2q_0}(Q_T)$ we obtain that (4.1) holds for $u = \hat{u}$, $n = 2$, $r = \frac{1}{2q_0}$, $s = \frac{1}{2q_0}$.

Step 2. For any function $u \in C^{\alpha,\alpha/2}(\bar{Q}_T) \cap L^2(0,T; \mathring{W}_2^1 \cap W_2^2(\Omega))$ the following inequality holds; see [18, Chap. V, eqs. (3.15) and (4.6)]:

$$\|\nabla u\|^4_{4,\Omega_\rho(x_0)\times(0,T)} \le C\|u\|^2_{C^{\alpha,\frac{\alpha}{2}}(\bar{Q}_T)} \rho^{2\alpha} \left\{ \|\nabla^2 u\|^2_{2,\Omega_{2\rho}\times(0,T)} + \frac{1}{\rho^2}\|\nabla u\|^2_{2,\Omega_{2\rho}\times(0,T)} \right\},$$

where $\Omega_\rho(x_0) = \Omega \cap B_\rho(x_0)$ and $x_0 \in \bar{\Omega}$ is arbitrary. The proof of this relation is based on a simple integration-by-parts trick,

$$\int_\Omega \zeta^2 |\nabla u|^4 \, dx - \int_\Omega \operatorname{div}(\zeta^2 |\nabla u|^2 \nabla u)(u(x) - u(x_0)) \, dx,$$

where $x_0 \in \bar{\Omega}$ and $\zeta \in C_0^\infty(B_{2\rho}(x_0))$ is a cut-off function; see [18] for details.

From the assumption of the smoothness of $\partial\Omega$, the existence of numbers N_0, ρ_0 follows such that for any $\rho \le \rho_0$ there is a finite covering of Ω by the sets of type $\Omega_\rho(x_i)$, $x_i \in \bar{\Omega}$, such that the total number of intersections of different $\Omega_{2\rho}(x_i)$ does not increase N_0. Hence we have the estimate

$$\|\nabla u\|^4_{4,Q_T} \le C\|u\|^2_{C^{\alpha,\alpha/2}(\bar{Q}_T)} \rho^{2\alpha} \left\{ \|\nabla^2 u\|^2_{2,Q_T} + \frac{1}{\rho^2}\|\nabla u\|^2_{2,Q_T} \right\}. \tag{4.3}$$

This inequality is valid for any $\rho < \rho_0$, where ρ_0 depends only on curvature of $\partial\Omega$, and C does not depend on ρ.

Step 3. Let (u, θ, f) satisfy (1.1), (1.2), (1.4), and (1.5). From (1.1), taking into account (1.3), we can estimate the second derivatives of u:

$$\sup_{t\in(0,T)} \|\nabla u\|^2_{2,\Omega} + \|\nabla^2 u\|^2_{2,Q_T} \le C(\|\nabla u\|^4_{4,Q_T} + \|\nabla\theta\|^4_{4,Q_T}) + C_{f,u_0}. \tag{4.4}$$

Step 4. To estimate $\|\nabla\theta\|_{4,Q_T}$ by $\|\nabla u\|_{4,Q_T}$ we take advantage of the special structure (3.7) of the term $\mu(\theta)|\nabla u|^2$. From (1.2) we have

$$\begin{gathered} \partial_t\theta - \Delta\theta = \operatorname{div}(\mu(\theta)u\nabla u) + fu - u\partial_t u, \\ \theta|_{\partial\Omega} = 0, \quad \theta|_{t=0} = \theta_0. \end{gathered}$$

Denote $v := \theta + \frac{1}{2}u^2$, $G := (\mu(\theta) + 1)u\nabla u$, $g := fu$, $v_0 \equiv \theta_0 + \frac{1}{2}u_0^2$. Hence

$$\begin{aligned} &\partial_t v - \Delta v = \operatorname{div} G + g, \\ &v|_{\partial\Omega} = 0, \quad v|_{t=0} = v_0. \end{aligned}$$

We split $v = v^{(1)} + v^{(2)}$, where $v^{(1)}$ and $v^{(2)}$ are solutions of problems

$$\begin{aligned} \partial_t v^{(1)} - \Delta v^{(1)} &= g, \\ v^{(1)}|_{\partial\Omega} = 0, \quad v^{(1)}|_{t=0} &= v_0, \end{aligned} \tag{4.5}$$

$$\begin{aligned} \partial_t v^{(2)} - \Delta v^{(2)} &= \operatorname{div} G, \\ v^{(2)}|_{\partial\Omega} = 0, \quad v^{(2)}|_{t=0} &= 0, \end{aligned} \tag{4.6}$$

respectively. For $v^{(1)}$ the well-known estimate

$$\|v^{(1)}\|_{W_2^{2,1}(Q_T)} \le C(\|g\|_{2,Q_T} + \|v_0\|_{W_2^1(\Omega)}) \tag{4.7}$$

holds, provided the corresponding compatibility condition $v_0|_{\partial\Omega} = 0$ for initial and boundary data holds. Using two-dimensional parabolic imbedding $W_2^{2,1}(Q_T) \hookrightarrow W_4^{1,0}(Q_T)$ (see [18, Chap. II, Lem. 3.3]) we arrive at the estimate

$$\begin{aligned} \|\nabla v^{(1)}\|_{4,Q_T} &\le \|v^{(1)}\|_{W_2^{2,1}(Q_T)} \le C\big(\|g\|_{2,Q_T} + \|v_0\|_{W_2^1(\Omega)}\big) \\ &\le C\big(\|u\|_{\infty,Q_T}\|f\|_{2,Q_T} + \|u_0^2\|_{W_2^1(\Omega)} + \|\theta_0\|_{W_2^1(\Omega)}\big). \end{aligned}$$

It is well known (see [14]) that for any $G \in L^4(Q_T;\mathbb{R}^2)$ the problem (4.6) has the unique solution belonging to the class $v^{(2)} \in W_4^{1,1/2}(Q_T)$ and, moreover, the estimate

$$\|\nabla v^{(2)}\|_{4,Q_T} \le C\|G\|_{4,Q_T} \le C\|u\|_{\infty,Q_T}\|\nabla u\|_{4,Q_T} \tag{4.8}$$

holds. In [14] the analogous problem was studied for the much more difficult case of the Stokes operator. For solutions of the heat equation, the estimate (4.8) can be justified by the simple duplication of all considerations from [14] with the obvious change of hydrodynamical potentials for heat potentials. Note that in our case no convexity condition on Ω is necessary.

Remark 4.1. Note that (4.8) also can be obtained by interpolation of the corresponding $L^2 - L^2$ and $L^\infty - BMO$ estimates; see [12, Chap. IV, Thm. 4.6] or [19, Chap. VII, section 2, Thm. 7.2]. Precisely, the following two estimates hold:

$$\|\nabla v^{(2)}\|_{L^2(Q_T)} \le C\|G\|_{L^2(Q_T)}, \qquad \|\nabla v^{(2)}\|_{BMO(Q_T)} \le C\|G\|_{L^\infty(Q_T)}.$$

The first estimate is obvious, and the proof of the second one is actually contained in [19, Chap. VII]; also see arguments there in the proof of Theorem 7.17 in sections 3–5.

So, we get

$$\begin{aligned} \|\nabla\theta\|_{4,Q_T} &\le \|\nabla v\|_{4,Q_T} + \left\|\frac{1}{2}\nabla u^2\right\|_{4,Q_T} \\ &\le \|\nabla v^{(1)}\|_{4,Q_T} + \|\nabla v^{(2)}\|_{4,Q_T} + \left\|\frac{1}{2}\nabla u^2\right\|_{4,Q_T} \\ &\le C\|u\|_{\infty,Q_T}(\|\nabla u\|_{4,Q_T} + \|f\|_{2,Q_T}) + C\|u_0^2\|_{W_2^1(\Omega)} + C\|\theta_0\|_{W_2^1(\Omega)} \end{aligned}$$

or taking into account (4.2)

$$\|\nabla\theta\|_{4,Q_T} \le C_{f,u_0,\theta_0}\|\nabla u\|_{4,Q_T} + C'_{f,u_0,\theta_0}. \tag{4.9}$$

Step 5. Gathering (4.2), (4.3), (4.4), and (4.9) together we obtain the estimate

$$\|\nabla u\|^4_{4,Q_T} \le C_{f,u_0,\theta_0}\rho^{2\alpha}\left\{\|\nabla u\|^4_{4,Q_T} + \frac{1}{\rho^2}\|\nabla u\|^2_{2,Q_T}\right\} + C'_{f,u_0,\theta_0},$$

which is true for any $\rho < \rho_0$. Choosing $\rho^{2\alpha} < (1/2C_{f,u_0,\theta_0})$ we get the inclusion (2.8) for $u\hat{u}$. The inclusion (2.8) for $\hat{\theta}$ follows from (4.9).

Step 6. From (2.8) we obtain that the right-hand side of the heat equation (1.2) belongs to $L^2(Q_T)$, and hence $\theta \in W_2^{2,1}(Q_T)$. The inclusion $u \in W_2^{2,1}(Q_T)$ follows from (4.4). So, (2.9) is proved.

Step 7. Assume conditions (2.6) hold. Differentiating (1.1)–(1.2) with respect to t, multiplying them by $\partial_t u$, $\partial_t\theta$, respectively, and integrating over Ω using the two-dimensional multiplicative inequality (see [16]),

$$\|v\|^4_{4,\Omega} \le C\|v\|^2_{2,\Omega}\|\nabla v\|^2_{2,\Omega} \quad \forall v \in \mathring{W}_2^1(\Omega), \tag{4.10}$$

and after routine calculations we arrive at the inequality

$$\frac{d}{dt}\big(\|\partial_t u\|^2_{2,\Omega} + \|\partial_t\theta\|^2_{2,\Omega}\big) + \|\partial_t\nabla u\|^2_{2,\Omega} + \|\partial_t\nabla\theta\|^2_{2,\Omega} \le C\|\nabla u\|^4_{4,\Omega}\big(\|\partial_t u\|^2_{2,\Omega} + \|\partial_t\theta\|^2_{2,\Omega}\big) + C_{f,u_0,\theta_0}.$$

Applying the Gronwall lemma we obtain (2.10).

Step 8. From (2.10) and (4.10) we get $\partial_t\theta \in L^4(Q_T)$, which together with (2.8) and the Sobolev imbedding theorem $W_4^1(Q_T) \hookrightarrow C^{1/4}(\bar{Q}_T)$ (as $Q_T \subset \mathbb{R}^3$) provides the inclusion (2.11). Moreover, $\partial_t\nabla u, \nabla^2 u \in L^2(Q_T)$, and by the three-dimensional Sobolev imbedding theorem $W_2^1(Q_T) \hookrightarrow L^6(Q_T)$ we also have $\nabla u \in L^6(Q_T)$. This means that the right-hand side of the heat equation (1.2) belongs to $L^3(Q_T)$ and, due to the coercive estimates for the heat operator, we obtain $\theta \in W_3^{2,1}(Q_T)$. As $q_0 \le \frac{3}{2}$, the inclusion (2.12) for θ is proved. The same inclusion for u follows now from the assumption $f \in L^{2q_0}(Q_T)$ and the usual coercive estimates for parabolic equations with smooth coefficients; see [18, Chap. IV, Thm. 9.1] (see also Lemma 4.2 below; here in our case $s = 2q_0$ and $r = \frac{4q_0}{2-q_0} > s$). Therefore, Theorem 2.2 is proved.

Now we turn to the proof of Theorem 2.3. Assume $n = 2$ and $q \in (1, 2)$. The two-dimensional parabolic imbedding theorems (see [18, Chap. II, Lem. 3.3]) provide the estimates

$$\|v\|_{W^{1,0}_{\frac{4q}{2-q}}(Q_T)} \le C\|v\|_{W^{2,1}_{2q}(Q_T)}, \quad v \in W^{2,1}_{2q}(Q_T), \tag{4.11}$$

and

$$\|v\|_{C^{\lambda,\lambda/2}(\bar{Q}_T)} \le C\|v\|_{W^{2,1}_{2q}(Q_T)}, \quad v \in W^{2,1}_{2q}(Q_T). \tag{4.12}$$

From (4.11) and (1.3) inclusions $\mu'(\theta)\nabla\theta\cdot\nabla u, |\nabla u|^2 \in L^{\frac{2q}{2-q}}(Q_T) \subset L^{2q}(Q_T)$ for $q > 1$ follow, and hence F really maps $\mathcal{W}\times\mathcal{W}$ into $\mathcal{H}\times\mathcal{H}$. By the definition of the Gâteaux derivative of F,

$$\delta F(u,\theta)(w,e) \equiv \frac{d}{ds}F(u+sw, \theta+se)\Big|_{s=0},$$

and hence for all $u, \theta, w, e \in \mathcal{W}$,

$$\delta F(u,\theta)(w,e) = \begin{pmatrix} \partial_t w - \operatorname{div}(\mu(\theta)\nabla w) - \operatorname{div}(\mu'(\theta)e\nabla u), & \gamma_0 w \\ \partial_t e - \Delta e - \mu'(\theta)e|\nabla u|^2 - 2\mu(\theta)\nabla u\cdot\nabla w, & \gamma_0 e \end{pmatrix}.$$

Because of (1.3), (4.11), (4.12), and continuity of μ'', it is easy to see that for each given $(u,\theta) \in \mathcal{W}\times\mathcal{W}$ the operator $\delta F(u,\theta)$ is linear and bounded as a map from $\mathcal{W}\times\mathcal{W}$ into $\mathcal{H}\times\mathcal{H}$. Indeed, for example,

$$\begin{aligned} \delta_u F_1(u,\theta)(w,e) &\equiv \partial_t w - \operatorname{div}(\mu(\theta)\nabla w) - \operatorname{div}(\mu'(\theta)e\nabla u) \\ &= \partial_t w - \mu(\theta)\Delta w - \mu'(\theta)\nabla\theta\cdot\nabla w \\ &\quad - \mu'(\theta)e\Delta u - \mu'(\theta)\nabla e\nabla u - \mu''(\theta)e\nabla\theta\cdot\nabla u. \end{aligned}$$

Hence

$$\begin{aligned} \|\delta_u F_1(u,\theta)(w,e)\|_{2q,Q_T} &\le \|\partial_t w\|_{2q,Q_T} + \mu_1\|\Delta w\|_{2q,Q_T} + \mu_2\|\nabla\theta\cdot\nabla w\|_{2q,Q_T} \\ &\quad + \mu_2\|e\Delta u\|_{2q,Q_T} + \mu_2\|\nabla e\cdot\nabla u\|_{2q,Q_T} \\ &\quad + \|\mu''(\theta)\|_{\infty,Q_T}\|e\nabla\theta\cdot\nabla u\|_{2q,Q_T}. \end{aligned} \tag{4.13}$$

Using the Hölder inequality and (4.11)–(4.12) we get

$$\begin{aligned} \|e\nabla\theta\cdot\nabla u\|_{2q,Q_T} &\le \|e\|_{\infty,Q_T}\|\nabla\theta\cdot\nabla u\|_{2q,Q_T} \\ &\le \|e\|_{\infty,Q_T}\|\nabla\theta\|_{\frac{4q}{2-q},Q_T}\|\nabla u\|_{4,Q_T} \\ &\le C\|\theta\|_{\mathcal{W}}\|u\|_{\mathcal{W}}\|e\|_{\mathcal{W}} \end{aligned}$$

and similar relations for all other terms in the right-hand side of (4.13). From the continuity of μ'' and (4.12), existence of the nondecreasing majorants $\mathcal{F}_1, \mathcal{F}_2$ follows such that

$$\|\mu''(\theta)\|_{\infty,Q_T} \le \mathcal{F}_1(\|\theta\|_{C(\bar{Q}_T)}) \le \mathcal{F}_2(\|\theta\|_{\mathcal{W}}).$$

Hence we obtain the final majorant $\mathcal{F}$ such that

$$\|\delta F(u,\theta)(w,e)\|_{\mathcal{H}\times\mathcal{H}} \le \mathcal{F}(\|u\|_{\mathcal{W}}, \|\theta\|_{\mathcal{W}})(\|w\|_{\mathcal{W}} + \|e\|_{\mathcal{W}}).$$

So the first part of Theorem 2.3 is proved.

Let us prove (2.13), i.e., for any given (g,a) and $(h,b) \in \mathcal{H}$ the system

$$\left.\begin{aligned} \partial_t w - \operatorname{div}(\mu(\hat{\theta})\nabla w) - \operatorname{div}(\mu'(\hat{\theta})e\nabla\hat{u}) &= g, \\ \partial_t e - \Delta e - \mu'(\hat{\theta})e|\nabla\hat{u}|^2 - 2\mu(\hat{\theta})\nabla\hat{u}\cdot\nabla w &= h, \\ w|_{\partial\Omega} = 0, \quad w|_{t=0} &= a, \\ e|_{\partial\Omega} = 0, \quad e|_{t=0} &= b \end{aligned}\right\} \tag{4.14}$$

is uniquely solvable on $\mathcal{W}\times\mathcal{W}$. This statement follows directly from properties (2.12) of the optimal solution $(\hat{u},\hat{\theta})$ and the following lemma on solutions to the linear parabolic systems we have already used several times (see [18, Thm. 9.1 of Chap. IV and Thm. 10.4 of Chap. VII]; see also [22]).

LEMMA 4.2 (coercive estimate for linear parabolic systems). *Assume $Q_T = \Omega\times(0,T)$, $\Omega \subset \mathbb{R}^n$, is a bounded domain with a boundary of class C^2, and assume also*

$$A_{ijkl} \in C(\bar{Q}_T), \quad b_{ijk} \in L^r(Q_T), \quad c_{ij} \in L^{r/2}(Q_T)$$

with

$$r > n + 2, \tag{4.15}$$

and A satisfying the strong ellipticity condition

$$\exists \nu_0 > 0 : \quad A_{ijkl}(x) B_{ij} B_{kl} \geq \nu_0 |B|^2 \quad \forall B \in \mathbb{M}^{n\times n},$$

$$A_{ijkl} A_{klij} = A_{jikl} = A_{ijlk}.$$

Then for any $s \in (1, r)$, $s \neq \frac{3}{2}$, and for arbitrary functions $f \in L^s(Q_T; \mathbb{R}^N)$, $u_0 \in \mathring{W}_s^{2-2/s}(\Omega; \mathbb{R}^N)$ there is a unique solution $u \in W_s^{2,1}(Q_T; \mathbb{R}^N)$ of the problem

$$\partial_t u_i - A_{ijkl}(x) u_{k,jl} + b_{ijk}(x) u_{j,k} + c_{ij}(x) u_j = f_i(x),$$
$$u|_{t=0} = u_0, \quad u|_{\partial\Omega} = 0.$$

Moreover, the estimate

$$\|u\|_{W_s^{2,1}(Q_T)} \leq C \left(\|f\|_{s,Q_T} + \|u_0\|_{W_s^{2-\frac{2}{s}}(\Omega)} \right)$$

holds for some constant C depending only on n, Ω, T, ν_0, and norms of the coefficients.

Indeed, we can rewrite the system (4.14) in the form

$$\left.\begin{aligned} \partial_t w - \mu(\hat\theta)\Delta w - \mu'(\hat\theta)\nabla\hat\theta \cdot \nabla w - \mu'(\hat\theta) e \Delta \hat u \\ - \mu'(\hat\theta)\nabla e \cdot \nabla \hat u - \mu''(\hat\theta) e \nabla\hat\theta \cdot \nabla \hat u = g, \\ \partial_t e - \Delta e - \mu'(\hat\theta) e |\nabla \hat u|^2 - 2\mu(\hat\theta)\nabla\hat u \cdot \nabla w = h, \\ w|_{\partial\Omega} = 0, \quad w|_{t=0} = a, \\ e|_{\partial\Omega} = 0, \quad e|_{t=0} = b. \end{aligned}\right\}$$

Due to (2.11) and (2.12), it is easy to see that its coefficients

$$\mu'(\hat\theta)\nabla\hat\theta, \quad \mu'(\hat\theta)\nabla\hat u, \quad \mu(\hat\theta)\nabla\hat u \in L^{4q_0}(Q_T), \tag{4.16}$$

$$\mu'(\hat\theta)\Delta\hat u, \quad \mu''(\hat\theta)\nabla\hat\theta \cdot \nabla\hat u, \quad \mu'(\hat\theta)|\nabla\hat u|^2 \in L^{2q_0}(Q_T) \tag{4.17}$$

satisfy conditions (4.15) with $n = 2$, $r = 4q_0$. Moreover, due to (2.11) the coefficient $\mu(\hat\theta(\cdot))$ is Hölder continuous. Hence the system (4.14) is uniquely solvable on $\mathcal{W} \times \mathcal{W}$ for any data in $\mathcal{H} \times \mathcal{H}$ as $q \in (1, q_0)$. The statement (2.13) and Theorem 2.3 are proved.

5. Proof of Theorem 2.4. We consider our cost functional J as a map

$$J : \mathcal{W} \times \mathcal{W} \times \mathcal{V} \to \mathbb{R},$$

and (1.1)–(1.2) as a restriction

$$F_*(u, \theta, f) = 0, \qquad F_* : \mathcal{W} \times \mathcal{W} \times \mathcal{V} \to \mathcal{H} \times \mathcal{H},$$

where

$$F_*(u, \theta, f) = \begin{pmatrix} \partial_t u - \operatorname{div}(\mu(\theta)\nabla u) - f, & \gamma_0 u - u_0 \\ \partial_t \theta - \Delta\theta - \mu(\theta)|\nabla u|^2, & \gamma_0\theta - \theta_0 \end{pmatrix}$$

(note that F_* differs from the map F in section 2 only by the additional argument f). Hence

$$\delta F_*(u,\theta,f)(w,\eta,h) = \begin{pmatrix} \partial_t w - \operatorname{div}(\mu(\theta)\nabla w) - \operatorname{div}(\mu'(\theta)\eta\nabla u) - h, & \gamma_0 w \\ \partial_t \eta - \Delta\eta - \mu'(\theta)\eta|\nabla u|^2 - 2\mu(\theta)\nabla u\cdot\nabla w, & \gamma_0\eta \end{pmatrix}.$$

Let us show that

$$\text{image of } \ \delta F_* = \mathcal{H}\times\mathcal{H}. \tag{5.1}$$

Let $(g,a),(d,b)\in\mathcal{H}$ be arbitrary. To find $(w,e,h)\in\mathcal{W}\times\mathcal{W}\times\mathcal{V}$ such that

$$\delta F_*(u,\theta,f)(w,e,h) = \begin{pmatrix} g,a \\ d,b \end{pmatrix}$$

it is enough to take $h\equiv 0$ and use the property (2.13) of δF. For any $((p,a),(e,b))\in \mathcal{H}'\times\mathcal{H}'$ consider Lagrangian

$$\mathcal{L}(u,\theta,f,p,e,a,b) \equiv J(u,\theta,f) + \left\langle F_*(u,\theta,f), \begin{pmatrix} p,a \\ e,b \end{pmatrix}\right\rangle,$$

where $\langle\cdot,\cdot\rangle$ is a duality relation between $\mathcal{H}$ and $\mathcal{H}'$.

From Theorem 2.3 and (5.1) it follows that J and F_* satisfy all conditions of Theorem 2.1.5 of [11]; see also [1, Chap. III, section 3.2, Thm. 3.2.2]. Hence if $(\hat u,\hat\theta,\hat f)$ is an optimal solution of Problem P, then there are Lagrange multipliers $((\hat p,\hat a),(\hat e,\hat b))\in\mathcal{H}'\times\mathcal{H}'$ satisfying

$$\delta_{(u,\theta,f)}\mathcal{L}(\hat u,\hat\theta,\hat f,\hat p,\hat e,\hat a,\hat b)(w,\eta,h) = 0 \quad \forall (w,\eta,h)\in\mathcal{W}\times\mathcal{W}\times\mathcal{V},$$

where $\delta_{(u,\theta,f)}\mathcal{L}$ is the Gâteaux derivative of $\mathcal{L}$ with respect to (u,θ,f). The last relation is equivalent to the following system:

$$\begin{aligned} &\textstyle\int_{Q_T}((\hat u - U)w + (\hat\theta-\Theta)\eta + 2q_0\beta_1|\hat f|^{2q_0-2}\hat f h + \beta_2\partial_t\hat f\partial_t h)\,dz \\ &\quad + \textstyle\int_{Q_T}(\partial_t w - \operatorname{div}(\mu(\hat\theta)\nabla w) - \operatorname{div}(\mu'(\hat\theta)\eta\nabla u) - h)\hat p\,dz \\ &\quad + \textstyle\int_{Q_T}(\partial_t\eta - \Delta\eta - 2\mu(\hat\theta)\nabla\hat u\cdot\nabla w - \mu'(\hat\theta)\eta|\nabla\hat u|^2)\hat e\,dz \\ &\quad + \langle\gamma_0 w,\hat a\rangle + \langle\gamma_0\eta,\hat b\rangle = 0 \quad \forall(w,\eta,h)\in\mathcal{W}\times\mathcal{W}\times\mathcal{V}, \end{aligned}$$

or equivalently

$$\begin{aligned} &c\textstyle\int_{Q_T}((\hat u - U)w + \partial_t w\hat p - \operatorname{div}(\mu(\hat\theta)\nabla w)\hat p - 2\mu(\hat\theta)\nabla\hat u\cdot\nabla w\hat e)\,dxdt \\ &+ \textstyle\int_{Q_T}((\hat\theta-\Theta)\eta - \operatorname{div}(\mu'(\hat\theta)\eta\nabla\hat u)\hat p + \partial_t\eta\hat e - \Delta\eta\hat e - \mu'(\hat\theta)\eta|\nabla\hat u|^2\hat e)\,dxdt \\ &+ \textstyle\int_{Q_T}(2q_0\beta_1|\hat f|^{2q_0-2}\hat f h + \beta_2\partial_t\hat f\partial_t h - \hat p h)dxdt + \langle\gamma_0 w,\hat a\rangle + \langle\gamma_0\eta,\hat b\rangle \\ &= 0 \quad \forall(w,\eta,h)\in\mathcal{W}\times\mathcal{W}\times\mathcal{V}. \end{aligned} \tag{5.2}$$

Define functions $(p,e)\in W_2^{2,1}(Q_T)\times W_2^{2,1}(Q_T)$ as a solution to the following system:

$$\left.\begin{aligned} \partial_t p + \operatorname{div}(\mu(\hat\theta)\nabla p) - \operatorname{div}(2\mu(\hat\theta)e\nabla\hat u) &= \hat u - U, \\ \partial_t e + \Delta e - \mu'(\hat\theta)\nabla\hat u\cdot\nabla p + \mu'(\hat\theta)|\nabla\hat u|^2 e &= \hat\theta - \Theta, \\ p|_{\partial\Omega} = 0, \quad e|_{\partial\Omega} = 0, \quad p|_{t=T} = 0, \quad e|_{t=T} &= 0. \end{aligned}\right\} \tag{5.3}$$

Existence and uniqueness of solutions to this backward parabolic system follow from Lemma 4.2, and conditions (2.1) and inclusions (4.17) and (4.16) hold for the coefficients of the system (5.3). Taking $h \equiv 0$ in (5.2), multiplying (5.3) by $(w, \eta) \in \mathcal{W} \times \mathcal{W}$, integrating by parts, and taking a difference with (5.2), we arrive at the identity

$$\begin{aligned} &\int_{Q_T} (\partial_t w - \operatorname{div}(\mu(\hat\theta)\nabla w) - \operatorname{div}(\mu'(\hat\theta)\eta\nabla\hat u))(p - \hat p)\, dxdt \\ &+ \int_{Q_T} (\partial_t \eta - \Delta\eta - \mu'(\hat\theta)\eta|\nabla\hat u|^2 - 2\mu(\hat\theta)\nabla\hat u \cdot \nabla w)(e - \hat e)\, dxdt \\ &+ \langle \gamma_0 w, \gamma_0 \hat p - \hat a\rangle + \langle \gamma_0\eta, \gamma_0\hat e - \hat b\rangle = 0 \quad \forall (w,\eta) \in \mathcal{W}\times\mathcal{W}. \end{aligned} \tag{5.4}$$

Take in (5.4) the test functions $(w, \eta) \in \mathcal{W} \times \mathcal{W}$ satisfying the system

$$\begin{gathered} \partial_t w - \operatorname{div}(\mu(\hat\theta)\nabla w) - \operatorname{div}(\mu'(\hat\theta)\eta\nabla\hat u)\operatorname{sign}(p - \hat p), \\ \partial_t \eta - \Delta\eta - \mu'(\hat\theta)\eta|\nabla\hat u|^2 - 2\mu(\hat\theta)\nabla\hat u\cdot\nabla w = \operatorname{sign}(e - \hat e), \\ \gamma_0 w = 0, \quad \gamma_0\eta = 0. \end{gathered}$$

Note that the right-hand sides of this system belong to $L^\infty(Q_T)$, and hence existence of the solution to this system in the class $W_{2q}^{2,1}(Q_T) \times W_{2q}^{2,1}(Q_T)$ follows for $q < q_0$ from Lemma 4.2 and conditions (4.17) and (4.16). Then from (5.4) the identities $p = \hat p$, $e = \hat e$ follow, and hence inclusions $\hat p, \hat\theta \in W_2^{2,1}(Q_T)$ hold. From (5.4) we also get $\hat a = \gamma_0\hat p$, $\hat b = \gamma_0\hat e$. Finally, taking in (5.2) $w = \eta = 0$ we get a relation

$$\int_{Q_T} (2q_0\beta_1|\hat f|^{2q_0-2}\hat f h + \beta_2\partial_t\hat f\partial_t h - \hat p h)dxdt = 0 \quad \forall h \in \mathcal{V},$$

which is the weak form of (2.16). Theorem 2.4 is proved.

REFERENCES

[1] V. M. Alekseev, V. M. Tikhomirov, and S. V. Fomin, *Optimal Control*, Nauka, Moscow, 1979 (in Russian).

[2] S. N. Antontsev and M. Chipot, *The thermistor problem: Existence, smoothness, uniqueness, blowup*, SIAM J. Math. Anal., 25 (1994), pp. 1128–1156.

[3] A. Bensoussan, J. L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, North-Holland, Amsterdam, 1978.

[4] S. Campanato, *L^p-regularity and partial Hölder continuity for solutions of second order parabolic systems with strictly controlled growth*, Ann. Mat. Pura Appl. (4), 128 (1981), pp. 287–316.

[5] G. Cimatti, *On the problem of the electrical heating of a conductor*, Riv. Mat. Univ. Parma (4), 14 (1988), pp. 53–59.

[6] G. Cimatti and M. Chipot, *A uniqueness result for the thermistor problem*, European J. Appl. Math., 2 (1991), pp. 97–103.

[7] G. Cimatti, *A remark on the thermistor problem with rapidly growing conductivity*, Appl. Anal., 80 (2001), pp. 133–140.

[8] G. Cimatti, *Stability and multiplicity of solutions for the thermistor problem*, Ann. Mat. Pura Appl. (4), 181 (2002), pp. 181–212.

[9] E. De Giorgi, *Sulla differenziabilita e l'analiticita delle estremali degli integrali multipli regolari*, Mem. Accad. Sci. Torino Cl. Sci. Fis. Mat. Natur. (3), 3 (1957), pp. 25–43.

[10] J. Frehse, *A discontinuous solution of mildly nonlinear elliptic systems*, Math. Z., 134 (1973), pp. 229–230.

[11] A. V. Fursikov, *Optimal Control of Distributed Systems. Theory and Applications*, AMS, Providence, RI, 2000.

[12] M. Giaquinta, *Introduction to Regularity Theory for Nonlinear Elliptic Systems*, Lectures Math. ETH Zürich, Birkhäuser Verlag, Basel, 1993.

[13] S. D. Howison, J. F. Rodrigues, and M. Shillor, *Stationary solutions to the thermistor problem*, J. Math. Anal. Appl., 174 (1993), pp. 573–588.

[14] H. Koch and V. A. Solonnikov, *L_p-estimates of solutions of the nonstationary Stokes problem*, J. Math. Sci. (New York), 106 (2001), pp. 3042–3072.
[15] Q. Han and F.-H. Lin, *Elliptic Partial Differential Equations*, Courant Lect. Notes Math. 1, Courant Institute of Mathematical Sciences, New York; AMS, Providence, RI, 1997.
[16] O. A. Ladyzhenskaya, *The Mathematical Theory of Viscous Incompressible Flow*, 2nd English ed., revised and enlarged, Math. Appl. 2, Gordon and Breach, New York, London, Paris, 1969.
[17] O. A. Ladyzhenskaya and N. N. Uraltseva, *Linear and Quasi-linear Equations of Elliptic Type*, Academic Press, New York, 1968.
[18] O. A. Ladyzhenskaya, V. A. Solonnikov, and N. N. Uraltseva, *Linear and Quasi-linear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1967.
[19] G. Lieberman, *Second Order Parabolic Differential Equations*, World Scientific, River Edge, NJ, 1996.
[20] J. Nash, *Continuity of solutions of parabolic and elliptic equations*, Amer. J. Math., 80 (1958), pp. 931–954.
[21] J. F. Rodrigues, *A nonlinear parabolic system arising in thermo mechanics and in thermo magnetism*, Math. Models Methods Appl. Sci., 2 (1992), pp. 271–281.
[22] V. A. Solonnikov, *On boundary value problems for linear parabolic systems of differential equations of general form*, Trudy Mat. Inst. Steklov., 83 (1965), pp. 3–163 (in Russian).

SIAM J. CONTROL OPTIM.
Vol. 44, No. 1, pp. 283–312

STABILITY OF STOCHASTIC APPROXIMATION UNDER VERIFIABLE CONDITIONS*

CHRISTOPHE ANDRIEU†, ÉRIC MOULINES‡, AND PIERRE PRIOURET§

Abstract. In this paper we address the problem of the stability and convergence of the stochastic approximation procedure

$$\theta_{n+1} = \theta_n + \gamma_{n+1}[h(\theta_n) + \xi_{n+1}].$$

The stability of such sequences $\{\theta_n\}$ is known to heavily rely on the behavior of the mean field h at the boundary of the parameter set and the magnitude of the stepsizes used. The conditions typically required to ensure convergence, and in particular the boundedness or stability of $\{\theta_n\}$, are either too difficult to check in practice or not satisfied at all. This is the case even for very simple models. The most popular technique for circumventing the stability problem consists of constraining $\{\theta_n\}$ to a compact subset $\mathcal{K}$ in the parameter space. This is obviously not a satisfactory solution, as the choice of $\mathcal{K}$ is a delicate one. In this paper we first prove a "deterministic" stability result, which relies on simple conditions on the sequences $\{\xi_n\}$ and $\{\gamma_n\}$. We then propose and analyze an algorithm based on projections on adaptive truncation sets, which ensures that the aforementioned conditions required for stability are satisfied. We focus in particular on the case where $\{\xi_n\}$ is a so-called Markov state-dependent noise. We establish both the stability and convergence with probability 1 (w.p. 1) of the algorithm under a set of simple and verifiable assumptions. We illustrate our results with an example related to adaptive Markov chain Monte Carlo algorithms.

Key words. stochastic approximation, state-dependent noise, randomly varying truncation, adaptive Markov chain Monte Carlo

AMS subject classifications. 62L20, 90C15

DOI. 10.1137/S0363012902417267

1. Introduction.

In many contexts it is of interest to find the roots of possibly nonlinear equations of the form

$$h(\theta) = 0, \qquad \theta \in \Theta, \tag{1.1}$$

for some mapping $h : \Theta \to \mathbb{R}^{n_\theta}$, where $\Theta \subset \mathbb{R}^{n_\theta}$ for some integer n_θ. Most of the methods for solving the previous equation are iterative, i.e., produce a sequence of iterates $\{\theta_n, n \geq 0\}$, which eventually converges to the set of solutions of (1.1),

$$\mathcal{S} := \{\theta \in \Theta, h(\theta) = 0\}. \tag{1.2}$$

Stochastic approximation (SA) is a class of algorithms for solving (1.1) in the situation where only noisy measurements of h are available. In its simplest form, the Robbins–Monro algorithm produces a sequence $\{\theta_n, n \geq 0\}$ defined recursively as

$$\theta_0 \in \Theta, \qquad \theta_{n+1} = \theta_n + \gamma_{n+1}\zeta_{n+1}, \quad n \geq 1, \tag{1.3}$$

*Received by the editors November 6, 2002; accepted for publication (in revised form) August 4, 2004; published electronically July 18, 2005.
http://www.siam.org/journals/sicon/44-1/41726.html
†University of Bristol, School of Mathematics, University Walk, BS8 1TW, UK (c.andrieu@bris.ac.uk).
‡École Nationale Supérieure des Télécommunications, URA CNRS 820, 46, rue Barrault, F 75634 Paris 13, France (moulines@tsi.enst.fr).
§Université Pierre & Marie Curie, Laboratoire de Probabilités et Modélisation Aléatoire, URA CNRS 224, F 75252 Paris 05, France (priouret@ccr.jussieu.fr).

where $\{\gamma_n, n \geq 1\}$ is a sequence of stepsizes satisfying standard conditions (say, $\gamma_n \downarrow 0$ and $\sum_{n\geq 1} \gamma_n = \infty$) and, for any $n \geq 1$, ζ_n is a noisy measurement of $h(\theta_n)$. It is useful to introduce the sequence $\{\xi_n, n \geq 1\}$ defined as

$$\zeta_{n+1} = h(\theta_n) + \xi_{n+1}, \tag{1.4}$$

which will be referred to as the *noise sequence*. Convergence of SA has been studied under various sets of assumptions for the mean field h and the noise sequence $\{\xi_n, n \geq 1\}$ since the early work of [22]; see also [5], [17], [23], [16], and the references therein. Essentially, convergence of the SA sequence can be established toward an *attractive* subset provided the sequence $\{\theta_n, n \geq 0\}$ is with probability 1 (w.p. 1) in a compact subset of Θ and is w.p. 1 infinitely often in the *domain of attraction* of this attractive subset. Showing in practice that $\{\theta_n, n \geq 0\}$ satisfies these boundedness and recurrence conditions proves to be a difficult task. The available results hold under conditions which are still restrictive, despite recent advances (see [1], [7], [6], and references therein). This major drawback has motivated the design of modified Robbins–Monro recursions. Probably the most widely used method in practice consists of constraining the sequence $\{\theta_n, n \geq 0\}$ to some compact set $\mathcal{K} \subset \Theta$ by means of a reprojection onto $\mathcal{K}$. This method has been thoroughly investigated in [23] (see also [8] and the references therein). Although relatively easy to implement, and appropriate when constraints about the system considered are available a priori, this approach becomes impractical and questionable in many situations.

Our contributions toward solving the stability and convergence problems are twofold. First we establish and prove in section 2 a general result of stability, Theorem 2.2, for deterministic sequences of the form given by (1.3)–(1.4). This key deterministic result assumes the existence of a global Lyapunov function for the mean field h and mild general assumptions about the noise and stepsize sequences. In contrast with previous results, the conditions required on the growth of the Lyapunov functions and the mean field h when θ approaches the boundaries of the parameter set Θ are minimal. As a consequence the result is applicable to quite general settings. We then show that, under the conditions that guarantee stability, the convergence of the deterministic sequence (1.3)–(1.4) is ensured (see Theorem 2.3).

Our second contribution here consists of proposing an SA algorithm (section 3) for which the aforementioned noise and stepsize conditions are satisfied w.p. 1. There are many different applications of stochastic approximations which imply markedly different types of assumptions on the noise sequence $\{\xi_n\}$. Whereas our deterministic stability and convergence results mentioned above can be applied quite generally, we focus in this paper on the subtle Markov state-dependent noise (see [23, Chapter 6, section 6.6] and section 3 in this paper) for which the availability of algorithms, whose convergence can be established under general but nevertheless verifiable assumptions, is still missing. The proposed algorithm is a modification of the classical Robbins–Monro procedure described in (1.3)–(1.4), based on truncations on adaptive truncation sets, in the spirit of the seminal works [11] and [10].

The convergence of SA with adaptive truncation sets has been considered under various conditions on the noise sequence $\{\xi_n\}$. These include state-independent noise conditions (see, for example, [12, section 2.4, pp. 42–44]) but also state-dependent martingale differences [30], [14], [9], [12, section 2.5, pp. 49–57] or state-dependent ϕ-mixing processes [9], [12, section 2.5, p. 49]. However, the application of this strategy to the Markovian state-dependent case requires even more care, and it is therefore not surprising to find that the results on the topic are scarce and have been obtained

under conditions that are more stringent than those considered in the present paper; see [31], [13] and, for the special case of ARMAX models, see [12, Chapter 6]. As we shall see, our procedure differs in some respects from the original procedure proposed in [11] and [10] and offers additional degrees of freedom. Our technique of proof for the stability relies on a novel approach and offers as a byproduct an explicit bound for the tail probability of the number of reprojections, which is found to be super-exponential under mild technical conditions.

In order to illustrate our findings and their applicability, we propose (see section 7) to analyze the convergence of an adaptive Markov chain Monte Carlo (MCMC) algorithm recently proposed in [19] and analyzed under more stringent conditions than those considered here. Other examples can be found in [2].

2. Key deterministic results. In this section we establish both stability and convergence results for deterministic recursions of the type described in (1.3)–(1.4). Before stating our first assumptions, some definitions and notation are needed. Let d be a positive integer. An element v of $\mathbb{R}^d$ is denoted by its column vector v and its transpose is denoted by v^{T}. For elements v, w of $\mathbb{R}^d$, we denote by $\langle v, w\rangle$ their inner product, so that $|v| = \sqrt{\langle v, v\rangle}$ denotes the norm of v. Our first assumption is the existence of a global Lyapunov function w for the mean field h. Denoting $\mathcal{W}_M := \{\theta \in \Theta, w(\theta) \le M\} \subset \Theta$ we assume the following:

(A1) Θ is an open subset of $\mathbb{R}^{n_\theta}$, $h : \Theta \to \mathbb{R}^{n_\theta}$ is continuous, and there exists a continuously differentiable function $w : \Theta \to [0, \infty)$ such that

(i) there exists $M_0 > 0$ such that

$$\mathcal{L} := \left\{\theta \in \Theta, \left\langle \nabla w(\theta), h(\theta)\right\rangle = 0\right\} \subset \{\theta \in \Theta, w(\theta) < M_0\};$$

(ii) there exists $M_1 \in (M_0, \infty]$ such that $\mathcal{W}_{M_1}$ is a compact set;

(iii) for any $\theta \in \Theta \setminus \mathcal{L}$, $\langle \nabla w(\theta) h(\theta)\rangle < 0$;

(iv) the closure of $w(\mathcal{L})$ has an empty interior.

If h is a gradient field, i.e., $h = -\nabla J$ for some lower bounded real valued and differentiable function $\theta \mapsto J(\theta)$, then the choice $w = J$ is appropriate, provided that J is continuously differentiable. Note that, in situations where the set of stationary points cannot be characterized explicitly, one might use Sard's theorem from differential geometry in order to check (A1)(iv). Indeed, Sard's theorem states that if w is n_θ-times continuously differentiable, then $w(\{\nabla w = 0\})$ has an empty interior.

Our approach to proving our stability and convergence results can be decomposed into two distinct steps. In the first step (this section), we establish deterministic conditions on a noise sequence $\{\xi_n\}$ and a stepsize sequence $\{\rho_n\}$, upon which a deterministic sequence $\{\theta_n\}$ defined as

$$\theta_0 \in \Theta \quad \theta_{n+1} = \theta_n + \rho_{n+1}[h(\theta_n) + \xi_{n+1}] \quad \text{for} \quad n \ge 0, \tag{2.1}$$

has the following properties: (i) It remains in a compact subset of Θ (see Theorem 2.2) and (ii) it converges to $\mathcal{L}$ (Theorem 2.3) provided that $\{\theta_n\}$ remains in a compact subset of Θ. In a second step—which is probabilistic in nature and depends on how the noise is generated—we develop a general algorithm for the case where $\{\xi_n\}$ follows a Markovian state-dependent dynamic, which allows one to show that the required condition on $\{\xi_n\}$ is satisfied w.p. 1 (sections 3–6).

Before proving Theorems 2.2 and 2.3 we prove in the following lemma a fundamental contraction property of the Lyapunov function w. This result is the crux of both the proof of stability and the proof of convergence.

LEMMA 2.1. *Assume* (A1). *Then*

(i) *Let* $\mathcal{K} \subset \Theta$ *be a compact subset such that* $0 < \inf_{\theta\in\mathcal{K}} |\langle \nabla w, h\rangle|$. *For any* $0 < \delta < \inf_{\theta\in\mathcal{K}} |\langle \nabla w, h\rangle|$, *there exist* $\lambda > 0$ *and* $\beta > 0$ *such that, for any* ρ, $0 \le \rho \le \lambda$, ζ, $|\zeta| \le \beta$, *and* $\theta \in \mathcal{K}$, $w(\theta + \rho h(\theta) + \rho\zeta) \le w(\theta) - \rho\delta$.

(ii) *For any* $M \in (M_0, M_1]$ (*where* M_0 *is defined in* (A1)(i) *and* M_1 *is defined in* (A1)(ii)), *there exist* $\lambda > 0$ *and* $\beta > 0$ *such that, for any* ρ, $0 \le \rho \le \lambda$, ζ, $|\zeta| \le \beta$, *and* $\theta \in \mathcal{W}_M$, $\theta + \rho h(\theta) + \rho\zeta \in \mathcal{W}_M$.

Proof. We first prove (i).For any $0 < \delta < \inf_{\theta\in\mathcal{K}} |\langle \nabla w, h\rangle|$, there exist $\lambda > 0$ and $\beta > 0$ such that, for all ρ, $0 \le \rho \le \lambda$, ζ, $|\zeta| \le \beta$, and t, $0 \le t \le 1$, we have for all $\theta \in \mathcal{K}$, $\theta + \rho t h(\theta) + \rho t\zeta \in \Theta$ and

$$\left|\left\langle \nabla w(\theta), h(\theta)\right\rangle - \left\langle \nabla w(\theta + \rho t h(\theta) + \rho t\zeta), h(\theta) + \zeta\right\rangle\right| \le \inf_{\theta\in\mathcal{K}} |\left\langle \nabla w, h\right\rangle| - \delta.$$

Then for any ρ, $0 \le \rho \le \lambda$, and ζ, $|\zeta| \le \beta$,

$$\begin{aligned} w(\theta + \rho h(\theta) + \rho\zeta) - w(\theta) &= \rho\left\langle \nabla w(\theta), h(\theta)\right\rangle + \rho \int_0^1 \left(\left\langle \nabla w(\theta + t\rho h(\theta) + t\rho\zeta), h(\theta) + \zeta\right\rangle \right. \\ &\qquad \left. - \left\langle \nabla w(\theta), h(\theta)\right\rangle\right) dt \\ &\le -\rho \inf_{\theta\in\mathcal{K}} |\left\langle \nabla w, h\right\rangle| + \rho\left(\inf_{\theta\in\mathcal{K}} |\left\langle \nabla w, h\right\rangle| - \delta\right) = -\rho\delta. \end{aligned}$$

We now prove (ii). Consider $M' \in (M_0, M)$. Since $\mathcal{W}_{M'}$ is compact and w continuous, there exists $\lambda_0 > 0$ and $\beta_0 > 0$ such that, for all $0 \le \rho \le \lambda_0$, $|\zeta| \le \beta_0$, and $\theta \in \mathcal{W}_{M'}$, $\theta + \rho h(\theta) + \rho\zeta \in \mathcal{W}_M$. We can apply (i) to the set $\mathcal{K} = \{\theta \in \Theta, M' \le w(\theta) \le M\}$ to show that there exists λ_1, β_1 such that, for all ρ, $0 \le \rho \le \lambda_1$, ζ, $|\zeta| \le \beta_1$, and $\theta \in \mathcal{K}$, $w(\theta + \rho h(\theta) + \rho\zeta) \le w(\theta) \le M$, showing that $\theta + \rho h(\theta) + \rho\zeta \in \mathcal{W}_M$. □

2.1. Boundedness. In this section, we show that, under (A1) and mild additional conditions on $\{\xi_n\}$ and $\{\rho_n\}$, the sequence defined in (2.1) remains in a compact subset of Θ.

THEOREM 2.2. *Assume* (A1). *For any* $M \in (M_0, M_1]$ *there exist* $\delta_0 > 0$ *and* $\lambda_0 > 0$ *such that, for all* $n \ge 1$, *all* $\theta_0 \in \mathcal{W}_{M_0}$, *all sequences* $\{\rho_k\}$ *of nonnegative integers, and all sequences* $\{\xi_k\}$ *of* n_θ*-dimensional vectors satisfying*

$$\sup_{1\le k\le n} \rho_k \le \lambda_0 \quad \textit{and} \quad \sup_{1\le k\le n} \left|\sum_{j=1}^{k} \rho_j \xi_j\right| \le \delta_0,$$

we have for $k \in \{1, \dots, n\}$, $w(\theta_k) \le M$, *where* $\theta_k = \theta_{k-1} + \rho_k h(\theta_{k-1}) + \rho_k \xi_k$.

Proof. Let M' be such that $M' \in (M_0, M)$. Lemma 2.1 shows that there exists $\lambda_0 > 0$, $\beta_0 > 0$ such that, for all θ, ρ, and ζ satisfying $w(\theta) \le M'$, $0 \le \rho \le \lambda_0$, and $|\zeta| \le \beta_0$,

$$w(\theta + \rho h(\theta) + \rho\zeta) \le M'. \tag{2.2}$$

By continuity of h and w there exists $\delta_0 \in (0, \beta_0]$ such that for all $(\theta, \bar\theta) \in \Theta \times \Theta$ satisfying $w(\theta) \le M$ and $|\theta - \bar\theta| \le \delta_0$, we have

$$|h(\bar\theta) - h(\theta)| \le \beta_0 \quad \text{and} \quad |w(\bar\theta) - w(\theta)| \le M - M'. \tag{2.3}$$

We will now prove by induction that, for all $k \in \{1, \dots, n\}$, $w(\bar{\theta}_k) \leq M'$, and $w(\theta_k) \leq M$, where the sequence $\{\bar{\theta}_k\}$ is defined recursively as $\bar{\theta}_0 = \theta_0$ and for all $k \in \{1, \dots, n\}$,

$$\bar{\theta}_k = \bar{\theta}_{k-1} + \rho_k h(\theta_{k-1}).$$

Under the stated assumptions $w(\theta_0) = w(\bar{\theta}_0) \leq M_0$ and since $0 \leq \rho_1 \leq \lambda_0$ and $|\theta_1 - \bar{\theta}_1| = |\rho_1 \xi_1| \leq \delta_0$, on the one hand Lemma 2.1 shows that $w(\bar{\theta}_1) = w(\bar{\theta}_0 + \rho_1 h(\bar{\theta}_0)) \leq M'$ and on the other hand $w(\theta_1) = w(\theta_0 + \rho_1 h(\theta_0) + \rho_1 \xi_1) \leq M$, which proves the result for $n = 1$. Assume now that the result holds up to $1 \leq k \leq n-1$ for $n > 1$. By construction, for $j \in \{1, \dots, k\}$, $\theta_j - \bar{\theta}_j = \theta_{j-1} - \bar{\theta}_{j-1} + \rho_j \xi_j$, which implies that

$$\theta_j - \bar{\theta}_j = \sum_{i=1}^{j} \rho_i \xi_i. \tag{2.4}$$

Under the stated assumptions and (2.3), for $j \in \{1, \dots, k\}$, $|\theta_j - \bar{\theta}_j| \leq \delta_0$ and $|h(\theta_j) - h(\bar{\theta}_j)| \leq \beta_0$. On the other hand,

$$\bar{\theta}_{k+1} = \bar{\theta}_k + \rho_{k+1} h(\theta_k) = \bar{\theta}_k + \rho_{k+1} h(\bar{\theta}_k) + \rho_{k+1} \left(h(\theta_k) - h(\bar{\theta}_k) \right).$$

Since $0 \leq \rho_{k+1} \leq \lambda_0$ and $w(\bar{\theta}_k) \leq M'$, Lemma 2.1 shows that $w(\bar{\theta}_{k+1}) \leq M'$. Using again that $|\theta_{k+1} - \bar{\theta}_{k+1}| \leq \delta_0$, (2.3) implies that $w(\theta_{k+1}) \leq M$, which concludes the proof. □

2.2. Convergence. Theorem 2.2 provides us with conditions on $\{\xi_n\}$ and $\{\rho_n\}$ upon which a sequence as defined in (2.1) stays within a compact subset of Θ. In the next theorem we show that, whenever $\{\theta_k\}$ stays in a compact subset of Θ, under mild additional assumptions it converges to $\mathcal{L}$. The key result of this section is the following theorem, adapted here from [14, Theorem 2] (see [12] for a similar result). This theorem states that whenever $\{\theta_i\}$ stays in a compact subset of Θ, under mild additional assumptions it converges to $\mathcal{L}$. For an integer d and A a subset of $\mathbb{R}^d$, we define $d(x, A) = \inf\{y \in A, |x - y|\}$. For any set $A \subset \Theta$ and any $\delta > 0$, we define $A_\delta := \{\theta \in \Theta, d(\theta, A) \leq \delta\}$; for any function $\phi : \Theta \to \mathbb{R}$, we define $\|\phi\|_A := \sup_{\theta \in A} |\phi(\theta)|$.

THEOREM 2.3. *Assume* (A1). *Let $\mathcal{K}$ be a compact subset of Θ such that $\mathcal{L} \cap \mathcal{K} \neq \emptyset$. Let $\{\rho_k\}$ be a monotone nonincreasing sequence of positive numbers such that $\rho_0 \leq \lambda_0$ (where λ_0 is given in Theorem* 2.2),

$$\sum_{k=1}^{\infty} \rho_k = \infty, \quad \textit{and} \quad \lim_{k \to \infty} \rho_k = 0.$$

Let $\{\xi_n\}$ be a sequence in $\mathbb{R}^{n_\theta}$ satisfying $\limsup_{k \to \infty} \sup_{l \geq k} |\sum_{i=k}^{l} \rho_i \xi_i| = 0$. Assume that the sequence defined by $\theta_k = \theta_{k-1} + \rho_k h(\theta_{k-1}) + \rho_k \xi_k$ is such that $\{\theta_k\} \subset \mathcal{K}$. Then, $\limsup_{k \to \infty} d(\theta_k, \mathcal{L} \cap \mathcal{K}) = 0$.

We preface the proof of this theorem with two lemmas. Lemmas 2.4 and 2.5 are proved under the assumptions of Theorem 2.3.

LEMMA 2.4. *Let $\mathcal{N} \subset \Theta$ be an open neighborhood of $\mathcal{L} \cap \mathcal{K}$. There exist positive constants δ, ε, and λ (depending only on the sets $\mathcal{N}$ and $\mathcal{K}$) such that for any $\delta' \in (0, \delta]$, $\lambda' \in (0, \lambda]$, and $\eta > 0$, one can find an integer N and a sequence $\{\bar{\theta}_j\}_{j \geq N}$ satisfying*

$$\sup_{j \geq N} |\theta_j - \bar{\theta}_j| \leq \delta', \quad \sup_{j \geq N} \rho_j \leq \lambda', \quad \textit{and} \quad \sup_{j \geq N} |w(\theta_j) - w(\bar{\theta}_j)| \leq \eta, \tag{2.5}$$

$$w(\bar{\theta}_j) \leq w(\bar{\theta}_{j-1}) - \rho_j \varepsilon + (\eta + \rho_j \varepsilon) \, \mathbb{1}_{\mathcal{N}}(\bar{\theta}_{j-1}) \quad \textit{for} \quad j \geq N + 1. \tag{2.6}$$

Proof. Let us choose $\delta_0 > 0$ such that the compact set $\mathcal{K}_{\delta_0} \subset \Theta$. The set $\mathcal{K}_{\delta_0} \setminus \mathcal{N}$ is compact and $\sup_{\mathcal{K}_{\delta_0}\setminus\mathcal{N}} \langle \nabla w, h\rangle < 0$. By Lemma 2.1, for any $\varepsilon > 0$ such that $\sup_{\theta\in\mathcal{K}_{\delta_0}\setminus\mathcal{N}} \langle \nabla w(\theta), h(\theta)\rangle < -\varepsilon$, one may choose $\lambda > 0$ and $\beta > 0$ small enough so that for any $\rho \in [0, \lambda]$, $|\zeta| \le \beta$, and $\theta \in \mathcal{K}_{\delta_0} \setminus \mathcal{N}$,

$$w(\theta + \rho h(\theta) + \rho\zeta) \le w(\theta) - \rho\varepsilon. \tag{2.7}$$

Using the uniform continuity of continuous functions on compact sets, for any $\eta > 0$ one may choose $\delta \in (0, \lambda\|h\|_{\mathcal{K}}]$ small enough so that for all $(\theta, \bar\theta) \in \mathcal{K}_{\delta_0} \times \mathcal{K}_{\delta_0}$ satisfying $|\theta - \bar\theta| \le \delta \le \lambda\|h\|_{\mathcal{K}}$,

$$|h(\theta) - h(\bar\theta)| \le \beta \quad \text{and} \quad |w(\theta) - w(\bar\theta)| \le \eta. \tag{2.8}$$

Under the stated conditions for all $\delta' \in (0, \delta]$ and $\lambda' \in (0, \lambda]$ there exists an integer N such that for any $n \ge N + 1$, $\rho_n \le \lambda'$ and $\left|\sum_{k=N+1}^{n} \rho_k\xi_k\right| \le \delta'$. Define recursively for $j \ge N$ the sequence $\{\bar\theta_j\}_{j\ge N}$ as $\bar\theta_N := \theta_N$ and for $j \ge N + 1$,

$$\bar\theta_j = \bar\theta_{j-1} + \rho_j h(\theta_{j-1}). \tag{2.9}$$

By construction, for $j \ge N+1$, $\bar\theta_j - \theta_j = \sum_{i=N+1}^{j} \rho_i\xi_i$ which implies that $\sup_{j\ge N} |\bar\theta_j - \theta_j| \le \delta'$. On the other hand, for $j \ge N + 1$,

$$\bar\theta_j = \bar\theta_{j-1} + \rho_j h(\bar\theta_{j-1}) + \rho_j \left(h(\theta_{j-1}) - h(\bar\theta_{j-1})\right), \tag{2.10}$$

and since $|\bar\theta_{j-1} - \theta_{j-1}| \le \delta' \le \delta$, (2.8) shows that $|h(\theta_{j-1}) - h(\bar\theta_{j-1})| \le \beta$. Thus, (2.7) implies that, whenever $\bar\theta_{j-1} \in \mathcal{K}_\delta \setminus \mathcal{N}$, $w(\bar\theta_j) \le w(\bar\theta_{j-1}) - \rho_j\varepsilon$. Now (2.8) implies that $|w(\bar\theta_j) - w(\bar\theta_{j-1})| \le \eta$ for any $\bar\theta_{j-1} \in \mathcal{K}_\delta$ and $|w(\theta_j) - w(\bar\theta_j)| \le \eta$ for any $\theta_j \in \mathcal{K}$, which concludes the proof. □

LEMMA 2.5. *Let ε be real constants, n be an integer, and let $-\infty < a_1 < b_1 < \cdots < a_n < b_n < \infty$ be real numbers. Let $\{u_j\}$ be a bounded real sequence such that, for any $\eta > 0$, there exists an integer J such that for all $j \ge J$,*

$$u_j \le u_{j-1} - \rho_j\varepsilon + (\eta + \rho_j\varepsilon)\mathbb{1}_A(u_{j-1}) \quad A = \bigcup_{i=1}^{n} [a_i, b_i]. \tag{2.11}$$

Then, the limit points of the sequence $\{u_j\}$ are included in A.

Proof. As $\{u_j\}$ is bounded, it has at least one limit point from the Bolzano–Weierstrass theorem. Let us denote by $\check a$ one of these limit points; since $\{u_j\}$ is bounded and satisfies (2.11), $\check a \ge a_1$. Now let us proceed by contradiction and assume that there exists $l \in \{1, 2, \ldots, n\}$ such that $\check a \in (b_l, a_{l+1})$, with the convention that $a_{n+1} = \infty$. For any $\epsilon > 0$ sufficiently small $[\check a - \epsilon, \check a + \epsilon] \subset A^c$. Now, for any integer j and any set $B \subset \mathbb{R}$, we define

$$\tau_B(j) = \inf\{k \ge j : u_k \in B\},$$

with the convention $\inf\emptyset = \infty$. Since $\sum_{k=1}^{\infty} \rho_k = \infty$ and $\{u_k\}_{k\ge0}$ is bounded, (2.11) implies that for any $\eta > 0$ and $j \ge J$, $\sigma(j) := \tau_A(j) < \infty$. Note also that for $k = j, \ldots, \sigma(j)$, $u_k \le u_j$. Since $\check a \in (b_l, a_{l+1})$ is a limit point, for any integer j, $\kappa(j) := \tau_{(b_l,\infty)}(j) < \infty$. Let $\eta > 0$ be such that, for any $j \ge J$, $0 < \eta < (\check a - \epsilon - b_l)/2$. Then for $j \ge J$, $u_{\kappa[\sigma(j)]} < (\check a - \epsilon + b_l)/2$ and for $k = \kappa[\sigma(j)], \ldots, \kappa(\sigma(\kappa[\sigma(j)])) - 1$, $u_k \le u_{\kappa[\sigma(j)]}$, which implies that for any $i \ge \kappa(\sigma(J))$, $u_i \le (\check a - \epsilon + b_l)/2$, which

contradicts the fact that $\check{a}$ is a limit point. Now using the same type of argument, one can show that if an accumulation point $\check{a} \in [a_k, b_k]$ for some $k \in 1, \ldots, n-1$, then there cannot be any accumulation point in $[a_l, b_l]$ for $n \geq l > k$. As a consequence there cannot be an accumulation point in an interval other than $[a_k, b_k]$. □

Proof of Theorem 2.3. We first prove that $\lim_{j\to\infty} w(\theta_j)$ exists. For any $\alpha > 0$, define the set $[w(\mathcal{L} \cap \mathcal{K})]_\alpha := \{x \in \mathbb{R} : d(x, w(\mathcal{L} \cap \mathcal{K})) \leq \alpha\}$. Since $\|w\|_{\mathcal{K}} < \infty$, $[w(\mathcal{L} \cap \mathcal{K})]_\alpha$ is a finite union of disjoint intervals of length at least equal to 2α. By Lemma 2.4, there exist positive constants δ, ε, λ such that for any $\delta' \in (0, \delta]$, $\lambda' \in (0, \lambda]$, and $\eta > 0$, one may find an integer N and a sequence $\{\bar{\theta}_j\}_{j\geq N}$ such that

$$\sup_{j\geq N} |\theta_j - \bar{\theta}_j| \leq \delta' \quad \text{and} \quad \sup_{j\geq N} |w(\theta_j) - w(\bar{\theta}_j)| \leq \eta$$

and

$$w(\bar{\theta}_j) \leq w(\bar{\theta}_{j-1}) - \rho_j\varepsilon + (\eta + \rho_j\varepsilon)\mathbb{1}_{[w(\mathcal{L}\cap\mathcal{K})]_\alpha}(w(\bar{\theta}_{j-1})) \quad \text{for any} \quad j \geq N+1,$$

where we have chosen $\mathcal{N} = w^{-1}(\text{int}([w(\mathcal{L}\cap\mathcal{K})]_\alpha))$ and used $\mathbb{1}_{\mathcal{N}}(\theta) \leq \mathbb{1}_{[w(\mathcal{L}\cap\mathcal{K})]_\alpha}(w(\theta))$. By Lemma 2.5, the limit points of the sequence $\{w(\bar{\theta}_j)\}$ are in $[w(\mathcal{L}\cap\mathcal{K})]_\alpha$ and, since $\sup_{j\geq N} |\theta_j - \bar{\theta}_j| \leq \delta'$, the limit points of the sequence $\{w(\theta_j)\}_{j\geq 0}$ are in $[w(\mathcal{L}\cap\mathcal{K})]_{\alpha'}$ for $\alpha' = \alpha + \eta$. Since α and η can be chosen arbitrarily small, this implies that the limit points of the sequence $\{w(\theta_j)\}_{j\geq 0}$ are included in $\bigcap_{\alpha>0}[w(\mathcal{L}\cap\mathcal{K})]_\alpha$. Because $\mathcal{L}\cap\mathcal{K}$ is a compact subset of $\mathbb{R}^{n_\theta}$ and w is continuous, $w(\mathcal{L}\cap\mathcal{K})$ is a compact subset of $\mathbb{R}$, which implies that $w(\mathcal{L}\cap\mathcal{K}) = \bigcap_{\alpha>0}[w(\mathcal{L}\cap\mathcal{K})]_\alpha$. Thus, the limit points of $\{w(\theta_j)\}$ belong to the set $w(\mathcal{L}\cap\mathcal{K})$.

On the other hand, $\limsup_{j\to\infty} |w(\theta_j) - w(\theta_{j-1})| = 0$, which implies that the set of limit points of $\{w(\theta_j)\}$ is an interval. Because $w(\mathcal{L})$ has an empty interior, the only intervals included in $w(\mathcal{L}\cap\mathcal{K})$ are isolated points, which shows that the limit $\lim_{j\to\infty} w(\theta_j)$ exists.

We now prove that $\limsup_{j\to\infty} d(\theta_j, \mathcal{L}\cap\mathcal{K}) = 0$. Let $\mathcal{N} \subset \mathcal{K}$ be an arbitrary neighborhood of $\mathcal{L}\cap\mathcal{K}$. From Lemma 2.4 there exist constants $\delta > 0$, $\varepsilon > 0$, $\lambda > 0$ such that for any $\delta' \in (0, \delta]$, $\lambda' \in (0, \lambda]$, and $\eta > 0$ one may find an integer N and a sequence $\{\bar{\theta}_j\}_{j\geq N}$ such that

$$\sup_{j\geq N} |\theta_j - \bar{\theta}_j| \leq \delta', \quad \sup_{j\geq N} \rho_j \leq \lambda', \quad \text{and} \quad \sup_{j\geq N} |w(\theta_j) - w(\bar{\theta}_j)| \leq \eta$$

and

$$w(\bar{\theta}_j) \leq w(\bar{\theta}_{j-1}) - \rho_j\varepsilon + (\eta + \rho_j\varepsilon)\mathbb{1}_{\mathcal{N}}(\bar{\theta}_{j-1}) \quad \text{for any} \quad j \geq N+1.$$

For $j \geq N$, define $\tau(j) := \inf\{k \geq 0, \bar{\theta}_{k+j} \in \mathcal{N}\}$. For any integer p, define $\tau^p(j) := \tau(j) \wedge p$, where $a \wedge b = \min(a, b)$. We have

$$w(\bar{\theta}_{j+\tau^p(j)}) - w(\bar{\theta}_j) = \sum_{i=j+1}^{j+\tau^p(j)} \{w(\bar{\theta}_i) - w(\bar{\theta}_{i-1})\} \leq -\varepsilon \sum_{i=j+1}^{j+\tau^p(j)} \rho_i, \tag{2.12}$$

with the convention that, for any sequence $\{a_i\}$ and any integer l, $\sum_{i=l+1}^{l} a_i = 0$. Therefore,

$$\begin{aligned} w(\theta_{j+\tau^p(j)}) - w(\theta_j) &= w(\theta_{j+\tau^p(j)}) - w(\bar{\theta}_{j+\tau^p(j)}) + w(\bar{\theta}_{j+\tau^p(j)}) - w(\bar{\theta}_j) + w(\bar{\theta}_j) - w(\theta_j) \\ &\leq 2\eta - \varepsilon \sum_{i=j+1}^{j+\tau^p(j)} \rho_i. \end{aligned}$$

Since $\{w(\theta_j)\}$ converges, for any $\varepsilon' > 0$ there exists $N' > N$ such that, for all $j \geq N'$,

$$-\varepsilon' < w(\theta_{j+\tau^p(j)}) - w(\theta_j) \leq 2\eta - \varepsilon \sum_{i=j+1}^{j+\tau^p(j)} \rho_i. \tag{2.13}$$

This implies that, for all $j \geq N'$ and all integer $p \geq 0$,

$$\sum_{i=j+1}^{j+\tau^p(j)} \rho_i \leq C(\varepsilon', \eta) := \varepsilon^{-1} (\varepsilon' + 2\eta). \tag{2.14}$$

Since $\sum_{i=j+1}^{j+\tau(j)} \rho_i = \lim_{p\to\infty} \sum_{i=j+1}^{j+\tau^p(j)} \rho_i$ and $\sum_{i=1}^{\infty} \rho_i = \infty$, the previous relation implies that, for all $j \geq N'$, $\tau(j) < \infty$, and $\sum_{i=j+1}^{j+\tau(j)} \rho_i \leq C(\varepsilon', \eta)$. For any integer p, $\theta_{j+p} - \theta_j = \sum_{i=j+1}^{j+p} \rho_i h(\theta_{i-1}) + \sum_{i=j+1}^{j+p} \rho_i \xi_i$, which implies that

$$|\theta_{j+p} - \theta_j| \leq \|h\|_{\mathcal{K}} \sum_{i=j+1}^{j+p} \rho_i + \left| \sum_{i=j+1}^{j+p} \rho_i \xi_i \right|.$$

Applying this inequality for $j \geq N'$ and $p = \tau(j)$ and using that, by definition, $\bar{\theta}_{j+\tau(j)} \in \mathcal{N}$,

$$d(\theta_j, \mathcal{N}) \leq |\bar{\theta}_{j+\tau(j)} - \theta_{j+\tau(j)}| + |\theta_{j+\tau(j)} - \theta_j| \leq \delta' + \|h\|_{\mathcal{K}}\, C(\varepsilon', \eta) + \left| \sum_{i=j+1}^{j+\tau(j)} \rho_i \xi_i \right|.$$

Since η, δ', and ε' can be chosen arbitrarily small, and $\limsup_{k\to\infty} \sup_{l\geq k} |\sum_{i=k}^{l} \rho_i \xi_i| = 0$, the latter inequality shows that $\lim_{j\to\infty} d(\theta_j, \mathcal{N}) = 0$. Since $\mathcal{N}$ is arbitrary, we thus have $\lim_{j\to\infty} d(\theta_j, \mathcal{L} \cap \mathcal{K}) = 0$. □

Note that the boundedness is here one of the required assumptions. It is therefore natural to try to apply Theorem 2.2. This is what motivates the next section, where we describe a modification of the stochastic approximation algorithm, which ensures that the conditions of Theorem 2.2 are satisfied. We consider here the Markov state-dependent noise, as it covers many applications of interest, encompasses the exogenous scenario and, as we shall see, leads to general and verifiable conditions.

3. Markov state-dependent noise. In this section, we describe our stochastic approximation procedure with adaptive truncation sets and introduce the relevant notation required in the Markovian state-dependent noise scenario (see [23, section 6.6, p. 159] for a detailed description and numerous examples). We first introduce a version without truncations of the algorithm in this setting (subsection 3.2) and describe our adaptive procedure in terms of this plain algorithm in subsection 3.1. This will prove extremely useful when proving that our procedure is stable in section 4 and particularly in section 5.

It is assumed hereafter that the state-space X and the parameter space Θ are equipped with a countably generated σ-field, $\mathcal{B}(\mathsf{X})$, and $\mathcal{B}(\Theta)$ (and measurability will always be defined w.r.t. these σ-fields).

3.1. Nonhomogeneous chain. Let $\boldsymbol{\rho} = \{\rho_n\}$ be a monotone nonincreasing sequence with $\rho_0 \leq 1$, define the product space $\bar{\mathsf{X}} := \mathsf{X} \cup \{x_c\} \times \bar{\Theta} := \Theta \cup \{\theta_c\}$, where $\theta_c \notin \Theta$ and $x_c \notin \mathsf{X}$ are two arbitrary cemetery points, and define the *nonhomogeneous*

Markov chain $\{Y_n^{\boldsymbol{\rho}} := (X_n, \theta_n)\}$ on $\bar{\mathsf{X}} \times \bar{\Theta}$ as follows. Set $\theta_0 = \theta \in \Theta$, $X_0 = x \in \mathsf{X}$, and for $n \geq 0$,

$$\theta_{n+1} = \begin{cases} \theta_n + \rho_{n+1} H(\theta_n, X_{n+1}) & \text{and} \quad X_{n+1} \sim P_{\theta_n}(X_n, \cdot) \quad \text{if} \quad \theta_n \in \Theta, \\ \theta_c & \text{and} \quad X_{n+1} = x_c \quad \text{if} \quad \theta_n \notin \Theta, \end{cases} \tag{3.1}$$

where it is assumed that the family of Markov transition probabilities $\{P_\theta, \theta \in \Theta\}$ and the field H satisfy the following conditions:

(A2) For any $\theta \in \Theta$, the Markov kernel P_θ has a single stationary distribution π_θ, $\pi_\theta P_\theta = \pi_\theta$. In addition $H : \Theta \times \mathsf{X} \to \Theta$ is measurable for all $\theta \in \Theta$, $\int_{\mathsf{X}} |H(\theta, x)| \pi_\theta(dx) < \infty$.

The existence and uniqueness of the invariant distribution can be guaranteed under classical irreducibility and recurrence conditions (see, e.g., [25, Chapters 9, 10]). We denote by $h(\theta) := \int_{\mathsf{X}} H(\theta, x) \pi_\theta(dx)$ the mean-field associated to this stochastic approximation procedure and define the noise sequence $\{\xi_n = H(\theta_{n-1}, X_n) - h(\theta_{n-1})\}$. Following [5], we will often write $H_\theta(x)$ as an equivalent expression for $H(\theta, x)$, h_θ for $h(\theta)$, etc.

We denote by $\mathcal{F} = \{\mathcal{F}_n, n \geq 0\}$ the natural filtration of this Markov chain, with $\mathcal{F}_n := \sigma((X_l, \theta_l), l \in \{0, \ldots, n\})$ and $\mathbb{P}_{x,\theta}^{\boldsymbol{\rho}}$ the probability measure on the canonical space $\left((\mathsf{X} \times \Theta)^{\mathbb{N}}, (\mathcal{B}(\mathsf{X}) \otimes \mathcal{B}(\Theta))^{\otimes \mathbb{N}}\right)$ generated by the nonhomogeneous Markov chain $\{Y_n^{\boldsymbol{\rho}}\}$ started from the initial conditions $(X_0, \theta_0) = (x, \theta) \in \mathsf{X} \times \Theta$ and using the sequence $\boldsymbol{\rho}$. Finally, it will be useful in what follows to introduce $\{Q_{\rho_n}\}$, the sequence of transition probabilities that generates the inhomogeneous Markov chain $\{Y_n^{\boldsymbol{\rho}}\}$, where for $\rho \geq 0$, Q_ρ is defined for any $(x, \theta) \in \mathsf{X} \times \Theta$, $A \in \mathcal{B}(\bar{\mathsf{X}})$, and $B \in \mathcal{B}(\bar{\Theta})$,

$$\begin{aligned} Q_\rho(x, \theta; A \times B) = & \int_A P_\theta(x, dy) \mathbb{1}\{\theta + \rho H(\theta, y) \in B\} \\ & + \delta_{\theta_c}(B) \int_A P_\theta(x, dy) \mathbb{1}\{\theta + \rho H(\theta, y) \notin \Theta\} \, . \end{aligned}$$

3.2. Homogeneous chain. Let $\{\mathcal{K}_q, q \geq 0\}$ be a sequence of compact subsets of Θ such that

$$\bigcup_{q \geq 0} \mathcal{K}_q = \Theta, \quad \text{and} \quad \mathcal{K}_q \subset \text{int}(\mathcal{K}_{q+1}), \quad q \geq 0, \tag{3.2}$$

where $\text{int}(A)$ denotes the interior of set A. Let $\boldsymbol{\gamma} = \{\gamma_k\}$ and $\boldsymbol{\epsilon} = \{\epsilon_k\}$ be two monotone nonincreasing sequences of positive numbers and let K be a subset of X. Let $\Phi : \mathsf{X} \times \Theta \to \mathsf{K} \times \mathcal{K}_0$ be a measurable function and $\phi : \mathbb{Z}^+ \to \mathbb{Z}$ be a function such that $\phi(k) > -k$ for any k. Our stochastic approximation algorithm with adaptive truncation sets is defined as a *homogeneous* Markov chain on $\mathsf{Z} := \mathsf{X} \times \Theta \times \mathbb{N} \times \mathbb{N} \times \mathbb{N}$,

$$\{Z_n := (X_n, \theta_n, \kappa_n, \varsigma_n, \nu_n)\} \in \mathsf{Z}^{\mathbb{N}}, \tag{3.3}$$

with the following transition at iteration $n + 1$:

- If $\nu_n = 0$, then draw $(X_{n+1}, \theta_{n+1}) \sim Q_{\gamma_{\varsigma_n}}(\Phi(X_n, \theta_n); \cdot)$; otherwise draw $(X_{n+1}, \theta_{n+1}) \sim Q_{\gamma_{\varsigma_n}}(X_n, \theta_n; \cdot)$.
- If $|\theta_{n+1} - \theta_n| \leq \epsilon_{\varsigma_n}$ and $\theta_{n+1} \in \mathcal{K}_{\kappa_n}$, then set $\kappa_{n+1} = \kappa_n$, $\varsigma_{n+1} = \varsigma_n + 1$, and $\nu_{n+1} = \nu_n + 1$; otherwise, set $\nu_{n+1} = 0$, $\kappa_{n+1} = \kappa_n + 1$, $\varsigma_{n+1} = \varsigma_n + \phi(\nu_n)$.

In other words, κ, ς, and ν are counters: κ is the index of the current active truncation set; ν counts the number of iterations since the last reinitialization; ς is the current

index in the sequences $\{\gamma_n\}$ and $\{\epsilon_n\}$, and therefore defines the current proposal kernel Q_γ. The event $\{\nu_n = 0\}$ means that a reinitialization occurs and the condition on ϕ ensures that the algorithm is reinitialized with a value for γ_{ς_n} smaller than that used the last time such an event occurred. This algorithm is reminiscent of the algorithm with adaptive truncation sets proposed in [11], [10]. When the current iterate wanders outside the active truncation set or when the difference between two successive values of the parameter is larger than a time-dependent threshold, then the algorithm is reinitialized with a smaller initial value of the stepsize and a larger truncation set. Various choices for the function ϕ can be considered. For example, the choice $\phi(k) = 1$ for all $k \in \mathbb{N}$ coincides with the procedure proposed in [10]: in this case $\varsigma_n = n$. Another sensible choice consists of setting $\phi(k) = 1 - k$ for all $k \in \mathbb{N}$, in which case the number of iterations between two successive reinitializations is not taken into account.

The intuitive motivation for this modification of the original stochastic approximation recursion lies in Theorem 2.2. Indeed, in order to ensure the stability of the algorithm it is required that the stepsizes not be too large and that the average effect of the noise be small in order for the drift $h(\theta)$ to dominate, and confine the recursion to a compact set. The reprojections act as a drastic drift toward the center of Θ when $\{\theta_n\}$ grows too rapidly and allow one to reinitialize the algorithm with a smaller stepsize and weaker noise inside a "ring" of the type $\{\theta \in \Theta : w(\theta) \in (M_0, M_1]\}$ (M_0 and M_1 are defined in (A1)) where the drift is strictly positive. The fact that M_0 and M_1 are unknown a priori is the reason for the adaptive truncations, which ensure that one eventually selects $\mathcal{K}_q$ large enough in order to have $\mathcal{L} \cap \mathcal{K}_q \neq \varnothing$. As we shall see, the limitation imposed on the increments of the sequence $\{\theta_n\}$ is required in order to ensure some type of homogeneity of the chain $\{\xi_n\}$, and therefore ergodicity properties of the noise sequence $\{\xi_n\}$.

In light of this heuristic, one can naturally propose many variations on this theme. We suggest here two possible extensions. First, one can suggest other strategies in order to adapt the magnitude of the stepsizes. Let $\{\gamma_{n,l}, n \geq 0, l \geq 0\}$ be an array of stepsizes. Then, when a reprojection occurs, instead of jumping forward in a unique sequence of stepsizes, it is possible to simply change the sequence of stepsizes from, say, l to $l + 1$. Another interesting variant of the proposed scheme consists of reinitializing the algorithm when $|\theta_n - \theta_{n-1}| > \epsilon_{\varsigma_{n-1}}$ without changing the truncation set. In either case the proof of convergence follows using the same types of arguments as those presented in this paper.

We now introduce some further notation and briefly state our main result. For μ a probability on Z, we denote $\bar{\mathbb{P}}_\mu$ (resp., $\bar{\mathbb{E}}_\mu$) the probability (resp., the expectation) on the canonical space $(\mathsf{Z}^{\mathbb{N}}, \mathcal{B}(\mathsf{Z})^{\otimes\mathbb{N}})$ associated to the Markov chain $\{Z_n\}$ with initial distribution μ. For $z \in \mathsf{Z}$ we set $\bar{\mathbb{P}}_z := \bar{\mathbb{P}}_{\delta_z}$, $\bar{\mathbb{E}}_z := \bar{\mathbb{E}}_{\delta_z}$ and for $(x, \theta) \in \mathsf{X} \times \Theta$,

$$\bar{\mathbb{P}}_{x,\theta} := \bar{\mathbb{P}}_{x,\theta,0,0,0} \quad \text{and} \quad \bar{\mathbb{E}}_{x,\theta} := \bar{\mathbb{E}}_{x,\theta,0,0,0}. \tag{3.4}$$

This probability measure depends upon the deterministic sequences $\boldsymbol{\gamma} = \{\gamma_n\}$ and $\boldsymbol{\epsilon} = \{\epsilon_n\}$; this will be implicit hereafter in order to alleviate notation. We define recursively $\{T_n, n \geq 0\}$ the sequence of successive reinitialization times

$$T_{n+1} = \inf\{k \geq T_n + 1,\ \nu_k = 0\}, \quad \text{with} \quad T_0 = 0, \tag{3.5}$$

where, by convention, $\inf\{\emptyset\} = \infty$. The following results hold under (A1), some regularity conditions on the family of transition probabilities $\{P_\theta, \theta \in \Theta\}$, and conditions

on the sequences $\boldsymbol{\gamma}$ and $\boldsymbol{\epsilon}$:

$$\inf_{(x,\theta)\in \mathsf{K}\times\mathcal{K}_0} \bar{\mathbb{P}}_{x,\theta}\left(\sup_{n\geq 0}\kappa_n<\infty\right) = \inf_{(x,\theta)\in \mathsf{K}\times\mathcal{K}_0} \bar{\mathbb{P}}_{x,\theta}\left(\bigcup_{n=0}^{\infty}\{T_n=\infty\}\right)=1;$$

i.e., the number of reinitializations of the procedure described above is finite $\bar{\mathbb{P}}_{x,\theta}$-a.s. for every $(x,\theta)\in \mathsf{K}\times\mathcal{K}_0$. Convergence will then follow using Theorem 2.3, for example.

4. Bound on $\bar{\mathbb{P}}_{x,\theta}(T_n<\infty)$. In this section we establish in Proposition 4.2 a bound on $\bar{\mathbb{P}}_{x,\theta}(T_n<\infty)$ in terms of the fluctuations of the noise sequence of the algorithm between successive reinitializations. Let $\mathcal{K}$ be a compact subset of Θ and let $\boldsymbol{\epsilon}=\{\epsilon_n\}$ be a nonincreasing sequence of positive numbers. We introduce

$$\sigma(\mathcal{K},\epsilon)=\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\epsilon}),\quad \sigma(\mathcal{K})=\inf\{k\geq 1,\theta_k\notin\mathcal{K}\},\quad \nu(\boldsymbol{\epsilon})=\inf\{k\geq 1,|\theta_k-\theta_{k-1}|\geq\epsilon_k\},$$

and for a sequence $\mathbf{a}=\{a_k\}$ and an integer l, we define $\mathbf{a}^{\leftarrow l}=\{a_k^{\leftarrow l}\}$ as $a_k^{\leftarrow l}=a_{k+l}$. We now prove the following lemma, which relates the expectation of the homogeneous Markov chain $\{Z_n\}$, defined in subsection 3.2, to the expectation of a nonhomogeneous Markov chain $\{Y_n^{\boldsymbol{\rho}}\}$, defined in subsection 3.1, for a particular $\boldsymbol{\rho}$.

LEMMA 4.1. *For any $m\geq 1$, for any nonnegative measurable function $\Psi_m:(\mathsf{X}\times\Theta)^m\to\mathbb{R}^+$, for any integers p and q, for any $x,\theta\in\mathsf{X}\times\Theta$,*

$$\begin{aligned}\bar{\mathbb{E}}_{x,\theta,p,q,0}\left[\Psi_m(X_1,\theta_1,\ldots,X_m,\theta_m)\mathbb{1}_{\{T_1\geq m\}}\right]\\ =\mathbb{E}^{\gamma^{\leftarrow q}}_{\Phi(x,\theta)}\left[\Psi_m(X_1,\theta_1,\ldots,X_m,\theta_m)\mathbb{1}_{\{\sigma(\mathcal{K}_p,\epsilon^{\leftarrow q})\geq m\}}\right].\end{aligned}\tag{4.1}$$

Proof. We proceed by induction. Let $\Psi_1:\mathsf{X}\times\Theta\to\mathbb{R}^+$. We notice that $\mathbb{1}_{\{T_1\geq 1\}}=1$. This, combined with the definition of $\bar{\mathbb{E}}_{x,\theta,p,q,0}$, leads to

$$\bar{\mathbb{E}}_{x,\theta,p,q,0}[\Psi_1(X_1,\theta_1)\mathbb{1}_{\{T_1\geq 1\}}]=Q_{\gamma_q}(\Phi(x,\theta);\Psi_1),\tag{4.2}$$

and by definition,

$$\mathbb{E}^{\gamma^{\leftarrow q}}_{\Phi(x,\theta)}[\Psi_1(X_1,\theta_1)\mathbb{1}_{\{\sigma(\mathcal{K}_p,\epsilon^{\leftarrow q})\geq 1\}}]=Q_{\gamma_q}(\Phi(x,\theta);\Psi_1).\tag{4.3}$$

Now assume that the property is true for some $m\geq 1$. It is sufficient to prove the induction for functions Ψ_{m+1} of the form

$$\Psi_{m+1}(x_1,\theta_1,\ldots,x_{m+1},\theta_{m+1})=\psi_{m+1}(x_{m+1},\theta_{m+1})\Psi_m(x_1,\theta_1,\ldots,x_m,\theta_m),\tag{4.4}$$

with $\psi_{m+1}:\mathsf{X}\times\Theta\to\mathbb{R}^+$. In order to alleviate notation we will often write Ψ_m (resp., ψ_m) for $\Psi_m(x_1,\theta_1,\ldots,x_m,\theta_m)$ (resp., $\psi_m(x_m,\theta_m)$) in what follows. Consider

$$\begin{aligned}\bar{\mathbb{E}}_{x,\theta,p,q,0}[\Psi_{m+1}(X_1,\theta_1,\ldots,X_{m+1},\theta_{m+1})\mathbb{1}_{\{T_1\geq m+1\}}]\\ =\bar{\mathbb{E}}_{x,\theta,p,q,0}[\psi_{m+1}\Psi_m\mathbb{1}_{\{T_1\geq m\}}\mathbb{1}_{\{\theta_m\in\mathcal{K}_{\kappa_m}\}}\mathbb{1}_{\{|\theta_m-\theta_{m-1}|<\epsilon_{\varsigma_m}\}}].\end{aligned}\tag{4.5}$$

Now, by definition of the stopping time T_1, we have

$$\begin{aligned}\mathbb{1}_{\{\theta_m\in\mathcal{K}_{\kappa_m}\}}\mathbb{1}_{\{|\theta_m-\theta_{m-1}|<\epsilon_{\varsigma_m}\}}\mathbb{1}_{\{T_1\geq m\}}&=\mathbb{1}_{\{\theta_m\in\mathcal{K}_{\kappa_0}\}}\mathbb{1}_{\{|\theta_m-\theta_{m-1}|<\epsilon_{\varsigma_0+m}\}}\mathbb{1}_{\{T_1\geq m\}}\\ &=\mathbb{1}_{\{\theta_m\in\mathcal{K}_{\kappa_0}\}}\mathbb{1}_{\{|\theta_m-\theta_{m-1}|<\epsilon_{\varsigma_0+m}\}}\mathbb{1}_{\{\sigma(\mathcal{K}_{\kappa_0},\epsilon^{\leftarrow\varsigma_0})\geq m\}},\end{aligned}$$

from which we may deduce, using the induction assumption, that

$$
\begin{aligned}
&\bar{\mathbb{E}}_{x,\theta,p,q,0}[\Psi_{m+1}\mathbb{1}_{\{T_1\geq m+1\}}] \\
&\quad= \bar{\mathbb{E}}_{x,\theta,p,q,0}\left[\bar{\mathbb{E}}_{x,\theta,p,q,0}[\psi_{m+1}|X_m,\theta_m,\kappa_m,\varsigma_m,\nu_m]\right. \\
&\qquad\qquad \left.\cdot\mathbb{1}_{\{\theta_m\in\mathcal{K}_{\kappa_m}\}}\mathbb{1}_{\{|\theta_m-\theta_{m-1}|<\epsilon_{\varsigma_m}\}}\Psi_m\mathbb{1}_{\{T_1\geq m\}}\right] \\
&\quad= \bar{\mathbb{E}}_{x,\theta,p,q,0}\left[Q_{\gamma_{q+m+1}}(X_m,\theta_m;\psi_{m+1})\mathbb{1}_{\{\theta_m\in\mathcal{K}_p\}}\mathbb{1}_{\{|\theta_m-\theta_{m-1}|<\epsilon_{q+m}\}}\Psi_m\mathbb{1}_{\{T_1\geq m\}}\right] \\
&\quad= \mathbb{E}^{\gamma^{\leftarrow q}}_{\Phi(x,\theta)}\left[Q_{\gamma_{q+m+1}}(X_m,\theta_m;\psi_{m+1})\mathbb{1}_{\{\theta_m\in\mathcal{K}_p\}}\mathbb{1}_{\{|\theta_m-\theta_{m-1}|<\epsilon_{q+m}\}}\Psi_m\mathbb{1}_{\{\sigma(\mathcal{K}_p,\epsilon^{\leftarrow q})\geq m\}}\right] \\
&\quad= \mathbb{E}^{\gamma^{\leftarrow q}}_{\Phi(x,\theta)}\left[Q_{\gamma_{q+m+1}}(X_m,\theta_m;\psi_{m+1})\Psi_m\mathbb{1}_{\{\sigma(\mathcal{K}_p,\epsilon^{\leftarrow q})\geq m+1\}}\right] \\
&\quad= \mathbb{E}^{\gamma^{\leftarrow q}}_{\Phi(x,\theta)}\left[\Psi_{m+1}\mathbb{1}_{\{\sigma(\mathcal{K}_p,\epsilon^{\leftarrow q})\geq m+1\}}\right],
\end{aligned}
$$

which concludes the proof. □

Define, for any compact set $\mathcal{K}\subset\Theta$, $\boldsymbol{\epsilon}=\{\epsilon_k\}$, $\boldsymbol{\rho}=\{\rho_k\}$, and $1\leq l\leq n$, the partial sum

$$
S_{l,n}(\boldsymbol{\epsilon},\boldsymbol{\rho},\mathcal{K}) := \mathbb{1}_{\{\sigma(\mathcal{K},\boldsymbol{\epsilon})\geq n\}}\sum_{k=l}^{n}\rho_k(H(\theta_{k-1},X_k)-h(\theta_{k-1})), \tag{4.6}
$$

and for any $\delta\geq 0$ and any $M\in(M_0,M_1]$,

$$
A(\delta,\boldsymbol{\epsilon},M,\boldsymbol{\rho}) := \sup_{\theta\in\mathcal{K}_0}\sup_{x\in\mathsf{K}}\left\{\mathbb{P}^{\boldsymbol{\rho}}_{\Phi(x,\theta)}\left[\sup_{k\geq 1}|S_{1,k}(\boldsymbol{\epsilon},\boldsymbol{\rho},\mathcal{W}_M)|>\delta\right]+\mathbb{P}^{\boldsymbol{\rho}}_{\Phi(x,\theta)}[\nu(\boldsymbol{\epsilon})<\sigma(\mathcal{W}_M)]\right\}, \tag{4.7}
$$

where $\mathcal{K}_0$ is defined in (3.2) and $\mathcal{W}_M$, M_0, and M_1 are defined in (A1).

PROPOSITION 4.2. *Assume* (A1) *and that* $\mathcal{K}_0\subset\mathcal{W}_{M_0}$ (*where* M_0 *is defined in* (A1)). *Then for any* $M\in(M_0,M_1]$ *there exist an integer* n_0 *and a constant* $\delta_0>0$ *such that, for any* $n>n_0$, *we have*

$$
\sup_{(x,\theta)\in\mathsf{K}\times\mathcal{K}_0}\bar{\mathbb{P}}_{x,\theta}[T_n<\infty]\leq\prod_{l=n_0}^{n-1}\sup_{q\geq l}A(\delta_0,\boldsymbol{\epsilon}^{\leftarrow q},M,\boldsymbol{\gamma}^{\leftarrow q}),
$$

where T_n *is defined in* (3.5).

Proof. By Theorem 2.2, for any $M\in(M_0,M_1]$ there exist constants $\delta_0>0$ and $\lambda_0>0$ such that, for all $\theta_0\in\mathcal{W}_{M_0}$ (where M_0 is defined in (A1)), all integer $m\geq 1$, all sequences $\{\rho_k\}$ of nonnegative real numbers, and all sequences $\{\xi_k\}$ of n_θ-dimensional vectors satisfying $\sup_{1\leq k\leq m}\rho_k\leq\lambda_0$ and $\sup_{1\leq k\leq m}\left|\sum_{j=1}^{k}\rho_j\xi_j\right|\leq\delta_0$, we have $\sup_{1\leq k\leq m}w(\theta_k)\leq M$, where $\theta_k=\theta_{k-1}+\rho_k h(\theta_k)+\rho_k\xi_k$.

Now, choose n_0 such that $\mathcal{W}_M\subset\mathcal{K}_{n_0}$ and $\gamma_{n_0}\leq\lambda_0$, where λ_0 is given in Theorem 2.2. The existence of such a n_0 follows from (i) for all $M\in(M_0,M_1]$, the level set $\mathcal{W}_M$ is compact and $\bigcup_{p=0}^{\infty}\mathcal{K}_p$ is an increasing covering of Θ, and (ii) $\gamma_p\downarrow 0$ as $p\to\infty$. We notice that for any $l\geq 0$,

$$
T_{l+1}=T_l+T_1\circ\tau^{T_l}, \tag{4.8}
$$

where τ denotes the shift operator on the canonical space associated to the chain $\{Z_n\}$. Consequently, by the strong Markov property,

$$
\bar{\mathbb{P}}_{x,\theta}[T_{l+1}<\infty]=\bar{\mathbb{E}}_{x,\theta}\left[\mathbb{1}_{\{T_l<\infty\}}\bar{\mathbb{P}}_{Z_{T_l}}(T_1<\infty)\right]. \tag{4.9}
$$

Using Lemma 4.1, we have

$$\bar{\mathbb{P}}_{Z_{T_l}}\{T_1 < \infty\}\mathbb{1}_{\{T_l<\infty\}} = C(X_{T_l}, \theta_{T_l}, l, \varsigma_{T_l})\mathbb{1}_{\{T_l<\infty\}},$$

where, for any $x, \theta \in \mathsf{X} \times \Theta$ and any integers p and q,

$$C(x, \theta, p, q) = \mathbb{P}^{\boldsymbol{\gamma}^{\leftarrow q}}_{\Phi(x,\theta)}\left(\sigma(\mathcal{K}_p, \boldsymbol{\epsilon}^{\leftarrow q}) < \infty\right).$$

Now, for $p \geq n_0$, we have $\mathcal{W}_M \subset \mathcal{K}_{n_0} \subset \mathcal{K}_p$, showing that for any $x, \theta \in \mathsf{X} \times \Theta$ and integers $p, q \geq n_0$,

$$\begin{aligned} C(x, \theta, p, q) &\leq \mathbb{P}^{\boldsymbol{\gamma}^{\leftarrow q}}_{\Phi(x,\theta)}\left[\sigma(\mathcal{W}_M) \wedge \nu(\boldsymbol{\epsilon}^{\leftarrow q}) < \infty\right] \\ &\leq \mathbb{P}^{\boldsymbol{\gamma}^{\leftarrow q}}_{\Phi(x,\theta)}\left[\sigma(\mathcal{W}_M) < \infty, \sigma(\mathcal{W}_M) \leq \nu(\boldsymbol{\epsilon}^{\leftarrow q})\right] + \mathbb{P}^{\boldsymbol{\gamma}^{\leftarrow q}}_{\Phi(x,\theta)}\left[\nu(\boldsymbol{\epsilon}^{\leftarrow q}) < \sigma(\mathcal{W}_M)\right]. \end{aligned}$$

By Theorem 2.2, for any integer $m \geq 0$ and any integer $q \geq n_0$, we have

$$\{\sigma(\mathcal{W}_M) = m, m \leq \nu(\boldsymbol{\epsilon}^{\leftarrow q})\} \subset \left\{\sup_{k\in\{1,\ldots,m\}} |S_{1,k}(\boldsymbol{\epsilon}^{\leftarrow q}, \boldsymbol{\gamma}^{\leftarrow q}, \mathcal{W}_M)| > \delta_0\right\},$$

which implies that for any $x, \theta \in \mathsf{X} \times \Theta$, any $l \geq n_0$, and any $q \geq n_0$

$$\begin{aligned} C(x, \theta, l, q) \leq \mathbb{P}^{\boldsymbol{\gamma}^{\leftarrow q}}_{\Phi(x,\theta)}\left(\sup_{k\geq 1} |S_{1,k}(\boldsymbol{\epsilon}^{\leftarrow q}, \boldsymbol{\gamma}^{\leftarrow q}, \mathcal{W}_M)| > \delta_0\right) + \mathbb{P}^{\boldsymbol{\gamma}^{\leftarrow q}}_{\Phi(x,\theta)}(\nu(\boldsymbol{\epsilon}^{\leftarrow q}) \\ < \sigma(\mathcal{W}_M)) \leq A(\delta_0, \boldsymbol{\epsilon}^{\leftarrow q}, M, \boldsymbol{\gamma}^{\leftarrow q}). \end{aligned}$$

Combining the results above, we have, noting that $\varsigma_{T_l} \geq l$,

$$\begin{aligned} \bar{\mathbb{P}}_{Z_{T_l}}\left[T_1 < \infty\right]\mathbb{1}_{\{T_l<\infty\}} &\leq A(\delta_0, \boldsymbol{\epsilon}^{\leftarrow \varsigma_{T_l}}, M, \boldsymbol{\gamma}^{\leftarrow \varsigma_{T_l}})\mathbb{1}_{\{T_l<\infty\}} \\ &\leq \sup_{q\geq l} A(\delta_0, \boldsymbol{\epsilon}^{\leftarrow q}, M, \boldsymbol{\gamma}^{\leftarrow q})\mathbb{1}_{\{T_l<\infty\}}; \end{aligned}$$

the proof now follows from a straightforward backward induction using (4.9) for $l = n_0, \ldots, n-1$ and $n > n_0$. □

COROLLARY 4.3. *Assume* (A1) *and that* $\mathcal{K}_0 \subset \mathcal{W}_{M_0}$ (*where* M_0 *is defined in* (A1)). *Then for any* $M \in (M_0, M_1]$ *and* $n \geq n_0$, *there exists a constant* $C < \infty$ *such that for any* $m \geq n$,

$$\bar{\mathbb{P}}_{x,\theta}\left[\sup_{k\geq 1} \kappa_k \geq m\right] \leq C\left(\sup_{q\geq n} A(\delta_0, \boldsymbol{\epsilon}^{\leftarrow q}, M, \boldsymbol{\gamma}^{\leftarrow q})\right)^m,$$

where $\{\kappa_k\}$ *is the counter corresponding to the number of reinitializations defined in* (3.3).

Proof. We have

$$\left\{\sup_{k\geq 1} \kappa_k \geq m\right\} \subset \{T_m < \infty\},$$

and consequently,

(4.10)
$$\begin{aligned}\bar{\mathbb{P}}_{x,\theta}\left(\sup_{k\geq 1}\kappa_k \geq m\right) &\leq \bar{\mathbb{P}}_{x,\theta}\left(T_m < \infty\right) \leq \prod_{l=n_0}^{m-1} \sup_{q\geq l} A(\delta_0, \boldsymbol{\epsilon}^{\leftarrow q}, M, \boldsymbol{\gamma}^{\leftarrow q}) \\ &\leq \prod_{l=n_0}^{n-1} \sup_{q\geq n_0} A(\delta_0, \boldsymbol{\epsilon}^{\leftarrow q}, M, \boldsymbol{\gamma}^{\leftarrow q}) \prod_{l=n}^{m-1} \sup_{q\geq n} A(\delta_0, \boldsymbol{\epsilon}^{\leftarrow q}, M, \boldsymbol{\gamma}^{\leftarrow q}) \\ &\leq C\left(\sup_{q\geq n} A(\delta_0, \boldsymbol{\epsilon}^{\leftarrow q}, M, \boldsymbol{\gamma}^{\leftarrow q})\right)^m. \qquad \square\end{aligned}$$

In the next section we derive conditions on the family of Markov kernels $\{P_\theta, \theta \in \Theta\}$ and on the sequences $\boldsymbol{\epsilon} = \{\epsilon_k\}$ and $\boldsymbol{\gamma} = \{\gamma_k\}$, which ensure that $\sup_{q\geq n} A(\delta_0, \boldsymbol{\epsilon}^{\leftarrow q}, M, \boldsymbol{\gamma}^{\leftarrow q}) < 1$ for n large enough. It should be emphasized here that this involves studying only the fluctuations of the canonical "interprojections" processes, i.e., $\{Y_n^{\boldsymbol{\rho}}\}$ for $\boldsymbol{\rho} = \boldsymbol{\gamma}^{\leftarrow \varsigma_{T_0}}, \boldsymbol{\gamma}^{\leftarrow \varsigma_{T_1}}, \boldsymbol{\gamma}^{\leftarrow \varsigma_{T_2}}, \ldots$.

5. Control of the fluctuations. Our aim now is to find a bound for $A(\delta, \boldsymbol{\epsilon}, M, \boldsymbol{\rho})$ defined in (4.7), which requires the following conditions to hold. Define, for $V : \mathsf{X} \to [1, \infty)$ and $g : \mathsf{X} \to \mathbb{R}^{n_\theta}$, the norm

$$\|g\|_V = \sup_{x\in\mathsf{X}} \frac{|g(x)|}{V(x)}. \tag{5.1}$$

Consider the following assumptions:

(A3) For any $\theta \in \Theta$, the Poisson equation $g - P_\theta g = H_\theta - \pi_\theta(H_\theta)$ has a solution g_θ. There exist a function $W : \mathsf{X} \to [1, \infty]$ such that $\{x \in \mathsf{X}, W(x) < \infty\} \neq \emptyset$, constants $\alpha \in (0, 1]$, $p \geq 2$ such that for any compact subset $\mathcal{K} \subset \Theta$,

(i) the following holds:

$$\sup_{\theta\in\mathcal{K}} \|H_\theta\|_W < \infty, \tag{5.2}$$

$$\sup_{\theta\in\mathcal{K}} \left(\|g_\theta\|_W + \|P_\theta g_\theta\|_W\right) < \infty, \tag{5.3}$$

$$\sup_{(\theta,\theta')\in\mathcal{K}} |\theta - \theta'|^{-\alpha} \left\{\|g_\theta - g_{\theta'}\|_W + \|P_\theta g_\theta - P_{\theta'} g_{\theta'}\|_W\right\} < \infty. \tag{5.4}$$

(ii) there exist constants $\{C_k, k \geq 0\}$ such that, for any $k \in \mathbb{N}$, for any sequence $\boldsymbol{\rho} = \{\rho_k\}$, and for any $x \in \mathsf{X}$,

$$\sup_{\theta\in\mathcal{K}} \mathbb{E}_{x,\theta}^{\boldsymbol{\rho}}[W^p(X_k)\mathbb{1}_{\{\sigma(\mathcal{K})\geq k\}}] \leq C_k W^p(x). \tag{5.5}$$

(iii) there exist $\epsilon > 0$ and a constant C such that for any sequence $\boldsymbol{\rho} = \{\rho_k\}$ and for any $x \in \mathsf{X}$,

$$\sup_{\theta\in\mathcal{K}} \mathbb{E}_{x,\theta}^{\boldsymbol{\rho}}[W^p(X_k)\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu_\epsilon\geq k\}}] \leq C W^p(x), \tag{5.6}$$

where

$$\nu_\epsilon = \inf\{k \geq 1, |\theta_k - \theta_{k-1}| > \epsilon\}. \tag{5.7}$$

Assumption (A3) states the existence and the regularity of the solutions of Poisson's equation for the family of transition kernels $\{P_\theta, \theta \in \Theta\}$. The conditions stated above are nonprimitive; a set of more tractable conditions implying (A3) is given in

section 6. Poisson's equation has proven to be fundamental in the analysis of additive functionals of Markov chains, in particular for establishing limit theorems such as the (functional) central limit theorem (see, e.g., [5], [26], [25, Chapter 17], [18], [16]); the existence of solutions to Poisson's equation is well established for geometrically ergodic Markov chains (see [26], [25, Chapter 17]); it has been more recently proven under assumptions weaker than geometric ergodicity (see [18, Theorem 2.3]); the regularity of the solution of Poisson's equation has been studied, under various ergodicity and regularity conditions on the mapping $\theta \mapsto P_\theta$, in [5], [4]. We stress here that the function W is global but that the bounds in (5.2), (5.3), (5.4), (5.5), and (5.6) depend on the particular compact $\mathcal{K}$ under consideration. We have the following.

LEMMA 5.1. *Assume* (A3). *Let $\mathcal{K}$ be a compact subset of Θ and $s \in \mathbb{N}$. There exists a constant C such that for any sequence $\boldsymbol{\epsilon} = \{\epsilon_k\}$ satisfying $0 < \epsilon_k \leq \epsilon$ for all $k \geq s$ (where ϵ is defined in* (A3)(iii)*), for any sequence $\boldsymbol{\rho} = \{\rho_k\}$ and for any $x \in \mathsf{X}$,*

$$\sup_{\theta \in \mathcal{K}} \sup_{k \geq 0} \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta} \left[W^p(X_k) \mathbb{1}_{\{\sigma(\mathcal{K},\boldsymbol{\epsilon}) \geq k\}} \right] \leq C W^p(x).$$

Proof. Under (A3), there exists a constant C such that, for any sequence $\boldsymbol{\rho} = \{\rho_k\}$ and any $x \in \mathsf{X}$, we have

$$\sup_{\theta \in \mathcal{K}} \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}[W^p(X_k) \mathbb{1}_{\{\sigma(\mathcal{K}) \wedge \nu_\epsilon \geq k\}}] \leq C W^p(x),$$

where ν_ϵ is defined in (5.7). For any sequence $\boldsymbol{\epsilon} = \{\epsilon_k\}$ such that $\epsilon_k \leq \epsilon$ for any $k \geq s$,

$$\begin{aligned}
&\mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}[W^p(X_{k+s}) \mathbb{1}_{\{\sigma(\mathcal{K}) \wedge \nu(\boldsymbol{\epsilon}) \geq k+s\}}] \\
&\quad = \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta} \left[\mathbb{E}^{\boldsymbol{\rho}^{\leftarrow s}}_{X_s,\theta_s}[W^p(X_k) \mathbb{1}_{\{\sigma(\mathcal{K}) \wedge \nu(\boldsymbol{\epsilon}^{\leftarrow s}) \geq k\}} \mathbb{1}_{\{\sigma(\mathcal{K}) \wedge \nu(\boldsymbol{\epsilon}) \geq s\}}] \right] \\
&\quad \leq \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta} \left[\sup_{\theta \in \mathcal{K}} \mathbb{E}^{\boldsymbol{\rho}^{\leftarrow s}}_{X_s,\theta}[W^p(X_k) \mathbb{1}_{\{\sigma(\mathcal{K}) \wedge \nu_\epsilon \geq k\}} \mathbb{1}_{\{\sigma(\mathcal{K}) \geq s\}}] \right] \\
&\quad \leq C \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}[W^p(X_s) \mathbb{1}_{\{\sigma(\mathcal{K}) \geq s\}}],
\end{aligned}$$

and the proof is concluded by (A3). □

PROPOSITION 5.2. *Assume* (A3). *Let $\mathcal{K}$ be a compact subset of Θ and let $\boldsymbol{\rho} = \{\rho_k\}$ and $\boldsymbol{\epsilon} = \{\epsilon_k\}$ be two nonincreasing sequences of positive numbers such that $\lim_{k \to \infty} \epsilon_k = 0$. Then, for p as defined in* (A3),

1. *there exists a constant C such that, for any $(x, \theta) \in \mathsf{X} \times \mathcal{K}$, any integer l, and any $\delta > 0$*

(5.8)
$$\mathbb{P}^{\boldsymbol{\rho}}_{x,\theta} \left(\sup_{n \geq l} |S_{l,n}(\boldsymbol{\epsilon}, \boldsymbol{\rho}, \mathcal{K})| \geq \delta \right) \leq C \delta^{-p} \left\{ \left(\sum_{k=l}^{\infty} \rho_k^2 \right)^{p/2} + \left(\sum_{k=l}^{\infty} \rho_k \epsilon_k^\alpha \right)^p \right\} W^p(x).$$

2. *there exists a constant C such that, for any $(x, \theta) \in \mathsf{X} \times \mathcal{K}$,*

$$\mathbb{P}^{\boldsymbol{\rho}}_{x,\theta} \left(\nu(\boldsymbol{\epsilon}) < \sigma(\mathcal{K}) \right) \leq C \left\{ \sum_{k=1}^{\infty} (\epsilon_k^{-1} \rho_k)^p \right\} W^p(x). \tag{5.9}$$

The proof is in Appendix A. We finally need a condition on the stepsize sequences, which will ensure that $A(\delta, \boldsymbol{\epsilon}^{\leftarrow q}, M, \boldsymbol{\rho}^{\leftarrow q}) \to 0$ when $q \to \infty$.

(A4) The sequences $\boldsymbol{\gamma} = \{\gamma_k\}$ and $\boldsymbol{\epsilon} = \{\epsilon_k\}$ are nonincreasing, positive and satisfy $\sum_{k=0}^{\infty} \gamma_k = \infty$, $\lim_{k\to\infty} \epsilon_k = 0$, and

$$\sum_{k=1}^{\infty} \left\{\gamma_k^2 + \gamma_k \epsilon_k^{\alpha} + (\epsilon_k^{-1}\gamma_k)^p\right\} < \infty,$$

where p and α are defined in (A3).

For instance, we may assume that $\sum_{k\geq 0} \gamma_k = \infty$ and $\sum_{k=0}^{\infty} \gamma_k^{\delta} < \infty$ for some $1 < \delta \leq p(1+\alpha)/(p+\alpha)$. Then, (A4) is verified by setting $\epsilon_k = C\gamma_k^{\eta}$ for some constant C and some η such that

$$\frac{\delta - 1}{\alpha} \leq \eta \leq \frac{p - \delta}{p}.$$

It is now straightforward to establish the following results.

PROPOSITION 5.3. *Assume* (A3) *and* (A4). *Then, for any subset* $\mathsf{K} \subset \mathsf{X}$ *such that* $\sup_{x\in\mathsf{K}} W(x) < \infty$, *any* $M \in (M_0, M_1]$, *and any* $\delta > 0$, *we have* $\lim_{k\to\infty} A(\delta, \boldsymbol{\epsilon}^{\leftarrow k}, M, \boldsymbol{\gamma}^{\leftarrow k}) = 0$, *where* $A(\delta, \boldsymbol{\epsilon}, M, \boldsymbol{\rho})$ *is given by* (4.7).

We may now summarize the discussion above to obtain the following stability result.

THEOREM 5.4. *Assume* (A1)–(A4). *Then, for any subset* $\mathsf{K} \subset \mathsf{X}$ *such that* $\sup_{x\in\mathsf{K}} W(x) < \infty$, $\mathcal{K}_0 \subset \mathcal{W}_{M_0}$ *(where* M_0 *is defined in* (A1)) *and any* $\rho \in (0,1)$, *there exists a constant* $C < \infty$ *such that, for all* $(x,\theta) \in \mathsf{X} \times \Theta$,

$$\bar{\mathbb{P}}_{x,\theta}\left[\sup_{n\geq 1} \kappa_n \geq k\right] \leq C\rho^k.$$

Hence, under the stated conditions, the tail probability of the number of reinitializations decreases faster than any exponential and $\sup_{n\geq 1} \kappa_n$ is finite $\bar{\mathbb{P}}_{x,\theta}$-a.s. Combining this result with Theorem 2.3, it is possible to obtain the following global convergence result.

THEOREM 5.5. *Assume* (A1)–(A4). *Let* $\mathsf{K} \subset \mathsf{X}$ *be such that* $\sup_{x\in\mathsf{K}} W(x) < \infty$ *and that* $\mathcal{K}_0 \subset \mathcal{W}_{M_0}$ *(where* M_0 *is defined in* (A1)), *and let* $\{Z_n\}$ *be as defined by* (3.3). *Then, for all* $(x,\theta) \in \mathsf{X} \times \Theta$, *we have* $\lim_{k\to\infty} d(\theta_k, \mathcal{L}) = 0$, $\bar{\mathbb{P}}_{x,\theta}$*-a.s.*

Proof. Define, for $k \geq 1$,

$$B_k = \limsup_{l\to\infty} \sup_{T_{k-1}+l\leq n} \left| \mathbb{1}_{\{n<T_k\}} \sum_{j=T_{k-1}+l}^{n} \gamma_{\varsigma_j} (H(\theta_{j-1}, X_j) - h(\theta_{j-1})) \right| \mathbb{1}_{\{T_{k-1}<\infty\}},$$

where ς_j and T_k are defined in section 3. We first show that, for any k and any $\delta > 0$, $\bar{\mathbb{P}}_{x,\theta}(|B_k| \geq \delta) = 0$ for all $(x,\theta) \in \mathsf{X} \times \Theta$. We have, by the strong Markov property and (4.6) that for $l \geq 1$

$$\bar{\mathbb{P}}_{x,\theta}\left(\sup_{T_{k-1}+l\leq n} \left| \mathbb{1}_{\{n<T_k\}} \sum_{j=T_{k-1}+l}^{n} \gamma_{\varsigma_j} (H(\theta_{j-1}, X_j) - h(\theta_{j-1})) \right| \mathbb{1}_{\{T_{k-1}<\infty\}} \geq \delta \right)$$
$$\leq \bar{\mathbb{E}}_{x,\theta}\left\{ C_l(X_{T_{k-1}}, \theta_{T_{k-1}}, \delta, \mathcal{K}_{k-1}, \varsigma_{T_{k-1}}) \mathbb{1}_{\{T_{k-1}<\infty\}} \right\},$$

where for any $x, \theta \in \mathsf{X} \times \Theta$, any $\delta > 0$, any set $\mathcal{K} \subset \Theta$, and any integer q,

$$C_l(x,\theta,\delta,\mathcal{K},q) = \mathbb{P}^{\boldsymbol{\gamma}^{\leftarrow q}}_{\Phi(x,\theta)}\left(\sup_{n\geq l} |S_{l,n}(\boldsymbol{\epsilon}^{\leftarrow q}, \boldsymbol{\gamma}^{\leftarrow q}, \mathcal{K})| \geq \delta \right).$$

By Proposition 5.2, for any compact subset $\mathcal{K}$, there exists a constant C such that, for all $q \geq 0$,

$$\sup_{(x,\theta)\in \mathsf{X}\times\Theta} C_l(x,\theta,\delta,\mathcal{K},q) \leq C\delta^{-p}\left\{\left(\sum_{j=l}^{\infty}\gamma_j^2\right)^{p/2} + \left(\sum_{j=l}^{\infty}\gamma_j\epsilon_j^{\alpha}\right)^{p}\right\},$$

which implies that, for all $k \geq 0$, $\bar{\mathbb{P}}_{x,\theta}(|B_k| \geq \delta) = 0$. Corollary 4.3 and Proposition 5.3 show that, for all $(x,\theta) \in \mathsf{K} \times \mathcal{K}_0$, $\kappa = \sup_k \kappa_k < \infty$ $\bar{\mathbb{P}}_{x,\theta}$-a.s. Set, for $k \geq 0$, $\bar{\theta}_k = \theta_{k+T_{\kappa-1}}$, $\bar{\gamma}_k = \gamma_{k+\varsigma_{T_{\kappa-1}}}$ and

$$\xi_k = H(\bar{\theta}_{k-1}, X_{k+T_{\kappa-1}}) - h(\bar{\theta}_{k-1}), \quad k \geq 1.$$

Then, $\bar{\theta}_k = \bar{\theta}_{k-1} + \bar{\gamma}_k h(\bar{\theta}_{k-1}) + \bar{\gamma}_k\xi_k$ and, since $T_\kappa = \infty$, for all $(x,\theta) \in \mathsf{K} \times \mathcal{K}_0$,

$$\limsup_{l\to\infty}\sup_{n\geq l}\left|\sum_{k=l}^{n}\bar{\gamma}_k\xi_k\right| = B_\kappa = 0, \quad \bar{\mathbb{P}}_{x,\theta}\text{-a.s.}$$

The proof follows from Theorem 2.3. □

6. Drift conditions. In this section, we give conditions which imply (A3) in terms of a minorization of the Markov kernel on a small set and a drift condition toward this small set (see [25] for definitions and main results). Denote, for $V : \mathsf{X} \to [1,\infty)$, $\mathcal{L}_V := \{g : \mathsf{X} \to \mathbb{R}^{n_\theta}, \sup_{x\in\mathsf{X}} \|g\|_V < \infty\}$, where $\|\cdot\|_V$ is defined in (5.1).

(DRI) For any $\theta \in \Theta$, P_θ is ψ-irreducible and aperiodic[1]. In addition there exist a function $V : \mathsf{X} \to [1,\infty)$ and constants $p \geq 2$ and $\beta \in [0,1]$ such that for any compact subset $\mathcal{K} \subset \Theta$,

(DRI1) there exist an integer m, constants $0 < \lambda < 1$, b, κ, $\delta > 0$, and a probability measure ν such that

$$\sup_{\theta\in\mathcal{K}} P_\theta^m V^p(x) \leq \lambda V^p(x) + b\mathbb{1}_{\mathsf{C}}(x), \tag{6.1}$$

$$\sup_{\theta\in\mathcal{K}} P_\theta V^p(x) \leq \kappa V^p(x) \qquad \forall x \in \mathsf{X}, \tag{6.2}$$

$$\inf_{\theta\in\mathcal{K}} P_\theta^m(x,A) \geq \delta\nu(A) \qquad \forall x \in \mathsf{C}, \quad \forall A \in \mathcal{B}(\mathsf{X}). \tag{6.3}$$

(DRI2) there exists C such that, for all $x \in \mathsf{X}$,

$$\sup_{\theta\in\mathcal{K}} |H_\theta(x)| \leq CV(x),$$
$$\sup_{(\theta,\theta')\in\mathcal{K}} |\theta-\theta'|^{-\beta}|H_\theta(x) - H_{\theta'}(x)| \leq CV(x).$$

(DRI3) there exists C such that, for all $(\theta,\theta') \in \mathcal{K} \times \mathcal{K}$,

$$\|P_\theta g - P_{\theta'} g\|_V \leq C\,\|g\|_V\,|\theta-\theta'|^\beta \qquad \forall g \in \mathcal{L}_V, \tag{6.4}$$

$$\|P_\theta g - P_{\theta'} g\|_{V^p} \leq C\,\|g\|_{V^p}\,|\theta-\theta'|^\beta, \qquad \forall g \in \mathcal{L}_{V^p}. \tag{6.5}$$

[1]We use the standard terminology and notation introduced in [25, Chapters 4, 5].

Assumption (DRI1) is classical in the Markov chain literature; it implies the existence of a stationary distribution π_θ for all $\theta \in \Theta$ and V^p-uniform ergodicity; i.e., for each $\theta \in \Theta$ there exist constants $C_\theta < \infty$ and $\rho_\theta \in [0,1)$ such that for any function $f \in \mathcal{L}_{V^p}$ and any integer $k > 0$,

$$\|P_\theta^k f - \pi_\theta(f)\|_{V^p} \le C_\theta \rho_\theta^k \|f\|_{V^p}.$$

Note that the constants C_θ and ρ_θ may be bounded over the compact sets of Θ; i.e., for each $\mathcal{K} \subset \Theta$, there exists $\bar{C} < \infty$ and $\bar{\rho} \in [0,1)$ such that $\sup_{\theta\in\mathcal{K}} C_\theta \le \bar{C}$ and $\sup_{\theta\in\mathcal{K}} \rho_\theta \le \bar{\rho}$. The regularity of the kernels $\theta \to P_\theta$ expressed in V and V^p norms is naturally less classical. The main result of this section follows.

PROPOSITION 6.1. *Assume* (DRI). *Then* (A2) *and* (A3) *are satisfied and for any* $0 < \alpha < \beta$,

$$\sup_{(\theta,\theta')\in\mathcal{K}\times\mathcal{K}} |\theta - \theta'|^{-\alpha} \, |h(\theta) - h(\theta')| < \infty. \tag{6.6}$$

The proof is in Appendix B.

7. Controlled MCMC algorithm. Markov chain Monte Carlo (MCMC), introduced in [24], is a popular computational method for generating samples from virtually any distribution π defined on a space $\mathsf{X} \subset \mathbb{R}^{n_x}$ (for some integer n_x). The method consists of simulating an ergodic Markov chain $\{X_n, n \ge 0\}$ on X with transition probability P such that π is a *stationary* distribution for this chain, i.e., $\pi P = \pi$. Let ψ be some π-integrable function $\psi : \mathsf{X} \to \mathbb{R}^{n_\psi}$ for some integer n_ψ. One can use the samples produced by the Markov chain to estimate integrals

$$\pi(\psi) := \int_{\mathsf{X}} \psi(x) \, \pi(dx)$$

with estimators of the type

$$S_n(\psi) = \frac{1}{n} \sum_{k=1}^{n} \psi(X_k). \tag{7.1}$$

In general the transition probability P of the Markov chain depends on some tuning parameter, say θ, defined on some space $\Theta \subset \mathbb{R}^{n_\theta}$ for some integer n_θ, and the convergence properties of the Monte Carlo averages in (7.1) might highly depend on a proper choice of this parameter.

We illustrate this here with the classical Metropolis–Hastings (MH) update, but it should be stressed at this point that the results presented in this paper apply to much more general settings. The MH algorithm requires the choice of a *proposal distribution* q. In order to simplify the discussion, we will assume that π and q admit densities with respect to the Lebesgue measure λ^{Leb}, denoted, with an abuse of notation, by π and q hereafter. The role of the distribution q consists of proposing potential transitions y for the Markov chain $\{X_n\}$. Given that the chain is currently at x, a candidate y is accepted with probability $\alpha(x,y)$ defined as

$$\alpha(x,y) = \begin{cases} 1 \wedge \frac{\pi(y)}{\pi(x)} \frac{q(y,x)}{q(x,y)} & \text{if } \pi(x)q(x,y) > 0 \\ 1 & \text{otherwise,} \end{cases}$$

where $a \wedge b := \min(a,b)$. Otherwise it is rejected and the Markov chain stays at its current location x. The transition kernel P of this Markov chain takes the form for

$x, A \in \mathsf{X} \times \mathcal{B}(\mathsf{X})$ of

$$P(x,A) = \int_A \alpha(x,y) q(x,y) \lambda^{\mathrm{Leb}}(dy) + \mathbb{1}_A(x) \int_{\mathsf{X}} (1 - \alpha(x,y)) q(x,y) \lambda^{\mathrm{Leb}}(dy). \tag{7.2}$$

The Markov chain P is reversible with respect to π and therefore admits π as an invariant distribution. Conditions on the proposal distribution q that guarantee irreducibility and positive recurrence are mild and many satisfactory choices are possible.

7.1. Symmetric random walk MH. We focus here on the symmetric increments random-walk MH algorithm (SRWM), which corresponds to the case where $q(x,y) = q(x-y)$ for some symmetric probability density q. Other examples are considered in [2]. The transition kernel of the SRWM algorithm is then given for $x, A \in \mathsf{X} \times \mathcal{B}(\mathsf{X})$ by

$$\begin{aligned} P_q^{\mathrm{SRW}}(x,A) = {} & \int_{A-x} \left(1 \wedge \frac{\pi(x+z)}{\pi(x)}\right) q(z)\, \lambda^{\mathrm{Leb}}(dz) \\ & + \mathbb{1}_A(x) \int_{\mathsf{X}-x} \left(1 - \left(1 \wedge \frac{\pi(x+z)}{\pi(x)}\right)\right) q(z)\, \lambda^{\mathrm{Leb}}(dz), \quad x \in \mathsf{X}, A \in \mathcal{B}(\mathsf{X}), \end{aligned} \tag{7.3}$$

where $A - x := \{z \in \mathsf{X}, x + z \in A\}$. A classical choice for the proposal distribution is $q = \phi_{0,\Gamma}$, where $\phi_{\mu,\Gamma}$ is the density of a multivariate normal distribution with mean μ and covariance matrix Γ. We will later on refer to this algorithm as the N-SRWM. It is well known that either too small or too large a covariance matrix will result in highly positively correlated Markov chains and therefore estimators $S_n(\psi)$ with large variance. In practice this covariance matrix Γ is determined by trial and error using several realizations of the Markov chain. This hand-tuning requires some expertise and can be time consuming.

In order to circumvent this problem, in the context of the N-SRWM update described above, the authors of [19] have proposed to "learn Γ on the fly." Their algorithm can be summarized as (see [19])

$$\begin{aligned} \mu_{n+1} &= \mu_n + \gamma_{n+1}(X_{n+1} - \mu_n), && n \geq 0, \\ \Gamma_{n+1} &= \Gamma_n + \gamma_{n+1}((X_{n+1} - \mu_n)(X_{n+1} - \mu_n)^{\intercal} - \Gamma_n), \end{aligned} \tag{7.4}$$

where

- X_{n+1} is drawn from $P_{\theta_n}(X_n, \cdot)$, where for $\theta = (\mu, \Gamma)$, $P_\theta := P^{\mathrm{SRW}}_{\phi_{0,\lambda\Gamma}}$ with $\lambda > 0$, a constant scaling factor depending only on the dimension of the state-space n_x and kept constant across the iterations.
- $\boldsymbol{\gamma} = \{\gamma_n\}$ is a nonincreasing sequence of positive stepsizes such that $\sum_{n=1}^{\infty} \gamma_n = \infty$ and $\sum_{n=1}^{\infty} \gamma_n^{1+\delta} < \infty$ for some $\delta > 0$ ([19] suggests the choice $\gamma_n = 1/n$).

It was realized in [3] that such a scheme is a particular case of a more general framework akin to stochastic control, combined with the use of the Robbins–Monro procedure. More precisely, let $\theta = (\mu, \Gamma) \in \Theta$, where $\Theta := \mathbb{R}^{n_x} \times \mathcal{C}_+^{n_x}$ and $\mathcal{C}_+^{n_x}$ is the cone of positive $n_x \times n_x$ matrices; then

$$H(x;\theta) = (x - \mu, (x-\mu)(x-\mu)^{\intercal} - \Gamma)^{\intercal}. \tag{7.5}$$

With this notation, the recursion in (7.4) may be written in the standard Robbins–Monro form as

$$\theta_{n+1} = \theta_n + \gamma_{n+1} H(X_{n+1}, \theta_n), \quad n \geq 0, \tag{7.6}$$

with $X_{n+1} \sim P_{\theta_n}(X_n, \cdot)$. For the present example, assuming that $\int_{\mathsf{X}} |x|^2 \pi(dx) < \infty$, one can easily check that

$$h(\theta) = \int_{\mathsf{X}} H(x,\theta)\pi(dx) = (\mu_\pi - \mu, (\mu_\pi - \mu)(\mu_\pi - \mu)^{\mathrm{T}} + \Gamma_\pi - \Gamma)^{\mathrm{T}}, \tag{7.7}$$

with μ_π and Γ_π the mean and covariance of the target distribution. It is assumed in the sequel that Γ_π is positive definite. We now analyze the corresponding homogeneous Markov chain $\{Z_n, n \geq 0\}$ as defined in section 3, i.e., prove under mild conditions on π that (A1)–(A3) are satisfied.

7.2. Condition (A1). In the algorithm described above the parameter estimates μ_n and Γ_n take the form of maximum likelihood estimates under the i.i.d. multivariate normal model. It therefore comes as no surprise if the appropriate Lyapunov function is

$$w(\mu,\Gamma) = -\int_{\mathsf{X}} \log\left(\frac{\pi(x)}{\phi_{0,\Gamma}(x)}\right)\pi(dx), \tag{7.8}$$

the Kullback–Leibler divergence between the target density π and a normal density $\phi_{0,\Gamma}$.

PROPOSITION 7.1. *Let h be as defined in* (7.7) *where π satisfies (M). Then* (A1) *is satisfied with w as in* (7.8). *Furthermore $\mathcal{L}$ is reduced to a single point, $\theta_\pi := (\mu_\pi, \Gamma_\pi)$.*

Proof. h is naturally continuous (and, as we shall see later, is in fact Lipschitz continuous under (DRI) (or when (DRI) is assumed) since Proposition 6.1 holds under (DRI). Now w is equal, up to multiplicative and additive constants, to

$$\log\det\Gamma + (\mu - \mu_\pi)^{\mathrm{T}}\Gamma^{-1}(\mu - \mu_\pi) + \mathrm{Tr}(\Gamma^{-1}\Gamma_\pi). \tag{7.9}$$

Using straightforward algebra, one can show that there exists a constant $C > 0$ such that

$$\begin{aligned} C\Big\langle \nabla w(\mu,\Gamma), h(\mu,\Gamma)\Big\rangle &= -2(\mu - \mu_\pi)^{\mathrm{T}}\Gamma^{-1}(\mu - \mu_\pi) \\ &\quad - \mathrm{Tr}(\Gamma^{-1}(\Gamma - \Gamma_\pi)\Gamma^{-1}(\Gamma - \Gamma_\pi)) - \big((\mu - \mu_\pi)^{\mathrm{T}}\Gamma^{-1}(\mu - \mu_\pi)\big)^2, \end{aligned} \tag{7.10}$$

that is, $\langle \nabla w(\theta), h(\theta)\rangle \leq 0$ for any $\theta = (\mu,\Gamma) \in \Theta$, with equality if and only if $\Gamma = \Gamma_\pi$ and $\mu = \mu_\pi$. As $w(\Theta) = [w(\mu_\pi,\Gamma_\pi), \infty)$ and w is continuous, any $w(\mu_\pi,\Gamma_\pi) < M_0 < M_1 < \infty$ satisfy (A1)(i) and (A1)(ii), and (A1)(iii) is automatically satisfied. Now as the set of stationary points $\mathcal{L}$ is reduced to a single point, (A1)(iv) is also satisfied. ∎

7.3. Condition (A3). In order to check (A3) in this case, we check (DRI). The geometric ergodicity of the random walk Metropolis–Hastings (RWMH) kernel has been studied by [27] and refined in [21]; the regularity of the RWMH has, to the best of our knowledge, not been considered in the literature. The geometric ergodicity of the RWMH kernel mainly depends on the tail properties of the target distribution π. We will therefore restrict our discussion to target distributions that satisfy the following set of conditions. These are not minimal but are easy to check in practice (see [21] for details).

(M) The probability density π has the following properties:

(i) It is bounded, bounded away from zero on every compact set, and continuously differentiable.

(ii) It is superexponential, i.e.,

$$\lim_{|x|\to+\infty} \left\langle \frac{x}{|x|}, \nabla \log \pi(x) \right\rangle = -\infty.$$

(iii) The contours $\partial \mathsf{A}(x) = \{y : \pi(y) = \pi(x)\}$ are asymptotically regular, i.e.,

$$\lim_{|x|\to+\infty} \sup \left\langle \frac{x}{|x|}, \frac{\nabla \pi(x)}{|\nabla \pi(x)|} \right\rangle < 0.$$

Note that this condition implies the existence and finiteness of μ_π and Γ_π. We now establish uniform minorization and drift conditions for P_q^{SRW} defined in (7.3). Let $\mathcal{M}(\mathsf{X})$ denote the set of probability densities w.r.t. the Lebesgue measure λ^{Leb}. Let $\varepsilon > 0$ and $\delta > 0$ and define the subset $\mathcal{K}_{\delta,\varepsilon} \subset \mathcal{M}(\mathsf{X})$,

$$\mathcal{K}_{\delta,\varepsilon} = \{q \in \mathcal{M}(\mathsf{X}), q(z) = q(-z) \quad \text{and} \quad |z| \le \varepsilon \Rightarrow q(z) \ge \delta\}. \tag{7.11}$$

PROPOSITION 7.2. *Assume* (M). *For any* $\eta \in (0,1)$, *set* $W = \pi^{-\eta}/(\inf_{\mathsf{X}} \pi^{-\eta})$. *Then,*

1. *any nonempty compact set* $\mathsf{C} \subset \mathsf{X}$ *is a* $(1,\delta)$*-small set for some* $\delta > 0$ *and some measure* ν,

$$\forall (x, A) \in \mathsf{C} \times \mathcal{B}(\mathsf{X}) \qquad \inf_{q \in \mathcal{K}_{\delta,\varepsilon}} P_q^{\mathrm{SRW}}(x, A) \ge \delta \nu(A). \tag{7.12}$$

2. *furthermore, for any* $\delta > 0$ *and* $\varepsilon > 0$,

$$\sup_{q \in \mathcal{K}_{\delta,\varepsilon}} \limsup_{|x|\to+\infty} \frac{P_q^{\mathrm{SRW}} W(x)}{W(x)} < 1, \tag{7.13}$$

$$\sup_{(x,q) \in \mathsf{X} \times \mathcal{K}_{\delta,\varepsilon}} \frac{P_q^{\mathrm{SRW}} W(x)}{W(x)} < +\infty. \tag{7.14}$$

3. *let* $q, q' \in \mathcal{M}(\mathsf{X})$ *be two symmetric probability distributions. Then, for any* $r \in [0,1]$ *and any* $g \in \mathcal{L}_{W^r}$ *we have*

$$\left\| P_q^{\mathrm{SRW}} g - P_{q'}^{\mathrm{SRW}} g \right\|_{W^r} \le 2 \|g\|_{W^r} \int_{\mathsf{X}} |q(z) - q'(z)| \lambda^{\mathrm{Leb}}(dz). \tag{7.15}$$

Proof. For any $x \in \mathsf{X}$, define the acceptance region $\mathsf{A}(x) = \{z \in \mathsf{X} - x; \pi(x+z) \ge \pi(x)\}$ and the rejection region $\mathsf{R}(x) = \{z \in \mathsf{X} - x; \pi(x+z) < \pi(x)\}$. From the definition, (7.11) of $\mathcal{K}_{\delta,\varepsilon}$ [27, Theorem 2.2] applies for any $q \in \mathcal{K}_{\delta,\varepsilon}$ and we can conclude that (7.12) is satisfied. Noting that the two sets $\mathsf{A}(x)$ and $\mathsf{R}(x)$ do not depend on the proposal distribution q, and using the conclusion of the proof of Theorem 4.3 of [21], we have

$$\inf_{q \in \mathcal{K}_{\delta,\varepsilon}} \liminf_{|x|\to+\infty} \int_{\mathsf{A}(x)} q(z) \lambda^{\mathrm{Leb}}(dz) > 0,$$

so that from the conclusion of the proof of Theorem 4.1 of [21],

$$\sup_{q \in \mathcal{K}_{\delta,\varepsilon}} \limsup_{|x|\to+\infty} \frac{P_q^{\mathrm{SRW}} W(x)}{W(x)} = 1 - \inf_{q \in \mathcal{K}_{\delta,\varepsilon}} \liminf_{|x|\to+\infty} \int_{\mathsf{A}(x)} q(z) \lambda^{\mathrm{Leb}}(dz) < 1,$$

which proves (7.13). Finally, for any $q \in \mathcal{K}_{\delta,\varepsilon}$,

$$\begin{aligned}\frac{P_q^{\mathrm{SRW}} W(x)}{W(x)} &= \int_{\mathsf{A}(x)} \frac{\pi(x+z)^{-\eta}}{\pi(x)^{-\eta}} q(z) \lambda^{\mathrm{Leb}}(dz) \\ &\quad + \int_{\mathsf{R}(x)} \left(1 - \frac{\pi(x+z)}{\pi(x)} + \frac{\pi(x+z)^{1-\eta}}{\pi(x)^{1-\eta}}\right) q(z) \lambda^{\mathrm{Leb}}(dz) \\ &\leq \sup_{0 \leq u \leq 1} (1 - u + u^{1-\eta}),\end{aligned}$$

which proves (7.14). Now notice that

$$\begin{aligned}P_q^{\mathrm{SRW}} g(x) - P_{q'}^{\mathrm{SRW}} g(x) &= \int_{\mathsf{X}} \alpha(x, x+z)(q(z) - q'(z)) g(x+z) \lambda^{\mathrm{Leb}}(dz) \\ &\quad + g(x) \int_{\mathsf{X}} \alpha(x, x+z)(q'(z) - q(z)) \lambda^{\mathrm{Leb}}(dz).\end{aligned}$$

We therefore focus, for $r \in [0,1]$ and $g \in \mathcal{L}_{W^r}$, on the term

$$\begin{aligned}&\frac{\left|\int_{\mathsf{X}} \alpha(x, x+z)(q(z) - q'(z)) g(x+z) \lambda^{\mathrm{Leb}}(dz)\right|}{\|g\|_{W^r} W^r(x)} \\ &\leq \frac{\int_{\mathsf{X}} \alpha(x, x+z) |q(z) - q'(z)| W^r(x+z) \lambda^{\mathrm{Leb}}(dz)}{W^r(x)} \\ &= \int_{\mathsf{A}(x)} \frac{\pi(x+z)^{-r\eta}}{\pi(x)^{-r\eta}} |q(z) - q'(z)| \lambda^{\mathrm{Leb}}(dz) \\ &\qquad + \int_{\mathsf{R}(x)} \frac{\pi(x+z)^{1-r\eta}}{\pi(x)^{1-r\eta}} |q(z) - q'(z)| \lambda^{\mathrm{Leb}}(dz) \\ &\leq \int_{\mathsf{X}} |q(z) - q'(z)| \lambda^{\mathrm{Leb}}(dz).\end{aligned}$$

We now conclude that for any $x \in \mathsf{X}$ and any $g \in \mathcal{L}_{W^r}$,

$$\frac{|P_q^{\mathrm{SRW}} g(x) - P_{q'}^{\mathrm{SRW}} g(x)|}{W^r(x)} \leq 2 \|g\|_{W^r} \int_{\mathsf{X}} |q(z) - q'(z)| \lambda^{\mathrm{Leb}}(dz). \qquad \square$$

One can specialize the regularity property (7.15) to the N-SRWM, where the proposal distribution q_θ is a zero-mean normal distribution with covariance matrix Γ, and for simplicity we set $q_\Gamma := \phi_{0,\Gamma}$.

LEMMA 7.3. *Let $\mathcal{K}$ be a convex compact subset of $\mathcal{C}_+^{n_x}$ and set $W = \pi^{-\eta}/(\inf_{\mathsf{X}} \pi^{-\eta})$ for some $\eta \in (0,1)$. For any $r \in [0,1]$, any $\Gamma, \Gamma' \in \mathcal{K} \times \mathcal{K}$, $g \in \mathcal{L}_{W^r}$, we have*

$$\left\| P_{q_\Gamma}^{\mathrm{SRW}} g - P_{q_{\Gamma'}}^{\mathrm{SRW}} g \right\|_{W^r} \leq \frac{2 n_x}{\lambda_{\min}(\mathcal{K})} \|g\|_{W^r} |\Gamma - \Gamma'|, \tag{7.16}$$

where $\lambda_{\min}(\mathcal{K})$ is the minimum possible eigenvalue for matrices in $\mathcal{K}$.

Proof. We have

$$\int_{\mathsf{X}} |q_\Gamma(z) - q_{\Gamma'}(z)| dz = \int_{\mathsf{X}} \left| \int_0^1 \frac{d}{dv} q_{\Gamma + v(\Gamma' - \Gamma)}(z) dv \right| dz$$

and let $\Gamma_v = \Gamma + v(\Gamma' - \Gamma)$, so that

$$\frac{d}{dv} \log q_{\Gamma + v(\Gamma' - \Gamma)}(z) = -\frac{1}{2} \mathrm{Tr} \left[\Gamma_v^{-1} (\Gamma' - \Gamma) + \Gamma_v^{-1} z z^{\intercal} \Gamma_v^{-1} (\Gamma' - \Gamma) \right],$$

and consequently,

$$\int_{\mathsf{X}} \left| \int_0^1 \frac{d}{dv} q_{\Gamma + v(\Gamma' - \Gamma)}(z) dv \right| dz \le \frac{n_x}{\lambda_{\min}(\mathcal{K})} |\Gamma' - \Gamma|,$$

where we have used the inequality

$$|\mathrm{Tr}[\Gamma_v^{-1} z z^{\intercal} \Gamma_v^{-1} (\Gamma' - \Gamma)]| \le |\Gamma' - \Gamma| \mathrm{Tr}[\Gamma_v^{-1} \Gamma_v^{-1} z z^{\intercal}]. \qquad \square$$

COROLLARY 7.4. *For any compact subset $\mathcal{K}$ of $\mathcal{C}_+^{n_x}$, there exists $C < \infty$ such that*

$$\left\| P_{q_\Gamma}^{\mathrm{SRW}} g - P_{q_{\Gamma'}}^{\mathrm{SRW}} g \right\|_{W^r} \le C \, \|g\|_{W^r} \, |\Gamma - \Gamma'|. \tag{7.17}$$

7.4. Convergence of the adaptive MCMC algorithm. The main result of this section is the following.

THEOREM 7.5. *Let $\pi \in \mathcal{M}(\mathsf{X})$ satisfying* (M). *Let $\{Z_n\}$ be the homogeneous Markov chain defined as in section* 3, *with H as in* (7.5), *$P_\theta := P_{\phi_{0,\lambda\Gamma}}^{\mathrm{SRW}}$ for some $\lambda > 0$ with $\theta = (\mu, \Gamma) \in \Theta = \mathbb{R}^{n_x} \times \mathcal{C}_+^{n_x}$, K a compact set, and $\boldsymbol{\gamma} = \{\gamma_n\}$ and $\boldsymbol{\epsilon} = \{\epsilon_n\}$ satisfying* (A4). *Then,* (A1)–(A3) *are satisfied for any $\mathcal{K}_0$ and $\theta_n \to \theta_\pi$ w.p.* 1, *where $\theta_\pi := (\mu_\pi, \Gamma_\pi)$ is the unique stationary point of $\{\theta_n\}$.*

Proof. (A1) is implied by Proposition 7.1. (A2) is satisfied by construction of P_θ and from (M) and the definition of H in (7.5). Now we prove that (DRI) is satisfied. Choose $V^p = W = \pi^{-\eta} / \inf_{\mathsf{X}} \pi^{-\eta}$ for $p \ge 2$ in Proposition 7.2; then (DRI1) and (DRI3) are satisfied. Now (DRI2) is satisfied since, from (7.5),

$$|H_\theta(x) - H_{\theta'}(x)| \le |\mu - \mu'| \{1 + |\mu + \mu'| + 2|x|\} + |\Gamma - \Gamma'|, \tag{7.18}$$

and $\|x\|_W + \||x|^2\| < \infty$ from (M). Theorem 5.5 now applies. $\square$

This result is an important step for the study of the asymptotic properties of $\{S_n\}$ in [2], in particular the proof that $\{S_n\}$ satisfies a central limit theorem.

Appendix A. Proof of Proposition 5.2. Denote

$$D(\boldsymbol{\epsilon}, \boldsymbol{\rho}, \mathcal{K}, x) = \sup_{k \ge 1} \sup_{\theta \in \mathcal{K}} \mathbb{E}_{x,\theta}^{\boldsymbol{\rho}} [W^p(X_k) \mathbb{1}_{\{\sigma(\mathcal{K}) \wedge \nu(\boldsymbol{\epsilon}) \ge k\}}].$$

We first consider the case $l = 1$. Denote

$$T_n = \sum_{k=1}^{n} \rho_k \left(g_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}} g_{\theta_{k-1}}(X_k) \right) \mathbb{1}_{\{\sigma(\mathcal{K}) \wedge \nu(\boldsymbol{\epsilon}) \ge k\}},$$

Using $\mathbb{1}_{\{\sigma(\mathcal{K}) \wedge \nu(\boldsymbol{\epsilon}) \ge k\}} = \mathbb{1}_{\{\sigma(\mathcal{K}) \wedge \nu(\boldsymbol{\epsilon}) \ge k+1\}} + \mathbb{1}_{\{\sigma(\mathcal{K}) \wedge \nu(\boldsymbol{\epsilon}) = k\}}$, we may write $T_n = \sum_{i=1}^{5} T_n^{(i)}$,

where

$$T_n^{(1)} = \sum_{k=1}^{n} \rho_k \left(g_{\theta_{k-1}}(X_k) - P_{\theta_{k-1}} g_{\theta_{k-1}}(X_{k-1}) \right) \mathbb{1}_{\{\sigma(\mathcal{K}) \wedge \nu(\boldsymbol{\epsilon}) \geq k\}}, \tag{A.1}$$

$$T_n^{(2)} = \sum_{k=1}^{n-1} \rho_{k+1} \left(P_{\theta_k} g_{\theta_k}(X_k) - P_{\theta_{k-1}} g_{\theta_{k-1}}(X_k) \right) \mathbb{1}_{\{\sigma(\mathcal{K}) \wedge \nu(\boldsymbol{\epsilon}) \geq k+1\}}, \tag{A.2}$$

$$T_n^{(3)} = \sum_{k=1}^{n-1} (\rho_{k+1} - \rho_k) P_{\theta_{k-1}} g_{\theta_{k-1}}(X_k) \mathbb{1}_{\{\sigma(\mathcal{K}) \wedge \nu(\boldsymbol{\epsilon}) \geq k+1\}}, \tag{A.3}$$

$$T_n^{(4)} = \rho_1 P_{\theta_0} g_{\theta_0}(X_0) \mathbb{1}_{\{\sigma(\mathcal{K}) \wedge \nu(\boldsymbol{\epsilon}) \geq 1\}} - \rho_n P_{\theta_{n-1}} g_{\theta_{n-1}}(X_n) \mathbb{1}_{\{\sigma(\mathcal{K}) \wedge \nu(\boldsymbol{\epsilon}) \geq n\}}, \tag{A.4}$$

$$T_n^{(5)} = -\sum_{k=1}^{n-1} \rho_k P_{\theta_{k-1}} g_{\theta_{k-1}}(X_k) \mathbb{1}_{\{\sigma(\mathcal{K}) \wedge \nu(\boldsymbol{\epsilon}) = k\}}. \tag{A.5}$$

We now evaluate bounds for $T_n^{(i)}$, $i = 1, \ldots, 4$. In what follows, sequel C denotes a constant which depends only upon the compact set $\mathcal{K}$ through the quantities defined in the assumptions and whose value may change upon each appearance. We have

$$\sup_{\theta \in \mathcal{K}} \mathbb{E}^{\boldsymbol{\rho}}_{\theta,x} \left[\sup_{n \geq 0} \left| T_n^{(1)} \right|^p \right] \leq C \left(\sum_{k=0}^{\infty} \rho_k^2 \right)^{p/2} \sup_{\theta \in \mathcal{K}} \sup_k \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta} \left[W^p(X_k) \mathbb{1}_{\{\sigma(\mathcal{K}) \wedge \nu(\boldsymbol{\epsilon}) \geq k\}} \right], \tag{A.6}$$

$$\sup_{\theta \in \mathcal{K}} \mathbb{E}^{\boldsymbol{\rho}}_{\theta,x} \left[\sup_{n \geq 0} \left| T_n^{(2)} \right|^p \right] \leq C \left(\sum_{k=1}^{\infty} \rho_k \epsilon_k^{\alpha} \right)^{p} \sup_{\theta \in \mathcal{K}} \sup_k \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta} \left[W^p(X_k) \mathbb{1}_{\{\sigma(\mathcal{K}) \wedge \nu(\boldsymbol{\epsilon}) \geq k\}} \right], \tag{A.7}$$

$$\sup_{\theta \in \mathcal{K}} \mathbb{E}^{\boldsymbol{\rho}}_{\theta,x} \left[\sup_{n \geq 0} \left| T_n^{(3)} \right|^p \right] \leq C \rho_1^p \sup_{\theta \in \mathcal{K}} \sup_k \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta} \left[W^p(X_k) \mathbb{1}_{\{\sigma(\mathcal{K}) \wedge \nu(\boldsymbol{\epsilon}) \geq k\}} \right], \tag{A.8}$$

$$\sup_{\theta \in \mathcal{K}} \mathbb{E}^{\boldsymbol{\rho}}_{\theta,x} \left[\sup_{n \geq 0} \left| T_n^{(4)} \right|^p \right] \leq C \left(\sum_{k=1}^{\infty} \rho_k^2 \right)^{p/2} \sup_{\theta \in \mathcal{K}} \sup_k \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta} \left[W^p(X_k) \mathbb{1}_{\{\sigma(\mathcal{K}) \wedge \nu(\boldsymbol{\epsilon}) \geq k\}} \right]. \tag{A.9}$$

The proof of these inequalities can be adapted from [5, Part II, section 3.2]; see also [4, Chapter 6, Lemmas 6.2–6.4].

Proof of (A.6). Under (A3),

$$\sup_{\theta \in \mathcal{K}} \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta} \left[(|g_{\theta_k}(X_{k+1})|^p + |P_{\theta_k} g_{\theta_k}(X_{k+1})|^p) \mathbb{1}_{\{\sigma(\mathcal{K}) \wedge \nu(\boldsymbol{\epsilon}) \geq k+1\}} \right] \leq C D(\boldsymbol{\epsilon}, \boldsymbol{\rho}, \mathcal{K}, x).$$

Since

$$\begin{aligned} \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta} \left[(g_{\theta_k}(X_{k+1}) - P_{\theta_k} g_{\theta_k}(X_k)) \mathbb{1}_{\{\sigma(\mathcal{K}) \wedge \nu(\boldsymbol{\epsilon}) \geq k+1\}} \middle| \mathcal{F}_k \right] \\ = \left(P_{\theta_k} g_{\theta_k}(X_k) - P_{\theta_k} g_{\theta_k}(X_k) \right) \mathbb{1}_{\{\sigma(\mathcal{K}) \wedge \nu(\boldsymbol{\epsilon}) \geq (k+1)\}} = 0, \end{aligned}$$

$T_n^{(1)}$ is an ($\mathbb{R}^d$-valued) martingale. Using the Burkholder inequality [20, Theorem 2.10], we have

(A.10)
$$\mathbb{E}_{x,\theta}^{\boldsymbol{\rho}}\left[\left|T_n^{(1)}\right|^p\right] \le C_p \mathbb{E}_{x,\theta}^{\boldsymbol{\rho}}\left(\sum_{k=0}^{n-1} \rho_{k+1}^2 |g_{\theta_k}(X_{k+1}) - P_{\theta_k} g_{\theta_k}(X_k)|^2 \mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\epsilon})\ge k+1\}}\right)^{p/2},$$

where C_p is a universal constant. Using Minkowski's inequality, $\sup_{\theta\in\mathcal{K}}(\|g_\theta\|_V + \|P_\theta g_\theta\|_V) < \infty$, and (A3), we have

$$\mathbb{E}_{x,\theta}^{\boldsymbol{\rho}}\left[\left|T_n^{(1)}\right|^p\right] \le C\left(\sum_{k=1}^{\infty}\rho_k^2\right)^{p/2} D(\boldsymbol{\epsilon},\boldsymbol{\rho},\mathcal{K},x).$$

Since $T_n^{(1)}$ is a martingale in $\mathcal{L}^p$, then $|T_n^{(1)}|$ is a nonnegative submartingale in $\mathcal{L}^p$ and Doob's $\mathcal{L}^p$ inequality implies that

$$\mathbb{E}_{x,\theta}^{\boldsymbol{\rho}}\left[\sup_{n\ge1}\left|T_n^{(1)}\right|^p\right] \le C\left(\sum_{k=1}^{\infty}\rho_k^2\right)^{p/2} D(\boldsymbol{\epsilon},\boldsymbol{\rho},\mathcal{K},x),$$

which concludes the proof of (A.6).

Proof of (A.7). Under (A3), we have

$$\begin{aligned}
&\sup_{n\ge1}\left|\sum_{k=1}^{n-1}\rho_{k+1}\left(P_{\theta_k}g_{\theta_k}(X_k) - P_{\theta_{k-1}}g_{\theta_{k-1}}(X_k)\right)\ \mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\epsilon})\ge k+1\}}\right| \\
&\le C\sum_{k=0}^{\infty}\rho_{k+1}W(X_k)\,|\theta_k-\theta_{k-1}|^\alpha\ \mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\epsilon})\ge k+1\}}, \\
&\le C\sum_{k=0}^{\infty}\rho_{k+1}\epsilon_k^\alpha\ W(X_k)\ \mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\epsilon})\ge k+1\}}.
\end{aligned}$$

We conclude the proof by applying Minkowski's inequality.

Proof of (A.8). Under (A3),

$$\begin{aligned}
&\sup_{n\ge1}\left|\sum_{k=1}^{n-1}(\rho_{k+1}-\rho_k)P_{\theta_{k-1}}g_{\theta_{k-1}}(X_k)\ \mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\epsilon})\ge k+1\}}\right| \\
&\qquad\le C\sum_{k=1}^{\infty}(\rho_k-\rho_{k+1})W(X_k)\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\epsilon})\ge k+1\}},
\end{aligned}$$

and the proof follows from Minkowski's inequality.

Proof of (A.9). Under (A3),

$$\begin{aligned}
&\sup_{n\ge1}\left|\rho_1 P_{\theta_0}g_{\theta_0}(X_0)\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\epsilon})\ge1\}} - \rho_n P_{\theta_{n-1}}g_{\theta_{n-1}}(X_n)\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\epsilon})\ge n\}}\right|^p \\
&\qquad\le C\left(\rho_1^p W^p(X_0)\ \mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\epsilon})\ge1\}} + \sup_{n\ge1}\rho_n^p W^p(X_n)\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\epsilon})\ge n\}}\right) \\
&\qquad\le C\sum_{k=1}^{\infty}\rho_k^p W^p(X_k)\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\epsilon})\ge k\}}.
\end{aligned}$$

The proof follows from (A3) and the inequality $\sum_{k=1}^n \rho_k^p \le \left(\sum_{k=1}^n \rho_k^2\right)^{p/2}$ for $p \ge 2$.

Since $T_n^{(5)} \mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\epsilon})\ge n\}} = 0$, we have

$$S_{1,n}(\boldsymbol{\epsilon},\boldsymbol{\rho},\mathcal{K}) = T_n \mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\epsilon})\ge n\}} = \sum_{i=1}^4 T_n^{(i)} \mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\epsilon})\ge n\}}.$$

The Markov inequality and Lemma 5.1 imply that

(A.11)
$$\mathbb{P}_{x,\theta}^{\boldsymbol{\rho}}\left(\sup_{n\ge1}|S_{1,n}(\boldsymbol{\epsilon},\boldsymbol{\rho},\mathcal{K})|\ge\delta\right) \le C\delta^{-p}\left\{\left(\sum_{k=1}^\infty \rho_k^2\right)^{p/2} + \left(\sum_{k=1}^\infty \rho_k \epsilon_k^\alpha\right)^p\right\} W^p(x).$$

The proof for all l then follows from the Markov property: for all $(x,\theta)\in\mathsf{X}\times\mathcal{K}$,

$$\begin{aligned}
&\mathbb{P}_{x,\theta}^{\boldsymbol{\rho}}\left(\sup_{n\ge1}|S_{l+1,n}(\boldsymbol{\epsilon},\boldsymbol{\rho},\mathcal{K})|\ge\delta\right)\\
&\quad= \mathbb{E}_{x,\theta}^{\boldsymbol{\rho}}\left(\mathbb{P}_{X_l,\theta_l}^{\boldsymbol{\rho}^{\leftarrow l}}\left(\sup_{n\ge1}|S_{1,n}(\boldsymbol{\epsilon}^{\leftarrow l},\boldsymbol{\rho}^{\leftarrow l},\mathcal{K})|\ge\delta\right)\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\epsilon})\ge l\}}\right)\\
&\quad\le \mathbb{E}_{x,\theta}^{\boldsymbol{\rho}}\left(\sup_{\theta\in\mathcal{K}}\mathbb{P}_{X_l,\theta}^{\boldsymbol{\rho}^{\leftarrow l}}\left(\sup_{n\ge1}|S_{1,n}(\boldsymbol{\epsilon}^{\leftarrow l},\boldsymbol{\rho}^{\leftarrow l},\mathcal{K})|\ge\delta\right)\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\epsilon})\ge l\}}\right).
\end{aligned}$$

Since the sequence $\boldsymbol{\epsilon}$ is nonincreasing, there exists an integer s such that for all l and all $k\ge s$, $\epsilon_k^{\leftarrow l}\le\epsilon$, for all $k\ge s$ (where ϵ is defined in (A3)) and Lemma 5.1 shows that there exists a constant C such that for any l, for any $x\in\mathsf{X}$, and any monotone nonincreasing sequence $\boldsymbol{\rho}$,

$$\sup_{\theta\in\mathcal{K}}\sup_{k\ge0}\mathbb{E}_{x,\theta}^{\boldsymbol{\rho}^{\leftarrow l}}[W^p(X_k)\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\epsilon}^{\leftarrow l})\ge k\}}] \le CW^p(x).$$

The proof follows from (A.11).

It remains to bound $\mathbb{P}_{x,\theta}^{\boldsymbol{\rho}}(\nu(\boldsymbol{\epsilon})<\sigma(\mathcal{K})) \le \mathbb{P}_{x,\theta}^{\boldsymbol{\rho}}(\nu(\boldsymbol{\epsilon})\le\sigma(\mathcal{K}))$.

$$\begin{aligned}
\mathbb{P}_{x,\theta}^{\boldsymbol{\rho}}(\nu(\boldsymbol{\epsilon})\le\sigma(\mathcal{K})) &= \sum_{k=1}^\infty \mathbb{P}_{x,\theta}^{\boldsymbol{\rho}}(\nu(\boldsymbol{\epsilon})=k,\sigma(\mathcal{K})\ge k)\\
&= \sum_{k=1}^\infty \mathbb{P}_{x,\theta}^{\boldsymbol{\rho}}\left(|H(\theta_{k-1},X_k)|\ge\epsilon_k\rho_k^{-1},\sigma(\mathcal{K})\ge k,\nu(\boldsymbol{\epsilon})=k\right)\\
&\le C\sum_{k=1}^\infty(\epsilon_k^{-1}\rho_k)^p \sup_{k\ge0}\sup_{\theta\in\mathcal{K}}\mathbb{E}_{x,\theta}^{\boldsymbol{\rho}}[W^p(X_k)\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\epsilon})\ge k\}}].
\end{aligned}$$

The proof follows from Lemma 5.1. □

Appendix B. Proof of Proposition 6.1. The following proposition is a partial restatement of [25, Theorem 2.3].

PROPOSITION B.1. *Suppose that P is irreducible and aperiodic and that $P^m(x,\cdot)\ge\mathbb{1}_C(x)\delta\nu(\cdot)$ for a set $C\in\mathcal{B}(\mathsf{X})$, some integer m and $\delta>0$ and that there is a drift to C in the sense that, for some $\lambda<1$, b, and a function $V:\mathsf{X}\to[1,\infty)$,*

(B.1) $$PV(x)\le\lambda V(x)\quad\forall x\notin C\quad\text{and}\quad\sup_{x\in C}(V(x)+PV(x))\le b.$$

Then, there exist constants K and $\rho < 1$, depending only upon m, δ, λ, b, such that, for all $x \in \mathsf{X}$ and all $g \in \mathcal{L}_V$,

$$\|P^k g - \pi(g)\|_V \leq K\rho^n \, \|g\|_V. \tag{B.2}$$

In addition, $u = \sum_{n\geq 0}(P^k g - \pi(g))$ is a solution of the Poisson equation $u - Pu = g - \pi(g)$.

We state [25, Theorem 2.3] in the strongly aperiodic case, i.e., where C is a $(1, \delta)$ small set. Explicit but intricate expressions for K and ρ (in terms of the constants m, δ, λ, b) are given in this reference. Partial extensions to the general aperiodic case are considered in [25, Theorem 2.4], based on splitting and regeneration techniques. Sharper and simpler bounds have been recently obtained in [15] using coupling technique. This result extends to V-norm results obtained earlier for the total variation distance by [29] (see also [28]). These results have been derived in the strongly aperiodic case; extensions to the general aperiodic case can be considered in the same framework.

PROPOSITION B.2. *Assume* (DRI1)–(DRI3). *Then, there exist a constant C and $\rho < 1$ such that, for all $g \in \mathcal{L}_{V^q}$, with $q = 1$ or $q = p$ and any $k \geq 0$,*

$$\sup_{\theta\in\mathcal{K}} \|P_\theta^k g - \pi_\theta(g)\|_{V^q} \leq C\rho^k \|g\|_{V^q}, \tag{B.3}$$

$$\sup_{(\theta,\theta')\in\mathcal{K}\times\mathcal{K}} |\theta - \theta'|^{-\beta} \|P_\theta^k g - P_{\theta'}^k g\|_{V^q} \leq C\|g\|_{V^q}. \tag{B.4}$$

Proof. Equation (B.3) follows from Proposition B.1. To prove (B.4) write, for all $(\theta, \theta') \in \Theta \times \Theta$, all $n \in \mathbb{N}$, and all $g \in \mathcal{L}_{V^q}$,

$$\begin{aligned} P_\theta^n g(x) - P_{\theta'}^n g(x) &= \sum_{j=0}^{n-1} P_\theta^j (P_\theta - P_{\theta'}) P_{\theta'}^{n-j-1} g(x) \\ &= \sum_{j=0}^{n-1} P_\theta^j (P_\theta - P_{\theta'})(P_{\theta'}^{n-j-1} g(x) - \pi_{\theta'}(g)). \end{aligned}$$

Equation (B.3) shows that there exists a constant C such that, for any $l \geq 0$,

$$\sup_{\theta\in\mathcal{K}} \|P_\theta^l g - \pi_\theta(g)\|_{V^q} \leq C\|g\|_{V^q} \, \rho^l \quad \text{and} \quad \sup_{j\geq 0} \sup_{\theta\in\mathcal{K}} \|P_\theta^j V^q\|_{V^q} < \infty.$$

Under assumption (DRI3) we thus have, for any $l \geq 0$,

$$\|(P_\theta - P_\theta')(P_{\theta'}^l g(x) - \pi_{\theta'}(g))\|_{V^q} \leq C|\theta - \theta'|^\beta \|(P_{\theta'}^l g(x) - \pi_{\theta'}(g))\|_{V^q} \leq C|\theta - \theta'|^\beta \|g\|_{V^q} \rho^l,$$

which concludes the proof. □

Proof of Proposition 6.1. Under (DRI1), P_θ is positive recurrent and admits a single stationary measure π_θ, which verifies $\sup_{\theta\in\mathcal{K}} \pi_\theta(V^p) < \infty$, which implies that $\sup_{\theta\in\mathcal{K}} |h(\theta)| < \infty$.

Proof of (6.6). Let $x_0 \in \mathsf{X}$ and $k \in \mathbb{N}$. Write $h(\theta) - h(\theta') = A(\theta, \theta') + B(\theta, \theta') + C(\theta, \theta')$, where

$$A(\theta, \theta') = \left(h(\theta) - P_\theta^k H_\theta(x_0)\right) + \left(P_{\theta'}^k H_{\theta'}(x_0) - h(\theta')\right), \tag{B.5}$$

$$B(\theta, \theta') = P_\theta^k H_\theta(x_0) - P_{\theta'}^k H_\theta(x_0), \tag{B.6}$$

$$C(\theta, \theta') = P_{\theta'}^k H_\theta(x_0) - P_{\theta'}^k H_{\theta'}(x_0). \tag{B.7}$$

Propositions B.1 and B.2 show that there exist constants C and $\rho < 1$ such that, for all $(\theta, \theta') \in \mathcal{K} \times \mathcal{K}$,

$$
\begin{aligned}
&|A(\theta,\theta')| \le C\,\rho^k \sup_{\theta\in\mathcal{K}} \|H_\theta\|_V\, V(x_0),\\
&|B(\theta,\theta')| \le C \sup_{\theta\in\mathcal{K}} \|H_\theta\|_V\, |\theta-\theta'|^\beta\, V(x_0),\\
&|C(\theta,\theta')| \le \int_{\mathsf{X}} P^k_{\theta'}(x_0,dy)|H_\theta(y)-H_{\theta'}(y)| \le C|\theta-\theta'|^\beta \int_{\mathsf{X}} P^k_{\theta'}(x_0,dy)V(y)\\
&\qquad\qquad \le C|\theta-\theta'|^\beta\, V(x_0).
\end{aligned}
$$

Hence, there exists a constant C such that, for all $(\theta, \theta') \in \mathcal{K} \times \mathcal{K}$,

$$
|h(\theta) - h(\theta')| \le C\, V(x_0)\, (\rho^k + |\theta-\theta'|^\beta). \tag{B.8}
$$

The proof is concluded by setting $k = [\beta \log|\theta-\theta'|/\log(\rho)]$ (where $[x]$ is the integer part of x) if $|\theta-\theta'| \le \delta < 1$ and $k = 1$ otherwise. □

Proof of (5.4). Using (6.6) and Propositions B.1 and B.2, there exists a constant C such that, for all $(\theta, \theta') \in \mathcal{K}$, we have

$$
\begin{aligned}
&|(P^k_\theta H_\theta(x) - h(\theta)) - (P^k_{\theta'}(x)H_{\theta'}(x) - h(\theta'))|\\
&\quad \le |P^k_\theta H_\theta(x) - P^k_\theta H_{\theta'}(x)| + |P^k_\theta H_{\theta'}(x) - P^k_{\theta'} H_{\theta'}(x)| + |h(\theta)-h(\theta')|\\
&\quad \le C|\theta-\theta'|^\beta V(x).
\end{aligned}
$$

On the other hand, by Proposition B.1, there exist constants $\rho < 1$ and C such that, for all $(\theta, \theta') \in \mathcal{K} \times \mathcal{K}$,

$$
|(P^k_\theta H_\theta(x) - h(\theta)) - (P^k_{\theta'} H_{\theta'}(x) - h(\theta'))| \le C\rho^k V(x).
$$

Hence, for any s and $N \ge s$, we have

$$
\begin{aligned}
|P^s_\theta g_\theta(x) - P^s_{\theta'} g_{\theta'}(x)| &\le \sum_{k=s}^{\infty} |(P^k_\theta H_\theta(x) - h(\theta)) - (P^k_{\theta'}(x) - h(\theta'))|\\
&\le CV(x)\, \left\{ N|\theta-\theta'|^\beta + \frac{\rho^{N+s}}{1-\rho} \right\}.
\end{aligned}
$$

The proof follows by setting $N = [\beta \log|\theta-\theta'|/\log\rho]$ for $|\theta-\theta'| \le \delta < 1$, $\theta \ne \theta'$, $N = s$ otherwise, and using the fact that for any $0 < \alpha < \beta$, $|\theta-\theta'|^\beta \log|\theta-\theta'|$ $= o(|\theta-\theta'|^\alpha)$. □

Proof of (5.6). Let $\boldsymbol{\rho} = \{\rho_k, k \ge 0\}$ be a nonincreasing sequence of positive numbers and let $\mathcal{K}$ be a compact subset of Θ. (DRI1) and (6.2) shows that, for all $k \ge 0$, $l \ge 0$, and all $x \in \mathsf{X}$,

$$
\sup_{\theta\in\mathcal{K}} \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\left[V^p(X_{k+l})\mathbb{1}_{\{\sigma(\mathcal{K})\ge k+l\}} \mid \mathcal{F}_k\right] \le \kappa^l V^p(X_k)\mathbb{1}_{\{\sigma(\mathcal{K})\ge k\}}. \tag{B.9}
$$

We will show that there exist constants $\epsilon > 0$, $0 < \rho < 1$, and C such that, for all k,

$$
\mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\left[V^p(X_{k+m})\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu_\epsilon\ge k+m\}} \mid \mathcal{F}_k\right] \le \rho V^p(X_k)\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu_\epsilon\ge k\}} + C. \tag{B.10}
$$

For $n \in \mathbb{N}$, write $n = um + v$, where $v \in \{0, \dots, m-1\}$. Equation (B.10) shows that

$$
\mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\left[V^p(X_{um+v})\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu_\epsilon\ge um+v\}}\right] \le \rho^u \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}[V^p(X_v)\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu_\epsilon\ge v\}}] + \frac{C}{1-\rho}
$$

and the proof follows from (B.9). It remains to prove (B.10). We repeatedly use the following lemma adapted from [5, Lemma 3, p. 292].

LEMMA B.3. *Assume* (DRI). *Let* $\psi : \Theta \times \mathsf{X} \to \mathbb{R}$ *be a function verifying* $\sup_{\theta\in\mathcal{K}} \|\psi_\theta\|_{V^p} < \infty$. *Then, for any* $\epsilon > 0$ *and for any* $l \geq 1$ *there exists a constant* C *such that, for all* $k \geq 0$,

$$\mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\left[\psi_{\theta_k}(X_{k+l})\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu_\epsilon\geq k+l\}} \mid \mathcal{F}_k\right] \leq \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\left[P_{\theta_k}\psi_{\theta_k}(X_{k+l-1})\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu_\epsilon\geq k+l-1\}} \mid \mathcal{F}_k\right] + C\kappa^l\epsilon^\alpha \sup_{\theta\in\mathcal{K}} \|\psi_\theta\|_{V^p}\, V^p(X_k)\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu_\epsilon\geq k\}}.$$

Proof.

$$\begin{aligned}\mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\left[\psi(\theta_k, X_{k+l})\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu_\epsilon\geq k+l\}} \mid \mathcal{F}_k\right] &= \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\left[P_{\theta_{k+l-1}}\psi_{\theta_k}(X_{k+l-1})\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu_\epsilon\geq k+l\}} \mid \mathcal{F}_k\right] \\ &= \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\left[P_{\theta_k}\psi_{\theta_k}(X_{k+l-1})\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu_\epsilon\geq k+l\}} \mid \mathcal{F}_k\right] + R_l,\end{aligned}$$

where

$$R_l = \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\left[(P_{\theta_{k+l-1}} - P_{\theta_k})\psi_{\theta_k}(X_{k+l-1})\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu_\epsilon\geq k+l\}} \mid \mathcal{F}_k\right].$$

Under (DRI3), there exists a constant C such that for all $x \in \mathsf{X}$,

$$|(P_{\theta_{k+l-1}} - P_{\theta_k})\psi_{\theta_k}(x)\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu_\epsilon\geq k+l\}}| \leq C \sup_{\theta\in\mathcal{K}} \|\psi_\theta\|_{V^p}\, V^p(x)(l\epsilon)^\alpha \mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu_\epsilon\geq k+l\}}.$$

Finally, (DRI1) implies that

$$\mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\left[V^p(X_{k+l-1})\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu_\epsilon\geq k+l\}} \mid \mathcal{F}_k\right] \leq \kappa^l V^p(X_k)\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu_\epsilon\geq k\}},$$

which implies

$$|R_l| \leq C\kappa^l\,(l\epsilon)^\alpha \sup_{\theta\in\mathcal{K}} \|\psi_\theta\|_{V^p}\, V^p(X_k)\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu_\epsilon\geq k\}}. \qquad \square$$

Using repeatedly the lemma above, we may write

$$\begin{aligned}&\mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\left[V^p(X_{k+m})\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu_\epsilon\geq k+m\}} \mid \mathcal{F}_k\right] \\ &\quad\leq \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\left[P_{\theta_k}V^p(X_{k+m-1})\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu_\epsilon\geq k+m-1\}} \mid \mathcal{F}_k\right] \\ &\qquad\quad + C_m\epsilon^\alpha V^p(X_k)\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu_\epsilon\geq k\}} \\ &\quad\leq \mathbb{E}^{\boldsymbol{\rho}}_{x,\theta}\left[P^2_{\theta_k}V^p(X_{k+m-2})\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu_\epsilon\geq k+m-2\}} \mid \mathcal{F}_k\right] \\ &\qquad\quad + (C_m + C_{m-1}\kappa)\epsilon^\alpha V^p(X_k)\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu_\epsilon\geq k\}} \\ &\quad\vdots \\ &\quad\leq P^m_{\theta_k}V^p(X_k)\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu_\epsilon\geq k\}} + \left(\sum_{i=0}^{m-1} C_{m-i}\kappa^i\right)\epsilon^\alpha V^p(X_k)\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu_\epsilon\geq k\}}.\end{aligned}$$

The proof follows from (DRI) for ϵ sufficiently small. $\square$

Acknowledgments. The authors would like to thank Stas Volkov and Sumeetpal Singh for their careful reading of parts of the paper and their helpful comments. C. Andrieu would like to thank the CNRS/Royal Society Partnership and the Nuffield Foundation. É. Moulines would like to thank the CNRS/Royal Society Partnership.

REFERENCES

[1] J. Abounadi, D. P. Bertsekas, and V. Borkar, *Stochastic approximation for nonexpansive maps: Application to Q-learning algorithms*, SIAM J. Control Optim., 41 (2002), pp. 1–22.

[2] C. Andrieu and E. Moulines, *On the ergodicity properties of some Markov chain Monte Carlo algorithms*, Ann. Appl. Probab., to appear.

[3] C. Andrieu and C. P. Robert, *Controlled MCMC for Optimal Sampling*, Cahiers du Cérémade 0125, Université de Paris Dauphine, Paris, 2001.

[4] J. Bartusek, *Stochastic Approximation and Optimization of Markov Chains*, Ph.D. thesis, The Institute for Systems Research, University of Maryland, College Park, MD, 2000.

[5] A. Benveniste, M. Métivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, New York, 1990.

[6] V. Borkar, *Stability of annealing schemes and related processes*, Systems Control Lett., 41 (2000), pp. 325–331.

[7] V. S. Borkar and S. P. Meyn, *The o.d.e. method for convergence of stochastic approximation and reinforcement learning*, SIAM J. Control Optim., 38 (2000), pp. 447–469.

[8] R. Buche and H. J. Kushner, *Rate of convergence for constrained stochastic approximation algorithms*, SIAM J. Control Optim., 40 (2001), pp. 1011–1041.

[9] H. Chen, *Stochastic approxmation with state-dependent noise*, Sci. China Ser. E, 43 (2000), pp. 531–541.

[10] H. Chen, L. Guo, and A. Gao, *Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds*, Stochastic Process. Appl., 27 (1988), pp. 217–231.

[11] H. Chen and Y.-M. Zhu, *Stochastic approximation procedures with randomly varying truncations*, Sci. Sinica, 29 (1986), pp. 914–926.

[12] H.-F. Chen, *Stochastic Approximation and Its Applications*, Nonconvex Optimization and Its Applications 64, Kluwer Academic Publishers, Dordrecht, 2002.

[13] B. Delyon, *Stochastic Approximation with Decreasing Gain: Convergence and Asymptotic Theory*, Tech. report, Université de Rennes, Rennes, France, 2000.

[14] B. Delyon, M. Lavielle, and E. Moulines, *Convergence of a stochastic approximation version of the EM algorithm*, Ann. Stat., 27 (1999), pp. 94–128.

[15] R. Douc, E. Moulines, and J. Rosenthal, *Quantitative bounds on convergence of time-inhomogeneous Markov chains*, Ann. Appl. Probab., 14 (2004), pp. 1643–1665.

[16] M. Duflo, *Random Iterative Systems*, Appl. Math. 34, Springer-Verlag, Berlin, 1997.

[17] L. Gerencsér and S. S. Wilson, *Rate of convergence of recursive estimators*, SIAM J. Control Optim., 30 (1992), pp. 1200–1227.

[18] P. W. Glynn and S. P. Meyn, *A Liapounov bound for solutions of the Poisson equation*, Ann. Probab., 24 (1996), pp. 916–931.

[19] H. Haario, E. Saksman, and J. Tamminen, *An adaptive Metropolis algorithm*, Bernoulli, 7 (2001), pp. 223–242.

[20] P. Hall and C. Heyde, *Martingale Limit Theory and Its Application*, Academic Press, New York, London, 1980.

[21] S. Jarner and E. Hansen, *Geometric ergodicity of Metropolis algorithms*, Stochastic Process. Appl., 85 (2000), pp. 341–361.

[22] H. Kushner and D. Clark, *Stochastic Approximation for Constrained and Unconstrained Systems*, Springer-Verlag, Berlin, Heidelberg, 1978.

[23] H. Kushner and G. Yin, *Stochastic Approximation Algorithms and Applications*, Appl. Math. 35, Springer-Verlag, New York, 1997.

[24] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and M. Teller, *Equations of state calculations by fast computing machines*, J. Chem. Phys., 21 (1953), pp. 1087–1091.

[25] S. Meyn and R. Tweedie, *Markov Chains and Stochastic Stability*, Communication and Control Engineering Series, Springer-Verlag, London, 1993.

[26] E. Nummelin, *On the Poisson equation in the potential theory of a single kernel*, Math. Scand., 68 (1991), pp. 59–82.

[27] G. Roberts and R. Tweedie, *Geometric convergence and central limit theorem for multidimensional Hastings and Metropolis algorithms*, Biometrika, 83 (1996), pp. 95–110.

[28] G. Roberts and R. Tweedie, *Bounds on regeneration times and convergence rates for Markov chains*, Stochastic Process. Appl., 80 (1999), pp. 211–229.

[29] J. Rosenthal, *Minorization conditions and convergence rates for Markov chain Monte Carlo*, J. Amer. Statis. Assoc., 90 (1995), pp. 558–566.

[30] V. Tadic, *Stochastic gradient with random truncations*, European J. Oper. Res., 101 (1997), pp. 261–284.

[31] V. Tadic, *Stochastic approximations with random truncations, state dependent noise and discontinuous dynamics*, Stochastics Stochastics Rep., 64 (1998), pp. 283–326.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 1, pp. 313–327

A STUDY ON THE SPECTRUM OF THE SAMPLED-DATA TRANSFER OPERATOR WITH APPLICATION TO ROBUST EXPONENTIAL STABILITY PROBLEMS*

TOMOMICHI HAGIWARA†

Abstract. This paper begins by studying some spectral properties of the transfer operators of sampled-data systems described by applying the lifting technique. Through a "nonasymptotic" characterization of the transfer operator, its spectrum is determined in terms of finite-dimensional eigenvalue problems. Then, it is shown that a close connection with such eigenvalue problems and the exponential stability condition can be exploited to study the robust internal (exponential) stability problem of sampled-data systems. Since the transfer operator is relevant to input-output characteristics, the relationship between input-output stability and internal stability is also discussed in the context of sampled-data systems.

Key words. sampled-data system, spectral analysis, robust stability, L_2-stability, exponential stability

AMS subject classifications. 47A10, 47N70, 93C57, 93D09, 93D20

DOI. 10.1137/S036301290343682X

1. Introduction.

The widespread use of digital controllers has stimulated the study of sampled-data systems with their intersample behavior taken into account, and a lot of important results have been obtained since the late 1980s. Among them are the studies on the H_∞ control problem [1, 2, 3, 4] and robust stability problem [5, 6, 7, 8, 9] of sampled-data systems, as well as the continuous-time lifting technique [1, 2, 10] and the frequency response theory [11, 12]. In the studies of the H_∞ control and robust stability problems, L_2-stability [13], L_2-induced norm [14], and H_∞ norm [15] of sampled-data systems play important roles, and these notions can be dealt with also in the frequency domain with the transfer operator [1, 2, 10, 11, 15] of sampled-data systems. Also, another frequency-domain study has been conducted in [16, 17] by introducing the notion of positive-real sampled-data systems, and some phase properties of sampled-data systems were discussed. This study has been extended in [18, 19], which lead to the positive-realness approach (or the passivity approach) to the robust stability analysis of sampled-data systems. Transfer operators play an important role also in such an approach.

Thus, it is important to study the properties of the transfer operators so that the scope of the frequency-domain studies of sampled-data systems can be extended further. In this paper, we focus on the spectrum of transfer operators, and clarify some useful spectral properties. More specifically, we show that the spectrum of the transfer operator can be characterized by means of finite-dimensional eigenvalue problems. Then, it is demonstrated that such spectral analysis is indeed useful in the study of sampled-data systems by applying it to the robust internal (exponential) stability analysis of sampled-data systems. Furthermore, since the transfer operator is relevant

*Received by the editors October 26, 2003; accepted for publication (in revised form) January 15, 2005; published electronically August 22, 2005.
http://www.siam.org/journals/sicon/44-1/43682.html
†Department of Electrical Engineering, Kyoto University, Kyotodaigaku-Katsura, Nishikyo-ku, Kyoto 615-8510, Japan (hagiwara@kuee.kyoto-u.ac.jp).

to the input-output characteristics, we apply the results of the spectral analysis to relate (robust) L_2-stability and (robust) internal stability in the context of sampled-data systems. The consequent result is not surprising, but to the best knowledge of the author, studies relating robust L_2-stability and robust internal stability in the context of sampled-data systems are rare, with the study by the author and a colleague in [9] being an exception. Nevertheless, that study in [9] is limited to the case of additive and multiplicative perturbations, unfortunately, and it is quite hard to extend it to the general case. The present paper shows that our spectral analysis provides a simple and rigorous proof to relate these two stability notions. Here, it would be worth mentioning that a similar problem of relating L_2-stability and exponential stability has been studied, e.g., in [20, 21] for a class of (ordinary type of) infinite-dimensional systems, but those studies do not cover the present setting nor do our developments here follow similar techniques to those employed therein.

The contents of this paper are as follows. In section 2, we review the notion of the transfer operator $\widehat{G}(z)$ of sampled-data systems with a slight but crucial extension (i.e., its nonasymptotic characterization). This characterization allows us to introduce some appropriate nonzero initial states to the study of the mapping defined by the transfer operator, and makes it fairly easy to carry out the following discussions (e.g., the derivations of Theorem 5 and Proposition 7). In section 3, we study the spectral properties of the transfer operators, and show that they are nearly as amenable as those of compact normal operators, even though the transfer operators are generally noncompact and nonnormal. Based on these properties, we further show that the spectrum of the transfer operator $\widehat{G}(z)$ can be characterized with finite-dimensional eigenvalue problems for each z such that $\widehat{G}(z)$ is well-defined. Section 4 applies the spectral study in section 3 to the study of robust internal (exponential) stability of sampled-data systems against perturbations. More specifically, subsection 4.1 studies the case where the perturbations are identities up to a real scalar constant, and the basic result for this case is applied in subsection 4.2 to the study of robust internal stability of sampled-data systems with general perturbations. In particular, we give a rigorous proof to the equivalence of robust L_2-stability and robust internal stability when the nominal sampled-data system and perturbations are internally stable; roughly speaking, we show that whatever robust internal stability/robust performance problems we may consider in the sampled-data setting, the conditions in the L_2-stability context are enough to guarantee the robust stability/performance in the internal stability context, provided that the perturbations belong to the class of internally stable finite-dimensional LTI (linear time-invariant) or h-periodic systems (where h is the sampling period). Section 5 concludes the paper with some remarks.

We use the following notation in this paper: $\lambda(\cdot)$ denotes the set of the eigenvalues of a finite-dimensional matrix, while $\sigma(\cdot)$ denotes the spectrum of an operator. $\sigma_{le}(\cdot)$, $\sigma_{re}(\cdot)$, and $\sigma_e(\cdot)$ denote the left essential, right essential, and essential spectrum, respectively, [22]. Furthermore, whenever we refer to internal stability in what follows, it means exponential stability.

2. Transfer operators of sampled-data systems. In this paper, we deal with the sampled-data system Σ_0 shown in Figure 1, where P denotes the continuous-time generalized plant, Ψ the discrete-time controller, $\mathcal{H}$ the zero-order hold, and $\mathcal{S}$ the ideal sampler. Solid lines represent continuous-time (vector) signals, while dashed lines discrete-time (vector) signals. The underlying sampling period will be denoted by h. We assume that the state-space representations of P and Ψ are given,

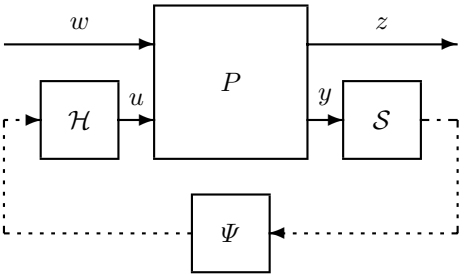


FIG. 1. *Open-loop sampled-data system* Σ_0.

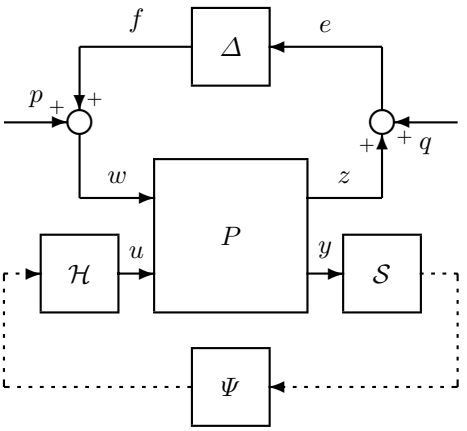


FIG. 2. *Closed-loop sampled-data system* Σ_Δ.

respectively, by

$$\begin{aligned} \frac{dx}{dt} &= Ax + B_1 w + B_2 u \\ z &= C_1 x + D_{11} w + D_{12} u \\ y &= C_2 x \end{aligned} \tag{1}$$

and

$$\begin{aligned} \xi_{k+1} &= A_\Psi \xi_k + B_\Psi y_k \\ u_k &= C_\Psi \xi_k + D_\Psi y_k, \end{aligned} \tag{2}$$

where $y_k = y(kh)$ and $u(t) = u_k$ $(kh \le t < (k+1)h)$. In the following arguments, we assume that Σ_0 is internally (exponentially) stable [13, 23]. For lack of better terminologies, we call Σ_0 an open-loop sampled-data system, while if w is given as $w = \Delta z$ with some causal mapping Δ, then we call the resulting system a closed-loop sampled-data system, which we denote by Σ_Δ. Also, the corresponding input-output mapping from $[p^T,\ q^T]^T$ to $[f^T,\ z^T]^T$ in Figure 2 will be denoted by $\mathcal{G}_\Delta$ in the following, when it is well-defined. If $\mathcal{G}_\Delta$ maps L_2 into L_2, and if its L_2-induced norm is bounded, then $\mathcal{G}_\Delta$ is said to be L_2-stable. In subsection 2.1, we review the lifted description and transfer operator [1, 2, 10, 11, 15] of the open-loop sampled-data system Σ_0. Then, in subsection 2.2, a slightly different interpretation of the transfer operators of sampled-data systems is given.

2.1. Lifted description and transfer operators of sampled-data systems. With a slight abuse of notation, the Hilbert space of square integrable vector functions over the time interval $[0, h)$ with the standard inner product will be denoted by $\mathcal{K}$, whatever the dimension of the vector may be. The Euclidean space with dimension $\dim(x)$ will be denoted by $\mathcal{F}_x$. We also define $\mathcal{F}_u$ and $\mathcal{F}_\xi$ in a similar way, and we further define $\mathcal{F} := \mathcal{F}_x \oplus \mathcal{F}_\xi$. Now, introduce the following matrices A_d, B_{d2}, and C_{d2}, and the operators $\mathbf{B}_1$, $\mathbf{C}_1$, $\mathbf{D}_{11}$, and $\mathbf{D}_{12}$:

$$A_d := \exp(Ah), \quad B_{d2} := \int_0^h \exp(A\sigma)B_2 d\sigma, \quad C_{d2} := C_2 \tag{3}$$

$$\mathbf{B}_1 : \ \mathcal{K} \ni w \mapsto \int_0^h \exp\left(A(h-\sigma)\right)B_1 w(\sigma)d\sigma \in \mathcal{F}_x \tag{4}$$

$$\mathbf{C}_1 : \ \mathcal{F}_x \ni x \mapsto z \in \mathcal{K}, \quad z(\theta) = C_1 \exp\left(A\theta\right)x \tag{5}$$

$$\mathbf{D}_{11} : \ \mathcal{K} \ni w \mapsto z \in \mathcal{K}, \quad z(\theta) = \int_0^\theta C_1 \exp\left(A(\theta-\sigma)\right)B_1 w(\sigma)d\sigma + D_{11}w(\theta) \tag{6}$$

$$\mathbf{D}_{12} : \ \mathcal{F}_u \ni u \mapsto z \in \mathcal{K}, \quad z(\theta) = \int_0^\theta C_1 \exp\left(A(\theta-\sigma)\right)B_2 d\sigma\, u + D_{12}u. \tag{7}$$

Then, the lifted description of the sampled-data system Σ_0 is given by

$$\chi_{k+1} = \mathcal{A}\chi_k + \mathcal{B}\widehat{w}_k, \quad \widehat{z}_k = \mathcal{C}\chi_k + \mathcal{D}\widehat{w}_k, \tag{8}$$

where $\chi_k := [x(kh)^T,\ \xi_k^T]^T$, and the associated transfer operator $\widehat{G}(z)$ is defined by

$$\widehat{G}(z) := \mathcal{C}(zI - \mathcal{A})^{-1}\mathcal{B} + \mathcal{D}, \tag{9}$$

where

$$\mathcal{A} := \begin{bmatrix} A_d + B_{d2}D_\Psi C_{d2} & B_{d2}C_\Psi \\ B_\Psi C_{d2} & A_\Psi \end{bmatrix} : \ \mathcal{F} \to \mathcal{F}, \quad \mathcal{B} := \begin{bmatrix} \mathbf{B}_1 \\ 0 \end{bmatrix} : \ \mathcal{K} \to \mathcal{F}$$

$$\mathcal{C} := \begin{bmatrix} \mathbf{C}_1 & \mathbf{D}_{12} \end{bmatrix} \begin{bmatrix} I & 0 \\ D_\Psi C_{d2} & C_\Psi \end{bmatrix} : \ \mathcal{F} \to \mathcal{K}, \quad \mathcal{D} := \mathbf{D}_{11} : \ \mathcal{K} \to \mathcal{K}. \tag{10}$$

In (8) above, $\widehat{w}$ and $\widehat{z}$ denote, respectively, the lifted representations of w and z (see the subsequent subsection for details). Note that $\mathcal{A}$ is a finite-dimensional matrix, and that $\widehat{G}(z)$ takes a value on the class of linear bounded operators on $\mathcal{K}$ for each z unless z is an eigenvalue of $\mathcal{A}$. The importance of $\widehat{G}(z)$ lies in that it captures all the intersample behavior (i.e., the aliasing phenomena) in Σ_0 [1, 2, 10, 15, 11].

In the following, we assume $\dim(w) = \dim(z)$ so that D_{11} is square, unless otherwise stated explicitly. Also, with a slight abuse of notation[1], the operator of multiplication by the matrix D_{11} that maps $w(\cdot) \in \mathcal{K}$ to $z(\cdot) = D_{11}w(\cdot) \in \mathcal{K}$ is also denoted by D_{11}. Then, the operator $\mathbf{D}_{11}$ given in (6), known as the compression operator, can be rewritten as $\mathbf{D}_{11} = \mathbf{D}_{110} + D_{11}$ with an obvious definition of $\mathbf{D}_{110}$, and accordingly, $\mathcal{D}$ can also be rewritten as $\mathcal{D} = \mathcal{D}_0 + D_{11}$. Then, $\mathcal{D}_0$ is compact, so that $\mathcal{D}$ (and thus $\widehat{G}(z)$) is compact if and only if $D_{11} = 0$ (see, e.g., [11]).

[1]It will be clear from the context whether D_{11} refers to the operator of multiplication or the underlying matrix. However, it would be worthwhile mentioning that whenever we refer to $\sigma_{le}(D_{11})$, $\sigma_{re}(D_{11})$, and $\sigma_e(D_{11})$, we are talking about D_{11} viewed as an operator, because otherwise these spectra are always empty.

2.2. Nonasymptotic input-output relations about EMP-signals. In this subsection, we aim at giving a "nonasymptotic" characterization of the transfer operator $\widehat{G}(z)$. That interpretation is not surprising and only a slight modification of the well known "asymptotic" interpretation of the transfer operator, and is largely a review of the preliminary part of the arguments in [24], but does play an important role in the subsequent arguments. As such, we review somewhat detailed descriptions for this interpretation. To this end, let us begin by reviewing the lifting technique used in the derivation of the lifted description of sampled-data systems. Given a (vector) signal w over the nonnegative time interval $[0, \infty)$, the lifting operation of w is defined as

$$w \mapsto \{\widehat{w}_k\}_{k=0}^{\infty}, \tag{11}$$

where $\widehat{w}_k$ is given by

$$\widehat{w}_k(\theta) = w(kh + \theta) \quad (0 \le \theta < h,\ k = 0, 1, 2, \cdots). \tag{12}$$

The signal w is called an EMP-signal of characteristic multiplier ζ [25] if its lifted representation satisfies

$$\widehat{w}_k(\theta) = \widehat{w}_0(\theta)\zeta^k \quad (0 \le \theta < h) \tag{13}$$

for some $\widehat{w}_0 \in \mathcal{K}$ and a complex number ζ, where EMP stands for "exponentially modulated periodic." In this case, let us denote the "initial function" $\widehat{w}_0$ of the EMP-signal w by $\widehat{w}_0 = \mathrm{INI}(w)$. Conversely, let us denote by $w = \mathrm{EMP}(\widehat{w}_0)$ the operation of constructing an EMP-signal w from the initial function $\widehat{w}_0$ according to (13) and then (12). Note in these notations that we suppress the underlying characteristic multiplier ζ for simplicity, and that $\mathrm{INI}(w)(\theta)$ is nothing but $w(\theta)$ for $0 \le \theta < h$.

It is a fact [11] that the output z of Σ_0 to the EMP-signal $w = \mathrm{EMP}(\widehat{w}_0)$ with characteristic multiplier $\zeta \notin \lambda(\mathcal{A})$ tends to some EMP-signal $z^\star$ of the same characteristic multiplier ζ and that the initial function $\widehat{z}_0 = \mathrm{INI}(z^\star)$ of the asymptotic response $z^\star$ is given by

$$\widehat{z}_0 = \widehat{G}(\zeta)\widehat{w}_0. \tag{14}$$

Note carefully that $\widehat{z}_0(\theta)$ $(0 \le \theta < h)$ in (14) is generally different from the actual response $z(t)$ $(0 \le t < h)$ of Σ_0 for the *zero initial state* (given by $z = \mathcal{D}\widehat{w}_0$), because the actual response z is not exactly an EMP-signal over the entire nonnegative time interval but it just tends to the EMP signal $z^\star$ as t goes to infinity.

However, given any $\widehat{w}_0 \in \mathcal{K}$, let us take $\widehat{z}_0$ given by (14) for $\zeta \notin \lambda(\mathcal{A})$, and let us construct the EMP-signals $w = \mathrm{EMP}(\widehat{w}_0)$ and $z = \mathrm{EMP}(\widehat{z}_0)$ with characteristic multiplier ζ. Then, it is easy to show that there exists an appropriate initial state χ_0 of Σ_0 (to be more precise, χ_0 is given by $(\zeta I - \mathcal{A})^{-1}\mathcal{B}\widehat{w}_0$) such that this EMP-signal input w *together with the initial state* χ_0 yields exactly the above-constructed EMP-signal output z over the entire nonnegative time interval $[0, \infty)$. Conversely, it is also easy to show that if under some initial state χ_0, the output z of Σ_0 to some EMP-signal input w with characteristic multiplier ζ is exactly an EMP-signal with the same characteristic multiplier over the whole nonnegative time interval $[0, \infty)$, then $\widehat{w}_0 = \mathrm{INI}(w)$ and $\widehat{z}_0 = \mathrm{INI}(z)$ are related by (14).

Now, let us introduce the following definition.

DEFINITION 1. *The EMP-signals w and z with the same characteristic multiplier are said to be consistent with the sampled-data system Σ_0 if there exists an initial state*

χ_0 of Σ_0 such that the input w yields exactly the output z over the entire nonnegative time interval $[0, \infty)$.

Then, the above arguments can be summarized as follows.

LEMMA 2. *Suppose that $\zeta \notin \lambda(\mathcal{A})$. The relation* (14) *holds if and only if the EMP-signals $w = EMP(\widehat{w}_0)$ and $z = EMP(\widehat{z}_0)$ with characteristic multiplier ζ are consistent with the sampled-data system Σ_0.*

3. Characterization of the spectrum of the transfer operator. The purpose of this section is to give a method for determining the spectrum of $\widehat{G}(\zeta)$ for $\zeta \notin \lambda(\mathcal{A})$. To give such a method, it is helpful to begin by studying some spectral properties of $\widehat{G}(\zeta)$. This is done in subsection 3.1, while in subsection 3.2 we give a method to determine $\sigma(\widehat{G}(\zeta))$.

3.1. Preliminary considerations on the spectrum. Let $\lambda(D_{11})$ denote the set of the eigenvalues of the matrix D_{11}. Then, it is easy to show that

$$\begin{aligned} \sigma_{le}(\widehat{G}(\zeta)) = \sigma_{re}(\widehat{G}(\zeta)) = \sigma_e(\widehat{G}(\zeta)) = \sigma_{le}(D_{11}) \\ = \sigma_{re}(D_{11}) = \sigma_e(D_{11}) = \sigma(D_{11}) = \lambda(D_{11}) \end{aligned} \tag{15}$$

(see, e.g., [22], in particular Proposition XI.4.2, and [26], in particular Corollary XXIII.2.5). Since the essential spectrum is a subset of the spectrum, it follows that $\lambda(D_{11})$ is a subset of $\sigma(\widehat{G}(\zeta))$ for any $\zeta \notin \lambda(\mathcal{A})$. Hence, to find all the points in the spectrum of $\widehat{G}(\zeta)$, it is enough for us to construct a method to check if $\gamma \notin \lambda(D_{11})$ is a point in the spectrum of $\widehat{G}(\zeta)$. Thus, we assume $\gamma \notin \lambda(D_{11})$ without loss of generality.

Since $\mathcal{D} = \mathcal{D}_0 + D_{11}$, we have

$$\gamma I - \widehat{G}(\zeta) = (\gamma I - D_{11})(I - \widehat{G}_\gamma(\zeta)), \tag{16}$$

where

$$\widehat{G}_\gamma(z) := (\gamma I - D_{11})^{-1} \left(\mathcal{C}(zI - \mathcal{A})^{-1}\mathcal{B} + \mathcal{D}_0 \right). \tag{17}$$

Hence, by (16), it is obvious that $\gamma I - \widehat{G}(\zeta)$ is invertible if and only if $I - \widehat{G}_\gamma(\zeta)$ is. Since $\widehat{G}_\gamma(\zeta)$ is a compact operator because $\mathcal{D}_0$ is, it follows that $\gamma \notin \lambda(D_{11})$ is a point in the spectrum of $\widehat{G}(\zeta)$ if and only if $\widehat{G}_\gamma(\zeta)$ has an *eigenvalue* at 1. This is an important step for the following discussion, while the following result will also be useful.

LEMMA 3. *If $\gamma_1 \in \partial\sigma(\widehat{G}(\zeta))$ and $\gamma_1 \notin \lambda(D_{11})$, then γ_1 is an isolated point of $\sigma(\widehat{G}(\zeta))$.*

Proof. By (15), the assertion follows immediately from Theorem XI.6.8 of [22]. □

Now, we are in a position to show the following result.

THEOREM 4. *$\sigma(\widehat{G}(\zeta)) \setminus \sigma_e(\widehat{G}(\zeta))$ coincides with $\sigma_p(\widehat{G}(\zeta)) \setminus \sigma_e(\widehat{G}(\zeta))$, where $\sigma_p(\cdot)$ denotes the point spectrum (i.e., the set of the eigenvalues of an operator). Furthermore, every $\gamma \in \sigma_p(\widehat{G}(\zeta)) \setminus \sigma_e(\widehat{G}(\zeta))$ is an isolated point of $\sigma(\widehat{G}(\zeta))$, and has finite multiplicity.*

Remark 3.1. The above assertion is well known for a compact operator and also for a normal operator ([22, Proposition XI.4.6]), but $\widehat{G}(\zeta)$ is generally noncompact and nonnormal. It is not hard to see that the assertion in particular implies that the accumulation points of $\sigma(\widehat{G}(\zeta))$ can exist only at $\sigma_e(\widehat{G}(\zeta)) = \lambda(D_{11})$, which consists

of finitely many points, and that $\sigma(\widehat{G}(\zeta)) \setminus \sigma_e(\widehat{G}(\zeta))$ (and thus $\sigma(\widehat{G}(\zeta))$, too) is a countable set. Hence, this proposition suggests that the properties of the spectrum of $\widehat{G}(\zeta)$ are almost as amenable as that of a compact normal operator.

Proof of Theorem 4. Since $\sigma_e(\widehat{G}(\zeta)) = \lambda(D_{11})$, for the first assertion it is enough to show that $\gamma \notin \lambda(D_{11})$ belongs to $\sigma(\widehat{G}(\zeta))$ only if it is an eigenvalue of $\widehat{G}(\zeta)$. To show this, suppose that $\gamma \notin \lambda(D_{11})$ is a point in $\sigma(\widehat{G}(\zeta))$. Then, by the arguments preceding Lemma 3, $\widehat{G}_\gamma(\zeta)$ has an eigenvalue at 1. This implies that there exists some nonzero $\widehat{w} \in \mathcal{K}$ such that $(I - \widehat{G}_\gamma(\zeta))\widehat{w} = 0$. Hence, it follows from (16) that $(\gamma I - \widehat{G}(\zeta))\widehat{w} = 0$, which implies that γ is an eigenvalue of $\widehat{G}(\zeta)$. This completes the proof for the first assertion.

As for the second assertion, it is a direct consequence from [22, Corollary XI.2.4] that $\gamma \in \sigma_p(\widehat{G}(\zeta)) \setminus \sigma_e(\widehat{G}(\zeta))$ (which we abbreviate as $\sigma_p \setminus \sigma_e$ in what follows) has finite multiplicity, since for $\gamma \notin \lambda(D_{11}) = \sigma_e(\widehat{G}(\zeta))$, $\gamma I - \widehat{G}(\zeta)$ is Fredholm [22]. Thus, it remains only to show that every $\gamma \in \sigma_p \setminus \sigma_e$ is an isolated point of $\sigma(\widehat{G}(\zeta))$. Let $\gamma \in \sigma_p \setminus \sigma_e$. If $\gamma \in \partial\sigma(\widehat{G}(\zeta))$, then the assertion follows immediately from Lemma 3. If $\gamma \notin \partial\sigma(\widehat{G}(\zeta))$, on the other hand, then $\gamma \in \sigma(\widehat{G}(\zeta))$ is an interior point of $\sigma(\widehat{G}(\zeta))$, and thus there exists some ε-neighborhood of γ contained in $\sigma(\widehat{G}(\zeta))$. This, together with the compactness of $\sigma(\widehat{G}(\zeta))$ means that we can take some number γ_1 such that $\gamma_1 \in \partial\sigma(\widehat{G}(\zeta))$ and at the same time γ_1 is not an isolated point of $\sigma(\widehat{G}(\zeta))$, where such γ_1 can always be taken so that $\gamma_1 \notin \lambda(D_{11})$ since $\lambda(D_{11})$ is only a finite set. This contradicts Lemma 3, and hence $\gamma \notin \partial\sigma(\widehat{G}(\zeta))$ cannot occur. This completes the proof. □

3.2. Reduction to a finite-dimensional eigenvalue problem. Theorem 4 tells us that in essence we have only to find the *eigenvalues* of the operator $\widehat{G}(\zeta)$ to determine its spectrum. The purpose of this subsection is to give a result with which we can characterize the eigenvalues of $\widehat{G}(\zeta)$ through a *finite-dimensional* eigenvalue problem, and this is facilitated by the "nonasymptotic" characterization of the transfer operator.

To this end, let us consider the closed-loop sampled-data system $\Sigma_{1/\gamma}$ (i.e., Σ_Δ with Δ set to $\frac{1}{\gamma}I$ in Figure 2), where γ is a nonzero complex number. Let $\zeta \notin \lambda(\mathcal{A})$, and suppose that the responses of w and z in this closed-loop sampled-data system under the input $p = 0$, $q = 0$ and some appropriate initial state $\chi_0 = [x(0)^T,\ \xi_0^T]^T$ are exactly EMP-signals of characteristic multiplier ζ over the entire nonnegative time interval. Then, it follows from Lemma 2 that (14) holds for $\widehat{w}_0 = \mathrm{INI}(w)$ and $\widehat{z}_0 = \mathrm{INI}(z)$. On the other hand, from Figure 2 (recall that $\Delta = \frac{1}{\gamma}I$), it is obvious that $\widehat{w}_0 = \frac{1}{\gamma}\widehat{z}_0$. Hence we are led to

$$(\gamma I - \widehat{G}(\zeta))\widehat{w}_0 = 0. \tag{18}$$

Thus, we can conclude that γ is an eigenvalue of $\widehat{G}(\zeta)$ if $\widehat{w}_0 \neq 0$. This suggests that we can determine the eigenvalues of the transfer operator of the open-loop sampled-data system Σ_0 by considering the responses of the closed-loop sampled-data system $\Sigma_{1/\gamma}$.

Now, when $q = 0$, the continuous-time part of $\Sigma_{1/\gamma}$ is described by

$$\frac{dx}{dt} = A_\gamma x + B_{1\gamma} p + B_{2\gamma} u, \quad z = C_{1\gamma} x + D_{11\gamma} p + D_{12\gamma} u, \quad y = C_2 x, \tag{19}$$

where

$$A_\gamma := A+B_1(\gamma I-D_{11})^{-1}C_1, \quad B_{1\gamma} := \gamma B_1(\gamma I-D_{11})^{-1}, \quad B_{2\gamma} := B_2+B_1(\gamma I-D_{11})^{-1}D_{12},$$

$$C_{1\gamma} := \gamma(\gamma I - D_{11})^{-1}C_1, \quad D_{11\gamma} := \gamma(\gamma I - D_{11})^{-1}D_{11}, \quad D_{12\gamma} := \gamma(\gamma I - D_{11})^{-1}D_{12}. \tag{20}$$

Hence, by also letting $p = 0$, the lifted description of this closed-loop "autonomous" sampled-data system $\Sigma_{1/\gamma}$ (i.e., without an external input) is given by

$$\chi_{k+1} = \mathcal{A}_\gamma \chi_k, \quad \widehat{z}_k = \mathcal{C}_\gamma \chi_k, \tag{21}$$

where $\mathcal{A}_\gamma$ and $\mathcal{C}_\gamma$ are, respectively, given by $\mathcal{A}$ and $\mathcal{C}$ in (10) with A, B_2, C_1, and D_{12} replaced by A_γ, $B_{2\gamma}$, $C_{1\gamma}$, and $D_{12\gamma}$ in (20), respectively. Note that $\mathcal{A}_\gamma$ is nothing but the state transition matrix of the "discrete-time equivalent" of $\Sigma_{1/\gamma}$.

We are in a position to state the following theorem.

THEOREM 5. *Given a complex number $\gamma \notin \lambda(D_{11})$ and a complex number $\zeta \notin \lambda(\mathcal{A})$, the operator $\widehat{G}(\zeta)$ has an eigenvalue at γ if and only if $\zeta I - \mathcal{A}_\gamma$ is not invertible.*

Proof. We first establish the assertion assuming $\gamma \neq 0$. Let us first prove the sufficiency. Suppose that $\zeta I - \mathcal{A}_\gamma$ is not invertible.

Then, by the first equation in (21), the system $\Sigma_{1/\gamma}$ has a nontrivial solution of the form

$$\chi_k = \chi_0 \zeta^k \tag{22}$$

for some nonzero initial state χ_0. Hence, by the second equation in (21), we can see that z is an EMP-signal with characteristic multiplier ζ. Since $w = \frac{1}{\gamma} z$, it follows that w is also an EMP-signal with the same characteristic multiplier (carefully note that w and z could be both zero at this stage of our discussion). Thus, by the arguments preceding this theorem, we are led to (18). Therefore, it remains only to show that $\widehat{w}_0 \neq 0$. To show this, suppose the contrary. Then, $w = 0$ so that $\Sigma_{1/\gamma}$ is essentially nothing but Σ_0 with $w = 0$. Then, the existence of the nontrivial solution (22) contradicts the assumption that $\zeta \notin \lambda(\mathcal{A})$.

To prove the necessity, suppose that $\widehat{G}(\zeta)\widehat{w}_0 = \gamma \widehat{w}_0$ for some $\widehat{w}_0 \neq 0$. Then, letting $\widehat{z}_0 := \gamma \widehat{w}_0$, Lemma 2 implies that the EMP-signals $w = \mathrm{EMP}(\widehat{w}_0)$ and $z = \mathrm{EMP}(\widehat{z}_0) = \gamma w$ with characteristic multiplier ζ are consistent with the open-loop sampled-data system Σ_0, so that these two EMP-signals can be represented as the responses of w and z in the closed-loop autonomous (i.e., $p = 0$, $q = 0$) system $\Sigma_{1/\gamma}$ for some appropriate initial state χ_0. Furthermore, since $w \neq 0$ is an EMP-signal, it follows readily that the discrete-time signal $\widehat{w}_k(\theta)$ is represented as $\widehat{w}_0(\theta)\zeta^k$, which is not identically zero as a sequence in k at least for some $\theta \in [0, h)$ (note that the "sampling" of $\widehat{w}_k(\cdot)$ at θ is well defined since the signal w is well behaved as a response of the *autonomous* sampled-data system $\Sigma_{1/\gamma}$). Thus, from a basic property of discrete-time systems, it must be true that ζ is an eigenvalue of the transition matrix of the discrete-time equivalent of $\Sigma_{1/\gamma}$ viewed at every sampling period h, which is given by $\mathcal{A}_\gamma$. This completes the proof for the case of $\gamma \neq 0$.

Finally, let us consider the case of $\gamma = 0$ (note that D_{11} is invertible in this case by the assumption $\gamma \notin \lambda(D_{11})$). In this case, it is enough to consider $\widehat{G}(\zeta) + \alpha I$ ($\alpha \neq 0$) and study the condition for it to have an eigenvalue at α. Noting that considering $\widehat{G}(\zeta) + \alpha I$ instead of $\widehat{G}(\zeta)$ is nothing but replacing D_{11} with $D_{11} + \alpha I$ (for which

$\alpha \notin \lambda(D_{11} + \alpha I)$), and observing the form of A_γ and $B_{2\gamma}$ given in (20), it is easy to see that the statement is valid even when $\gamma = 0$. □

Summarizing the arguments in this section, it follows that Theorems 4 and 5, together with (15), give a method to determine the spectrum of $\widehat{G}(\zeta)$ for each $\zeta \notin \lambda(\mathcal{A})$. That is, every point in $\lambda(D_{11})$ belongs to $\sigma(\widehat{G}(\zeta))$, and the remaining points in the spectrum can be found by searching for γ such that $\zeta I - \mathcal{A}_\gamma$ is not invertible[2]. Since the spectral radius of $\widehat{G}(\zeta)$ is no larger than $\|\widehat{G}(\zeta)\|$, it is enough to consider the disk $\{\gamma : |\gamma| \le \|\widehat{G}(\zeta)\|\}$ in such a search. An easily computable upper bound for $\|\widehat{G}(\zeta)\|$ can be obtained by an obvious extension of Theorem 1 of [27] (i.e., this theorem holds even if $|\zeta| \neq 1$) when $D_{11} = 0$; if $D_{11} \neq 0$, a simple upper bound is obtained from a triangle inequality in which the upper bound is increased by $\|D_{11}\|$. These considerations give a basis for the numerical computation of $\sigma(\widehat{G}(\zeta))$, but we do not pursue numerical studies in this paper. Instead, we advance our study to demonstrate the importance of our spectral analysis for theoretical studies such as the stability and robust stability problems of sampled-data systems.

Remark 3.2. We point out that most of the discussions in this section carries over, without essential difficulties, to the case where the generalized plant P is a finite-dimensional linear continuous-time h-periodic (FDLCP) system and Δ is an internally stable FDLCP system, where h is the sampling period; the only nontrivial point will be the treatment of the essential spectrum. However, it is not hard to see that (15) still holds with $\lambda(D_{11})$ replaced by

$$\lambda_{[0,h]}(D_{11}) := \{\lambda \,|\, \text{the set of } t \in [0,h] \text{ such that } |\det(\lambda I - D_{11}(t))| < \gamma \text{ has nonzero measure whenever } \gamma > 0\}, \tag{23}$$

which follows from section XXIII.2 of [26]. Thus, the arguments in this section still apply *mutatis mutandis*. The only point that requires some more careful arguments will be the isolatedness assertion in Theorem 4, since $\sigma_e(\widehat{G}(\zeta)) = \lambda_{[0,h]}(D_{11})$ can now form a closed curve; thus the arguments in the proof of Theorem 4 are not enough to establish the isolatedness of some eigenvalues within the essential spectrum radius. Fortunately, however, the isolatedness property is not relevant for Theorem 5, which will be used as a major tool in the following section which demonstrates the usefulness of the spectral analysis in this section.

4. Application to robust internal stability problems. In [16, 17], the positive-realness notion was introduced to sampled-data systems and some phase properties of sampled-data systems were also addressed. A more advanced study on the positive-realness of sampled-data systems was pursued, and the positive-realness gap index $\rho_{\min}$ was introduced in [18, 19]. It was also shown in [19] that this index plays an important role in the positive-realness approach (or the passivity approach) to the stability analysis of sampled-data systems. In this section, we first review the above-mentioned study in [19] briefly, and then show that our study in the preceding sections has an important application to such or more general stability and robust stability analysis.

In subsection 4.1, we deal with the gain margin analysis problem of sampled-data systems and derive some useful results by applying the spectral analysis in the

[2]It would be possible to determine the multiplicity of γ as an eigenvalue of $\widehat{G}(\zeta)$ by considering the geometric multiplicity of ζ as an eigenvalue of $\zeta I - \mathcal{A}_\gamma$, if we introduce some sort of controllability/observability conditions. However, we do not pursue this direction in this paper.

preceding section (in particular, Theorem 5). Then, such results will be applied in subsection 4.2 to give a result about robust stability of sampled-data systems.

4.1. Gain margin analysis for internal stability in the sampled-data context. The transfer operator $\widehat{G}(z)$ of the internally stable sampled-data system Σ_0 is said to be strongly positive-real [19] if there exists a positive number ε such that

$$\widehat{G}(z) + \widehat{G}(z)^* \geq \varepsilon I \quad (\forall |z| \geq 1). \tag{24}$$

The transfer operator $\widehat{G}(z)$ is not strongly positive-real, in general, but we can consider the following number:

$$\rho_{\min} := \inf_{\rho > 0} \left\{ \widehat{G}(z) + \rho I \text{ is strongly positive-real} \right\} \geq 0. \tag{25}$$

This number is called the positive-realness gap index for Σ_0, and plays an important role in the stability analysis as shown in [19]; one important result shown about this index is that for $k > 0$, the (negative feedback) closed-loop sampled-data system Σ_{-k} (i.e., Σ_Δ with Δ set to $-kI$), or to be more precise, the input-output mapping $\mathcal{G}_{-k}$ is L_2-stable if $0 < k < k_{\max}^{\mathrm{PR}}$, where

$$k_{\max}^{\mathrm{PR}} := 1/\rho_{\min}. \tag{26}$$

By a suitable construction of the generalized plant P, this "gain margin analysis problem" in the context of sampled-data control can represent a sort of stability-radius analysis problem with respect to the uncertainties in the physical parameters of the plant, and as such, to compute $k_{\max}^{\mathrm{PR}}$ is quite important. In [19], an efficient finite-dimensional state-space method for the computation of $\rho_{\min}$ and thus $k_{\max}^{\mathrm{PR}}$ was given. Furthermore, an iterative procedure was given to compute the number $k_{\max} (\geq k_{\max}^{\mathrm{PR}})$, which is defined as the largest $\bar{k}$ such that Σ_{-k} $(k > 0)$ is *internally* stable for all $k < \bar{k}$. However, in the derivation of that procedure, the following result was used *without proof*; we now give its proof by applying the spectral analysis results in the preceding section so that the procedure for computing $k_{\max}$ given in [19] is validated rigorously.

PROPOSITION 6. *If Σ_0 is internally stable, then Σ_{-k} is internally stable for all $k \in (0, k_{\max}^{\mathrm{PR}})$.*

Proof. Now, suppose the contrary. Then, by the definition of internal stability [13, 23], there exists some $k^\star \in (0, k_{\max}^{\mathrm{PR}})$ such that the state transition matrix of $\Sigma_{-k^\star}$ has an eigenvalue on or outside the unit circle, say at $\zeta^\star$. Therefore, it follows from the definition of $\mathcal{A}_\gamma$ that if we put $-\rho^\star := -1/k^\star$, then $\zeta^\star I - \mathcal{A}_{-\rho^\star}$ is not invertible where $|\zeta^\star| \geq 1$. Here, note that

$$\rho^\star > \rho_{\min} \geq 0 \tag{27}$$

since $0 < k^\star < k_{\max}^{\mathrm{PR}}$ by the assumption. Thus, by the properties of strongly positive-real transfer operators [19], we have $D_{11} + D_{11}^T + 2\rho^\star I > 0$. This in particular implies that $-\rho^\star \notin \lambda(D_{11})$. Summarizing the above arguments and applying Theorem 5, we are led to the conclusion that $\widehat{G}(\zeta^\star)$ has an eigenvalue at $-\rho^\star$. This in particular implies that $\widehat{G}(\zeta^\star) + \widehat{G}(\zeta^\star)^* + 2\rho^\star I \not> 0$. Since $|\zeta^\star| \geq 1$, it follows that $\widehat{G}(z) + \rho^\star I$ is not a strongly positive-real transfer operator, and hence $\rho^\star \leq \rho_{\min}$. This clearly contradicts (27). Hence, we have established that Σ_{-k} is internally stable for all $k \in (0, k_{\max}^{\mathrm{PR}})$. ∎

It would be worthwhile mentioning that the above proposition can be extended to Σ_k $(k > 0)$ by considering $-\widehat{G}(z)$ instead and redefining $\rho_{\min}$ and thus $k_{\max}^{\mathrm{PR}}$ accordingly.

Next, we claim Proposition 7 given below, which plays a crucial role in the robust stability analysis in the following subsection.

PROPOSITION 7. *Suppose that Σ_0 is internally stable. Then, for each fixed k, Σ_k is internally stable if and only if $\mathcal{G}_k$ is L_2-stable.*

Remark 4.1. This proposition does not say that an (open-loop) sampled-data system is internally stable if and only if it is L_2-stable, an assertion (the sufficiency part) that is obviously false; note that this proposition deals only with such (closed-loop) sampled-data systems that can arise from an internally stable sampled-data system. Note also that the condition $1/k \notin \lambda(D_{11})$, which is the well-posedness condition of Σ_k and at the same time the well-definedness condition for $\mathcal{G}_k$, is implicit in the proposition. To be more precise, if we say that Σ_k is internally stable or L_2-stable, it in particular means that $1/k \notin \lambda(D_{11})$. Further, note that this proposition shows in particular that $k_{\max,L_2} = k_{\max}$, where $k_{\max,L_2}$ denotes the largest number $\bar{k}$ such that $\mathcal{G}_k$ is L_2-stable whenever $0 < k < \bar{k}$, while $k_{\max}$ has been defined similarly but in terms of internal stability. It should be stressed, however, that the assertion of the above proposition holds even if $k > k_{\max}$, as long as Σ_k or $\mathcal{G}_k$ is stable for such k.

Proof of Proposition 7. Let us begin with the necessity part. It is not hard to see by the inspection of the procedure for computing $k_{\max}$ given in [19] that the set of k for which Σ_k is internally stable is a subset of those k for which Σ_k is L_2-stable[3]. Hence, the necessity assertion follows immediately. (Instead of this proof that is based on the transfer operator $\widehat{G}(z)$ and the spectral analysis results in the preceding section as a whole, an alternative proof is also possible in which we use the well-known fact that an internally stable sampled-data system is L_2-stable under mild conditions [13, 23].)

To show the sufficiency part, we assume that Σ_k is not internally stable for some k (i.e., $\zeta I - \mathcal{A}_\gamma$ with $\gamma := 1/k$ is not invertible for some $|\zeta| \geq 1$), and show that $\mathcal{G}_k$ is not L_2-stable. Here, it is enough to assume $1/k \notin \lambda(D_{11})$, because otherwise $\mathcal{G}_k$ is not L_2-stable as stated in Remark 4.1.

Since $\zeta \notin \lambda(\mathcal{A})$ by the internal stability assumption of Σ_0, it follows from Theorem 5 that $\widehat{G}(\zeta)\widehat{w}_0 = \gamma\widehat{w}_0$ for some $\widehat{w}_0 \neq 0$. Hence by Lemma 2, there exists an appropriate initial state $\chi_0 = \chi^\star$ of Σ_0 such that $w = \mathrm{EMP}(\widehat{w}_0)$ and $z = \gamma\mathrm{EMP}(\widehat{w}_0)$ are consistent with Σ_0 (the corresponding characteristic multiplier for the EMP signals is ζ throughout the proof). Now, let us denote by $z^\star$ the response of z in Σ_0 when its initial state is $\chi_0 = \chi^\star$ and its input is $w = 0$. Note that $z^\star \in L_2$ by the internal stability of Σ_0, which is well known, e.g., in the context of sampled-data H_2 problem [23]. Also, by linearity, it follows immediately that Σ_0 yields the response $z = \gamma\mathrm{EMP}(\widehat{w}_0) - z^\star$ when the initial state is $\chi_0 = 0$ and the input is $w = \mathrm{EMP}(\widehat{w}_0)$. Hence, it is easy to see that Σ_Δ with $\Delta = (1/\gamma)I$ yields $f = w = \mathrm{EMP}(\widehat{w}_0)$, $z = \gamma\mathrm{EMP}(\widehat{w}_0) - z^\star$, and $e = \gamma\mathrm{EMP}(\widehat{w}_0)$ when the initial state is $\chi_0 = 0$ and the inputs are $p = 0 \in L_2$ and

[3]In a sense, a direct application of the arguments in [19] is limited only to $k \in [0, k_{\max}]$. However, if Σ_k is internally stable for some $k = k^\star \notin [0, k_{\max}]$, then we can readily introduce a modified generalized plant $P_{k^\star}$ such that Σ_k can be viewed as $\Sigma_{k'}^\star$ (and thus $\Sigma_{k^\star}$ can be viewed as the open-loop sampled-data system $\Sigma_0^\star$), where $k' = k - k^\star$ and $\Sigma_k^\star$ denotes Σ_k with P replaced by $P_{k^\star}$ (this idea of introducing a modified generalized plant is quite similar to that employed in the procedure for computing $k_{\max}$; see [19] for details). Hence, we can repeat the same arguments on $\Sigma_k^\star$, which implies that the assertion can be established even for those k around $k^\star \notin [0, k_{\max}]$.

$q = z^\star \in L_2$. Noting that none of these f, w, z, and e belong to L_2 since $|\zeta| \geq 1$, we can conclude that $\mathcal{G}_k$ is not L_2-stable. □

Remark 4.2. Even though it can be seen that Proposition 6 is implied by Proposition 7, it should be noted that the independent proof of Proposition 6 is indispensable. This is because in the proof of Proposition 7, we have referred to the procedure for computing $k_{\max}$ stated in [19], which in turn has been validated rigorously by the very proof of Proposition 6.

4.2. Robust internal stability of sampled-data systems. Now, we are in a position to demonstrate the significance of Proposition 7, not merely in justifying the arguments of [19] about the computation of $k_{\max}$ through Proposition 6. More specifically, we give the following theorem about robust stability of the sampled-data system Σ_Δ, which clarifies the relationship between robust L_2-stability and robust internal (exponential) stability.

THEOREM 8. *Consider the closed-loop sampled-data system Σ_Δ shown in Figure* 2, *where we assume that D_{11} is possibly nonsquare and $\Delta \in \boldsymbol{\Delta}$ for some set $\boldsymbol{\Delta}$ of (possibly nonsquare) finite-dimensional linear time-invariant (FDLTI) internally stable systems, and suppose that Σ_0 is internally stable. Then, Σ_Δ is internally stable for all $\Delta \in \boldsymbol{\Delta}$ if and only if $\mathcal{G}_\Delta$ is L_2-stable for all $\Delta \in \boldsymbol{\Delta}$.*

Proof. The assertion is almost just a direct consequence from Proposition 7, but we need some careful arguments.

Let us begin with the sufficiency proof. Consider the (series-connected) open-loop sampled-data system $\Sigma_0\Delta$, and observe that it can be represented as the open-loop sampled-data system shown in Figure 1 with the generalized plant P replaced by $P_\Delta := P\,\mathrm{diag}[\Delta, I]$ (for which the "D_{11} matrix" is square and thus the preceding arguments apply). Also, by the internal stability assumptions, $\Sigma_0\Delta$ is internally stable. Thus, this reformation corresponds to replacing Σ_0 and Δ by $\Sigma_0\Delta$ and 1, respectively, if we interpret it in the closed-loop sampled-data system in Figure 2; let us denote by $\mathcal{G}'_\Delta$ the corresponding input-output mapping of thus reformed closed-loop sampled-data system. Since $\mathcal{G}_\Delta$ is L_2-stable for each $\Delta \in \boldsymbol{\Delta}$, it is straightforward to show that $\mathcal{G}'_\Delta$ is also L_2-stable for each $\Delta \in \boldsymbol{\Delta}$. Thus, applying Proposition 7 with $k = 1$ leads to the assertion.

To show the necessity, we also consider the closed-loop sampled-data system Σ_Δ with Σ_0 and Δ replaced by $\Delta\Sigma_0$ and 1, respectively, and denote by $\mathcal{G}''_\Delta$ the corresponding input-output mapping. Applying Proposition 7 with $k = 1$, it follows readily that both $\mathcal{G}'_\Delta$ and $\mathcal{G}''_\Delta$ are L_2-stable. In view of the linearity of Σ_0 and Δ (and thus the mapping $\mathcal{G}_\Delta$), it is not hard to show that L_2-stability of $\mathcal{G}'_\Delta$ and $\mathcal{G}''_\Delta$ implies that of $\mathcal{G}_\Delta$.

This completes the proof. □

Remark 4.3. If we recall Remark 3.2, it is not hard to see that Proposition 7 and Theorem 8 still hold even when P (and/or Δ) is a finite-dimensional linear continuous-time h-periodic system. Also, concerning Figure 2, a typical interpretation is that P_{22} (the subsystem from u to y) denotes the nominal plant, while the perturbed (actual) plant is represented by the upper LFT (linear fractional transformation) $\mathcal{F}_u(P, \Delta) =: P_{22\Delta}$. Since uncontrollable/unobservable modes do not affect internal (exponential) stability if and only if they are stable, this theorem in particular says that

(i) any robust stability condition such as the small-gain condition in particular guarantees that no unstable pole/zero cancellations can occur within $P_{22\Delta}$, irrespectively of Δ belonging to the perturbation set $\boldsymbol{\Delta}$ that the condition takes care of,

and it is obvious that

(ii) the discrete-time controller Ψ internally (exponentially) stabilizes any stabilizable detectable plants whose minimal realization coincides with that of $P_{22\Delta}$ for some Δ in the corresponding $\boldsymbol{\Delta}$.

In particular, (ii) implies that Ψ could internally stabilize also a lower-order plant than the generalized plant (because of stable pole/zero cancellations via Δ), even though this might not be necessarily clear in the above arguments where we treated Δ in such a way that it has an *independent state* from the nominal plant.

One important perturbation set $\boldsymbol{\Delta}$ is the set of norm-bounded LTI perturbations, for which the small-gain condition can be very conservative [6, 8]. To get around such conservatism, a necessary and sufficient condition has been derived [6, 8] *in the L_2-stability context*. The above theorem guarantees that such a necessary and sufficient condition is indeed necessary and sufficient for robust *internal* stability; this has already been shown for special perturbations (i.e., additive and multiplicative perturbations), including the case of unstable perturbations [9]. The above theorem in particular extends the previous results to the case of general perturbation structures but with stable perturbations.

To state the importance of this theorem more generally and precisely, we can rephrase it in the following way: whatever allowable perturbation sets/structures we may consider (for example, LTI/h-periodic perturbations, full-block/block-diagonal structures, real-parametric/dynamical perturbations, norm-bounded/unbounded perturbations, connected/nonconnected perturbation sets, convex/nonconvex perturbations with respect to the origin[4], and so on, as well as their arbitrary combinations), considering robust L_2-stability is enough to ensure robust *internal* stability; once a condition for robust L_2-stability under such FDLTI *stable* perturbations is established somehow, the condition automatically guarantees robust *internal* stability. It would also be worth stressing that the theorem applies even to such cases where some part of the perturbations are fictitious and introduced just for the robust performance analysis/synthesis (so that they do not affect internal stability) while the remaining part of the perturbations does represent the plant uncertainty, as in the main loop theorem. Robust internal stability is obviously guaranteed by robust L_2-stability even in such cases with robust performance taken into consideration.

5. Conclusion. In this paper, we first gave a nonasymptotic characterization of the transfer operator $\widehat{G}(z)$ of sampled-data systems, so that some appropriate nonzero initial states can be introduced into the study of the transfer operator $\widehat{G}(z)$. Based on such a characterization, we then studied the spectral properties of the transfer operator $\widehat{G}(z)$. More specifically, it was shown that the properties of the spectrum of $\widehat{G}(z)$ are nearly as amenable as those of compact normal operators, in spite of the generally noncompact and nonnormal nature of $\widehat{G}(z)$, and that the spectrum can be characterized with finite-dimensional eigenvalue problems. Exploiting a close relation with the eigenvalue problems and the condition for internal stability of sampled-data systems, we further extended our arguments on the spectral analysis of $\widehat{G}(z)$ to the robust internal stability analysis of sampled-data systems. To summarize the results very concisely, what we have shown is that robust L_2-stability and robust internal (exponential) stability are equivalent irrespective of the perturbation structures/sets to be considered, if the nominal sampled-data system is internally stable, if the

[4]We say here that $\boldsymbol{\Delta}$ is convex with respect to the origin if $\Delta \in \boldsymbol{\Delta}$ implies $k\Delta \in \boldsymbol{\Delta}$ for all $k \in [0, 1]$.

perturbations are either finite-dimensional LTI or h-periodic, and if the perturbations are internally stable. Although we confined our input-output stability notion to L_2-stability in this paper, it is not hard to see that Proposition 7 and thus Theorem 8 hold even if L_2-stability is replaced by L_p-stability, where $1 \leq p < \infty$. Hence, a solid theoretical basis is established for the robust stabilization/performance design for sampled-data systems even when it is carried out only under such input-output stability conditions as in [6]. Finally, it will be worthwhile mentioning that all the results in section 4 can be specialized to the continuous-time setting without any changes (since a continuous-time system can always be embedded into the class of sampled-data systems), and can readily be generalized to the discrete-time setting.

REFERENCES

[1] B. A. Bamieh and J. B. Pearson, *A general framework for linear periodic systems with applications to H_∞ sampled-data control*, IEEE Trans. Automat. Control, 37 (1992), pp. 418–435.

[2] H. T. Toivonen, *Sampled-data control of continuous-time systems with an H_∞ optimality criterion*, Automatica J. IFAC, 28 (1992), pp. 45–54.

[3] P. T. Kabamba and S. Hara, *Worst-case analysis and design of sampled-data control systems*, IEEE Trans. Automat. Control, 38 (1993), pp. 1337–1357.

[4] Y. Hayakawa, Y. Yamamoto, and S. Hara, *H_∞ type problem for sampled-data control systems—A solution via minimum energy characterization*, IEEE Trans. Automat. Control, 39 (1994), pp. 2278–2284.

[5] N. Sivashankar and P. P. Khargonekar, *Robust stability and performance analysis of sampled-data systems*, IEEE Trans. Automat. Control, 38 (1993), pp. 58–69.

[6] G. Dullerud and K. Glover, *Robust stabilization of sampled-data systems to structured LTI perturbations*, IEEE Trans. Automat. Control, 38 (1993), pp. 1497–1508.

[7] G. E. Dullerud and K. Glover, *Analysis of structured LTI uncertainty in sampled-data systems*, Automatica J. IFAC, 31 (1995), pp. 99–113.

[8] Y. Oishi, *Computation-oriented expression of a non-conservative condition for robust stability of sampled-data systems*, Internat J. Control, 62 (1995), pp. 1085–1104.

[9] T. Hagiwara and M. Araki, *Robust stability of sampled-data systems under possibly unstable additive/multiplicative perturbations*, IEEE Trans. Automat. Control, 43 (1998), pp. 1340–1346.

[10] Y. Yamamoto, *A function space approach to sampled data systems and tracking problems*, IEEE Trans. Automat. Control, 39 (1994), pp. 703–713.

[11] Y. Yamamoto and P. P. Khargonekar, *Frequency response of sampled-data systems*, IEEE Trans. Automat. Control, 41 (1996), pp. 166–176.

[12] M. Araki, Y. Ito, and T. Hagiwara, *Frequency response of sampled-data systems*, Automatica J. IFAC, 32 (1996), pp. 483–497.

[13] T. Chen and B. A. Francis, *Input-output stability of sampled-data systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 50–58.

[14] T. Chen and B. Francis, *On the L_2-induced norm of a sampled-data system*, Systems Control Lett. 15 (1990), pp. 211–219.

[15] Y. Yamamoto, *On the state space and frequency domain characterization of H_∞-norm of sampled-data systems*, Systems Control Lett. 21 (1993), pp. 163–172.

[16] K. Sugimoto and M. Suzuki, *On γ-positive real sampled-data control systems*, Proceedings of the 13th International Symposium on Mathematical Theory of Networks and Systems, Padova, Italy, 1998, pp. 409–412.

[17] K. Sugimoto and M. Suzuki, *On γ-positive real sampled-data control systems and their phase property*, (in Japanese), Transactions of the Society of Instrument and Control Engineers, 35 (1999), pp. 71–76.

[18] T. Hagiwara, *Nyquist stability criterion and positive-realness of sampled-data systems*, Systems Control Lett. 45 (2002), pp. 283–291.

[19] T. Hagiwara and T. Mugiuda, *Positive-realness analysis of sampled-data systems and its applications*, Automatica J. IFAC, 40 (2004), pp. 1043–1051.

[20] Y. Yamamoto and S. Hara, *Relationships between internal and external stability for infinite-dimensional systems with applications to a servo problem*, IEEE Trans. Automat. Control, 33 (1988), pp. 1044–1052.

[21] Y. Yamamoto and S. Hara, *Internal and external stability and robust stability condition for a class of infinite-dimensional systems*, Automatica J. IFAC, 28 (1992), pp. 81–93.
[22] J. B. Conway, *A Course in Functional Analysis*, 2nd ed., Springer-Verlag, New York, 1990.
[23] T. Chen and B. Francis, *Optimal Sampled-Data Control Systems*, Springer-Verlag, Berlin, 1995.
[24] T. Hagiwara, *Spectral analysis and singular value computations of the noncompact frequency response and compression operators in sampled-data systems*, SIAM J. Control Optim., 41 (2002), pp. 1350–1371.
[25] N. M. Wereley, *Analysis and Control of Linear Periodically Time Varying Systems*, Ph.D. thesis, Dept. of Aeronautics and Astronautics, MIT, Cambridge, MA, 1990.
[26] I. Gohberg, S. Goldberg, and M. A. Kaashoek, *Classes of Linear Operators, Vol.* II, Birkhäuser, Basel, Switzerland, 1993.
[27] T. Hagiwara, M. Suyama, and M. Araki, *Upper and lower bounds of the frequency response gain of sampled-data systems*, Automatica J. IFAC, 37 (2001), pp. 1363–1370.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 1, pp. 328–348

STOCHASTIC APPROXIMATIONS AND DIFFERENTIAL INCLUSIONS*

MICHEL BENAÏM†, JOSEF HOFBAUER‡, AND SYLVAIN SORIN§

Abstract. The dynamical systems approach to stochastic approximation is generalized to the case where the mean differential equation is replaced by a differential inclusion. The limit set theorem of Benaïm and Hirsch is extended to this situation. Internally chain transitive sets and attractors are studied in detail for set-valued dynamical systems. Applications to game theory are given, in particular to Blackwell's approachability theorem and the convergence of fictitious play.

Key words. stochastic approximation, differential inclusions, set-valued dynamical systems, chain recurrence, approachability, game theory, learning, fictitious play

AMS subject classifications. 62L20, 34G25, 37B25, 62P20, 91A22, 91A26, 93E35, 34F05

DOI. 10.1137/S0363012904439301

1. Introduction.

1.1. Presentation. A powerful method for analyzing stochastic approximations or recursive stochastic algorithms is the so-called ODE (ordinary differential equation) method, which allows us to describe the limit behavior of the algorithm in terms of the asymptotics of a certain ODE,

$$\frac{dx}{dt} = F(x),$$

obtained by suitable averaging.

This method was introduced by Ljung [24] and extensively studied thereafter (see, e.g., the books by Kushner and Yin [23] or Duflo [14] for a comprehensive introduction and further references). However, until recently most works in this direction have assumed the simplest dynamics for F, for example, that F is linear or given by the gradient of a cost function. While this type of assumption makes perfect sense in engineering applications (where algorithms are often designed to minimize a cost function), there are several situations, including models of learning or adaptive behavior in games, for which F may have more complicated dynamics.

In a series of papers Benaïm [2, 3] and Benaïm and Hirsch [5] have demonstrated that the asymptotic behavior of stochastic approximation processes can be described with a great deal of generality beyond gradients and other simple dynamics. One of their key results is that the limit sets of the process are almost surely *compact connected attractor free* (or *internally chain transitive* in the sense of Conley [13]) for the deterministic flow induced by F.

*Received by the editors January 6, 2004; accepted for publication (in revised form) November 23, 2004; published electronically August 22, 2005. This research was partially supported by the Austrian Science Fund P15281 and the Swiss National Science Foundation grant 200021-1036251/1.
http://www.siam.org/journals/sicon/44-1/43930.html

†Institut de Mathematiques, Université de Neuchâtel, Rue Emile-Argand 11, Neuchâtel, Switzerland (michel.benaim@unine.ch).

‡Department of Mathematics, University College London, London WC1E 6BT, UK, and Institut für Mathematik, Universität Wien, Nordbergstrasse 15, 1090 Wien, Austria (jhofb@math.ucl.ac.uk).

§Laboratoire d'Econométrie, Ecole Polytechnique, 1 rue Descartes, 75005 Paris, France, and Equipe Combinatoire et Optimisation, UFR 929, Université P. et M. Curie - Paris 6, 175 Rue du Chevaleret, 75013 Paris, France (sorin@math.jussieu.fr).

The purpose of this paper is to show that such a dynamical system approach easily extends to the situation where the mean ODE is replaced by a differential inclusion. This is strongly motivated by certain problems arising in economics and game theory. In particular, the results here allow us to give a simple and unified presentation of Blackwell's approachability theorem, Smale's results on the prisoner's dilemma, and convergence of fictitious play in potential games. Many other applications[1] will be considered in a forthcoming paper, by Benaïm, Hofbauer, and Sorin [7], the present one being mainly devoted to theoretical issues.

The organization of the paper is as follows. Part 1 introduces the different notions of solutions, perturbed solutions, and stochastic approximations associated with a differential inclusion. Part 2 is devoted to the presentation of two classes of examples. Part 3 is a general study of the dynamical system defined by a differential inclusion. The main result (Theorem 3.6) on the limit set of a perturbed solution being internally chain transitive is stated. Then related notions—invariant and attracting sets, attractors, and Lyapunov functions—are analyzed. Part 4 contains the proof of the limit set theorem. Finally, Part 5 applies the previous results to two adaptive processes in game theory: approachability and fictitious play.

1.2. The differential inclusion. Let F denote a set-valued function mapping each point $x \in \mathbb{R}^m$ to a set $F(x) \subset \mathbb{R}^m$. We suppose throughout that the following holds.

Hypothesis 1.1 (standing assumptions on F).

(i) F is a closed set-valued map. That is,

$$\mathrm{Graph}(F) = \{(x, y) : y \in F(x)\}$$

is a closed subset of $\mathbb{R}^m \times \mathbb{R}^m$.

(ii) $F(x)$ is a nonempty compact convex subset of $\mathbb{R}^m$ for all $x \in \mathbb{R}^m$.

(iii) There exists $c > 0$ such that for all $x \in \mathbb{R}^m$

$$\sup_{z \in F(x)} \|z\| \leq c(1 + \|x\|),$$

where $\|\cdot\|$ denotes any norm on $\mathbb{R}^m$.

Definition I. *A* solution *for the differential inclusion*

$$\frac{d\mathbf{x}}{dt} \in F(\mathbf{x}) \tag{I}$$

with initial point $x \in \mathbb{R}^m$ is an absolutely continuous mapping $\mathbf{x} : \mathbb{R} \to \mathbb{R}^m$ such that $\mathbf{x}(0) = x$ and

$$\frac{d\mathbf{x}(t)}{dt} \in F(\mathbf{x}(t))$$

for almost every $t \in \mathbb{R}$.

Under the above assumptions, it is well known (see Aubin and Cellina [1, Chapter 2.1] or Clarke et al. [12, Chapter 4.1]) that (I) admits (typically nonunique) solutions through every initial point.

[1] As pointed out to us by an anonymous referee, applications to resource sharing may be considered as in Buche and Kushner [11], where the dynamics are given by a differential inclusion. Possible applications to engineering include dry friction; see, e.g., Kunze [22].

Remark 1.2. Suppose that a differential inclusion is given on a compact convex set $C \subset \mathbb{R}^m$, of the form $F(x) = \Phi(x) - x$, such that $\Phi(x) \subset C$ for all $x \in C$ and Φ satisfies Hypothesis 1.1(i) and (ii), with $\mathbb{R}^m$ replaced by C. Then we can extend it to a differential inclusion defined on the whole space $\mathbb{R}^m$: For $x \in \mathbb{R}^m$ let $P(x) \in C$ denote the unique point in C closest to x, and define $F(x) = \Phi(P(x)) - x$. Then F satisfies Hypothesis 1.1.

1.3. Perturbed solutions. The main object of this paper is paths which are obtained as certain (deterministic or random) perturbations of solutions of (I).

DEFINITION II. *A continuous function* $\mathbf{y} : \mathbb{R}_+ = [0, \infty) \to \mathbb{R}^m$ *will be called a* perturbed solution *to* (I) *(we also say a perturbed solution to* F*) if it satisfies the following set of conditions* (II):

(i) $\mathbf{y}$ *is absolutely continuous.*

(ii) *There exists a locally integrable function* $t \mapsto U(t)$ *such that*

(a)
$$\lim_{t\to\infty} \sup_{0\le v\le T} \left\| \int_t^{t+v} U(s)\, ds \right\| = 0$$
for all $T > 0$*; and*

(b) $\frac{d\mathbf{y}(t)}{dt} - U(t) \in F^{\delta(t)}(\mathbf{y}(t))$ *for almost every* $t > 0$*, for some function* $\delta : [0, \infty) \to \mathbb{R}$ *with* $\delta(t) \to 0$ *as* $t \to \infty$*. Here* $F^\delta(x) := \{y \in \mathbb{R}^m : \exists z : \|z - x\| < \delta,\ d(y, F(z)) < \delta\}$ *and* $d(y, C) = \inf_{c\in C} \|y - c\|$.

The purpose of this paper is to investigate the long-term behavior of $\mathbf{y}$ and to describe its limit set
$$L(\mathbf{y}) = \bigcap_{t\ge 0} \overline{\{\mathbf{y}(s) : s \ge t\}}$$
in terms of the dynamics induced by F.

1.4. Stochastic approximations. As will be shown here, a natural class of perturbed solutions to F arises from certain stochastic approximation processes.

DEFINITION III. *A discrete time process* $\{x_n\}_{n\in\mathbb{N}}$ *living in* $\mathbb{R}^m$ *is a* solution for (III) *if it verifies a recursion of the form*

$$x_{n+1} - x_n - \gamma_{n+1} U_{n+1} \in \gamma_{n+1} F(x_n), \tag{III}$$

where the characteristics γ *and* U *satisfy*

- $\{\gamma_n\}_{n\ge 1}$ *is a sequence of nonnegative numbers such that*
$$\sum_n \gamma_n = \infty, \qquad \lim_{n\to\infty} \gamma_n = 0;$$
- $U_n \in \mathbb{R}^m$ *are (deterministic or random) perturbations.*

To such a process is naturally associated a continuous time process as follows.

DEFINITION IV. *Set*
$$\tau_0 = 0 \quad \textit{and} \quad \tau_n = \sum_{i=1}^{n} \gamma_i \quad \textit{for } n \ge 1,$$
and define the continuous time affine interpolated process $\mathbf{w} : \mathbb{R}_+ \to \mathbb{R}^m$ *by*

$$\mathbf{w}(\tau_n + s) = x_n + s\frac{x_{n+1} - x_n}{\tau_{n+1} - \tau_n}, \quad s \in [0, \gamma_{n+1}). \tag{IV}$$

1.5. From interpolated process to perturbed solutions. The next result gives sufficient conditions on the characteristics of the discrete process (III) for its interpolation (IV) to be a perturbed solution (II). If (U_i) are random variables, assumptions (i) and (ii) below have to be understood with probability one.

PROPOSITION 1.3. *Assume that the following hold:*

(i) *For all* $T > 0$

$$\lim_{n\to\infty} \sup \left\{ \left\| \sum_{i=n}^{k-1} \gamma_{i+1} U_{i+1} \right\| : k = n+1, \dots, m(\tau_n + T) \right\} = 0,$$

where

$$m(t) = \sup\{k \geq 0 : t \geq \tau_k\}; \tag{1.1}$$

(ii) $\sup_n \|x_n\| = M < \infty$.

Then the interpolated process $\mathbf{w}$ *is a perturbed solution of* F.

Proof. Let $\mathbf{U}, \gamma : \mathbb{R}_+ \to \mathbb{R}^m$ denote the continuous time processes defined by

$$\mathbf{U}(\tau_n + s) = U_{n+1}, \qquad \gamma(\tau_n + s) = \gamma_{n+1}$$

for all $n \in \mathbb{N}$, $0 \leq s < \gamma_{n+1}$.

Then, for any t,

$$\mathbf{w}(t) \in x_{m(t)} + (t - \tau_{m(t)})[\mathbf{U}(t) + F(x_{m(t)})];$$

hence

$$\dot{\mathbf{w}}(t) \in \mathbf{U}(t) + F(x_{m(t)}).$$

Let us set $\delta(t) = \|\mathbf{w}(t) - x_{m(t)}\|$. Then obviously

$$F(x_{m(t)}) \subset F^{\delta(t)}(\mathbf{w}(t)).$$

In addition,

$$\delta(t) \leq \gamma_{m(t)+1}[\|U_{m(t)+1}\| + c(1 + M)]$$

hence goes to 0, using hypothesis (i) of the statement of the proposition. It remains to check condition (ii)(a) of (II), but one has

$$\left\| \int_t^{t+v} \mathbf{U}(s)ds \right\| \leq \gamma_{m(t)+1} \|U_{m(t)+1}\| + \left\| \sum_{\ell=m(t)+1}^{m(t+v)-1} \gamma_{\ell+1} U_{\ell+1} \right\|$$
$$+ \gamma_{m(t+v)+1} \|U_{m(t+v)+1}\|,$$

and the result follows from condition (i). □

Sufficient conditions. Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\{\mathcal{F}_n\}_{n\geq 0}$ a filtration of $\mathcal{F}$ (i.e., a nondecreasing sequence of sub-σ-algebras of $\mathcal{F}$). We say that a stochastic process $\{x_n\}$ given by (III) satisfies the *Robbins–Monro condition with martingale difference noise* (Kushner and Yin [23]) if its characteristics satisfy the following:

(i) $\{\gamma_n\}$ is a deterministic sequence.

(ii) $\{U_n\}$ is *adapted* to $\{\mathcal{F}_n\}$. That is, U_n is measurable with respect to $\mathcal{F}_n$ for each $n \geq 0$.

(iii) $\mathsf{E}(U_{n+1} \mid \mathcal{F}_n) = 0$.

The next proposition is a classical estimate for stochastic approximation processes. Note that F does not appear. We refer the reader to (Benaïm [3, Propositions 4.2 and 4.4]) for a proof and further references.

PROPOSITION 1.4. *Let* $\{x_n\}$ *given by* (III) *be a Robbins–Monro equation with martingale difference noise process. Suppose that one of the following condition holds:*

(i) *For some* $q \geq 2$

$$\sup_n \mathsf{E}(\|U_n\|^q) < \infty$$

and

$$\sum_n \gamma_n^{1+q/2} < \infty.$$

(ii) *There exists a positive number* Γ *such that for all* $\theta \in \mathbb{R}^m$

$$\mathsf{E}(\exp(\langle \theta, U_{n+1} \rangle) \mid \mathcal{F}_n) \leq \exp\left(\frac{\Gamma}{2}\|\theta\|^2\right)$$

and

$$\sum_n e^{-c/\gamma_n} < \infty$$

for each $c > 0$.

Then assumption (i) *of Proposition* 1.3 *holds with probability* 1.

Remark 1.5. Typical applications are

(i) U_n uniformly bounded in L^2 and $\gamma_n = \frac{1}{n}$,

(ii) U_n uniformly bounded and $\gamma_n = o(\frac{1}{\log n})$.

2. Examples.

2.1. A multistage decision making model. Let A and B be measurable spaces, respectively called the *action space* and the *states of nature*; $E \subset \mathbb{R}^m$ a convex compact set called the *outcomes space*; and $H : A \times B \to E$ a measurable function, called the *outcome* function.

At discrete times $n = 1, 2 \ldots$ a decision maker (DM) chooses an action a_n from A and observes an outcome $H(a_n, b_n)$. We suppose the following.

(A) The sequence $\{a_n, b_n\}_{n \geq 0}$ is a random process defined on some probability space $(\Omega, \mathcal{F}, P)$ and adapted to some filtration $\{\mathcal{F}_n\}$. Here $\mathcal{F}_n$ has to be understood as the history of the process until time n.

(B) Given the history $\mathcal{F}_n$, DM and nature act independently:

$$\mathsf{P}((a_{n+1}, b_{n+1}) \in da \times db \mid \mathcal{F}_n) = \mathsf{P}(a_{n+1} \in da \mid \mathcal{F}_n)\mathsf{P}(b_{n+1} \in db \mid \mathcal{F}_n)$$

for any measurable sets $da \subset A$ and $db \subset B$.

(C) DM keeps track of only the cumulative average of the past outcomes,

$$x_n = \frac{1}{n}\sum_{i=1}^{n} H(a_i, b_i), \tag{2.1}$$

and his decisions are based on this average. That is,

$$\mathsf{P}(a_{n+1} \in da \mid \mathcal{F}_n) = Q_{x_n}(da),$$

where $Q_x(\cdot)$ is a probability measure over A for each $x \in E$, and $x \in E \mapsto Q_x(da) \in [0,1]$ is measurable for each measurable set $da \subset A$. The family $Q = \{Q_x\}_{x \in E}$ is called a *strategy* for DM.

Assumption (C) can be justified by considerations of limited memory and bounded rationality. It is partially motivated by Smale's approach to the prisoner's dilemma [27] (see also Benaïm and Hirsch [4, 5]), Blackwell's approachability theory ([8]; see also Sorin [28]), as well as fictitious play (Brown [10], Robinson [26]) and stochastic fictitious play (Benaïm and Hirsch [6], Fudenberg and Levine [15], Hofbauer and Sandholm [20]) in game theory (see the examples below).

For each $x \in E$ let

$$C(x) = \left\{ \int_{A \times B} H(a,b) Q_x(da) \nu(db) : \nu \in \mathcal{P}(B) \right\},$$

where $\mathcal{P}(B)$ denotes the set of probability measures over B. Then clearly

$$\mathsf{E}(H(a_{n+1}, b_{n+1}) \mid \mathcal{F}_n) \in C(x_n) \subset \overline{C}(x_n),$$

where $\overline{C}$ denote the smallest closed set-valued extension of C with convex values. More precisely, the graph of $\overline{C}$ is the intersection of all closed subsets $G \subset E \times E$ for which the fiber $G_x = \{y \in E : (x,y) \in G\}$ is convex and contains $C(x)$.

For $x \in \mathbb{R}^m$ let $P(x)$ denote the unique point in E closest to x. Extend $\overline{C}$ as in Remark 1.2 to a set-valued map on $\mathbb{R}^m$ by setting

$$\widehat{C}(x) = \overline{C}(P(x)).$$

Then the map

$$F(x) = -x + \overline{C}(P(x)) = -x + \widehat{C}(x) \tag{2.2}$$

clearly satisfies Hypothesis 1.1, and $\{x_n\}$ verifies the recursion

$$x_{n+1} - x_n = \frac{1}{n+1}(-x_n + H(a_{n+1}, b_{n+1})),$$

which can be rewritten as (see (III))

$$x_{n+1} - x_n \in \gamma_{n+1}[F(x_n) + U_{n+1}]$$

with $\gamma_n = \frac{1}{n}$ and $U_{n+1} = H(a_{n+1}, b_{n+1}) - \int_A H(a, b_{n+1}) Q_{x_n}(da)$. Hence, the conditions of Proposition 1.4 are satisfied and one deduces the following claim.

PROPOSITION 2.1. *The affine continuous time interpolated process* (IV) *of the process* $\{x_n\}$ *given by* (2.1) *is almost surely a perturbed solution of* F *defined by* (2.2).

Example 2.2 (Blackwell's approachability theory). A set $\Lambda \subset E$ is said to be *approachable* if there exists a strategy Q such that $x_n \to \Lambda$ almost surely. Blackwell [8] gives conditions ensuring approachability. We will show in section 5.1 how Blackwell's results can be partially derived from our main results and generalized (Corollary 5.2) in certain directions.

2.2. Learning in games. The preceding formalism is well suited to analyzing certain models of learning in games.

Consider the situation where m players are playing a game over and over. Let A^i (for $i \in I = \{1, \dots, m\}$) be a finite set representing the actions (pure strategies) available to player i, and let X^i be the finite dimensional simplex of probabilities over A^i (the set of mixed strategies for player i). For $i \in I$ we let A^{-i} and X^{-i} respectively denote the actions and mixed strategies available to the opponents of i. The payoff function to player i is given by a function $U^i : A^i \times A^{-i} \to \mathbb{R}$. As usual, we extend U^i to a function (still denoted U^i) on $X^i \times X^{-i}$, by multilinearity.

Example 2.3 (fictitious and stochastic fictitious play). Consider the game from the viewpoint of player i so that the DM is player i, and "nature" is given by the other players. In fictitious or stochastic fictitious play the outcome space is the space $X^i \times X^{-i}$ of mixed strategies, and the outcome function is the "identity" function $H : A^i \times A^{-i} \to X^i \times X^{-i}$ mapping every profile of actions a to the corresponding profile of mixed strategy δ_a.

Let

$$BR^i(x^{-i}) = \operatorname*{Argmax}_{a^i \in A^i} U^i(a^i, x^{-i}) \subset A^i$$

be the set of best actions that i can play in response to x^{-i}.

Both classical fictitious play (Brown [10], Robinson [26]) and stochastic fictitious play (Benaïm and Hirsch [6], Fudenberg and Levine [15], Hofbauer and Sandholm [20]) assume that the strategy of player i, $Q^i = \{Q^i_x\}$, can be written as

$$Q^i_x(a^i) = q^i(a^i, x^{-i}),$$

where $q^i : A^i \times X^{-i} \to [0, 1]$ is such that one of the following assumptions holds:

fictitious play assumption:

$$\sum_{a^i \in BR^i(x^{-i})} q^i(a^i, x^{-i}) = 1,$$

or *stochastic fictitious play assumption*, q^i is smooth in x^{-i} and

$$\sum_{a^i \in BR^i(x^{-i})} q^i(a^i, x^{-i}) \geq 1 - \delta$$

for some $0 < \delta \ll 1$.

In this framework, if a_ℓ denotes the profile of actions at stage ℓ, one has

$$x_n = \frac{1}{n} \sum_{\ell=1}^{n} a_\ell$$

and

$$x_{n+1} - x_n = \frac{1}{n+1}(a_{n+1} - x_n).$$

Thus for each i

$$\mathsf{E}(x^i_{n+1} - x^i_n \mid \mathcal{F}_n) \in \frac{1}{n+1}(\overline{BR}^i(x^{-i}_n) - x^i_n),$$

where $\overline{BR}^i(x^{-i}) \subset X^i$ is the convex hull of $BR^i(x^{-i})$ for the standard fictitious play, and $\overline{BR}^i(x^{-i}) = \sum_{a^i \in A^i} q^i(a^i, x^{-i})\delta_{a^i}$ for the stochastic fictitious play.

Thus the set-valued map F defined in (2.2) is given as

$$F^i(x) = -x + \overline{BR}^i(x^{-i}) \times X^{-i}.$$

Observe that if a subset $J \subset I$ of players plays a fictitious (or stochastic fictitious) play strategy, then F^i has to be replaced by

$$F^J(x) = \bigcap_{i \in J} F^i(x).$$

In particular, if all players play a fictitious play strategy, the differential inclusion induced by F is the best-response differential inclusion (Gilboa and Matsui [16], Hofbauer [19], Hofbauer and Sorin [21]), while if all play a stochastic fictitious play, F is a smooth best-response vector field (Benaïm and Hirsch [6], Fudenberg and Levine [15], Hofbauer and Sandholm [20]).

Example 2.4 (Smale approach to the prisoner's dilemma). We still consider the game from the viewpoint of player i, so that the DM is player i and nature the other players, but we take for H the payoff vector function

$$H : A^i \times A^{-i} \to E,$$
$$a \to U(a) = (U^1(a), \dots, U^m(a)),$$

where $E \subset \mathbb{R}^m$ is the convex hull of the payoff vectors $\{U(a)\}$.

This setting fits exactly with Smale's approach to the prisoner's dilemma [27] later revisited by Benaïm and Hirsch [4]. Details will be given in section 5.2, where Smale's approach will be reinterpreted in the framework of approachability.

3. Set-valued dynamical systems.

3.1. Properties of the trajectories of (I). Let $C^0(\mathbb{R}, \mathbb{R}^m)$ denote the space of continuous paths $\{\mathbf{z} : \mathbb{R} \to \mathbb{R}^m\}$ equipped with the topology of uniform convergence on compact intervals. This is a complete metric space for the distance $\mathbf{D}$ defined by

$$\mathbf{D}(\mathbf{x}, \mathbf{z}) = \sum_{k=1}^{\infty} \frac{1}{2^k} \min(\|\mathbf{x} - \mathbf{z}\|_{[-k,k]}, 1),$$

where $\|\cdot\|_{[-k,k]}$ stands for the supremum norm on $C^0([-k, k], \mathbb{R}^m)$.

Given a set $M \subset \mathbb{R}^m$, we let $S_M \subset C^0(\mathbb{R}, \mathbb{R}^m)$ denote the set of all solutions to (I) with initial conditions $x \in M$ ($S_M = \bigcup_{x \in M} S_x$), and $S_{M,M} \subset S_M$ the subset consisting of solutions $\mathbf{x}$ that remain in M (i.e., $\mathbf{x}(\mathbb{R}) \subset M$).

LEMMA 3.1. *Assume M compact. Then S_M is a nonempty compact set and $S_{M,M}$ is a compact (possibly empty) set.*

Proof. The first assertion follows from Aubin and Cellina [1, section 2.2, Theorem 1, p. 104]. The second easily follows from the first. □

3.2. Set-valued dynamical system induced by (I). The differential inclusion (I) induces a set-valued dynamical system $\{\Phi_t\}_{t \in \mathbb{R}}$ defined by

$$\Phi_t(x) = \{\mathbf{x}(t) : \mathbf{x} \text{ is a solution to (I) with } \mathbf{x}(0) = x\}.$$

The family $\Phi = \{\Phi_t\}_{t \in \mathbb{R}}$ enjoys the following properties:

(a) $\Phi_0(x) = \{x\}$;
(b) $\Phi_t(\Phi_s(x)) = \Phi_{t+s}(x)$ for all $t, s \geq 0$;
(c) $y \in \Phi_t(x) \Rightarrow x \in \Phi_{-t}(y)$ for all $x, y \in \mathbb{R}^m$, $t \in \mathbb{R}$;
(d) $(x,t) \mapsto \Phi_t(x)$ is a closed set-valued map with compact values (i.e., $\Phi_t(x)$ is a compact set for each t and x).

Properties (a), (b), (c) are immediate to verify, and property (d) easily follows from Lemma 3.1.

For subsets $T \subset \mathbb{R}$ and $A \subset \mathbb{R}^m$ we will define

$$\Phi_T(A) = \bigcup_{t \in T} \bigcup_{x \in A} \Phi_t(x).$$

Invariant sets.

DEFINITION V. *A set $A \subset \mathbb{R}^m$ is said to be*
(i) strongly invariant *(for Φ) if $A = \Phi_t(A)$ for all $t \in \mathbb{R}$;*
(ii) quasi-invariant *if $A \subset \Phi_t(A)$ for all $t \in \mathbb{R}$;*
(iii) semi-invariant *if $\Phi_t(A) \subset A$ for all $t \in \mathbb{R}$;*
(iv) invariant *(for F) if for all $x \in A$ there exists a solution $\mathbf{x}$ to* (I) *with $\mathbf{x}(0) = x$ and such that $\mathbf{x}(\mathbb{R}) \subset A$.*

We call a set A *strongly positive invariant* if $\Phi_t(A) \subset A$ for all $t > 0$.

At first glance (at least for those used to ordinary differential equations) the good notion might seem to be the one defined by strong invariance. However, this notion is too strong for differential inclusions, as shown by the simple example below (Example 3.2), and the main notions that will really be needed here are invariance and strong positive invariance. We have included the definition of quasi invariance mainly because some of our later results may be related to a paper by Bronstein and Kopanskii [9] making use of this notion.[2] Observe, however, that by Lemma 3.3 below, quasi invariance coincides with invariance for compact sets.

Example 3.2. (a) Let F be the set-valued map defined on $\mathbb{R}$ by $F(x) = -\operatorname{sgn}(x)$ if $x \neq 0$ and $F(0) = [-1, 1]$. Then $\Phi_t(0) = \{0\}$ for $t \geq 0$, and $\Phi_t(0) = [t, -t]$ for $t < 0$. Hence $\{0\}$ is invariant and strongly positively invariant but is not strongly invariant.

(b) Let now $F(x) = x$ for $x < 0$, $F(x) = 1$ for $x > 0$, and $F(0) = [0, 1]$. Then $\Phi_t(0) = \{0\}$ for $t \leq 0$, and $\Phi_t(0) = [0, t]$ for $t \geq 0$. Hence $\{0\}$ is invariant but not strongly positively invariant.

LEMMA 3.3. *Every invariant set is quasi-invariant. Every compact quasi-invariant set is invariant.*

Proof. Suppose that A is invariant. Let $x \in A$ and $\mathbf{x}$ be a solution to (I) with $\mathbf{x}(0) = x$ and $\mathbf{x}(\mathbb{R}) \subset A$. For all $t \in \mathbb{R}$ we have $x \in \Phi_t(\mathbf{x}(-t))$. Hence A is quasi-invariant.

Conversely suppose that A is quasi-invariant and compact. Choose $x \in A$ and fix $N \in \mathbb{N}$. Then for every $p \in \mathbb{N}$ there exists, by quasi invariance and by gluing pieces of solutions together, a solution $\mathbf{x}_{p,N}$ to (I) such that $\mathbf{x}_{p,N}(0) = x$ and for all $q \in \{-2^p, \ldots, 2^p\}$, $\mathbf{x}_{p,N}(\frac{qN}{2^p}) \in A$. By Lemma 3.1, the sequence $\{\mathbf{x}_{p,N}\}_{p \in \mathbb{N}}$ is relatively compact in $C^0([-N, N], \mathbb{R}^m)$. Let $\mathbf{x}_N$ be a limit point of this sequence. Then for each dyadic point $t = \frac{qN}{2^p}$, where $q \in \{-2^p, \ldots, 2^p\}$, $\mathbf{x}_N(t) \in \overline{A}$. Continuity of $\mathbf{x}_N$ implies $\mathbf{x}_N([-N, N]) \subset \overline{A}$. Now let $\mathbf{x}$ be a limit point of the sequence $\{\mathbf{x}_N\}_{N \in \mathbb{N}}$ in $C^0(\mathbb{R}, \mathbb{R}^m)$. Then $\mathbf{x}(\mathbb{R}) \subset \overline{A}$ and $\mathbf{x}$ is a solution to (I). □

[2] Invariant sets in Bronstein and Kopanskii [9] coincide with what we define here as strongly invariant sets.

Remark 3.4. A invariant together with strong positive invariance implies $\Phi_t(A) = A$ for $t > 0$.

3.3. Chain-recurrence and the limit set theorem. Given a set $A \subset \mathbb{R}^m$ and $x, y \in A$, we write $x \hookrightarrow_A y$ if for every $\varepsilon > 0$ and $T > 0$ there exists an integer $n \in \mathbb{N}$, solutions $\mathbf{x}_1, \dots, \mathbf{x}_n$ to (I), and real numbers $t_1, t_2, \dots, t_n$ greater than T such that

(a) $\mathbf{x}_i(s) \in A$ for all $0 \le s \le t_i$ and for all $i = 1, \dots, n$,
(b) $\|\mathbf{x}_i(t_i) - \mathbf{x}_{i+1}(0)\| \le \varepsilon$ for all $i = 1, \dots, n-1$,
(c) $\|\mathbf{x}_1(0) - x\| \le \varepsilon$ and $\|\mathbf{x}_n(t_n) - y\| \le \varepsilon$.

The sequence $(\mathbf{x}_1, \dots, \mathbf{x}_n)$ is called an (ε, T) chain (in A from x to y) for F.

DEFINITION VI. *A set $A \subset \mathbb{R}^m$ is said to be* internally chain transitive, *provided that A is compact and $x \hookrightarrow_A y$ for all $x, y \in A$.*

LEMMA 3.5. *An internally chain transitive set is invariant.*

Proof. Let A be such a set and $x \in A$. Let $(\mathbf{x}_1, \dots, \mathbf{x}_n)$ be an (ε, T) chain from x to x. Set $\mathbf{y}_{\varepsilon,T}(t) = \mathbf{x}_1(t)$ for $0 \le t \le T$ and $\mathbf{z}_{\varepsilon,T}(t) = \mathbf{x}_n(t_n + t)$ for $-T \le t \le 0$. By Lemma 3.1 we can extract from $(\mathbf{y}_{1/p,T})_{p\in\mathbb{N}}$ and $(\mathbf{z}_{1/p,T})_{p\in\mathbb{N}}$ some subsequences converging, respectively, to $\mathbf{y}_T$ and $\mathbf{z}_T$, where $\mathbf{y}_T$ and $\mathbf{z}_T$ are solutions to (I), $\mathbf{y}_T(0) = x = \mathbf{z}_T(0)$, $\mathbf{y}_T([0,T]) \subset A$, and $\mathbf{z}_T([-T,0]) \subset A$. The map $\mathbf{w}_T(t) = \mathbf{y}_T(t)$ for $t \ge 0$ and $\mathbf{w}_T(t) = \mathbf{z}_T(t)$ for $t \le 0$ is then a solution to (I) with initial condition x and such that $\mathbf{w}_T([-T,T]) \subset A$. By Lemma 3.1, again we extract from $(\mathbf{w}_T)_{T\ge 0}$ a subsequence converging to a solution $\mathbf{w}$ whose range lies in A and with initial condition x. □

This notion of recurrence due to Conley [13] for classical dynamical systems is well suited to the description of the asymptotic behavior of a perturbed solution to (I), as shown by the following theorem.

THEOREM 3.6. *Let $\mathbf{y}$ be a bounded perturbed solution to* (I). *Then, the limit set of $\mathbf{y}$,*

$$L(\mathbf{y}) = \bigcap_{t\ge 0} \overline{\{\mathbf{y}(s) : s \ge t\}},$$

is internally chain transitive.

This theorem is the set-valued version of the limit set theorem proved by Benaïm [2] for stochastic approximation and Benaïm and Hirsch [5] for asymptotic pseudotrajectories of a flow. We will deduce it from the more general results of section 4.

3.4. Limit sets. The set

$$\omega_\Phi(x) := \bigcap_{t\ge 0} \overline{\Phi_{[t,\infty)}(x)}$$

is the ω-limit set of a point $x \in \mathbb{R}^m$. Note that $\omega_\Phi(x)$ contains the limit sets $L(\mathbf{x})$ of all solutions $\mathbf{x}$ with $\mathbf{x}(0) = x$ but is in general larger than the union of these.

In contrast to the limit set of a solution, the ω-limit set of a point need not be internally chain transitive.

Example 3.7. Let F be the set-valued map defined on $\mathbb{R}$ by $F(x) = 1 - x$ for $x > 0$ and $F(0) = [0,1]$ and $F(x) = -x$ for $x < 0$. Then for every solution $\mathbf{x}$, one has $\lim_{t\to\infty} \mathbf{x}(t) = 0$ or 1. But $\omega_\Phi(0) = [0,1]$ is not internally chain transitive.

More generally one defines

$$\omega_\Phi(Y) := \bigcap_{t\ge 0} \overline{\Phi_{[t,\infty)}(Y)}.$$

DEFINITION VII. *A set Y is* forward precompact *if $\overline{\Phi_{[t,\infty)}(Y)}$ is compact for some $t > 0$.*

LEMMA 3.8. (i) *$\omega_\Phi(Y)$ is the set of points $p \in \mathbb{R}^m$ such that*

$$p = \lim_{n\to\infty} \mathbf{y}_n(t_n)$$

for some sequence $\{\mathbf{y}_n\}$ of solutions to (I) *with initial conditions $\mathbf{y}_n(0) \in Y$ and some sequence $\{t_n\} \in \mathbb{R}$ with $t_n \to \infty$.*

(ii) *$\omega_\Phi(Y)$ is a closed invariant (possibly empty) set. If Y is forward precompact, then $\omega_\Phi(Y)$ is nonempty and compact.*

Proof. Point (i) is easily seen from the definition.

(ii) Let $p = \lim_{n\to\infty} \mathbf{y}_n(t_n) \in \omega_\Phi(Y)$. Set $\mathbf{z}_n(s) = \mathbf{y}_n(t_n + s)$ for all $s \in \mathbb{R}$. By Lemma 3.1 we may extract from $(\mathbf{z}_n)_{n\geq 0}$ a subsequence converging to some solution $\mathbf{z}$ with $\mathbf{z}(0) = p$ and $\mathbf{z}(s) = \lim_{n_k\to\infty} \mathbf{y}_{n_k}(t_{n_k} + s) \in \omega_\Phi(Y)$. This proves invariance. The rest is clear. □

Note that the limit set $\omega_\Phi(Y)$ is in general not strongly positively invariant (e.g., in Example 3.7 for $x < 0$, $\omega_\Phi(x) = \{0\}$).

3.5. Attracting sets and attractors. For applications it is useful to characterize $L(\mathbf{y})$ in terms of certain compact invariant sets for Φ, namely, *the attractors,* as defined below.

Given a closed invariant set L, the induced set-valued dynamical system Φ^L is the family of (set-valued) mappings $\Phi^L = \{\Phi^L_t\}_{t\in\mathbb{R}}$ defined on L by

$$\Phi^L_t(x) = \{\mathbf{x}(t) : \mathbf{x} \text{ is a solution to (I) with } \mathbf{x}(0) = x \text{ and } \mathbf{x}(\mathbb{R}) \subset L\}.$$

Note that L is strongly invariant for Φ^L.

DEFINITION VIII. *A compact set $A \subset L$ is called an* attracting set *for Φ^L, provided that there is a neighborhood U of A in L (i.e., for the induced topology) with the property that for every $\varepsilon > 0$ there exists $t_\varepsilon > 0$ such that*

$$\Phi^L_t(U) \subset N^\varepsilon(A)$$

for all $t \geq t_\varepsilon$. Or, equivalently, $\Phi^L_{[t_\varepsilon,\infty)}(U) \subset N^\varepsilon(A)$. Here $N^\varepsilon(A)$ stands for the ε-neighborhood of A.

If, additionally, A is invariant, then A is called an attractor *for Φ^L.*

The set U is called a fundamental neighborhood *of A for Φ^L. If $A \neq L$ and $A \neq \emptyset$, then A is called a* proper attracting set *(or* proper attractor*) for Φ^L.*

Furthermore, an attracting set (respectively, attractor) for Φ is an attracting set (respectively, attractor) for Φ^L with $L = \mathbb{R}^m$.

Example 3.9. Let F be the set-valued map from Example 3.2(a), i.e., defined on $\mathbb{R}$ by $F(x) = -\operatorname{sgn}(x)$ if $x \neq 0$ and $F(0) = [-1, 1]$. Then $\{0\}$ is an attractor and every compact set $A \subset \mathbb{R}$ with $0 \in A$ is an attracting set.

PROPOSITION 3.10. *Let A be a nonempty compact subset of L, and U a neighborhood of A in L. Then the following hold:*

(i) *A is an attracting set for Φ^L with fundamental neighborhood U if and only if U is forward precompact and $\omega_{\Phi^L}(U) \subset A$. In this case $\omega_{\Phi^L}(U)$ is an attractor.*

(ii) *A is an attractor for Φ^L with fundamental neighborhood U if and only if U is forward precompact and $\omega_{\Phi^L}(U) = A$.*

Proof. (i) If A is an attracting set for Φ^L with fundamental neighborhood U, then $\omega_{\Phi^L}(U) \subset \bigcap_{\varepsilon>0} N^\varepsilon(A) \subset A$. Conversely, for t large enough $V_t = \overline{\Phi^L_{[t,\infty)}(U)}$ defines a

decreasing family of compact sets converging to $\omega_{\Phi^L}(U) \subset A$. Hence for any $\varepsilon > 0$ there exists t_ε with $V_{t_\varepsilon} \subset N^\varepsilon(A)$ and A is an attracting set. In particular, $\omega_{\Phi^L}(U)$ itself is an attracting set, invariant by Lemma 3.8(ii).

(ii) If $A = \omega_{\Phi^L}(U)$, then A is an attractor by (i). Conversely, if A is an attractor with fundamental neighborhood U, then $\omega_\Phi(U) \subset A$ by (i). Let $x \in A$. Since A is invariant, there exists a solution $\mathbf{y}$ to (I) with $\mathbf{y}(0) = x$ and $\mathbf{y}(\mathbb{R}) \subset A$. Set $\mathbf{y}_n(t) = \mathbf{y}(t-n)$. Then $\mathbf{y}_n(n) = x$, proving that $x \in \omega_{\Phi^L}(U)$ (by Lemma 3.8(i)). □

PROPOSITION 3.11. *Every attractor is strongly positively invariant. (Example* 3.2(a) *provides an attractor that is not strongly invariant.)*

Proof. By invariance, $A \subset \Phi_T^L(A)$ for all $T > 0$. Hence, given $t > 0$,

$$\Phi_t^L(A) \subset \Phi_{t+T}^L(A) \subset \Phi_{t+T}^L(U) \subset \Phi_{[t+T,\infty)}^L(U)$$

for all $T > 0$. Thus $\Phi_t^L(A) \subset N^\varepsilon(A)$ for all $\varepsilon > 0$, and hence $\Phi_t^L(A) \subset A$ for all $t > 0$. □

Remark 3.12. In the family of attracting sets A with a given fundamental neighborhood U, there exists a minimal one, which is in addition invariant, strongly positively invariant, and independent of the set U used to define the family. It is also the largest positively quasi-invariant set included in U.

Any attractor $A \subset L$ can be written as $A = \omega_{\Phi^L}(U)$ for some U. Hence any fundamental neighborhood uniquely determines the attractor A. This implies, as in Conley [13], that Φ^L can have at most countably many attractors.

3.6. Attractors and stability.

DEFINITION IX. *A set $A \subset L$ is* asymptotically stable *for Φ^L if it satisfies the following three conditions:*

(i) *A is invariant.*

(ii) *A is* Lyapunov stable; *i.e., for every neighborhood U of A there exists a neighborhood V of A such that $\Phi_{[0,\infty)}(V) \subset U$.*

(iii) *A is* attractive; *i.e., there is a neighborhood U of A such that for every $x \in U : \omega_\Phi(x) \subset A$.*

Alternatively, instead of (iii) *one could ask for the following weaker requirement:*

(iii′) *There is a neighborhood U of A such that for every solution $\mathbf{x}$ with $\mathbf{x}(0) \in U$ one has $L(\mathbf{x}) \subset A$.*

We show now that for compact sets the concepts of attractor and asymptotic stability are equivalent. The proof of Corollary 3.18 below shows that it makes no difference whether one uses (iii) or (iii′) in the definition of asymptotic stability.

We start with an upper bound for entry times.

LEMMA 3.13. *Let V be an open set and K compact such that for all solutions $\mathbf{x}$ with $\mathbf{x}(0) \in K$ there is $t > 0$ with $\mathbf{x}(t) \in V$. Then there exists $T > 0$ such that for every solution $\mathbf{x}$ with $\mathbf{x}(0) \in K$ there is $t \in [0, T]$ with $\mathbf{x}(t) \in V$.*

Proof. Suppose that there is no such upper bound T for the entry times into V. Then for each $n \in \mathbb{N}$ there is $\mathbf{x}_n(0) = x_n \in K$ and a solution $\mathbf{x}_n$ such that $\mathbf{x}_n(t) \notin V$ for $0 \le t \le n$. Since K is compact, we can assume that $x_n \to x \in K$. And by Lemma 3.1 a subsequence of $\mathbf{x}_n$ converges to a solution $\mathbf{x}$ with $\mathbf{x}(0) = x$ and $\mathbf{x}(t) \notin V$ for all $t > 0$. □

LEMMA 3.14. *If a closed set A is Lyapunov stable, then it is strongly positively invariant.*

Proof. A is the intersection of a family of strongly positively invariant neighborhoods. □

LEMMA 3.15. *If a compact set A satisfies* (ii) *and* (iii′), *it is attracting.*

Proof. Let B be a compact neighborhood of A, included in the fundamental neighborhood U, and let W be a neighborhood of A. A being Lyapunov stable, there exists an open neighborhood V of A with $\Phi^L_{[0,\infty)}(V) \subset W$. For any $x \in B$ and any solution $\mathbf{x}$ with $\mathbf{x}(0) = x$, there exists $t > 0$ with $\mathbf{x}(t) \in V$. Applying Lemma 3.13 implies $\Phi^L_T(B) \subset \Phi^L_{[0,T]}(V)$; hence $\Phi^L_{[T,\infty)}(B) \subset W$ and A is attracting. □

LEMMA 3.16. *If the set A is attracting and strongly positively invariant, then it is Lyapunov stable.*

Proof. Let A be attracting with fundamental neighborhood U, and V be any other (open) neighborhood of A. Then by definition there is $T > 0$ such that $\Phi^L_{[T,\infty)}(U) \subset V$. A being strongly positively invariant, $\Phi^L_{[0,T]}(A) \subset A$. Upper semicontinuity gives an $\varepsilon > 0$ such that $\Phi^L_{[0,T]}(N^\varepsilon(A)) \subset V$ and $N^\varepsilon(A) \subset U$. Hence $\Phi^L_{[0,\infty)}(N^\varepsilon(A)) \subset V$, which shows Lyapunov stability. □

COROLLARY 3.17. *For a compact set A, properties* (ii) *and* (iii′) *of Definition* IX, *together, are equivalent to attracting and strong positive invariance.*

COROLLARY 3.18. *A compact set A is an attractor if and only if it is asymptotically stable.*

We conclude with a simple useful condition ensuring that an open set contains an attractor.

PROPOSITION 3.19. *Let U be an open set with compact closure. Suppose that $\Phi_T(\overline{U}) \subset U$ for some $T > 0$. Then U is a fundamental neighborhood of some attractor A.*

Proof. Since Φ has a closed graph, $\Phi_T(\overline{U})$ is compact. Therefore $\Phi_T(\overline{U}) \subset V \subset \overline{V} \subset U$ for some open set V. By upper semicontinuity of Φ_T (which follows from property (d) of a set-valued dynamical system) there exists $\varepsilon > 0$ such that $\Phi_t(\overline{U}) \subset V$ for $T - \varepsilon \le t \le T + \varepsilon$. Let $t_0 = T(T+1)/\varepsilon$. For all $t \ge t_0$ write $t = kT + r$ with $k \in \mathbb{N}$ and $r < T$. Hence $t = k(T + r/k)$ with $0 \le r/k < \varepsilon$. Thus

$$\Phi_t(\overline{U}) = \Phi_{T+r/k} \circ \cdots \circ \Phi_{T+r/k}(\overline{U}) \subset V.$$

Hence $\omega_\Phi(U) = \bigcap_{t \ge t_0} \overline{\Phi_{[t,\infty)}(U)} \subset \overline{V} \subset U$ is an attractor with fundamental neighborhood U. □

3.7. Chain transitivity and attractors.

PROPOSITION 3.20. *Let L be internally chain transitive. Then L has no proper attracting set for Φ^L.*

Proof. Let $A \subset L$ be an attracting set. By definition, there exists a neighborhood U of A, and for all $\varepsilon > 0$ a number t_ε such that $\Phi^L_t(U) \subset N^\varepsilon(A)$ for all $t > t_\varepsilon$. Assume $A \ne L$ and choose ε small enough so that $N^{2\varepsilon}(A) \subset U$ and there exists $y \in L \setminus N^{2\varepsilon}(A)$. Then, for $T \ge t_\varepsilon$ and $x \in A$, there is no (ε, T) chain from x to y. In fact, $\mathbf{x}_1(0) \in N^{2\varepsilon}(A)$, and hence $\mathbf{x}_1(t_1) \in N^\varepsilon(A)$; by induction, $\mathbf{x}_i(t_i) \in N^\varepsilon(A)$ so that $\mathbf{x}_{i+1}(0) \in N^{2\varepsilon}(A)$ as well. Thus we arrive at a contradiction. □

Remark 3.21. This last proposition can also be deduced from Bronstein and Kopanskii [9, Theorem 1] combined with Lemma 3.1. Also the converse is true.

Recall that an attracting set (respectively, attractor) for Φ is an attracting set (respectively, attractor) for Φ^L with $L = \mathbb{R}^m$.

LEMMA 3.22. *Let A be an attracting set for Φ and L a closed invariant set. Assume $A \cap L \ne \emptyset$. Then $A \cap L$ is an attracting set for Φ^L.*

Proof. The proof follows from the definitions. □

If A is a set, then

$$B(A) = \{x \in \mathbb{R}^m : \omega_\Phi(x) \subset A\}$$

denotes its *basin of attraction*.

THEOREM 3.23. *Let A be an attracting set for Φ and L an internally chain transitive set. Assume $L \cap B(A) \neq \emptyset$. Then $L \subset A$.*

Proof. Suppose $L \cap B(A) \neq \emptyset$. Then there exists a solution $\mathbf{x}$ to (I) with $\mathbf{x}(0) = x \in B(A)$ and $\mathbf{x}(\mathbb{R}) \subset L$. Hence $d(\mathbf{x}(t), A) \to 0$ when $t \to \infty$, proving that L meets A. Proposition 3.20 and Lemma 3.22 imply that $L \subset A$. □

A *global attractor* for Φ is an attractor whose basin of attraction consists of all $\mathbb{R}^m$. If a global attractor exists, then it is unique and coincides with the maximal compact invariant set of Φ. The following corollary is an immediate consequence of Theorem 3.23 or even more easily of Lemma 3.5.

COROLLARY 3.24. *Suppose Φ has a global attractor A. Then every internally chain transitive set lies in A.*

3.8. Lyapunov functions.

PROPOSITION 3.25. *Let Λ be a compact set, $U \subset \mathbb{R}^m$ be a bounded open neighborhood of Λ, and $V : \overline{U} \to [0, \infty[$. Let the following hold:*

(i) *For all $t \geq 0$, $\Phi_t(U) \subset U$ (i.e., U is strongly positively invariant);*

(ii) *$V^{-1}(0) = \Lambda$;*

(iii) *V is continuous and for all $x \in U \setminus \Lambda$, $y \in \Phi_t(x)$ and $t > 0$, $V(y) < V(x)$;*

(iv) *V is upper semicontinuous, and for all $x \in \overline{U} \setminus \Lambda$, $y \in \Phi_t(x)$, and $t > 0$, $V(y) < V(x)$.*

(A) *Under* (i), (ii), *and* (iii), *Λ is a Lyapunov stable attracting set, and there exists an attractor contained in Λ whose basin contains U, and with $V^{-1}([0, r))$ as fundamental neighborhoods for small $r > 0$.*

(B) *Under* (i), (ii), *and* (iv), *there exists an attractor contained in Λ whose basin contains U.*

Proof. For the proof of (A), let $r > 0$ and $U_r = \{x \in U : V(x) < r\}$. Then $\{\overline{U}_r\}_{r>0}$ is a nested family of compact neighborhoods of Λ with $\bigcap_{r>0} \overline{U}_r = \Lambda$. Thus for $r > 0$ small enough, $\overline{U_r} \subset U$. Moreover, $\Phi_t(\overline{U}_r) \subset U_r$ for $t > 0$ by our hypotheses on U and V. Proposition 3.19 then implies the result.

For (B), let $A = \omega_\Phi(U)$, which is closed and invariant (by Lemma 3.8) and hence compact, since it is included in $\overline{U}$. Let $\alpha = \max_{y \in A} V(y)$ be reached at x, since V is upper semicontinuous. By invariance there exists a solution $\mathbf{x}$ and $t > 0$ with $z = \mathbf{x}(0) \in A$ and $\mathbf{x}(t) = x$. This contradicts (iv) unless $\alpha = 0$ and $A \subset \Lambda$. Thus U is a neighborhood of A, which is an attractor included in Λ. □

Remark 3.26. Given any attractor A, there exists a function V such that Proposition 3.25(iv) holds for $\Lambda = A$. Take $V(x) = \max\{d(y, A)g(t),\ y \in \Phi_t(x),\ t \geq 0\}$, where $d > g(t) > c > 0$ is any continuous strictly increasing function.

Let Λ be any subset of $\mathbb{R}^m$. A continuous function $V : \mathbb{R}^m \to \mathbb{R}$ is called a *Lyapunov function* for Λ if $V(y) < V(x)$ for all $x \in \mathbb{R}^m \setminus \Lambda$, $y \in \Phi_t(x)$, $t > 0$, and $V(y) \leq V(x)$ for all $x \in \Lambda$, $y \in \Phi_t(x)$, and $t \geq 0$. Note that for each solution $\mathbf{x}$, V is constant along its limit set $L(\mathbf{x})$.

The following result is similar to Benaïm [3, Proposition 6.4].

PROPOSITION 3.27. *Suppose that V is a Lyapunov function for Λ. Assume that $V(\Lambda)$ has empty interior. Then every internally chain transitive set L is contained in Λ and $V \mid L$ is constant.*

Proof. Let

$$v = \inf\{V(y) : y \in L\}.$$

Since L is compact and V is continuous, $v = V(x)$ for some point $x \in L$. Since L is invariant, there exists a solution $\mathbf{x}$ with $\mathbf{x}(t) \in L$ and $\mathbf{x}(0) = x$. Then $v = V(x) > V(\mathbf{x}(t))$, and thus is impossible for $t > 0$. Since $\mathbf{x}(t) \in \Phi_t(x)$, we conclude $x \in \Lambda$.

Thus v belongs to the range $V(\Lambda)$. Since $V(\Lambda)$ contains no interval, there is a sequence $v_n \notin V(\Lambda)$ decreasing to v. The sets $L_n = \{x \in L : V(x) < v_n\}$ satisfy $\Phi_t(\overline{L}_n) \subset L_n$ for $t > 0$. In fact, either $x \in \Lambda \cap \overline{L}_n$ and $V(y) \leq V(x) < v_n$ or $V(y) < V(x) \leq v_n$, for any $y \in \Phi_t(x)$, $t > 0$.

Thus, using Propositions 3.19 and 3.20, one obtains $L = \bigcap_n \overline{L}_n = \{x \in L : V(x) = v\}$. Hence V is constant on L. L being invariant, this implies, as above, $L \subset \Lambda$. □

COROLLARY 3.28. *Let V and Λ be as in Proposition* 3.27. *Suppose furthermore that V is C^m and Λ is contained in the critical points set of V. Then every internally chain transitive set lies in Λ and $V \mid L$ is constant.*

Proof. By Sard's theorem (Hirsch [18, p. 69]), $V(\Lambda)$ has empty interior and Proposition 3.27 applies. □

4. The limit set theorem.

4.1. Asymptotic pseudotrajectories for set-valued dynamics. The translation flow $\Theta : C^0(\mathbb{R}, \mathbb{R}^m) \times \mathbb{R} \to C^0(\mathbb{R}, \mathbb{R}^m)$ is the flow defined by

$$\Theta^t(\mathbf{x})(s) = \mathbf{x}(s + t).$$

A continuous function $\mathbf{z} : \mathbb{R}^+ \to \mathbb{R}^m$ is an asymptotic pseudotrajectory (APT) for Φ if

$$\lim_{t\to\infty} \mathbf{D}(\Theta^t(\mathbf{z}), S_{\mathbf{z}(t)}) = 0 \tag{4.1}$$

(or $\lim_{t\to\infty} \mathbf{D}(\Theta^t(\mathbf{z}), S) = 0$, where $S = \bigcup_{x\in\mathbb{R}^m} S_x$ denotes the set of all solutions of (I)).

Alternatively, for all T

$$\lim_{t\to\infty} \inf_{\mathbf{x}\in S_{\mathbf{z}}(t)} \sup_{0\leq s\leq T} \|\mathbf{z}(t + s) - \mathbf{x}(s)\| = 0.$$

In other words, for each fixed T, the curve

$$[0, T] \to \mathbb{R}^m : s \to \mathbf{z}(t + s)$$

shadows some Φ trajectory of the point $\mathbf{z}(t)$ over the interval $[0, T]$ with arbitrary accuracy for sufficiently large t. Hence $\mathbf{z}$ has a forward trajectory under Θ attracted by S. As usual, one extends $\mathbf{z}$ to $\mathbb{R}$ by letting $\mathbf{z}(t) = \mathbf{z}(0)$ for $t < 0$.

The next result is a natural extension of Benaïm and Hirsch [4], [5, Theorem 7.2].

THEOREM 4.1 (characterization of APT). *Assume $\mathbf{z}$ is bounded. Then there is equivalence between the following statements:*

(i) *$\mathbf{z}$ is an APT for Φ.*

(ii) *$\mathbf{z}$ is uniformly continuous, and any limit point of $\{\Theta^t(\mathbf{z})\}$ is in S.*

In both cases the set $\{\Theta^t(\mathbf{z});\ t \geq 0\}$ is relatively compact.

Proof. By hypothesis, $K = \overline{\{\mathbf{z}(t);\ t \geq 0\}}$ is compact.

For any $\varepsilon > 0$, there exists $\eta > 0$ such that $\|z - x\| < \varepsilon/2$, for any $x \in K$, any $z \in \Phi_s(x)$, and any $|s| < \eta$, using property (d) of the dynamical system.

$\mathbf{z}$ being an APT, there exists T such that $t > T$ implies

$$d(\mathbf{z}(t+s), \Phi_s(\mathbf{z}(t))) < \frac{\varepsilon}{2} \quad \forall |s| < \eta;$$

hence

$$\|\mathbf{z}(t+s) - \mathbf{z}(t)\| \leq \varepsilon$$

and $\mathbf{z}$ is uniformly continuous. Clearly any limit point belongs to S by the condition (4.1) above.

Conversely, if $\mathbf{z}$ is uniformly continuous, then the family of functions $\{\Theta^t(\mathbf{z});\ t \geq T\}$ is equicontinuous and hence (K being compact) relatively compact by Ascoli's theorem. Since any limit point belongs to S, property (4.1) follows. □

4.2. Perturbed solutions are APTs.

THEOREM 4.2. *Any bounded solution* $\mathbf{y}$ *of* (II) *is an APT of* (I).

Proof. Let us prove that $\mathbf{y}$ satisfies Theorem 4.1(ii). Set $v(t) = \dot{\mathbf{y}}(t) - U(t) \in F^{\delta(t)}(\mathbf{y}(t))$. Then,

$$\mathbf{y}(t+s) - \mathbf{y}(t) = \int_0^s v(t+\tau)d\tau + \int_t^{t+s} U(\tau)d\tau. \tag{4.2}$$

By assumption (iii) of (II), the second integral goes to 0 as $t \to \infty$. The boundedness of $\mathbf{y}$, $\mathbf{y}(\mathbb{R}) \subset M$, M compact (combined with the fact that F has linear growth) implies boundedness of v and shows that $\mathbf{y}$ is uniformly continuous. Thus the family $\Theta^t(\mathbf{y})$ is equicontinuous, and hence relatively compact. Let $\mathbf{z} = \lim_{t_n \to \infty} \Theta^{t_n}(\mathbf{y})$ be a limit point. Set $t = t_n$ in (4.2) and define $v_n(s) = v(t_n + s)$. Then, using the assumption (iii) on U, the second term in the right-hand side of this equality goes to zero uniformly on compact intervals when $n \to \infty$. Hence

$$\mathbf{z}(s) - \mathbf{z}(0) = \lim_{n \to \infty} \int_0^s v_n(\tau)d\tau.$$

Since (v_n) is uniformly bounded, it is bounded in $L^2[0,s]$, and by the Banach–Alaoglu theorem, a subsequence of v_n will converge weakly in $L^2[0,s]$ (or weak* in $L^\infty[0,s]$) to some function v with $v(t) \in F(\mathbf{z}(t))$, for almost every t, since $v_n(t) \in F^{\delta(t+t_n)}(\mathbf{y}(t+t_n))$ for every t. Here we use (ii) and that F is upper semicontinuous with convex values. In fact, by Mazur's theorem, a convex combination of $\{v_m, m \geq n\}$ converges almost surely to v and $\lim_{m\to\infty} \mathrm{Co}(\bigcup_{n \geq m} F^{\delta(t+t_n)}(\mathbf{y}(t+t_n))) \subset F(\mathbf{z}(t))$. Hence $\mathbf{z}(s) - \mathbf{z}(0) = \int_0^s v(\tau)d\tau$, proving that $\mathbf{z}$ is a solution of (I) and hence $\mathbf{z} \in S_{M,M}$. □

4.3. APTs are internally chain transitive.

THEOREM 4.3. *Let* $\mathbf{z}$ *be a bounded APT of* (I). *Then* $L(\mathbf{z})$ *is internally chain transitive.*

Proof. The set $\{\Theta^t(\mathbf{z}) : t \geq 0\}$ is relatively compact, and hence the ω-limit set of $\mathbf{z}$ for the flow Θ,

$$\omega_\Theta(\mathbf{z}) = \bigcap_{t \geq 0} \overline{\{\Theta^s(\mathbf{z}) : s \geq t\}},$$

is internally chain transitive. (By standard properties of ω-limit sets of bounded semiorbits, $\omega_\Theta(\mathbf{z})$ is a nonempty, compact, internally chain transitive set invariant under Θ; see Conley [13]; a short proof is also in Benaïm [3, Corollary 5.6].) By property (4.1), $\omega_\Theta(\mathbf{z}) \subset S$, the set of all solutions of (I).

Let $\Pi : (C^0(\mathbb{R}, \mathbb{R}^m), \mathbf{D}) \to (\mathbb{R}^m, \|\cdot\|)$ be the projection map defined by $\Pi(\mathbf{z}) = \mathbf{z}(0)$. One has $\Pi(\omega_\Theta(\mathbf{z})) = L(\mathbf{z})$. In fact if $p = \lim_{n\to\infty} \mathbf{z}(t_n)$, let $\mathbf{w}$ be a limit point of $\Theta^{t_n}(\mathbf{z})$. Then $\mathbf{w} \in \omega_\Theta(\mathbf{z})$ and $\Pi(\mathbf{w}) = p$.

It then easily follows that $L(\mathbf{z})$ is nonempty compact and invariant under Φ since $\omega_\Theta(\mathbf{z}) \subset S$. Since Π has Lipschitz constant 1, Π maps every (ε, T) chain for Θ to an (ε, T) chain for Φ. This proves that $L(\mathbf{z})$ is internally chain transitive for Φ. □

5. Applications.

5.1. Approachability. An application of Proposition 3.25 is the following result, which can be seen as a continuous asymptotic deterministic version of Blackwell's approachability theorem [8]. Note that one has no property on uniform speed of convergence.

Given a compact set $\Lambda \in \mathbb{R}^m$ and $x \in \mathbb{R}^m$, we let $\Pi_\Lambda(x) = \{y \in \Lambda : d^2(x, \Lambda) = \|x - y\|^2 = \langle x - y, x - y\rangle\}$.

COROLLARY 5.1. *Let $\Lambda \subset \mathbb{R}^m$ be a compact set, $r > 0$, and $U = \{x \in \mathbb{R}^m : d(x, \Lambda) < r\}$. Suppose that for all $x \in U \setminus \Lambda$ there exists $y \in \Pi_\Lambda(x)$ such that the affine hyperplane orthogonal to $[x, y]$ at y separates x from $x + F(x)$. That is,*

$$\langle x - y, x - y + v\rangle \le 0 \tag{5.1}$$

for all $v \in F(x)$. Then Λ contains an attractor for (I) *with fundamental neighborhood U.*

Proof. Set $V(x) = d(x, \Lambda)$. To apply Proposition 3.25 it suffices to verify condition (iii) of Proposition 3.25. Condition (i) will follow, and condition (ii) is clearly true.

Let $\mathbf{x}$ be a solution to (I) with initial condition $x \in U \setminus \Lambda$. Set $\tau = \inf\{t > 0 : \mathbf{x}(t) \in \Lambda\} \le \infty$, $g(t) = V(\mathbf{x}(t))$, and let $I \subset [0, \tau[$ be the set of $0 \le t < \tau$ such that $g'(t)$ and $\dot{\mathbf{x}}(t)$ exist and $\dot{\mathbf{x}}(t) \in F(\mathbf{x}(t))$. For all $t \in I$ and $y \in \Pi_\Lambda(\mathbf{x}(t))$

$$\begin{aligned} g(t + h) - g(t) &\le \|\mathbf{x}(t + h) - y\| - \|\mathbf{x}(t) - y\| \\ &= \|\mathbf{x}(t) + \dot{\mathbf{x}}(t)h - y\| - \|\mathbf{x}(t) - y\| + |h|\varepsilon(h), \end{aligned}$$

where $\lim_{h\to 0} \varepsilon(h) = 0$. Hence

$$\begin{aligned} g'(t) &\le \frac{1}{\|\mathbf{x}(t) - y\|}\langle \mathbf{x}(t) - y, \dot{\mathbf{x}}(t)\rangle \\ &= -g(t) + \frac{1}{\|\mathbf{x}(t) - y\|}\langle \mathbf{x}(t) - y, \mathbf{x}(t) - y + \dot{\mathbf{x}}(t)\rangle. \end{aligned}$$

Thus, $\dot{x} \in F(x)$ and (5.1) imply $g'(t) \le -g(t)$ for all $t \in I$. Since g and $\mathbf{x}$ are absolutely continuous, I has full measure in $[0, \tau[$. Hence $g(t) \le e^{-t}g(0)$ for all $t < \tau$. Therefore $V(\mathbf{x}(t)) < V(x)$ for all $0 < t < \tau$, which shows (iii). Finally, $V(\mathbf{x}(t)) \le e^{-t}V(x)$ shows that the sets $V^{-1}[0, r')$ (with $0 < r' \le r$) are fundamental neighborhoods of the attractor in Λ. □

In particular, if any point of E has a unique projection on Λ (for example, Λ convex), then $\overline{C} = C$, and one recovers exactly Blackwell's sufficient condition for approachability.

COROLLARY 5.2 (Blackwell's approachability theorem). *Consider the decision making process described in section* 2.1*, Example* 2.2*. Let $\Lambda \subset E$ be a compact set. Assume that there exists a strategy Q such that for all $x \in E \setminus \Lambda$ there exists $y \in \Pi_\Lambda(x)$ such that the hyperplane orthogonal to $[x, y]$ through y separates x from $\overline{C}(x)$. Then Λ is approachable.*

Proof. Let $L(x_n)$ denote the limit set of $\{x_n\}$. By Corollary 5.1, Λ is an attractor with fundamental neighborhood E, hence a global attractor. Thus Theorem 3.6 with Proposition 2.1 and Corollary 3.24 imply that $L(x_n)$ is almost surely contained in Λ. □

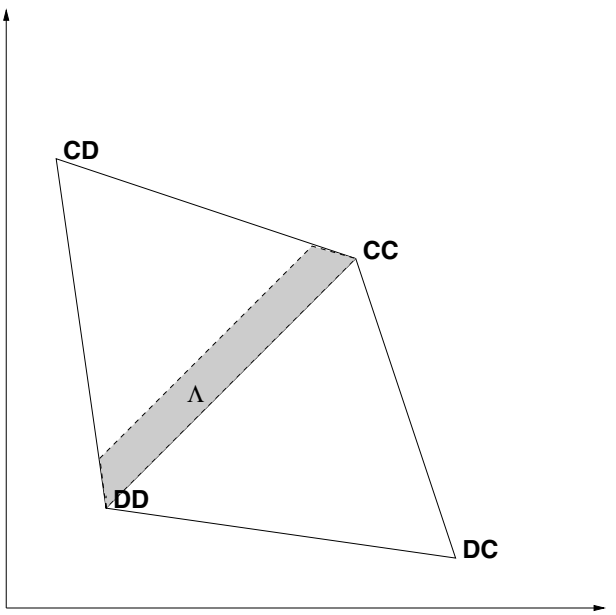
5.2. Smale's approach to the prisoner's dilemma. We develop here Example 2.4. Consider a 2×2 prisoner's dilemma game. Each player has two possible actions: cooperate (play C) or defect (play D). If both cooperate, each receives α; if both defect, each receives λ; if one cooperates and the other defects, the cooperator receives β and the defector γ. We suppose that $\gamma > \alpha > \lambda > \beta$, as is usual with a prisoner's dilemma game. We furthermore assume that

$$\gamma - \alpha < \alpha - \beta,$$

so that the outcome space E is the convex quadrilateral whose vertices are the payoff vectors

$$\mathbf{CD} = (\beta, \gamma), \qquad \mathbf{CC} = (\alpha, \alpha), \qquad \mathbf{DC} = (\gamma, \beta), \qquad \mathbf{DD} = (\lambda, \lambda);$$

see the figure below.



The outcome space E

Let δ be a nonnegative parameter. Adapting Smale [27] and Benaïm and Hirsch [4, 5], a *δ-good* strategy for player 1 is a strategy $Q^1 = \{Q^1_x\}$ (as defined in section 2.1) enjoying the following features:

$$Q^1_x(\text{play C}) = 1 \quad \text{if } x^1 > x^2$$

and

$$Q^1_x(\text{play C}) = 0 \quad \text{if } x^1 < x^2 - \delta.$$

The following result reinterprets the results of Smale [27] and Benaïm and Hirsch [4, 5] in the framework of approachability. It also provides some generalization (see Remark 5.4 below).

THEOREM 5.3. (i) *Suppose that player* 1 *plays a* δ*-good strategy. Then the set*

$$\Lambda = \{x \in E : x^2 - \delta \leq x^1 \leq x^2\}$$

is approachable.

(ii) *Suppose that both players play a* δ*-good strategy and that at least one of them is continuous (meaning that the corresponding function* $x \to Q^i_x$*(play* C*) is continuous). Then*

$$\lim_{n\to\infty} x_n = \mathbf{CC}$$

almost surely.

Proof. (i) Let $x \in E \setminus \Lambda$. If $x^1 > x^2$, then

$$C(x) = \overline{C}(x) = [\mathbf{CC}, \mathbf{CD}],$$

and the line $\{u \in \mathbb{R}^2 : u^1 = u^2\}$ separates x from $\overline{C}(x)$. Similarly if $x^1 < x^2 - \delta$, then

$$C(x) = \overline{C}(x) = [\mathbf{DD}, \mathbf{DC}],$$

which is separated from x by the line $\{u \in \mathbb{R}^2 : u^1 = u^2 - \delta\}$. Assertion (i) then follows from Corollary 5.2.

(ii) If both play a δ-good strategy, then (i) and its analogue for player 2 imply that the diagonal

$$\Delta = \{x \in E : x^1 = x^2\}$$

is approachable. Thus $L(x_n) \subset \Delta$. Also (by Proposition 2.1, Theorem 3.6, and Lemma 3.5) $L(x_n)$ is invariant under the differential inclusion induced by

$$F(x) = -x + \overline{C}(x),$$

where $C(x) = C^1(x) \cap C^2(x)$ and $C^i(x)$ is the convex set associated with Q^i (the strategy of player i). Suppose that one player, say 1, plays a continuous strategy. Then $\overline{C}(x) \subset \overline{C^1}(x) = C^1(x)$ and for all $x \in \Delta$, $C^1(x) = [\mathbf{CD}, \mathbf{CC}]$. Now, there is only one subset of Δ which is invariant under $\dot{x} \in -x + [\mathbf{CD}, \mathbf{CC}]$; this is the point $\mathbf{CC}$. This proves that $L(x_n) = \mathbf{CC}$. □

Remark 5.4. (i) In contrast to Smale [27] and Benaïm and Hirsch [4, 5], observe that assertion (i) makes no hypothesis on player 2's behavior. In particular, it is unnecessary to assume that player 2 has a strategy of the form defined by section 2.1.

(ii) The regularity assumptions (on strategies) are much weaker than in Benaïm and Hirsch [4, 5].

(iii) A 0-good strategy makes the diagonal Δ approachable. However, if both players play a 0-good strategy, then $\overline{C}(x) = E$ for all $x \in \Delta$, and we are unable to predict the long-term behavior of $\{x_n\}$ on Δ.

5.3. Fictitious play in potential games. Here we generalize the result of Monderer and Shapley [25]. They prove convergence of the classical discrete fictitious play process, as defined in Example 2.3, for n-linear payoff functions. Harris [17] studies the best-response dynamics in this case but does not derive convergence of fictitious play from it. Our limit set theorem provides the right tool for doing this, even in the following, more general setting.

Let X^i, $i = 1, \dots, n$, be compact convex subsets of Euclidean spaces and $U\colon X^1 \times \cdots \times X^n \to \mathbb{R}$ be a C^1 function which is concave in each variable. U is interpreted as the common payoff function for the n players. We write $x = (x^i, x^{-i})$ and define $BR^i(x^{-i}) := \operatorname{Argmax}_{x^i \in X^i} U(x)$ the set of maximizers. Then $x \mapsto BR(x) = (BR^1(x^{-1}), \dots, BR^n(x^{-n}))$ is upper semicontinuous (by Berge's maximum theorem, since U is continuous) with nonempty compact convex values. Consider the best response dynamics

$$\dot{\mathbf{x}} \in BR(\mathbf{x}) - \mathbf{x}. \tag{5.2}$$

Its constant solutions $\mathbf{x}(t) \equiv \hat{x}$ are precisely the Nash equilibria $\hat{x} \in BR(\hat{x})$; i.e., $U(\hat{x}) \geq U(x^i, \hat{x}^{-i})$ for all i and $x^i \in X^i$. Along a solution $\mathbf{x}(t)$ of (5.2), let $u(t) = U(\mathbf{x}(t))$. Then for almost all $t > 0$,

$$\dot{u}(t) = \sum_{i=1}^{n} \frac{\partial U}{\partial x^i}(\mathbf{x}(t))\dot{\mathbf{x}}^i(t) \tag{5.3}$$

$$\geq \sum_{i=1}^{n} [U(\mathbf{x}^i(t) + \dot{\mathbf{x}}^i(t), \mathbf{x}^{-i}(t)) - U(\mathbf{x}(t))] \tag{5.4}$$

$$= \sum_{i=1}^{n} \left[\max_{y^i \in X^i} U(y^i, \mathbf{x}^{-i}(t)) - U(\mathbf{x}(t)) \right] \geq 0, \tag{5.5}$$

where from (5.3) to (5.4) we use the concavity of U in x^i, and (5.5) follows from (5.2) and the definition of BR^i. Since the function $t \mapsto u(t)$ is locally Lipschitz, this shows that it is weakly increasing. It is constant in a time interval T, if and only if $\mathbf{x}^i(t) \in BR^i(\mathbf{x}^{-i}(t))$ for all $t \in T$ and $i = 1, \dots, n$, i.e., if and only if $\mathbf{x}(t)$ is a Nash equilibrium for $t \in T$ (but $\mathbf{x}(t)$ may move in a component of the set of Nash equilibria (NE) with constant U).

THEOREM 5.5. *The limit set of every solution of* (5.2) *is a connected subset of NE, along which U is constant. If, furthermore, the set $U(NE)$ contains no interval in $\mathbb{R}$, then the limit set of every fictitious play path is a connected subset of NE along which U is constant.*

Proof. The first statement follows from the above. The second statement follows from Theorem 3.6 together with Proposition 3.27 with $V = -U$ and $\Lambda = NE$. □

Remark 5.6. The assumption that the set $U(NE)$ contains no interval in $\mathbb{R}$ follows via Corollary 3.28 if U is smooth enough (e.g., in the n-linear case) and if each X^i has at most countably many faces, by applying Sard's lemma to the interior of each face.

Example 5.7 (2×2 coordination game). The global attractor of (5.2) consists of three equilibria and two line segments connecting them. The internally chain transitive sets are the three equilibria. Hence every fictitious play process converges to one of these equilibria.

The case of (continuous concave-convex) two-person zero-sum games was treated in Hofbauer and Sorin [21], where it is shown that the global attractor of (5.2) equals the set of equilibria. In this case the full strength of Theorem 3.6 and the notion of chain transitivity are not needed; the invariance of the limit set of a fictitious play path implies that it is contained in the global attractor; compare Corollary 3.24.

Acknowledgments. This research was started during visits of Josef Hofbauer in Paris in 2002. Josef Hofbauer thanks the Laboratoire d'Econométrie, Ecole Polytechnique, and the D.E.A. OJME, Université P. et M. Curie - Paris 6, for financial support

and Sylvain Sorin for his hospitality. Michel Benaïm thanks the Erwin Schrödinger Institute, and the organizers and participants of the 2004 Kyoto workshop on "game dynamics."

REFERENCES

[1] J.-P. Aubin and A. Cellina, *Differential Inclusions*, Springer, New York, 1984.
[2] M. Benaïm, *A dynamical system approach to stochastic approximations*, SIAM J. Control Optim., 34 (1996), pp. 437–472.
[3] M. Benaïm, *Dynamics of stochastic approximation algorithms*, in Séminaire de Probabilités XXXIII, Lecture Notes in Math. 1709, Springer, New York, 1999, pp. 1–68.
[4] M. Benaïm and M. W. Hirsch, *Stochastic Adaptive Behavior for Prisoner's Dilemma*, 1996, preprint.
[5] M. Benaïm and M. W. Hirsch, *Asymptotic pseudotrajectories and chain recurrent flows, with applications*, J. Dynam. Differential Equations, 8 (1996), pp. 141–176.
[6] M. Benaïm and M. W. Hirsch, *Mixed equilibria and dynamical systems arising from fictitious play in perturbed games*, Games Econom. Behav., 29 (1999), pp. 36–72.
[7] M. Benaïm, J. Hofbauer, and S. Sorin, *Stochastic Approximations and Differential Inclusions: Applications*, Cahier du Laboratoire d'Econometrie, Ecole Polytechnique, 2005-011.
[8] D. Blackwell, *An analog of the minmax theorem for vector payoffs*, Pacific J. Math., 6 (1956), pp. 1–8.
[9] I. U. Bronstein and A. Ya. Kopanskii, *Chain recurrence in dynamical systems without uniqueness*, Nonlinear Anal., 12 (1988), pp. 147–154.
[10] G. Brown, *Iterative solution of games by fictitious play*, in Activity Analysis of Production and Allocation, T. C. Koopmans, ed., Wiley, New York, 1951, pp. 374–376.
[11] R. Buche and H. J. Kushner, *Stochastic approximation and user adaptation in a competitive resource sharing system*, IEEE Trans. Automat. Control, 45 (2000), pp. 844–853.
[12] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth Analysis and Control Theory*, Springer, New York, 1998.
[13] C. C. Conley, *Isolated Invariant Sets and the Morse Index*, CBMS Reg. Conf. Ser. in Math. 38, AMS, Providence, RI, 1978.
[14] M. Duflo, *Algorithmes Stochastiques*, Springer, New York, 1996.
[15] D. Fudenberg and D. K. Levine, *The Theory of Learning in Games*, MIT Press, Cambridge, MA, 1998.
[16] I. Gilboa and A. Matsui, *Social stability and equilibrium*, Econometrica, 59 (1991), pp. 859–867.
[17] C. Harris, *On the rate of convergence of continuous time fictitious play*, Games Econom. Behav., 22 (1998), pp. 238–259.
[18] M. W. Hirsch, *Differential Topology*, Springer, New York, 1976.
[19] J. Hofbauer, *Stability for the Best Response Dynamics*, preprint, 1995.
[20] J. Hofbauer and W. H. Sandholm, *On the global convergence of stochastic fictitious play*, Econometrica, 70 (2002), pp. 2265–2294.
[21] J. Hofbauer and S. Sorin, *Best response dynamics for continuous zero-sum games*, in Cahier du Laboratoire d'Econometrie, Ecole Polytechnique, 22002-2028.
[22] M. Kunze, *Non-Smooth Dynamical Systems*, Lecture Notes in Math. 1744, Springer, New York, 2000.
[23] H. J. Kushner and G. G. Yin, *Stochastic Approximations Algorithms and Applications*, Springer, New York, 1997.
[24] L. Ljung, *Analysis of recursive stochastic algorithms*, IEEE Trans Automat. Control, 22 (1977), pp. 551–575.
[25] D. Monderer and L. S. Shapley, *Fictitious play property for games with identical interests*, J. Econom. Theory, 68 (1996), pp. 258–265.
[26] J. Robinson, *An iterative method of solving a game*, Ann. Math., 54 (1951), pp. 296–301.
[27] S. Smale, *The prisoner's dilemma and dynamical systems associated to non-cooperative games*, Econometrica, 48 (1980), pp. 1617–1633.
[28] S. Sorin, *A First Course on Zero-Sum Repeated Games*, Springer, New York, 2002.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 1, pp. 349–366

A SPILLOVER PHENOMENON IN THE OPTIMAL LOCATION OF ACTUATORS*

PASCAL HÉBRARD† AND ANTOINE HENROT‡

Abstract. In this paper, we are interested in finding the optimal location and shape of the actuators in a stabilization problem. Namely, we consider the one-dimensional wave equation damped by an internal feedback supported on a subdomain ω of given length. The criterion we want to optimize represents the rate of decay of the total energy of the system. It theoretically involves all the eigenmodes of the operator. From an engineering point of view, it seems more realistic to consider only a finite number of modes, say the N first ones. In that context, we are able to prove existence and uniqueness of an optimal domain ω_N^*: it is the better possible location for the actuators. We characterize this optimal domain and we point out the following strange phenomenon (at least for small lengths): the optimal domain ω_N^* which is the better one for the N first modes is actually the worse one for the $N+1$th mode. This looks like the well-known spillover phenomenon in control theory. At last, we will give some possible extension and open problems in higher dimension.

Key words. damped wave equation, optimal location, spillover, stabilization

AMS subject classifications. 49J20, 93B55, 93C20, 35L05

DOI. 10.1137/S0363012903436247

1. Introduction.

In control and stabilization problems the choice of the best location (and shape) of the actuators is a very important and practical question. Among criterion which can be studied, the *rate of decay* of the energy of the system is an important one since it does not depend on the initial conditions. In the one-dimensional problem that we are going to consider here (wave equation with internal distributed control), it is known that this rate of decay is precisely the opposite of the spectral abscissa of the corresponding operator (see [9]). Therefore, optimizing this rate of decay consists in pushing *all* the eigenvalues as far as possible to the left in the complex plane. Among works in this direction, we refer, e.g., to [8] where it is proved that a constant damping is a local maximizer of the rate of decay [11], which shows that the constant damping is not a global maximizer and [5] where they show that we can achieve an arbitrarily large rate of decay by considering damping of the kind $a(x) = 1/(x+b)$ (see below for the mathematical model). In this work we will restrict ourselves to a damping of the kind $k\chi_\omega(x)$ where k is a (small) positive constant and χ_ω is the characteristic function of a subdomain ω of the string which is our main unknown.

In higher dimension, this spectral abscissa is also an important component of the rate of decay, but we must also consider a geometric quantity describing the time each high frequency (or waves with a little wave length) stays in the zone of control; see [3], [18].

From an engineering point of view, it seems to be difficult (and perhaps useless) to take into account an *infinite* number of modes. So, a more reasonable version of this

*Received by the editors October 17, 2003; accepted for publication (in revised form) December 22, 2004; published electronically August 22, 2005.
http://www.siam.org/journals/sicon/44-1/43624.html

†Dassault Systems and Institut Élie Cartan Nancy UMR 7502 UHP-CNRS-INRIA 54506 Vandoeuvre-les-Nancy, France (pascal_hebrard@ds-fr.com).

‡(Corresponding author). École des Mines de Nancy-INPL, Institut Élie Cartan Nancy UMR 7502 UHP-CNRS-INRIA 54506 Vandoeuvre-les-Nancy, France (henrot@iecn.u-nancy.fr).

problem would be to consider only the N first modes. Indeed, the high frequencies are not too much penalizing for the vibrating structure. The aim of this paper is to show that if we choose the optimal domain, say ω_N^*, for the N first modes (we will prove that it exists and is unique), this domain behaves very poorly for the first mode that we have forgotten: the $(N+1)$th mode! More precisely, ω_N^* generally concentrates on the nodes of the $(N+1)$th mode, and therefore does not control it at all. Roughly speaking, *the best domain for the N first modes is the worse one for the $(N+1)$th mode!* In this paper, we are able to prove this result for dampers with *small* support, but actually we can observe numerically that, in general, the best damper for the N first modes behaves very poorly for the $N+1$th mode; see Figure 5.1 and section 5.2. Since the choice of N is generally arbitrary, it seems to be a very bad idea to look for the optimal zone of control for the N first modes. It is, somehow, as if we push the energy after the N first modes like in classical spillover phenomena, as described, for example, in [2]. In our paper, we choose to consider a wave packet constituted with all the low frequencies. In some sense, the phenomenon which occurs here seems similar to the one due to different group velocity as described in [25]; see also [26] for a survey, but in our case it is not only a question of low and high frequencies. Indeed, we could have chosen other wave packets, in particular with different group velocities. Actually, we can observe (at least numerically, we did not write proofs) the same phenomenon for any choice of wave packets. For example, if we want to damp at best the packet of eigenfrequencies $\lambda_1, \lambda_4, \lambda_5$, the optimal domain we get (which still exists and is unique) will mainly concentrate on the nodes of the second eigenfunction ϕ_2, which is the first frequency we have forgotten, and therefore will be unable to damp correctly this eigenmode. Nevertheless, E. Zuazua claims in [26] that *controlling a discrete version of a continuous wave model is often a bad way of controlling the continuous wave model itself.* We have another illustration of this phenomenon here.

The plan of this paper is the following. Section 2 deals with the mathematical model which is used and fix the notations. In section 3 we prove existence and uniqueness of the optimal domain and we characterize it. Section 4 is devoted to describe and prove the kind of spillover phenomenon that we have just described above. At last, in section 5, we will give some remarks and possible extensions to the two-dimensional case.

2. The mathematical model. Let us now give the model and the notations that we are going to use throughout this paper. We consider a string (a one-dimensional model), but it is essentially for technical reasons. We will say a few words, in section 5, about higher dimensional models, pointing out what has to be done to generalize our results.

So, let us denote by $\Omega = (0,1)$ the unit string that we suppose fixed at its extremities. We want to stabilize this string thanks to a damping acting only on a subdomain ω. More precisely, we consider the following modelling. The displacement u of the string in presence of viscous damping $2k\chi_\omega$ (where χ_ω denotes the characteristic function of the subdomain ω of positive length), satisfies the damped wave equation

$$\begin{cases} u_{tt}(x,t) - u_{xx}(x,t) + 2k\chi_\omega(x)u_t(x,t) = 0, \ x \in (0,1), t > 0 \\ u(0,t) = u(1,t) = 0, \ t > 0 \end{cases} \tag{2.1}$$

upon being set in motion by the initial disturbance

$$u(x,0) = u_0(x), \qquad u_t(x,0) = u_1(x) \qquad \forall x \in [0,1]. \tag{2.2}$$

The energy of the string at time t is defined by

$$E(t) = \int_0^1 \left[u_x^2(x,t) + u_t^2(x,t)\right] dx.$$

If ω has positive measure, this system is exponentially stable, i.e., its energy is known to obey (see, e.g., [6], [9])

$$E(t) \leq CE(0)e^{-2\tau t} \tag{2.3}$$

for some constants $C > 0$ and $\tau > 0$ independent of the initial data. We define the decay rate, as a function of k and ω, to be the largest such τ,

$$\tau(k,\omega) = \sup\{\tau' : \exists C(\tau') > 0 \; s.t. \; E(t) \leq CE(0)e^{-2\tau' t}, \\ \text{for every solution of (2.1) and (2.2)}\}.$$

Cox and Zuazua have shown in [9] that if χ_ω is of bounded variation, *i.e.,* ω is the union of a finite number of intervals, then $\tau(k,\omega)$ is equal to the opposite of the spectral abscissa of the operator A:

$$\tau(k,\omega) = -\mu = -\sup\{Re\,\lambda : \lambda \in sp(A)\},$$

where A denotes the linear operator associated to (2.1):

$$A = \begin{pmatrix} 0 & I \\ \frac{d^2}{dx^2} & -2k\chi_\omega(x) \end{pmatrix}, \qquad D(A) = (H^2(0,1) \cap H_0^1(0,1)) \times H_0^1(0,1) \tag{2.4}$$

and $sp(A)$ its spectrum. Therefore, a natural question would be to look for k and ω which minimize this spectral abscissa (or maximize $\tau(k,\omega)$).

In such a generality, looking for the maximizer of $(k,\omega) \mapsto \tau(k,\omega)$ is quite difficult. In [15], we explain (and we give justifications) how to simplify the problem by considering, instead of the decay rate, the quantity

$$J(\omega) = \inf_{n\in\mathbb{N}^*} \int_0^1 \chi_\omega(x)\phi_n^2(x)dx, \tag{2.5}$$

where $\mathbb{N}^*$ stands for the set of positive integers and $(\phi_n)_{n\in\mathbb{N}^*}$ denote the normalized eigenfunctions for the problem without damping, *i.e.,* $\phi_n = \sqrt{2}\sin n\pi x$. Actually, when k is not too large, we have $\tau(k,\omega) \simeq kJ(\omega)$, since $J(\omega)$ is nothing else but the derivative of τ with respect to k for $k = 0$. On the other hand, taking k large is not interesting at all, due to the classical *overdamping* phenomenon described, e.g., in [9], [12], [15]. Therefore, $J(\omega)$ gives a good approximation of the decay rate for small k. We also refer to [13] for a similar analysis.

As explained in the Introduction, it seems more realistic, at least from an engineering point of view, to take into consideration only a *finite* number of modes. It means that it seems reasonable to replace the functional J by the simpler J_N (where N is a given integer), defined by

$$J_N(\omega) = \min_{1\leq n\leq N} \int_0^1 \chi_\omega(x)\phi_n^2(x)dx. \tag{2.6}$$

Therefore, we are interested in solving the following problem:

$$\mathcal{P}_\omega \qquad \begin{cases} \text{Find } \omega^* \text{ subset of }]0,1[\text{ of measure, } l \text{ which maximizes} \\ J_N(\omega) = \min_{1\leq n\leq N} 2\int_0^1 \chi_\omega(x)\sin^2(n\pi x)\,dx. \end{cases}$$

In what follows, for each integer k and each function a, we will denote by $j_k(a)$ the quantity

$$j_k(a) = \int_0^1 a(x)\phi_k^2(x)dx\,.$$

We could also wonder whether J_N is a "good" approximation of J. Since it is proved in [15] that J has no maximizer in the class of characteristic functions (except for the particular case $l = 0.5$) while J_N has always a (unique) maximizer in this class (as proved below), it seems at first sight that it is not a good approximation. But, it becomes a good one if we accept working in the convex set $\mathcal{A}_l$ defined in (3.1). Actually, we can prove the following proposition.

PROPOSITION 2.1. *Let us consider the functionals J_N and J defined on the convex set $\mathcal{A}_l$ (defined in* (3.1)*), respectively, by $J_N(a) = \min_{1\le n\le N} j_n(a)$ and $J(a) = \inf_{n\in\mathbb{N}^*} j_n(a)$. Then J_N Γ-converge to J in the sense of De Giorgi. Moreover, let $\chi_{\omega_N^*}$ be the sequence of maximizers of J_N given by Theorem* 3.1*, then $\chi_{\omega_N^*}$ converges (up to a subsequence) weak-* to a maximizer of J and $\max_{\mathcal{A}_l} J = \lim_{N\to+\infty} J_N(\chi_{\omega_N^*})$.*

Proof. Since we are interested in a *maximization* problem, the definition of Γ-convergence reads here; see, e.g., [10]:

(i) For all sequence a_n in $\mathcal{A}_l$, which converge weak-* to a, $J(a) \ge \limsup J_N(a_N)$.
(ii) There exists one sequence a_n in $\mathcal{A}_l$, which converge weak-* to a, such that $J(a) \le \liminf J_N(a_N)$.

For (i), let us fix $\varepsilon > 0$ and choose an integer k_0 such that $J(a) \le j_{k_0}(a) \le J(a) + \varepsilon$. For every integer $N \ge k_0$, we have

$$J_N(a_N) \le j_{k_0}(a_N). \tag{2.7}$$

Now, $j_{k_0}(a_N) \to j_{k_0}(a)$ when $N \to +\infty$, therefore taking the lim-sup in both sides of (2.7) yields $\limsup J_N(a_N) \le \limsup j_{k_0}(a_N) = j_{k_0}(a) \le J(a) + \varepsilon$ which gives (i) since ε is arbitrary.

For (ii), it suffices to consider a constant sequence $a_n = a$ since $J(a) \le J_N(a)$.

Now, the last two claims come directly from the classical theorem of De Giorgi (see [10]) and the fact that the sequence $\chi_{\omega_N^*}$ is precompact in $\mathcal{A}_l$. □

The functional J may have several maxima but the characterization of ω_N^* which is given below (Theorem 4.1) shows that $\chi_{\omega_N^*}$ converges actually to the constant function $a(x) = l$ which is the more natural maximizer.

3. Existence, uniqueness, and characterization of the optimum. We begin by proving the following existence and uniqueness result for the optimal domain.

THEOREM 3.1. *The problem $\mathcal{P}_\omega$ has a unique solution ω_N^*. This solution is a union of at most N intervals. It is symmetric with respect to* $1/2$.

The proof of Theorem 3.1 will be done in several steps. First of all, we are going to use some kind of relaxation of the problem by introducing the convex hull $\mathcal{A}_l$ of the set of characteristic functions. Existence of an optimum in this set will be obtained easily. By characterization of this optimum thanks to the optimality conditions, we will be able to prove that it is indeed an extreme point of $\mathcal{A}_l$, i.e., the characteristic function of a subdomain. Uniqueness will then follow from the fact that the functional J is concave.

Step 1 (*relaxation*). The maximization problem is posed on the set of characteristic functions

$$\mathcal{L}_l = \left\{ a(x) \in L^\infty(0,1), a(x) = 0 \text{ or } 1 \, a.e., \int_0^1 a(x)\, dx = l \right\}.$$

This set is not very convenient for this maximization problem, since it is not closed for the natural topology associated to the functional J_N, namely the weak-star topology on $L^\infty(0,1)$. Indeed, J_N is clearly continuous for this topology. So, let us introduce the convex hull of $\mathcal{L}_l$,

$$\mathcal{A}_l = \left\{ a(x) \in L^\infty(0,1),\ 0 \le a(x) \le 1,\ \int_0^1 a(x)\, dx = l \right\} \tag{3.1}$$

which is also the closure of $\mathcal{L}_l$ for the weak-star topology on $L^\infty(0,1)$. The set $\mathcal{A}_l$ is compact for this topology, while $\mathcal{L}_l$ coincides with the set of extreme points of $\mathcal{A}_l$; see, e.g., [16]. Moreover, J_N has a natural extension (always denoted by J_N) to $\mathcal{A}_l$ defined by

$$\forall a \in \mathcal{A}_l, \qquad J_N(a) = \min_{1 \le n \le N} \int_0^1 a(x) \phi_n^2(x) dx. \tag{3.2}$$

It is clear that J_N is continuous on $\mathcal{A}_l$ for the weak-star topology. Therefore, J_N admits (at least) a maximum in $\mathcal{A}_l$.

Step 2 (*optimality conditions*). Let a^* such a maximum, and let us denote by $I(a^*)$ the active index-set

$$I(a^*) = \{k \in \{1, 2, \ldots, N\}, \text{ such that } j_k(a^*) = J_N(a^*)\}.$$

It is well known in nonsmooth analysis (see, e.g., [17]) that the subdifferential of J_N at a^* is given by

$$\partial J_N(a^*) := co\{\cup \partial j_k(a^*),\ k \in I(a^*)\}, \tag{3.3}$$

where *co* denotes the convex hull. Now, the j_k being linear, they are equal to their differential and the optimality condition reads

$$0 \in \partial J_N(a^*) + \lambda_0 L_0, \tag{3.4}$$

where λ_0 stands for a Lagrange multiplier taking into account the length constraint and L_0 is the linear form defined by $< L_0, h >= \int_0^1 h(x)\, dx$. What yields, thanks to (3.3) is

$$\left\{ \begin{array}{c} \exists (\lambda_k) \in [0,1],\ k \in I(a^*),\ \sum \lambda_k = 1,\ \exists \lambda_0 \in \mathbb{R} \text{ such that} \\ \forall h \in L^\infty([0,1]),\ h \text{ admissible}, \\ \displaystyle\sum_{k \in I(a^*)} \lambda_k \int_0^1 h(x) \phi_k^2(x)\, dx + \lambda_0 \int_0^1 h(x)\, dx = 0. \end{array} \right. \tag{3.5}$$

Step 3 (*maxima are characteristic functions*). Let us fix $\varepsilon \in (0, 1/2)$. We are going to prove that the set $A_\varepsilon = \{x \in \Omega \,|\, \varepsilon \le a^*(x) \le 1 - \varepsilon\}$ has zero measure for every $\varepsilon > 0$ which obviously implies that a^* is a characteristic function. Let us assume, for a contradiction, $|A_\varepsilon| > 0$ and let us use the optimality conditions (3.5). We choose h

with a support in A_ε: it is clearly admissible (we refer, e.g., to [7], [4] for the complete description of the cone of admissible functions for the set $\mathcal{A}_l$). This implies that

$$\sum_{k\in I(a^*)} \lambda_k \phi_k^2(x) + \lambda_0 = 0, \text{ for almost every } x \in A_\varepsilon \tag{3.6}$$

If A_ε has positive measure, this equality can be extended to the whole interval by analyticity of the eigenfunctions. But such an identity is impossible since the system of functions

$$\{1, \phi_1^2, \dots, \phi_k^2\} = \{1, 2\sin^2 n_1\pi x, \dots, 2\sin^2 n_k\pi x\} \text{ is linearly independent.} \tag{3.7}$$

Consequently, for all ε, A_ε has zero measure which proves the desired result.
Step 4 (*uniqueness and symmetry*). J_N is clearly a concave function as a minimum of linear functions. Therefore, if it would exist two distinct maxima a_1^* and a_2^*, all the points in the segment $[a_1^*, a_2^*]$ would also be maxima. But it is impossible since we have proved in step 3 that all the maxima were extreme points of the convex $\mathcal{A}_l$. Now, uniqueness implies symmetry of the minimizer with respect to 1/2 since $J_N(a(x)) = J_N(a(1-x))$.
Step 5 (*at most N connected components*). We recall that the maximum $a^* = \chi_{\omega^*}$ satisfies the optimality condition (3.5). Let us introduce

$$\Psi_\Lambda(x) = 2\sum_{k\in I(a^*)} \lambda_k \sin^2(k\pi x).$$

The Lagrangian of the maximization problem can be written as

$$L(a, \lambda_0, \Lambda) = \int_0^1 a(x)\Psi_\Lambda(x)dx + \lambda_0\left(\int_0^1 a(x)dx - l\right).$$

Now, for every admissible function h and every $\varepsilon > 0$ small enough, we have

$$L(\chi_{\omega^*} + \varepsilon h, \lambda_0, \Lambda) - L(\chi_{\omega^*}, \lambda_0, \Lambda) \le 0$$

which can be rewritten, thanks to the linearity of L with respect to its first variable,

$$L(h, \lambda_0, \Lambda) \le 0, \quad \text{for } h \text{ admissible.}$$

Now, we can choose as admissible h a function satisfying

$$\begin{array}{l} \forall x \in \omega^*,\ h(x) \le 0 \\ \forall x \in \omega^{*c},\ h(x) \ge 0 \\ \int_0^1 h(x)dx = 0. \end{array}$$

For such a choice we get,

$$\begin{array}{l} \forall x \in \omega^*,\ \Psi_\Lambda + \lambda_0 \ge 0 \\ \forall x \in \omega^{*c},\ \Psi_\Lambda + \lambda_0 \le 0. \end{array} \tag{3.8}$$

By continuity of ψ_Λ, equations (3.8) imply that for all $x \in \partial\omega^* \cap (0,1)$,

$$\psi_\Lambda(x) + \lambda_0 = 0. \tag{3.9}$$

Now

$$\psi_\Lambda(x) = \sum_{k\in I(a^*)} \lambda_k \left(1 - \cos(2k\pi x)\right) = 1 - \sum_{k\in I(a^*)} \lambda_k T_k(\cos 2\pi x),$$

where T_k is the kth Tchebyshev polynomial. Therefore, $\psi_\Lambda(x)$ is a polynomial in $\cos 2\pi x$ of degree less or equal to N and the equation $\psi_\Lambda(x) + \lambda_0 = 0$ has *at most* $2N$ solutions in $]0,1[$. Consequently, ω^* has at most N connected components unless if ω^* contains an interval of the kind $[0,\eta]$ (and also $[1-\eta,1]$ by symmetry). But this last case cannot happen since, for every small $\beta > 0$ and for every integer n, $\int_0^\eta \phi_n(x)dx < \int_\beta^{\beta+\eta} \phi_n(x)dx$. □

The equation (3.9) shows that the optimal domain ω^* is a level set of the function Ψ_Λ but, of course, it remains to find the Lagrange multipliers $\Lambda = (\lambda_k)_{k\in I(a^*)}$ and also to find $I(a^*)$. The following theorem gives the answer to the second question and also gives a practical way to determine ω^* *at least* for small l.

THEOREM 3.2. *For each integer N, there exists a real $l_N \le 1$ such that for $l \le l_N$, the optimal domain ω_N^* satisfies*

$$j_1(\omega_N^*) = j_2(\omega_N^*) = \cdots = j_N(\omega_N^*). \tag{3.10}$$

Remark 1. The relations (3.10) together with the description of ω_N^* as a symmetric union of at most N intervals yields a practical way to determine the optimum. Indeed, let us assume for example that $N = 2M$ is even. We write

$$\omega_N^* = \bigcup_{k=1}^{M} [a_k - l_k/2, a_k + l_k/2] \cup \bigcup_{k=1}^{M} [1 - a_k - l_k/2, 1 - a_k + l_k/2],$$

then the relations (3.10) with the supplementary equality $\sum_{k=1}^M l_k = l/2$ yields a $2M \times 2M$ nonlinear system whose (unique) solution gives the desired domain. We will use this remark later in section 4.

Remark 2. The relations (3.10) do not hold for any value of the constraint l as it is shown by the following (numerical) example. Take $N = 3$ and $l = 0.9$ then the optimal domain is

$$\omega_3^* = [0.0475707, 0.3417644] \cup [0.3441937, 0.6558063] \cup [0.6582356, 0.9524293]$$

which satisfies $j_1(\omega_3^*) > j_2(\omega_3^*) = j_3(\omega_3^*)$ and $J_3(\omega_3^*) = 0.987672$ while for the best domain satisfying $j_1 = j_2 = j_3$, we have only $J_3(\omega) = 0.987177$. Actually, we can see numerically that the constant l_N decreases when N increases.

Proof of Theorem 3.2. The first idea consists of transposing the problem in finite dimension thanks to the following trick. Let $\mathcal{K}_N^l$ be the subset of $\mathbb{R}^N$ defined by

$$\mathcal{K}_N^l = \{X = (x_1, x_2, \ldots, x_N), \text{ s.t. } \exists a \in \mathcal{A}_l \text{ with } x_i = j_i(a), i = 1, 2, \ldots, N\}. \tag{3.11}$$

We will also write $\mathcal{K}_N^l = \mathcal{K}$ when N and l are fixed, since no misunderstanding is possible. The set $\mathcal{K}$ is obviously convex and compact, since it is the image of $\mathcal{A}_l$ by the linear (continuous) functional $a \mapsto (j_1(a), j_2(a), \ldots, j_N(a))$. We will give in section 5 some supplementary properties of $\mathcal{K}$.

The first bissectrix $\Delta = \{X \in \mathbb{R}^N,\ x_1 = x_2 = \cdots = x_N\}$ meets $\mathcal{K}$ (take $a = l$ a constant: we have $j_i(a) = l$ for all i). Therefore, we can introduce the point X^*, the furthest point of $\Delta \cap \mathcal{K}$,

$$X^* = (x^*, x^*, \dots, x^*) \text{ with } x^* = \max\{x \text{ such that } X = (x, x, \dots, x) \in \Delta \cap \mathcal{K}\}.$$

Note that $x^* > l$ since $a = l$ cannot be the maximizer of J_N. The claim in Theorem 3.2 is equivalent to say that X^* solves the maximization problem

$$G(X^*) = \max_{X \in \mathcal{K}} G(X), \quad \text{where } G(X) = \min_{1 \le k \le N} \{x_k\}. \tag{3.12}$$

A geometrical interpretation of (3.12) consists in saying that there is no point of $\mathcal{K}$ in the quadrant

$$\mathcal{Q} = \{X = (x_1, x_2, \dots, x_N) \in \mathbb{R}^N;\ x_i > x^*, i = 1, 2, \dots, N\}.$$

To prove Theorem 3.2, we argue by contradiction. We will prove that if $\mathcal{Q} \cap \mathcal{K}$ is not empty, there exists a point of the bissectrix Δ in $\mathcal{Q} \cap \mathcal{K}$ (at least for l small enough) which contradicts the fact that X^* maximizes G on $\Delta \cap \mathcal{K}$. For that purpose, we introduce X_k to be the furthest point of $\mathcal{K}$ in the direction of x_k (the kth coordinate). Actually, we can determine X_k. Indeed, X_k is obtained by solving the maximization problem: find $a \in \mathcal{A}_l$ which maximizes $\int_0^1 a(x) \sin^2 k\pi x \, dx$. It follows from the proof of Theorem 3.1 (steps $2, 3, 5$, and, in particular relation (3.9)) that the maximizer is a characteristic function $\chi_{\omega_k^*}$, level set of the function $\sin^2 k\pi x$. Therefore,

$$\omega_k^* = \bigcup_{j=0}^{k-1} \left[\frac{2j+1}{2k} - \frac{l}{2k}, \frac{2j+1}{2k} + \frac{l}{2k} \right].$$

We can easily deduce the coordinates of X_k. For $k = 1$,

$$j_m(\omega_1^*) = l + (-1)^{m+1} \frac{\sin m\pi l}{m\pi} \tag{3.13}$$

while, for $k \ge 2$,

$$j_m(\omega_k^*) = l - \sum_{j=0}^{k-1} \frac{1}{m\pi} \sin \frac{m\pi l}{k} \cos \frac{2m\pi(2j+1)}{2k}.$$

Using $\sum_{j=0}^{k-1} \cos \frac{m\pi(2j+1)}{k} = 0$ for $m \neq k$, this yields

$$j_m(\omega_k^*) = l \text{ if } m \neq k \text{ and } j_k(\omega_k^*) = l + \frac{1}{\pi} \sin \pi l. \tag{3.14}$$

For simplicity, let us put the origin at X^*. In this case, to sum up the vertices of the set $\mathcal{K}$ in the direction of the coordinate axis are the points

$$X_1 = \begin{pmatrix} h \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_N \end{pmatrix} \quad X_2 = \begin{pmatrix} -\alpha \\ h \\ -\alpha \\ \vdots \\ -\alpha \end{pmatrix} \quad X_3 = \begin{pmatrix} -\alpha \\ -\alpha \\ h \\ -\alpha \\ \vdots \end{pmatrix} \quad \dots \quad X_N = \begin{pmatrix} -\alpha \\ -\alpha \\ \vdots \\ -\alpha \\ h \end{pmatrix},$$

where

$$\begin{array}{l}\alpha = x^* - l > 0,\\ h = l + \frac{\sin \pi l}{\pi} - x^* > 0,\\ \text{and } \beta_i = l + (-1)^{i+1}\frac{\sin i\pi l}{i\pi} - x^*, i = 2, \ldots, N.\end{array} \tag{3.15}$$

As explained above, let us assume (for a contradiction) that there is a point $X_0 = (x_1^0, \ldots, x_N^0)^T$ in $\mathcal{Q} \cap \mathcal{K}$. We are looking for a point X, convex combination of $X^* = O$, X_0, X_1, $X_2, \ldots, X_N$ (which ensures that X will be in $\mathcal{K}$) that we want to be in the set $\mathcal{Q}$ and on the bissectrix Δ. Existence of such a point would lead to a contradiction since X^* maximizes G on $\Delta \cap \mathcal{K}$. We write X as

$$X = (1-s)O + tX_0 + \sum_{i=1}^{N} \lambda_i X_i,$$

where

$$0 < s \leq 1,\ 0 \leq t \leq 1,\ 0 \leq \lambda_i \leq 1,\ t + \sum_{i=1}^{N} \lambda_i = s. \tag{3.16}$$

Let us first express that X must belong to Δ. Writing $x_1 = x_2$, $x_1 = x_3$, etc. this yields the following system:

$$\left\{\begin{array}{c} tx_1^0 + \lambda_1 h - \lambda_2 \alpha = tx_2^0 + \lambda_1 \beta_2 + \lambda_2 h\\ tx_1^0 + \lambda_1 h - \lambda_3 \alpha = tx_3^0 + \lambda_1 \beta_3 + \lambda_3 h\\ \vdots\\ tx_1^0 + \lambda_1 h - \lambda_N \alpha = tx_N^0 + \lambda_1 \beta_N + \lambda_N h. \end{array}\right. \tag{3.17}$$

Summing these relations and using (3.16) gives

$$(N-1)tx_1^0 + (N-1)\lambda_1 h - \alpha(s - t - \lambda_1) = t\sum_{i=2}^{N} x_i^0 + \lambda_1 \sum_{i=2}^{N} \beta_i + h(s - t - \lambda_1)$$

what can be written, thanks to (3.15) is

$$t\left[\left(Nx_1^0 - \sum_{i=1}^{N} x_i^0\right) + \frac{\sin \pi l}{\pi}\right] + \lambda_1 \left[\sum_{i=2}^{N}(h - \beta_i) + \frac{\sin \pi l}{\pi}\right] = \frac{\sin \pi l}{\pi}\, s. \tag{3.18}$$

Let us introduce

$$\Sigma = \sum_{i=2}^{N}(h - \beta_i) + \frac{\sin \pi l}{\pi} = N\frac{\sin \pi l}{\pi} + \sum_{i=2}^{N}(-1)^i \frac{\sin i\pi l}{i\pi}.$$

From now on, we will assume $l \leq 1/(2N)$. Then, Σ is clearly positive. We get λ_1 from (3.18),

$$\lambda_1 = \left(\frac{\sin \pi l}{\pi}\, s - t\left[\left(Nx_1^0 - \sum_{i=1}^{N} x_i^0\right) + \frac{\sin \pi l}{\pi}\right]\right)/\Sigma. \tag{3.19}$$

Replacing in each equation (3.17) gives the value of λ_k,

$$\lambda_k = \frac{t(x_1^0 - x_k^0)}{\frac{\sin \pi l}{\pi}} + \frac{\lambda_1(h - \beta_k)}{\frac{\sin \pi l}{\pi}}. \tag{3.20}$$

We now use $h - \beta_k = \frac{\sin \pi l}{\pi} + (-1)^k \frac{\sin k\pi l}{k\pi}$ (see (3.15)) and we introduce normalized coordinates: $y_k^0 := \frac{x_k^0}{\frac{\sin \pi l}{\pi}}$. Therefore, (3.19) and (3.20) can be rewritten

$$\begin{aligned} \lambda_1 &= \frac{\frac{\sin \pi l}{\pi}}{\Sigma} \left(s - t - t \left(N y_1^0 - \sum_{i=1}^N y_i^0 \right) \right) \\ \lambda_k &= t(y_1^0 - y_k^0) + \frac{\frac{\sin \pi l}{\pi} + (-1)^k \frac{\sin k\pi l}{k\pi}}{\Sigma} \left(s - t - t \left(N y_1^0 - \sum_{i=1}^N y_i^0 \right) \right). \end{aligned} \tag{3.21}$$

We also compute x_1 $(= x_2 = \cdots = x_N)$, and easily obtain

$$x_1 = s \left(\frac{(\frac{\sin \pi l}{\pi})^2}{\Sigma} - \alpha \right) + t \left(x_1^0 + \alpha - \frac{(\frac{\sin \pi l}{\pi})^2}{\Sigma} - \frac{\frac{\sin \pi l}{\pi}}{\Sigma} \left(N x_1^0 - \sum_{i=1}^N x_i^0 \right) \right). \tag{3.22}$$

We recall that we want $\lambda_1 \geq 0$. So, we have the *first necessary condition*,

$$Q_1 := s - t \left(1 + N y_1^0 - \sum_{i=1}^N y_i^0 \right) \geq 0. \tag{3.23}$$

We will now express that every λ_k has also to be nonnegative. For that purpose, we introduce the (nonpositive) number

$$-\xi := \min_{1 \leq k \leq N} (y_1^0 - y_k^0), \tag{3.24}$$

and we assume the *second necessary condition*,

$$\xi t \leq \frac{\frac{\sin \pi l}{\pi} - \frac{\sin 3\pi l}{3\pi}}{\Sigma} Q_1. \tag{3.25}$$

Equation (3.24) together with (3.25) imply

$$\forall k,\ t(y_1^0 - y_k^0) \geq -\xi t \geq \frac{\frac{\sin 3\pi l}{3\pi} - \frac{\sin \pi l}{\pi}}{\Sigma} Q_1.$$

Therefore, from (3.20), it comes that

$$\lambda_k \geq \frac{\frac{\sin 3\pi l}{3\pi} + (-1)^k \frac{\sin k\pi l}{k\pi}}{\Sigma} Q_1. \tag{3.26}$$

For $k = 2, 3, 4$, it is clear that (3.26) with (3.23) imply $\lambda_k \geq 0$. For higher values of k, it is also true. Indeed, since $x \mapsto \sin x/x$ is decreasing for $x < \pi/2$, we have $\sin(3\pi l)/(3\pi) \geq \sin(k\pi l)/(k\pi)$ for $3 \leq k \leq N$.

In conclusion, if we assume (3.23), (3.25), and $l < 1/(2N)$, then $\lambda_k \geq 0$, for all $k \leq N$. Let us set,

$$\gamma = \frac{t}{s} \quad Y = 1 + N y_1^0 - \sum_{i=1}^N y_i^0.$$

Then (3.23) can be rewritten as $1-\gamma Y \geq 0$ and is obviously true for γ small enough. In the same way, (3.25) is

$$\xi\gamma \leq \frac{\frac{\sin \pi l}{\pi} - \frac{\sin 3\pi l}{3\pi}}{\Sigma}(1-\gamma Y),$$

which is also true for γ small since $\dfrac{\sin \pi l}{\pi} - \dfrac{\sin 3\pi l}{3\pi} > 0$.

It remains to be checked whether x_1, given by (3.22), is positive. With the previous notations,

$$x_1 = \gamma\left(x_1^0 - Y\frac{(\frac{\sin \pi l}{\pi})^2}{\Sigma} + \alpha\right) + \frac{(\frac{\sin \pi l}{\pi})^2}{\Sigma} - \alpha.$$

therefore if we are able to prove that $\dfrac{(\frac{\sin \pi l}{\pi})^2}{\Sigma} - \alpha > 0$, it will imply that $x_1 > 0$ for γ small enough.
Now, when $l \to 0$

$$\frac{(\frac{\sin \pi l}{\pi})^2}{\Sigma} \sim \begin{cases} \frac{l}{N} \text{ if } N \text{ is odd} \\ \frac{l}{N+1} \text{ if } N \text{ is even.} \end{cases} \tag{3.27}$$

On the other hand, the proof of Theorem 4.1 shows that, when $l \to 0$

$$x^* \sim j_1\left(\bigcup_{k=1}^{N}\left[\frac{k}{N+1} - \frac{l}{2N}\,,\, \frac{k}{N+1} + \frac{l}{2N}\right]\right) = l - \frac{1}{\pi}\sin\frac{\pi l}{2N}\sum_{k=1}^{N}\cos\frac{2k\pi}{N+1}.$$

Since $\sum_{k=1}^{N}\cos\frac{2k\pi}{N+1} = -1$, this yields

$$\alpha = x^* - l \sim \frac{1}{\pi}\sin\frac{\pi l}{2N} \sim \frac{l}{2N}$$

and the result follows from the comparison with (3.27). This finishes the proof of Theorem 3.2. □

4. The spillover phenomenon. We are going to describe more precisely the optimal domain ω_N^* when the length constraint l goes to zero. According to Theorem 3.1, ω_N^* is symmetric and has, at most, N connected components. Therefore,
in the case N even we write $N = 2K$ and there exists $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_K < 1/2$ and $l_1, l_2, \ldots, l_K \geq 0$ such that

$$\omega_N^* = \left(\bigcup_{i=1}^{K}[\alpha_i - l_i/2,\, \alpha_i + l_i/2]\right)\bigcup\left(\bigcup_{i=1}^{K}[1-\alpha_i - l_i/2,\, 1-\alpha_i + l_i/2]\right)$$

in the case N odd we write $N = 2K+1$ and there exists $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_K < 1/2$ and $l_1, l_2, \ldots, l_{K+1} \geq 0$ such that

$$\omega_N^* = \left(\bigcup_{i=1}^{K}[\alpha_i - l_i/2,\, \alpha_i + l_i/2]\right)\bigcup[1/2 - l_{K+1}/2,\, 1/2 + l_{K+1}/2]\ \ldots$$

$$\ldots\ \bigcup\left(\bigcup_{i=1}^{K}[1-\alpha_i - l_i/2,\, 1-\alpha_i + l_i/2]\right).$$

The main result of this section is the following theorem.

THEOREM 4.1. *When l goes to 0, the optimal domain for the N first modes ω_N^* concentrates around the nodes of the $N+1$th eigenfunction.*
More precisely,when $l \to 0^+$,

$$\forall i, \quad \alpha_i(l) \to \frac{i}{N+1}, \quad l_i(l) \sim \frac{l}{N}.$$

For the sake of simplicity, we are going to prove this result in the case $N = 2K$ even, the case N odd is exactly similar. We now express that ω_N^* satisfies

$$j_1(\omega_N^*) = j_2(\omega_N^*) = \cdots = j_N(\omega_N^*). \tag{4.1}$$

This yields

$$\begin{cases} \displaystyle\sum_{i=1}^{K} \left[\frac{1}{n\pi} \sin n\pi l_i \cos 2n\pi\alpha_i - \frac{1}{(n-1)\pi} \sin(n-1)\pi l_i \cos(2n-2)\pi\alpha_i \right] = 0 \\ \qquad \text{for } n = 2, 3, \ldots, N. \end{cases} \tag{4.2}$$

Asymptotically, when l goes to 0, the (α_i)s and the (l_i)s are therefore solutions of the linearized system (in a neighborhood of $l = 0$),

$$\sum_{i=1}^{K} l_i(\cos 2n\pi\alpha_i - \cos(2n-2)\pi\alpha_i) = 0, \qquad \text{for } n = 2, 3, \ldots, N. \tag{4.3}$$

Now if (l_p) is a sequence which tends to 0, after extracting a finite number of subsequence, we can write $l_i^p = t_i(l_p).l_p$ with $t_i(l_p) \in [0, 1]$ converging to t_i and $a_i(l_p)$ which converge to a_i. The system (4.3) leads to the following system:

$$\begin{cases} t_1 \sin \pi\alpha_1 \sin 3\pi\alpha_1 + t_2 \sin \pi\alpha_2 \sin 3\pi\alpha_2 + \cdots + t_K \sin \pi\alpha_K \sin 3\pi\alpha_K = 0 \\ t_1 \sin \pi\alpha_1 \sin 5\pi\alpha_1 + t_2 \sin \pi\alpha_2 \sin 5\pi\alpha_2 + \cdots + t_K \sin \pi\alpha_K \sin 5\pi\alpha_K = 0 \\ \vdots \\ t_1 \sin \pi\alpha_1 \sin(2k+1)\pi\alpha_1 + t_2 \sin \pi\alpha_2 \sin(2k+1)\pi\alpha_2 + \cdots \\ \qquad\qquad + t_K \sin \pi\alpha_K \sin(2k+1)\pi\alpha_K = 0 \\ \vdots \\ t_1 \sin \pi\alpha_1 \sin(4K-1)\pi\alpha_1 + t_2 \sin \pi\alpha_2 \sin(4K-1)\pi\alpha_2 + \cdots \\ \qquad\qquad + t_K \sin \pi\alpha_K \sin(4K-1)\pi\alpha_K = 0, \end{cases} \tag{4.4}$$

where we have to add the supplementary equation

$$t_1 + t_2 + \ldots t_K = 1/2. \tag{4.5}$$

This new system can be viewed as a $(2K-1) \times (2K-1)$ linear system with unknowns t_1, t_2, …, t_K. Of course these unknowns cannot all be equal to zero because of the supplementary equation (4.5). Therefore the matrix A of system (4.4) must be of rank less or equal to $K-1$. It means that

$$\text{all determinants } K \times K \text{ extracted from } A \text{ are equal to 0.} \tag{4.6}$$

Let us denote by $\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_{2K-1}$, the lines of the matrix A,

(4.7)
$$A = \begin{pmatrix} \sin\pi\alpha_1 \sin 3\pi\alpha_1 & \ldots & \sin\pi\alpha_K \sin 3\pi\alpha_K \\ \sin\pi\alpha_1 \sin 5\pi\alpha_1 & \ldots & \sin\pi\alpha_K \sin 5\pi\alpha_K \\ \vdots & & \vdots \\ \sin\pi\alpha_1 \sin(4K-1)\pi\alpha_1 & \ldots & \sin\pi\alpha_K \sin(4K-1)\pi\alpha_K \end{pmatrix} \begin{matrix} \to \mathcal{L}_1 \\ \to \mathcal{L}_2 \\ \vdots \\ \to \mathcal{L}_{2K-1}. \end{matrix}$$

Now, let us compute the following $K \times K$ determinant:

$$D_0 = \det(\mathcal{L}_K + \mathcal{L}_K, \mathcal{L}_{K-1} + \mathcal{L}_{K+1}, \ldots, \mathcal{L}_1 + \mathcal{L}_{2K-1}).$$

According to (4.6), $D_0 = 0$.
On the other hand, for $k = 0, 1, \ldots, K-1$ and $i = 1, 2, \ldots, K$,

$$\sin(2(K-k)+1)\pi\alpha_i + \sin(2(K+k)+1)\pi\alpha_i = 2\sin(2K+1)\pi\alpha_i \cos 2k\pi\alpha_i.$$

Therefore

$$D_0 = 2^K \left(\prod_{i=1}^{K} \sin\pi\alpha_i\right) \left(\prod_{i=1}^{K} \sin(2K+1)\pi\alpha_i\right) D_0',$$

where D_0' is the determinant,

$$D_0' = \begin{vmatrix} 1 & 1 & \ldots & 1 \\ \cos 2\pi\alpha_1 & \cos 2\pi\alpha_2 & \ldots & \cos 2\pi\alpha_K \\ \vdots & \vdots & & \vdots \\ \cos 2(K-1)\pi\alpha_1 & \cos 2(K-1)\pi\alpha_2 & \ldots & \cos 2(K-1)\pi\alpha_K \end{vmatrix}.$$

This determinant D_0' can be computed thanks to Tchebyshev polynomials T_k, already introduced ($\cos kx = T_k(\cos x)$) with T_k of degree k and highest degree term is $2^{k-1}X^k$, $k \geq 1$. Since the family $(T_k)_{0 \leq k \leq K-1}$ is a basis of polynomial of degree less than $K-1$, we can write

$$D_0' = \left(\prod_{k=1}^{K-1} 2^{k-1}\right) \begin{vmatrix} 1 & 1 & \ldots & 1 \\ \cos 2\pi\alpha_1 & \cos 2\pi\alpha_2 & \ldots & \cos 2\pi\alpha_K \\ \vdots & \vdots & & \vdots \\ \cos^{K-1} 2\pi\alpha_1 & \cos^{K-1} 2\pi\alpha_2 & \ldots & \cos^{K-1} 2\pi\alpha_K \end{vmatrix}.$$

Now this last determinant is the so-called van der Mond determinant, so

$$\begin{aligned} D_0' &= 2^{\frac{(K-1)(K-2)}{2}} \prod_{1 \leq i < j \leq K} (\cos 2\pi\alpha_i - \cos 2\pi\alpha_j) \\ &= 2^{\frac{(K-1)(K-2)}{2}} \prod_{1 \leq i < j \leq K} 2(\cos^2 \pi\alpha_i - \cos^2 \pi\alpha_j) \\ &= 2^{(K-1)^2} \prod_{1 \leq i < j \leq K} (\cos^2 \pi\alpha_i - \cos^2 \pi\alpha_j). \end{aligned} \tag{4.8}$$

This last equality shows that D_0' cannot vanish (we recall that the (α_i) is an increasing sequence with $0 < \pi\alpha_i < \pi/2$). Moreover, the product $\prod_{i=1}^{K} \sin\pi\alpha_i$ is not zero

since, for all i, $\alpha_i \in]0,1/2[$. In conclusion, there exists $p \in \{1,2,\dots,K\}$ such that $\sin(2K+1)\pi\alpha_p = 0$, it means that there exists $q \in \{1,2,\dots,K\}$ such that $\alpha_p = \frac{q}{2K+1}$.

Let us now compute, in the same way, the following $K \times K$ determinant:

$$D_1 = \det(\mathcal{L}_K + \mathcal{L}_K, \mathcal{L}_{K-1} + \mathcal{L}_{K+1}, \dots, \mathcal{L}_2 + \mathcal{L}_{2K-2}, \mathcal{L}_1).$$

For the same reason as above, $D_1 = 0$. Grouping the sin terms yields

$$D_1 = 2^K \left(\prod_{i=1}^{K} \sin \pi\alpha_i \right) \left(\prod_{i=1,\,2\,,\dots,K\,;\,i\neq p} \sin(2K+1)\pi\alpha_i \right) D_1',$$

where D_1' is the determinant

$$D_1' = \begin{vmatrix} 1 & \dots & 0 & \dots & 1 \\ \cos 2\pi\alpha_1 & \dots & 0 & \dots & \cos 2\pi\alpha_K \\ \vdots & & \vdots & & \vdots \\ \cos 2(K-1)\pi\alpha_1 & \dots & \sin \frac{3\pi q}{2K+1} & \dots & \cos 2(K-1)\pi\alpha_K \end{vmatrix}.$$

Developing D_1' with respect to its pth column, it appears a $K-1 \times K-1$ determinant similar to D_0'. This shows that D_1' does not vanish. Therefore, there exists $r \neq p$ such that $\sin(2K+1)\pi\alpha_r = 0$, which means that there exists $s \in \{1,2,\dots,K\}$ such that $\alpha_r = \frac{s}{2K+1}$.

By computing successively, using the same method as for D_1, the determinants D_2, D_3, …, D_{K-1}, with

$$\begin{aligned} D_2 &= \det(\mathcal{L}_K + \mathcal{L}_K,\, \mathcal{L}_{K-1} + \mathcal{L}_{K+1}, \dots,\, \mathcal{L}_3 + \mathcal{L}_{2K-3},\, \mathcal{L}_2,\, \mathcal{L}_1 + \mathcal{L}_{2K-1}) \\ D_3 &= \det(\mathcal{L}_K + \mathcal{L}_K,\, \mathcal{L}_{K-1} + \mathcal{L}_{K+1}, \dots,\, \mathcal{L}_3,\, \mathcal{L}_2 + \mathcal{L}_{2K-2},\, \mathcal{L}_1 + \mathcal{L}_{2K-1}) \\ &\vdots \\ D_{K-1} &= \det(\mathcal{L}_K + \mathcal{L}_K,\, \mathcal{L}_{K-1},\, \mathcal{L}_{K-2} + \mathcal{L}_{K+2} \dots, \mathcal{L}_2 + \mathcal{L}_{2K-2},\, \mathcal{L}_1 + \mathcal{L}_{2K-1}), \end{aligned}$$

we can show that, for all $1 \le i \le K$, there exists p_i such that $\alpha_i = \frac{p_i\pi}{2K+1}$. Since the sequence (α_i) is increasing, we have

$$\forall i \in \{1,\, 2,\, \dots,\, K\} \quad \alpha_i = \frac{i\pi}{2K+1}.$$

Let us now show that the rank of the matrix A is exactly $K-1$. For that purpose, we compute the determinant D obtained from A by taking the $K-1$ first lines and removing the last column,

$$D = \left[\prod_{i=1}^{K-1} \sin \frac{i\pi}{2K+1} \right] \det \left(\left(\sin \frac{(2i+1)\pi j}{2K+1} \right)_{1 \le i,\, j \le K-1} \right).$$

Let us denote by U_n Tchebyshev's polynomial of second kind: $\sin n\theta = \sin\theta\, U_n(\cos\theta)$. We then obtain

$$D = \left[\prod_{i=1}^{K-1} \sin \frac{i\pi}{2K+1} \right] \det \left(\left(U_j \left(\sin \frac{(2i+1)\pi}{2K+1} \right) \right)_{1 \le i,\, j \le K-1} \right).$$

Since U_n has for first term $2^n X^n$, $n \geq 1$, we get

$$D = 2^{\frac{(K-2)(K-3)}{2}} \left(\prod_{i=1}^{K-1} \sin \frac{i\pi}{2K+1} \right) \left(\prod_{1 \leq i < j \leq K-1} \left(\cos \frac{(2i+1)\pi}{2K+1} - \cos \frac{(2j+1)\pi}{2K+1} \right) \right).$$

This proves that $D \neq 0$ and A has rank $K-1$. Therefore, its kernel is of dimension one.

Let us now show that $(1, 1, \ldots, 1)^t$ belongs to the kernel of A. For that purpose, we compute, for $1 \leq k \leq 2K-1$ the sum

$$S_k = 2 \sum_{l=1}^{K} \sin \frac{l\pi}{2K+1} \sin \frac{(2k+1)l\pi}{2K+1} = \sum_{l=1}^{K} \left[\cos \frac{2lk\pi}{2K+1} - \cos \frac{2l(k+1)\pi}{2K+1} \right].$$

Using $\sum_{l=1}^{K} \cos l\theta = \cos \frac{(K+1)\theta}{2} \cdot \frac{\sin \frac{K\theta}{2}}{\sin \frac{\theta}{2}}$, it comes that $S_k = 0$ if and only if

$$\cos \frac{(K+1)k\pi}{2K+1} \sin \frac{Kk\pi}{2K+1} \sin \frac{(k+1)\pi}{2K+1} = \cdots$$

$$\cdots = \cos \frac{(K+1)(k+1)\pi}{2K+1} \sin \frac{K(k+1)\pi}{2K+1} \sin \frac{k\pi}{2K+1}.$$

This last equality is easy to check.

In conclusion, since $(t_1, t_2, \ldots, t_K)^t$ belongs to the kernel of A and $t_1 + t_2 + \cdots + t_K = 1/2$, we get $t_1 = t_2 = \cdots = t_K = 1/(2K)$. Finally, for every subsequence (l_k) converging to 0, $(t_i(l_k))$ has a unique accumulation point; the whole sequence $t_i(l)$ converges to that point $1/K$. In the same way, the functions $\alpha_i(l)$ converge to $i/(K+1)$ when l goes to 0. □

5. Comments.

5.1. Possible remedies. To avoid the spillover phenomenon which is described here, we can imagine different possible strategies. The first one is obviously to take into account all eigenmodes, possibly with different weights for each (e.g., weights decreasing with the rank of the mode).

Another possibility could be inspired by the introduction of an artificial (numerical) viscosity like in papers [23], [24]. For the one-dimensional wave equation, these authors choose to introduce a semidiscrete term coming from $-h^2 u_{xxt}$. This has the great advantage to keep the decay properties of the discrete equation which are generally lost under the semidiscrete finite-differences scheme. It would be very interesting to see what is the impact of this viscosity term in the context of our paper.

5.2. Larger values of l. It is essentially for technical reasons that the "spillover" phenomenon is described in the case $l \to 0$. Actually, this phenomenon holds for most values of l. Figure 5.1 shows numerical results for $N = 3$. The left picture shows the optimal domain ω_3^* for each value of l. One obtains ω_3^* as the intersection of the vertical line $x = l$ with the interior of the three peeks. The right picture shows in boldface the graph of $l \mapsto J_3(\omega_3^*)$ and below, in medium, the graph of $l \mapsto j_4(\omega_3^*)$. We see that, in any case (unless for l close to 1), ω_3^* has a very poor behavior for the fourth eigenmode. One can find more examples in [14].

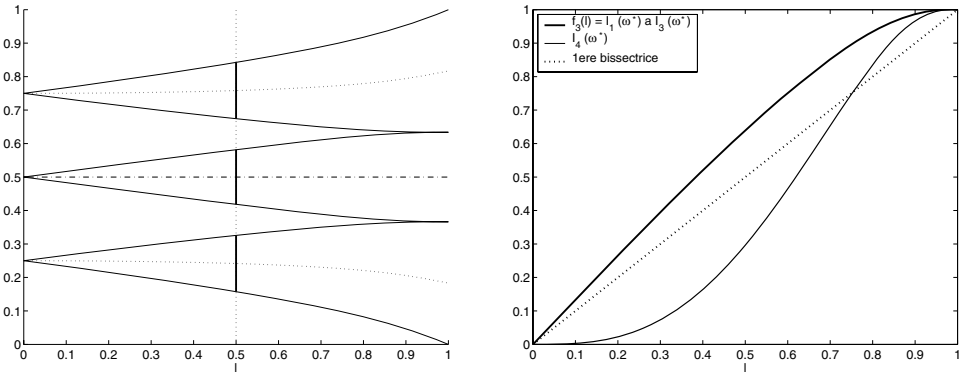


FIG. 5.1. Left: The optimal domain ω_3^* (read vertically). Right: Graphs of $J_3(\omega_3^*)$ (boldface) and $j_4(\omega_3^*)$

5.3. More about the set $\mathcal{K}$. For similar optimization problems, it is interesting to have a more precise description of the set $\mathcal{K}$ which is introduced in (3.11). In particular, a characterization of its boundary can be very useful. Let X be a point in $\mathcal{K}$. Then X belongs to $\partial\mathcal{K}$ if and only if there exists a unit vector $\mathbf{n}$ such that

$$\forall\, Y \in \mathcal{K} \quad (Y - X, \mathbf{n}) \leq 0.$$

Writing $\mathbf{n} = (n_1, n_2, \ldots, n_N)^T$, $X = (j_1(a), \ldots, j_N(a))^T$ and $Y = (j_1(b), \ldots, j_N(b))^T$, the latter reads

$$\forall b \in \mathcal{A}_l \quad \sum_{k=1}^{N} n_k j_k(b) \leq \sum_{k=1}^{N} n_k j_k(a).$$

In other terms, a is a maximizer of the functional $b \mapsto \sum_{k=1}^{N} n_k j_k(b)$. According to the proof of Theorem 3.1, it follows that a is necessarily the characteristic function of a set ω which is the union of at most $N+1$ intervals and is symmetric with respect to $1/2$. The fact that we can have $N+1$ intervals here is due to the fact that intervals of the kind $[0, \eta]$ or $[1-\eta, 1]$ are allowed here (see step 5 in the proof of Theorem 3.1). Figure 5.2 (left) shows the set $\mathcal{K}_2$ for $l = 0.3$. Its boundary is exactly the image of

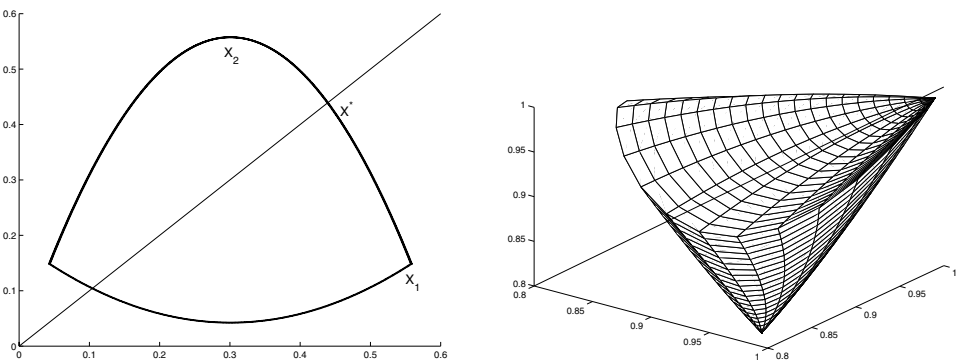


FIG. 5.2. Left: The set $\mathcal{K}_2$ for $l = 0.3$. Right: The set $\mathcal{K}_3$ for $l = 0.9$

characteristic functions χ_ω with ω as a symmetric (w.r.t. 1/2) set obtained as a union of one, two, or three intervals (in this last case two of the three intervals must touch 0 and 1). Figure 5.2 (right) shows the set $\mathcal{K}_3$ for $l = 0.9$. Its boundary is obtained with a symmetric union of one, two, three, or four intervals (in this last case two of the four intervals must touch 0 and 1). This picture shows a case where the first bissectrix does not cut the set $\mathcal{K}_3$ at the point which maximizes $\min(x_1, x_2, x_3)$; see Remark 2.

5.4. Generalization to the two-dimensional case. The existence and uniqueness part of Theorem 3.1 can be easily generalized to more general domains Ω in higher dimension if the following property holds:

$$\text{(5.1)} \quad \begin{cases} \text{Let } \varphi_1, \varphi_2, \dots, \varphi_N \text{ be the (normalized) eigenfunctions of the Laplace} \\ \text{operator on } \Omega \text{ with Dirichlet boundary conditions on } \partial\Omega, \text{ then} \\ \varphi_1^2, \varphi_2^2, \dots, \varphi_N^2 \text{ are linearly independent on } \omega. \end{cases}$$

Indeed, when we look at the proof of Theorem 3.1, we observe that it can easily be adapted to any dimension, the only technical point being (3.7).

The property (5.1) is obviously true for rectangles in two dimensions or, more generally, parallelepiped in N-dimension, but the authors do not know if it holds for every domain (even for a disc). For a related result in one dimension, see [19]. In this paper the authors prove that for a nonhomogeneous Sturm–Liouville eigenvalue problem, it happens very frequently that the N first eigenfunctions, with $N \geq 3$, have linearly dependent squares on some nontrivial interval. A transposition of this one-dimensional result to our case could lead to the following conjecture:

Open problem 1: Prove that for every domain Ω, there exists a domain $\tilde{\Omega}$ close to Ω such that the square of a given number of eigenfunctions of the Laplace–Dirichlet operator on $\tilde{\Omega}$ are linearly dependent. On the other hand, if the result (5.1) is wrong for some domain Ω, one can also imagine some genericity result in the spirit of [20], [21], [22] which could, for example, be stated like:

Open problem 2: Let Ω be an open set such that $\varphi_1^2, \varphi_2^2, \dots, \varphi_N^2$ are linearly dependent. Then, prove that there exists arbitrary small deformations of its boundary such that the square of the eigenfunctions of the perturbed domain become linearly independent.

Following step 5 of the proof of Theorem 3.1, if (5.1) is true (and therefore a unique optimal domain exists), this optimal domain can also be described as a level set of some linear combination of $\varphi_1^2, \varphi_2^2, \dots, \varphi_N^2$. Now, the other results of this paper, Theorems 3.2 and 4.1, seem more difficult to prove in the two-dimensional case, even if the authors believe that they are true.

REFERENCES

[1] A. Auslender, *Optimisation: Méthodes numériques*, Masson, Paris, New York, Barcelona, Milan, 1976.

[2] M. J. Balas, *Active control of flexible systems*, J. Optim. Theory Appl., 25 (1978), pp. 415–436.

[3] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[4] E. Bednarczuk, M. Pierre, E. Rouy, and J. Sokolowski, *Tangent sets in some functional spaces*, Nonlinear Anal., 42 (2000), pp. 871–876.

[5] C. Castro and S. Cox, *Achieving arbitrarily large decay in the damped wave equation*, SIAM J. Control Optim., 39 (2001), pp. 1748–1755.

[6] G. Chen, S. A. Fulling, F. J. Narcowich, and S. Sun, *Exponential decay of energy of evolution equations with locally distributed damping*, SIAM J. Appl. Math., 51 (1991), pp. 266–301.

[7] R. Cominetti and J.-P. Penot, *Tangent sets to unilateral convex sets*, C. R. Acad. Sci. Paris Sér. I Math, 321 (1995), pp. 1631–1636.
[8] S. Cox and M. Overton, *Perturbing the critically damped wave equation*, SIAM J. Appl. Math., 56 (1996), pp. 1353–1362.
[9] S. Cox and E. Zuazua, *The rate at which energy decays in a damped string*, Comm. Partial Differential Equations, 19 (1994), pp. 213–243.
[10] G. Dal Maso, *An Introduction to* Γ*-Convergence*, Birkhäuser, Boston, 1993.
[11] P. Freitas, *Optimizing the rate of decay of solutions of the wave equation using genetic algorithms: A counterexample to the constant damping conjecture*, SIAM J. Control Optim., 37 (1999), pp. 376–387.
[12] P. Freitas, *On some eigenvalue problems related to the wave equation with indefinite damping*, J. Differential Equations, 127 (1996), pp. 320–335.
[13] P. Freitas and E. Zuazua, *Stability results for the wave equation with indefinite damping*, J. Differential Equations, 132 (1996), pp. 338–352.
[14] P. Hébrard, Title Ph.D. thesis, Université Henri Poincaré Nancy, France, http://www.iecn.u-nancy.fr/hebrard/these_hebrard.ps.
[15] P. Hébrard and A. Henrot, *Optimal shape and position of the actuators for the stabilization of a string*, Systems Control Lett., 48 (2003), pp. 199–209.
[16] A. Henrot and M. Pierre, *Variation et Optimisation de formes*, Mathématiques et Applications, Vol. 48, Springer, Berlin, Heidelberg, New York, 2005.
[17] J. B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*, Springer, Berlin, 1993.
[18] G. Lebeau, *Équation des ondes amorties*, Algebraic and Geometric Methods in Mathematical Physics, Math. Phys. Stud., 19, Kluwer Acad. Publ., Dordrecht, 1996, pp. 73–109.
[19] T. J. Mahar and B. E. Willner, *Sturm-Liouville eigenvalue problems in which the squares of the eigenfunctions are linearly dependent*, Comm. Pure Appl. Math., 33 (1980), pp. 567–578.
[20] A. M. Micheletti, *Perturbazione dello spettro dell'operatore di Laplace, in relazione ad una variazione del campo*, Ann. Scuola Norm. Sup. Pisa, 3 26, (1972), pp. 151–169.
[21] J. Ortega and E. Zuazua, *Generic simplicity of the spectrum and stabilization for a plate model*, SIAM J. Control Optim., 39 (2000), pp. 1585–1614 (Addendum: 42 (2004), pp. 1905–1910).
[22] J. Ortega and E. Zuazua, *On a constrained approximate controllability problem for the heat equation*, J. Optim. Theory. Appl., 108 (2001), pp. 29–64 (Addendum: 118 (2003), pp. 183–190).
[23] L. R. Tcheugoué and E. Zuazua, *Uniform exponential long time decay for the space semi-discretization of a damped wave equation via an artificial numerical viscosity*, Numer. Math., 95 (2003), pp. 563–598.
[24] L. R. Tcheugoue and E. Zuazua, *Uniform boundary stabilization of the finite difference space discretization of the* $1-d$ *wave equation*, Advances in Computational Mathematics, to appear.
[25] L. N. Trefethen, *Group velocity in finite difference schemes*, SIAM Rev., 24 (1982), pp. 113–136.
[26] E. Zuazua, *Propagation, observation, and control of waves approximated by finite difference methods*, SIAM Rev., 47 (2005), pp. 197–243.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 1, pp. 367–387

DEGENERATE STOCHASTIC CONTROL PROBLEMS WITH EXPONENTIAL COSTS AND WEAKLY COUPLED DYNAMICS: VISCOSITY SOLUTIONS AND A MAXIMUM PRINCIPLE*

MINYI HUANG†, PETER E. CAINES‡, AND ROLAND P. MALHAMÉ§

Abstract. This paper considers a class of optimization problems arising in wireless communication systems. We analyze the optimal control and the associated Hamilton–Jacobi–Bellman (HJB) equations. It turns out that the value function is a unique viscosity solution of the HJB equation in a certain function class. To deal with the fast growth condition of the value function in establishing uniqueness, we construct particular semiconvex/semiconcave approximations for the viscosity sub/supersolutions, and obtain a maximum principle on unbounded domains. The localized envelope function technique introduced in this paper permits an analysis of the uniqueness of viscosity solutions defined on unbounded domains in cases with very general growth conditions when combined with appropriate system dynamics. The optimization problem with state constraints is also considered.

Key words. degenerate stochastic control, power control, HJB equations, dynamic programming, viscosity solutions

AMS subject classifications. 93E20, 93E03, 49L25, 49L20

DOI. 10.1137/S0363012902417644

1. Introduction.

This paper is concerned with a class of optimization problems arising in power control for wireless communication systems, and forms a mathematical foundation for the results in [7, 8]. We will first formulate a class of degenerate stochastic control problems, which take the approach of regulating the state of a controlled process where an exogenous random parameter process is involved in the performance function, and then we use a communications application example to give a background illustration for the general formulation.

The random parameter process and the controlled process are denoted by $x_t \in \mathbb{R}^n$ and $p_t \in \mathbb{R}^n$, $t \in \mathbb{R}_+$, respectively. Suppose that x is modeled by the stochastic differential equation

$$dx = f(t,x)dt + \sigma(t,x)dw, \quad t \geq 0, \tag{1.1}$$

where f and σ are the drift and diffusion coefficients, respectively; w is an n dimensional standard Wiener process with covariance $Ew_t w_t^\tau = tI$; and the initial state x_0 is independent of $\{w_t,\ t \geq 0\}$ with finite exponential moment, i.e., $Ee^{2|x_0|} < \infty$.

*Received by the editors November 8, 2002; accepted for publication (in revised form) October 5, 2004; published electronically August 22, 2005. This work was supported by NCE-MITACS program 1999–2002 and NSERC grants 1329-00 and 1361-00. This is the expanded version of a paper presented at the 40th IEEE Conference on Decision and Control, Orlando, FL, in 2001 (pp. 1031–1036 in the Conference Proceedings).

http://www.siam.org/journals/sicon/44-1/41764.html

†Department of Electrical and Computer Engineering, McGill University, 3480 University Street, Montreal, QC H3A 2A7, Canada. Current address: Department of Electrical and Electronic Engineering, University of Melbourne, Victoria 3010, Australia (m.huang@ee.mu.oz.au).

‡Corresponding author. Department of Electrical and Computer Engineering, McGill University, 3480 University Street, Montreal, QC H3A 2A7, Canada (peterc@cim.mcgill.ca). Also affiliated with GERAD.

§Department of Electrical Engineering, École Polytechnique de Montréal, 2900 Boul. Édouard Montpetit, Montreal, QC H3C 3A7, Canada (roland.malhame@polymtl.ca). Also affiliated with GERAD.

The process p is governed by the model

$$dp = g(t, x, p, u)dt, \quad t \geq 0, \tag{1.2}$$

where the component $g_i(t, x, p, u)$, $1 \leq i \leq n$, controls the size of the increment dp_i at the time instant t, $u \in \mathbb{R}^n$, $|u_i| \leq u_{i\max}$, $1 \leq i \leq n$. Without loss of generality we set $u_{i\max} = 1$, and we shall write

$$x = [x_1, \ldots, x_n]^\tau, \qquad p = [p_1, \ldots, p_n]^\tau, \qquad u = [u_1, \ldots, u_n]^\tau.$$

In the regulation of p, we introduce the following cost function:

$$J = E \int_0^T [p^\tau C(x)p + 2D^\tau(x)p]dt, \tag{1.3}$$

where $T < \infty$; $C(x)$ and $D(x)$ are an $n \times n$ positive definite matrix and an $n \times 1$ vector, respectively; and the components of $C(x)$ and $D(x)$ are exponential functions of linear combinations of x_i, $1 \leq i \leq n$. For simplicity, in this paper we take $C_{ij}(x) = c_{ij}e^{x_i + x_j}$, $D_i(x) = d_i e^{x_i} + s_i$ for $1 \leq i, j \leq n$. This particular structure of the weight coefficients indicates that in the cost function each p_i is directly associated with the parameter component x_i for $1 \leq i \leq n$. Specifically, an expansion of the cost integrand will produce entries in the form of $c_{ij}(e^{x_i}p_i)(e^{x_j}p_j)$, $d_i e^{x_i}p_i$, $s_i p_i$, $1 \leq i, j \leq n$. Intuitively, such a cost structure indicates that the relative weight of each p_i is influenced only by the process x_i. The more general case of expressing the components of $C(x)$ and $D(x)$ as exponential functions of general linear combinations of x_i, $1 \leq i \leq n$, can be considered without further difficulty. We will give the complete optimal control formulation in section 2, where the technical assumptions of weak coupling for the dynamics (1.1)–(1.2) will be introduced.

1.1. The stochastic power control example. We now briefly describe the motivating stochastic power control problem for lognormal fading channels. In an urban environment, due to long distance transmission and reflections, the power attenuations of wireless networks are described by lognormal random processes. Let $x_i(t)$, $1 \leq i \leq n$, denote the power attenuation (expressed in dBs and scaled to the natural logarithm basis) at the instant t of the ith mobile user, and let $\alpha_i(t) = e^{x_i(t)}$ denote the actual attenuation. Based upon the work in [1], the power attenuation dynamics are given as a special form of (1.1):

$$dx_i = -a_i(x_i + b_i)dt + \sigma_i dw_i, \quad t \geq 0,\ 1 \leq i \leq n. \tag{1.4}$$

In (1.4) the constants $a_i, b_i, \sigma_i > 0$, $1 \leq i \leq n$. See [1] for a physical interpretation of the parameters, and furthermore, an experimental justification of the lognormal attenuation modeling may be found in the communications literature [5] using discrete time measurements. In a network, at time t the ith mobile user sends its power $p_i(t)$, and the received power at the base station is $e^{x_i(t)}p_i(t)$. The mobile user has to adjust its power p_i in real time so that a certain communication quality of service (QoS) is maintained. In [6, 7] the adjustment of the (sent) power vector p for the n users is modeled by simply taking $g(t, x, p, u) = u$ in (1.2), which is called the rate adjustment model. Based upon the system signal-to-interference ratio (SIR) requirements, the following averaged integrated performance function,

$$J = E \int_0^T \left\{ \sum_{i=1}^n \left[e^{x_i}p_i - \mu_i \left(\sum_{j=1}^n e^{x_j}p_j + \eta \right) \right]^2 + \lambda \sum_{i=1}^n p_i \right\} dt, \tag{1.5}$$

was employed in [7, 8], where $\eta > 0$ is the system background noise intensity, $\lambda \geq 0$, and μ_i, $1 \leq i \leq n$, is a set of positive numbers determined from the SIR requirements. The resulting power control problem is to adjust u as a function of the system state (x, p) so that the above performance function is minimized.

1.2. The main contents and organization. The analysis in this paper treats a general class of performance functions that have an exponential growth rate with respect to x_i, $1 \leq i \leq n$; hence this analysis covers the cost function in (1.5) and differs from that appearing in most stochastic control problems in the literature, where linear or polynomial growth conditions usually pertain [3, 12]. Two novel features of the class of models (1.1)–(1.2) are (i) neither the drift nor the diffusion of the state subprocess x is subject to control, and hence x may be regarded as an exogenous signal, and (ii) further, the controlled state subprocess p has no diffusion part. Hence (1.1)–(1.2) gives rise to degenerate stochastic control systems. Optimization of such systems leads to degenerate Hamilton–Jacobi–Bellman (HJB) equations, which in general do not admit classical solutions [4, 12].

This paper deals with the mathematical control theoretic questions arising from the class of stochastic optimal control problems considered in [7, 8], where some approximation and numerical methods are proposed for implementation of the control laws. For the resulting degenerate HJB equations, we adopt viscosity solutions and show that the value function of the optimal control is a viscosity solution. To prove uniqueness of the viscosity solution, we develop a localized semiconvex/semiconcave approximation technique. Specifically, we introduce particular localized envelope functions on the unbounded domain to generate semiconvex/semiconcave approximations on any compact set. Compared to previous works [4, 12], by use of the set of envelope functions we can treat very rapid growth conditions, and we note that no Lipschitz or Hölder-type continuity assumption is required for the function class involved. It is worthwhile to note that the localized envelope functions may be applied to generate local semiconvex/semiconcave approximations for viscosity solutions in risk-sensitive stochastic control problems with degenerate diffusions in which the cost involves an exponential function and usually has a very rapid growth.

We also consider the optimal control subject to state constraints, which leads to the formulation of constrained viscosity solutions to the associated second order HJB equations; this part is parallel to [11], where a first order HJB equation is investigated. The paper is organized as follows: in section 2 we state the existence and uniqueness of the optimal control and show that the value function is a viscosity solution to a degenerate HJB equation; we then give two theorems as the main results about the solution of the HJB equation. Section 3 is devoted to introducing a class of semiconvex/semiconcave approximations for continuous functions; this technique enables us to treat viscosity solutions with rapid growth. In section 4, we analyze the HJB equation and prove a maximum principle by which it follows that the HJB equation has a unique viscosity solution in a certain function class. Section 5 considers the control problem subject to state constraints.

Finally, we remark that in the case when an additional control term is introduced to the state subprocess x to give mathematically more general dynamics, one can also derive an HJB equation for the corresponding optimal control problem, which is interesting in its own right, and the semiconvex/semiconcave approximations and uniqueness analysis procedure developed in sections 3 and 4 may still be carried out under appropriate conditions. However, without further conditions for the dynamics of x in the controlled case, in general the control problem needs to be formulated

in a weak solution framework, and the resulting analysis is not in the scope of the present paper.

2. The optimal control and HJB equations. We define

$$z = \begin{pmatrix} x \\ p \end{pmatrix}, \qquad \psi = \begin{pmatrix} f \\ g \end{pmatrix}, \qquad G = \begin{pmatrix} \sigma \\ 0_{n\times n} \end{pmatrix}.$$

We now write (1.1) and (1.2) together in the vector form

$$dz = \psi dt + G dw, \quad t \geq 0. \tag{2.1}$$

In the following analysis we will denote the state variable by (x, p) or z, or in a mixing form; as we do in section 4, we may also write the arguments for the functions in (1.1)–(1.2) in a unified way in terms of (t, z). We write the integrand in (1.3) as $l(z) = l(x, p) = p^\tau C(x)p + 2D^\tau(x)p$. For notational clarity, hereafter we use x_t with a real-valued subscript t to denote the value of the vector process x at time t, and x_i with an integer subscript i to denote the ith component of x; the interpretation of the notation should be clear from the context. This convention also holds for other vector processes involved in the analysis.

The admissible control set is specified as

$$\mathcal{U} = \{u(\cdot) \mid u_t \text{ is adapted to } \sigma(z_s, s \leq t) \text{ and } u_t \in U \stackrel{\triangle}{=} [-1, 1]^n \ \forall 0 \leq t \leq T\}.$$

As is stated in the introduction, the initial state vector is independent of the $n \times 1$ Wiener process w_t, $t \geq 0$; we make the additional assumption that p has a deterministic initial value p_0 at $t = 0$. Then it is easily verified that $\sigma(z_s,\ s \leq t) = \sigma(x_s,\ s \leq t)$. Define $\mathcal{L} = \{u(\cdot) \mid u \text{ is adapted to } \sigma(z_s,\ s \leq t), u_t \in \mathbb{R}^n \text{ and } E\int_0^T |u_s|^2 ds < \infty\}$. If we endow $\mathcal{L}$ with an inner product $\langle u, u'\rangle \stackrel{\triangle}{=} E\int_0^T u^\tau u' ds$ for $u, u' \in \mathcal{L}$, then $\mathcal{L}$ constitutes a Hilbert space with the induced norm $\|u\| = \langle u, u\rangle^{\frac{1}{2}} \geq 0$, $u \in \mathcal{L}$. Under this norm, $\mathcal{U}$ is a bounded, closed, and convex subset of $\mathcal{L}$. Finally, the cost associated with the system (2.1) and a control $u \in \mathcal{U}$ is specified to be

$$J(s, z, u) = E\left[\int_s^T l(z_t)dt | z_s = z\right], \quad z \in \mathbb{R}^{2n}, \tag{2.2}$$

where $s \in [0, T]$ is taken as the initial time of the system; further, we set the value function

$$v(s, z) = \inf_{u\in\mathcal{U}} J(s, z, u), \quad z \in \mathbb{R}^{2n},$$

and simply write $J(0, z, u)$ as $J(z, u)$. The following assumptions on the time interval $[0, T]$ will be used in our further analysis.

(H1) In (1.1)–(1.2), $f \in C([0, T]\times\mathbb{R}^n, \mathbb{R}^n)$, $\sigma \in C([0, T]\times\mathbb{R}^n, \mathbb{R}^{n\times n})$, $g \in C([0, T]\times\mathbb{R}^{3n}, \mathbb{R}^n)$ and f, σ, g satisfy a uniform Lipschitz condition; i.e., there exists a constant $C_0 > 0$ such that $|f(t, x)-f(s, y)| \leq C_0(|t-s|+|x-y|)$, $|\sigma(t, x)-\sigma(s, y)| \leq C_0(|t-s|+|x-y|)$, $|g(t, x, p, u)-g(s, x, q, u)| \leq C_0(|t-s|+|p-q|)$, and $|g(t, x, p, u)-g(t, 0, p, u)| \leq C_0$ for all $t, s \in [0, T]$, $u \in U$, and $x, y, p, q \in \mathbb{R}^n$. In addition, there exists a constant C_σ such that $|\sigma_{ij}(t, x)| \leq C_\sigma$ for $(t, x) \in [0, T] \times \mathbb{R}^n$ and $1 \leq i, j \leq n$.

(H2) For $1 \leq i \leq n$, $f_i(x)$ can be written as $f_i(x) = -a_i(t)x_i + f_i^0(t, x)$, where $a_i(t) \geq 0$ for $t \in [0, T]$, and $\sup_{[0,T]\times\mathbb{R}^n} |f_i^0(t, x)| \leq C_{f^0}$ for a constant $C_{f^0} > 0$.

Throughout this paper we assume that (H1) holds. (H2) is used in Theorems 2.5 and 2.6 for proving uniqueness of the viscosity solution. Clearly (H2) holds for the lognormal fading channel model in the power control example.

Remark 1. Assumption (H1) ensures existence and uniqueness of the solution to (2.1) for any fixed $u \in \mathcal{U}$, where the Lipschitz condition with respect to t will be used to obtain certain estimates in the proof of uniqueness of the viscosity solution. Here σ is assumed to be bounded so as to lead to a finite cost for any initial state and admissible control u.

From (H1)–(H2) it is seen that the system model has the following important features: first, in the diffusion process x the evolution of x_i does not receive strong influence from the other state component x_k, $k \neq i$, in the sense that the cross term $f_i^0(t,x)$ is bounded by a constant; second, an arbitrary increase of x alone in the function $g(t,x,p,u)$ does not result in an unbounded increase in the magnitude of g, and hence x imposes only a relatively weak impact on the evolution of p. Due to the above features, we shall refer to the model (1.1)–(1.2) analyzed in this paper as having *weakly coupled dynamics,* and (H2) will be conveniently referred to as the weak coupling condition for x, which will be used to establish uniqueness of the viscosity solution.

PROPOSITION 2.1 (see [7, 8]). *Assuming in the control model* (1.2) *that* $g(t,x,p,u)$ *is linear in* p *and* u*, i.e., that there exist continuous matrix functions* A_t*,* B_t *on* $[0,T]$ *such that* $g(t,x,p,u) = A_t p + B_t u$*, then there exists an optimal control* $\widehat{u} \in \mathcal{U}$ *such that* $J(x_0,p_0,\widehat{u}) = \inf_{u\in\mathcal{U}} J(x_0,p_0,u)$*, where* (x_0,p_0) *is the initial state at time* $s=0$*; if, in addition,* B_t *is invertible for all* $t \in [0,T]$*, then the optimal control* $\widehat{u}$ *is unique and uniqueness holds in the following sense: if* $\widetilde{u} \in \mathcal{U}$ *is another control such that* $J(x_0,p_0,\widetilde{u}) = J(x_0,p_0,\widehat{u})$*, then* $P_\Omega(\widetilde{u}_s \neq \widehat{u}_s) > 0$ *only on a set of times* $s \in [0,T]$ *of Lebesgue measure zero, where* Ω *is the underlying probability sample space.*

PROPOSITION 2.2. *Assuming* (H1)–(H2)*, the value function* v *is continuous on* $[0,T]\times\mathbb{R}^{2n}$ *and*

$$|v(s,z)| \le C\left[1 + \sum_{i=1}^{n} e^{4z_i} + \sum_{i=n+1}^{2n} z_i^4\right], \tag{2.3}$$

where $C > 0$ *is a constant independent of* (s,z)*.*

Proof. The continuity of v can be established by use of (H1) and the continuous dependence of the cost (2.2) on the initial condition for the system (2.1) when $u \in \mathcal{U}$ is fixed. For an initial state $z_s = z$ and any fixed $u \in \mathcal{U}$, using (H2), we express $z_i(t)$, $1 \le i \le n$, in terms of $z_i|_{t=0}$ with a bounded term involving $z_k(s)$, $0 \le s \le t$, $k \neq i$, and get

$$\sup_{0\le t\le T} E e^{4z_i(t)} \le C_1\big(1 + e^{4z_i|_{t=0}}\big),$$

where $C_1 > 0$. By use of the structure of $C(x)$ and $D(x)$ in the cost integrand l, we obtain the estimates in a straightforward way,

$$\begin{aligned} |J(s,z,u)| &\le E\int_s^T |l(z_t)|dt \le E\int_s^T C_2\left[1 + \sum_{i=1}^{n} e^{4z_i(t)} + \sum_{i=n+1}^{2n} z_i^4(t)\right] dt \\ &\le C_3\left[1 + \sum_{i=1}^{n} e^{4z_i} + \sum_{i=n+1}^{2n} z_i^4\right], \end{aligned}$$

for constants C_2, C_3 independent of (s,z), and (2.3) follows. ∎

We see that in (2.1) the noise covariance matrix GG^τ is not of full rank. In general, under such a condition the corresponding stochastic optimal control problem does not admit classical solutions due to the degenerate nature of the arising HJB equations. Here we analyze viscosity solutions.

DEFINITION 2.3. *$\underline{v}(t,z) \in C([0,T]\times\mathbb{R}^{2n})$ is called a* viscosity subsolution *to the HJB equation*

$$
\begin{aligned}
0 &= -\frac{\partial v}{\partial t} + \sup_{u\in U}\left\{-\frac{\partial^\tau v}{\partial z}\psi\right\} - \frac{1}{2}\mathrm{tr}\left(\frac{\partial^2 v}{\partial z^2}GG^\tau\right) - l,\\
v|_{t=T} &= h(z), \quad z\in\mathbb{R}^{2n},
\end{aligned}
\tag{2.4}
$$

if $\underline{v}|_{t=T}\le h$, and for any $\varphi(t,z)\in C^{1,2}([0,T]\times\mathbb{R}^{2n})$, whenever $\underline{v}-\varphi$ takes a local maximum at $(t,z)\in[0,T)\times\mathbb{R}^{2n}$, we have

$$
-\frac{\partial \varphi}{\partial t} + \sup_{u\in U}\left\{-\frac{\partial^\tau \varphi}{\partial z}\psi\right\} - \frac{1}{2}\mathrm{tr}\left(\frac{\partial^2 \varphi}{\partial z^2}GG^\tau\right) - l \le 0, \quad z\in\mathbb{R}^{2n},
\tag{2.5}
$$

at (t,z). Here $\overline{v}(t,z)\in C([0,T]\times\mathbb{R}^{2n})$ is called a viscosity supersolution *to* (2.4) *if $\overline{v}|_{t=T}\ge h$, and in* (2.5) *we have an opposite inequality at (t,z), whenever $\overline{v}-\varphi$ takes a local minimum at $(t,z)\in[0,T)\times\mathbb{R}^{2n}$. Additionally, $v(t,z)$ is called a* viscosity solution *if it is both a viscosity subsolution and a viscosity supersolution.*

THEOREM 2.4. *The value function v is a viscosity solution to the HJB equation*

$$
\begin{aligned}
0 &= -\frac{\partial v}{\partial t} + \sup_{u\in U}\left\{-\frac{\partial^\tau v}{\partial z}\psi\right\} - \frac{1}{2}\mathrm{tr}\left(\frac{\partial^2 v}{\partial z^2}GG^\tau\right) - l,\\
v(T,z) &= 0.
\end{aligned}
\tag{2.6}
$$

Proof. The value function v is continuous (by Proposition 2.2) and satisfies the boundary condition in (2.6). Now, for any $\varphi(t,z)\in C^{1,2}([0,T]\times\mathbb{R}^{2n})$, suppose $v-\varphi$ has a local maximum at (s,z_0), $s<T$. We denote by $z^{(1)}$, $z^{(2)}$ the first n and last n, respectively, components of z. In the following proof, we assume that $\varphi(t,z)=0$ for all $z^{(1)}$ such that $|z^{(1)}-z_0^{(1)}|\ge C$ for a constant $C>0$; otherwise we can multiply $\varphi(t,z)$ by a C^∞ function $\zeta(z^{(1)})$ with compact support and $\zeta(z^{(1)})=1$ for $|z^{(1)}-z_0^{(1)}|\le\frac{C}{2}$. We take a constant control $u\in[-1,1]$ on $[s,T]$ to generate z_u with initial state $z_s=z_0$ and write $\Delta(t,z)=v(t,z)-\varphi(t,z)$. Since (s,z_0) is a local maximum point of $\Delta(t,z)$, we can find $\epsilon>0$ such that $\Delta(s_1,z)\le\Delta(s,z_0)$ for $|s_1-s|+|z-z_0|\le\epsilon$. For $s_1\in(s,T]$, $z_s=z_0$, write $1_{A^\epsilon}=1_{(|s_1-s|+|z_{s_1}-z_0|\ge\epsilon)}$. Then we get the lower bound estimate

$$
\begin{aligned}
&E[\Delta(s,z_0)-\Delta(s_1,z_{s_1})]\\
&\quad= E[\Delta(s,z_0)-\Delta(s_1,z_{s_1})](1-1_{A^\epsilon}) + E[\Delta(s,z_0)-\Delta(s_1,z_{s_1})]1_{A^\epsilon}\\
&\quad\ge E[\Delta(s,z_0)-\Delta(s_1,z_{s_1})]1_{A^\epsilon} \stackrel{\triangle}{=} S_0,
\end{aligned}
$$

and using basic estimates for the change of the value function with respect to different initial states (see, e.g., [12, 3] for standard techniques), it follows that

$$
\begin{aligned}
|S_0| &= O\left(Ee^{2|z_{s_1}^{(1)}|}1_{A^\epsilon}\right)\\
&= O\left(Ee^{2|z_{s_1}^{(1)}|}1_{(|z_{s_1}^{(1)}-z_0^{(1)}|\ge\epsilon/2)}\right) \qquad (2.7)\\
&= O(|s-s_1|^2) \qquad (2.8)
\end{aligned}
$$

when $s_1 \downarrow s$. Here we obtain (2.7) by the fact that $z_{s_1}^{(2)} \to z_0^{(2)}$ uniformly as $s_1 \downarrow s$, which follows from the Lipschitz and boundedness (w.r.t. increment in x) conditions for $g(t,x,p,u)$, and obtain the bound (2.8) using basic moment estimates for $|z_{s_1}^{(1)} - z_0^{(1)}|^2$. It follows from (2.8) that

$$\lim_{s_1 \downarrow s} \frac{1}{s_1 - s} E[\Delta(s, z_0) - \Delta(s_1, z_{s_1})] \geq 0. \tag{2.9}$$

However, for $s_1 \in (s, T]$ we also have

$$\begin{aligned} &\frac{1}{s_1 - s} E[\Delta(s, z_0) - \Delta(s_1, z_{s_1})] \\ &\quad \leq \frac{1}{s_1 - s} E\left[\int_s^{s_1} l(z_t)dt - \varphi(s, z_0) + \varphi(s_1, z_{s_1})\right] \\ &\quad \to \left[l + \frac{\partial \varphi}{\partial s} + \frac{\partial^\tau \varphi}{\partial z}\psi\Big|_u + \frac{1}{2}\mathrm{tr}\left(\frac{\partial^2 \varphi}{\partial z^2} GG^\tau\right)\right]\Bigg|_{(s,z_0)} \qquad \forall u \in U \end{aligned} \tag{2.10}$$

as $s_1 \downarrow s$, where we get the inequality by the principle of optimality, and obtain (2.10) by using Ito's formula to express $\varphi(s_1, z_{s_1})$ near (s, z_0) and then taking expectations. In the above, since v satisfies the growth condition in Proposition 2.2, $\varphi(t,z) = 0$ for $|z^{(1)} - z_0^{(1)}| \geq C$, all the expectations are finite. Therefore, for $z \in \mathbb{R}^{2n}$, by (2.9) and (2.10),

$$\frac{\partial \varphi}{\partial s} + \min_{u \in U}\left\{\frac{\partial^\tau \varphi}{\partial z}\psi\right\} + \frac{1}{2}\mathrm{tr}\left(\frac{\partial^2 \varphi}{\partial z^2} GG^\tau\right) + l \geq 0$$

at (s, z_0). On the other hand, if $v - \varphi$ has a local minimum at (s, z_0), $s < T$, then for any small $\varepsilon > 0$ we can choose sufficiently small $s_1 \in (s, T]$ and find a control $u \in \mathcal{U}$ generating z_u such that

$$\begin{aligned} &E\{v(s, z_0) - \varphi(s, z_0) - v(s_1, z_{s_1}) + \varphi(s_1, z_{s_1})\} \\ &\qquad \geq E\left\{\int_s^{s_1} l(z_t)dt + \varphi(s_1, z_{s_1}) - \varphi(s, z_0)\right\} - \varepsilon(s_1 - s). \end{aligned} \tag{2.11}$$

Similar to (2.8), we also have

$$E[\Delta(s, z_0) - \Delta(s_1, z_{s_1})] \leq O(|s - s_1|^2),$$

which, together with (2.11) and Ito's formula, gives

$$\frac{\partial \varphi}{\partial s} + \min_{u \in U}\left\{\frac{\partial^\tau \varphi}{\partial z}\psi\right\} + \frac{1}{2}\mathrm{tr}\left(\frac{\partial^2 \varphi}{\partial z^2} GG^\tau\right) + l \leq 0$$

at (s, z_0), so that the value function v is a viscosity solution. □

To analyze uniqueness of the viscosity solution, we introduce the function class $\mathcal{G}$ such that each $W \in \mathcal{G}$ satisfies the following:

(i) $W \in C([0,T] \times \mathbb{R}^{2n})$, and

(ii) for any $W \in \mathcal{G}$, there exist $C, k_1, k_2 > 0$ such that $|W(t,z)| \leq C[\sum_{i=1}^n e^{k_1|z_i|} + \sum_{i=1}^{2n} |z_i|^{k_2}]$.

Notice that in condition (ii), the constants C, k_1, k_2 may take a different set of values for different $W \in \mathcal{G}$. By Proposition 2.2 and Theorem 2.4 it follows that the

value function v is a viscosity solution to the HJB equation (2.6) in the class $\mathcal{G}$. We now state the uniqueness result for the viscosity solutions.

THEOREM 2.5. *Assuming that* (H1)–(H2) *hold, there exists a unique viscosity solution to* (2.6) *in the class* $\mathcal{G}$.

Here we state a general maximum principle on an unbounded domain for the HJB equation (2.6). By considering two possibly distinct viscosity solutions v_1 and v_2 and setting, respectively, $(v_1, v_2) = (\underline{v}, \overline{v})$ and $(v_2, v_1) = (\underline{v}, \overline{v})$ in Theorem 2.6, we obtain Theorem 2.5 as a corollary. The proof of the maximum principle is postponed to section 4.

THEOREM 2.6. *Assuming that* (H1)–(H2) *hold, if* $\underline{v}, \overline{v} \in \mathcal{G}$ *are the viscosity subsolution and supersolution to* (2.6)*, respectively, and* $\sup_{\partial^* Q_0}(\underline{v} - \overline{v}) < \infty$*, then*

$$\sup_{Q_0}(\underline{v} - \overline{v}) = \sup_{\partial^* Q_0}(\underline{v} - \overline{v}), \tag{2.12}$$

where $Q_0 = [0, T] \times \mathbb{R}^{2n}$, $\partial^* Q_0 = \{(T, z) : z \in \mathbb{R}^{2n}\}$.

3. Semiconvex and semiconcave approximations on compact sets. To facilitate our analysis, write the Hamiltonian

$$\widetilde{H}(t, z, u, \xi, V) = -\xi^\tau \psi(t, z, u) - \frac{1}{2}\mathrm{tr}\{VG(t, z)G^\tau(t, z)\} - l(z), \tag{3.1}$$

$$H(t, z, \xi, V) = \sup_{u \in U} \widetilde{H}(t, z, u, \xi, V),$$

where $\xi \in \mathbb{R}^{2n}$, V is a $2n \times 2n$ real symmetric matrix, and the other terms are defined in section 2. Then the HJB equation (2.6) may be written as

$$0 = -v_t + H(t, z, v_z, v_{zz}), \tag{3.2}$$

$$v(T, z) = 0. \tag{3.3}$$

DEFINITION 3.1 (see [12]). *A real value function* $\varphi(x)$ *defined on a convex set* $Q \subset \mathbb{R}^m$ *is said to be semiconvex on* Q *if there exists a constant* $C > 0$ *such that* $\varphi(x) + C|x|^2$ *is convex;* $\varphi(x)$ *is semiconcave on* Q *if* $-\varphi(x)$ *is semiconvex on* Q.

DEFINITION 3.2. *A real value function* $\varphi(x)$ *defined on a convex set* $Q \subset \mathbb{R}^m$ *is said to be locally semiconvex on* Q *if for any* $y \in Q$ *there exists a convex neighborhood* N_y *(relative to* Q*) of* y *such that* $\varphi(x)$ *is semiconvex on* N_y.

PROPOSITION 3.3. *If* $\varphi(x)$ *is locally semiconvex on a convex compact set* Q*, then* $\varphi(x)$ *is semiconvex on* Q.

Proof. For any $y \in Q$, there exists a convex set N_y open relative to Q such that $y \in N_y$ and $\varphi(x)$ is semiconvex on N_y. Thus there exists $C_y > 0$ such that $\varphi(x) + C_y|x|^2$ is convex on N_y. Since $\{N_y,\ y \in Q\}$ is an open cover of Q, there exists a finite subcover denoted by $\{N_{y_i},\ 1 \le i \le k\}$. Take $C = \max_{1 \le i \le k} C_{y_i}$, and then obviously $\varphi(x) + C|x|^2 \stackrel{\triangle}{=} \widehat{\varphi}(x)$ is convex on each N_{y_i}, $1 \le i \le k$. Now for any $x_1, x_2 \in Q$, $0 \le \lambda \le 1$, we prove that $\widehat{\varphi}(\lambda x_1 + (1 - \lambda)x_2) \le \lambda\widehat{\varphi}(x_1) + (1 - \lambda)\widehat{\varphi}(x_2)$. It suffices to consider the case $0 < \lambda < 1$. First, from the collection $\{N_{y_i},\ 1 \le i \le k\}$ we select open sets, without loss of generality, denoted as $\mathcal{N} \stackrel{\triangle}{=} \{N_{y_i},\ i = 1, \ldots, m \le k\}$

such that $L \triangleq \{x : x = \alpha x_1 + (1-\alpha)x_2,\ 0 \le \alpha \le 1\} \subset \cup_{N_{y_i} \in \mathcal{N}} N_{y_i}$. For simplicity we consider the case $m = 2$ and $x_1 \in N_{y_1}$, $x_2 \in N_{y_2}$. The general case may be treated inductively. To avoid triviality, we assume that neither N_{y_1} nor N_{y_2} covers L individually, and then we can find $x_a \in L$, $x_a \neq x_\lambda \triangleq \lambda x_1 + (1-\lambda)x_2$ such that $x_a \in N_{y_1} \cap N_{y_2}$ and $x_a = c_1 x_1 + (1-c_1)x_2$, $0 < c_1 < 1$. Without loss of generality we assume that x_λ is between x_1 and x_a. Then we further choose $x_b \in N_{y_1} \cap N_{y_2}$ such that $x_b = c_2 x_1 + (1-c_2)x_2$ and x_b is between x_a and x_2. Now it is obvious that $0 < c_2 < c_1 < \lambda < 1$. It is straightforward to verify that

$$x_\lambda = \frac{\lambda - c_1}{1-c_1} x_1 + \frac{1-\lambda}{1-c_1} x_a, \qquad x_a = \frac{c_1 - c_2}{\lambda - c_2} x_\lambda + \frac{\lambda - c_1}{\lambda - c_2} x_b, \qquad x_b = \frac{c_2}{c_1} x_a + \frac{c_1 - c_2}{c_1} x_2.$$

Hence we have

$$\widehat{\varphi}(x_\lambda) \le \frac{\lambda - c_1}{1-c_1}\widehat{\varphi}(x_1) + \frac{1-\lambda}{1-c_1}\widehat{\varphi}(x_a),$$

$$\widehat{\varphi}(x_a) \le \frac{c_1 - c_2}{\lambda - c_2}\widehat{\varphi}(x_\lambda) + \frac{\lambda - c_1}{\lambda - c_2}\widehat{\varphi}(x_b),$$

$$\widehat{\varphi}(x_b) \le \frac{c_2}{c_1}\widehat{\varphi}(x_a) + \frac{c_1 - c_2}{c_1}\widehat{\varphi}(x_2),$$

where we get the first two inequalities and the last one by the convexity of $\widehat{\varphi}(x)$ on N_{y_1} and N_{y_2}, respectively. By a simple transformation with the above inequalities to eliminate $\widehat{\varphi}(x_a)$ and $\widehat{\varphi}(x_b)$, we get

$$\widehat{\varphi}(x_\lambda) \le \lambda\widehat{\varphi}(x_1) + (1-\lambda)\widehat{\varphi}(x_2).$$

By arbitrariness of x_1, x_2 in Q it follows that $\widehat{\varphi}(x)$ is convex on Q. This completes the proof. □

We adopt the semiconvex/semiconcave approximation technique of [12, 2, 9, 10], but due to the highly nonlinear growth condition of the class $\mathcal{G}$, we apply a particular localized technique to construct envelope functions to generate semiconvex/semiconcave approximations on any bounded domain. For any $W \in \mathcal{G}$, define the upper/lower envelope functions with $\eta \in (0, 1]$,

$$W^\eta(t,z) = \sup_{(s,w)\in B^\eta(t,z)} \left\{ W(s,w) - \frac{1}{2\eta^2}(|t-s|^2 + |z-w|^2) \right\}, \tag{3.4}$$

$$W_\eta(t,z) = \inf_{(s,w)\in B^\eta(t,z)} \left\{ W(s,w) + \frac{1}{2\eta^2}(|t-s|^2 + |z-w|^2) \right\}, \tag{3.5}$$

where $B^\eta(t,z)$ denotes the closed ball (relative to $[0,T]\times\mathbb{R}^{2n}$) centering (t,z) with radius η. As will be shown in the following lemma, our construction above will generate semiconvex/semiconcave approximations to a given continuous function on a compact set for small η.

LEMMA 3.4. *For any fixed $W \in \mathcal{G}$ and compact convex set $Q \subset [0,T]\times\mathbb{R}^{2n}$, there exists a positive constant $\eta_Q \le 1$ depending only on Q such that for all $\eta \in (0, \eta_Q]$, $W^\eta(t,z)$ is semiconvex on Q, and $W_\eta(t,z)$ is semiconcave on Q.*

Proof. Since any fixed $W \in \mathcal{G}$ is uniformly continuous and bounded on any compact set Q, there exists $\overline{\eta}_Q > 0$ depending only on Q, so that for all positive $\eta \le \overline{\eta}_Q$ and $(t,z) \in Q$,

$$W^\eta(t,z) = \sup_{(s,w)\in B^{\eta/2}(t,z)} \left\{ W(s,w) - \frac{1}{2\eta^2}[|t-s|^2 + |z-w|^2] \right\}. \tag{3.6}$$

Indeed, we can find $\overline{\eta}_Q > 0$ such that for all $\eta \le \overline{\eta}_Q$, $|W(s,w) - W(t,z)| \le \frac{1}{16}$ for $(s,w) \in B^\eta(t,z)$, where $(t,z) \in Q$. Then for any (s,w) satisfying $\frac{\eta^2}{4} \le |s-t|^2 + |w-z|^2 \le \eta^2$, we have

$$W(s,w) - \frac{1}{2\eta^2}(|s-t|^2 + |w-z|^2) \le W(t,z) + \frac{1}{16} - \frac{1}{2\eta^2}\frac{\eta^2}{4} < W(t,z),$$

and (3.6) follows. In the following we assume $\eta \le \overline{\eta}_Q$. Next we show that for any $(t_0,z_0) \in Q$, $W^\eta(t,z)$ is semiconvex on $B^{\eta/4}(t_0,z_0) \cap Q$. It suffices to show that $W^\eta(t,z) + \frac{1}{2\eta^2}(t^2 + |z|^2)$ is convex on $B^{\eta/4}(t_0,z_0) \cap Q$. Denote

$$R(s,w,t,z) = W(s,w) - \frac{1}{2\eta^2}(|t-s|^2 + |z-w|^2) + \frac{1}{2\eta^2}(t^2 + |z|^2).$$

If $(t_1,z_1), (t_2,z_2) \in B^{\eta/4}(t_0,z_0) \cap Q$, we have $(t_2,z_2) \in B^{\eta/2}(t_1,z_1)$. For any $\lambda \in [0,1]$, we denote $(t_\lambda, z_\lambda) = (\lambda t_1 + (1-\lambda)t_2, \lambda z_1 + (1-\lambda)z_2) \in Q$. It is obvious that $B^{\eta/2}(t_\lambda,z_\lambda) \subset B^\eta(t_1,z_1) \cap B^\eta(t_2,z_2)$. Then it follows that

$$\begin{aligned}
&W^\eta(t_\lambda,z_\lambda) + \frac{1}{2\eta^2}[t_\lambda^2 + |z_\lambda|^2] \\
&\quad = \sup_{(s,w)\in B^\eta(t_\lambda,z_\lambda)} R(s,w,t_\lambda,z_\lambda) = \sup_{(s,w)\in B^{\eta/2}(t_\lambda,z_\lambda)} R(s,w,t_\lambda,z_\lambda) \\
&\quad = \sup_{(s,w)\in B^{\eta/2}(t_\lambda,z_\lambda)} [\lambda R(s,w,t_1,z_1) + (1-\lambda)R(s,w,t_2,z_2)] \\
&\quad \le \sup_{(s,w)\in B^{\eta/2}(t_\lambda,z_\lambda)} \lambda R(s,w,t_1,z_1) + \sup_{(s,w)\in B^{\eta/2}(t_\lambda,z_\lambda)} (1-\lambda)R(s,w,t_2,z_2) \\
&\quad \le \sup_{(s,w)\in B^\eta(t_1,z_1)} \lambda R(s,w,t_1,z_1) + \sup_{(s,w)\in B^\eta(t_2,z_2)} (1-\lambda)R(s,w,t_2,z_2) \\
&\quad = \lambda\left[W^\eta(t_1,z_1) + \frac{1}{2\eta^2}(t_1^2 + |z_1|^2)\right] + (1-\lambda)\left[W^\eta(t_2,z_2) + \frac{1}{2\eta^2}(t_2^2 + |z_2|^2)\right].
\end{aligned}$$

Thus $W^\eta(t,z)$ is semiconvex on $B^{\eta/4}(t_0,z_0) \cap Q$. Further, by Proposition 3.3, $W^\eta(t,z)$ is semiconvex on Q. Similarly we can prove that $W_\eta(t,z)$ is semiconcave on Q for $\eta \in (0, \widetilde{\eta}_Q]$, where $\widetilde{\eta}_Q \le 1$ depends only on Q. The lemma follows by taking $\eta_Q = \min\{\overline{\eta}_Q, \widetilde{\eta}_Q\}$. ∎

We use an example to illustrate the construction of the semiconvex approximation to a given function.

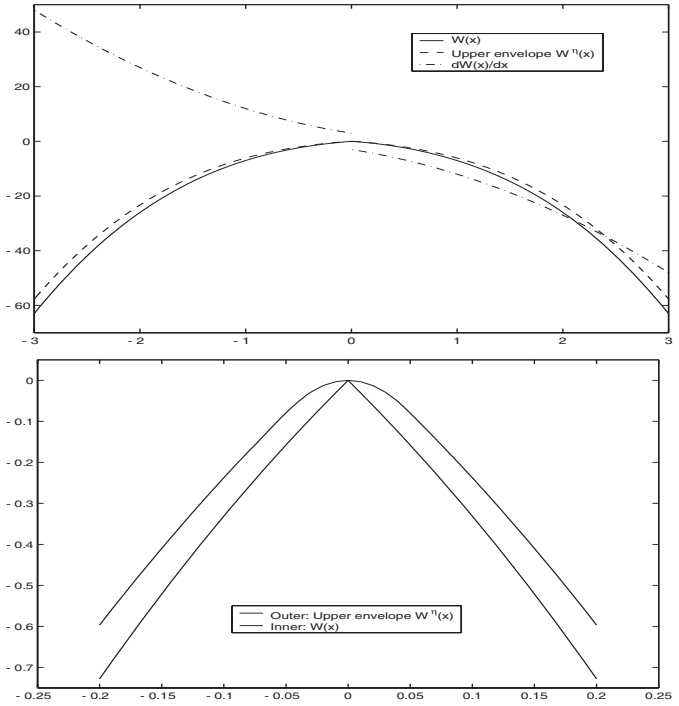


FIG. 3.1. *Semiconvex approximation with* $\eta = 0.125$. *Top: the curves in a large range. Bottom: the curves in the local region.*

Example 1. Consider a continuous function $W\colon \mathbb{R} \to \mathbb{R}$ defined as follows:

$$W(x) = \begin{cases} (x-1)^3 + 1 & \text{for } x \le 0, \\ -(x+1)^3 + 1 & \text{for } x > 0. \end{cases}$$

We take $0 < \eta \le 0.125$ and write

$$\theta(x) = 1 - x + \frac{1}{6\eta^2} - \sqrt{\left[1 - x + \frac{1}{6\eta^2}\right]^2 - (1-x)^2}, \quad x \le 0.$$

It is evident that the upper envelope function $W^\eta(x)$ is even on $\mathbb{R}$, and its value on $(-\infty, 0]$ is determined by

$$W^\eta(x) = \begin{cases} W(x+\eta) - \dfrac{1}{2} & \text{for } x \le 1 - \eta - \dfrac{1}{\sqrt{3\eta}}, \\ W(x+\theta(x)) - \dfrac{\theta^2(x)}{2\eta^2} & \text{for } 1 - \eta - \dfrac{1}{\sqrt{3\eta}} < x \le -3\eta^2, \\ W(0) - \dfrac{x^2}{2\eta^2} & \text{for } -3\eta^2 < x \le 0, \end{cases} \tag{3.7}$$

where $0 \le \theta(x) \le \eta \wedge |x|$ holds for $1 - \eta - \frac{1}{\sqrt{3\eta}} < x \le -3\eta^2$.

From Figure 3.1, it is seen that at $x = 0$ the first order derivative of $W(x)$ has a negative jump, which corresponds to a sharp turn at $x = 0$ on the function curve. After the semiconvexifying procedure, the sharp turn at $x = 0$ vanishes, as shown by the curve of $W^\eta(x)$.

We give a lemma which is parallel to the one in [12]. But here we do not make Lipschitz or Hölder-type continuity assumptions on W. For completeness we give the details.

LEMMA 3.5. *For $W \in \mathcal{G}$ and $\eta \in (0,1]$, W^η and W_η are equicontinuous (w.r.t. η) on any compact set $Q \subset [0,T] \times \mathbb{R}^{2n}$ and*

$$W^\eta(t,z) \le C\left[\sum_{i=1}^{n} e^{k_1|z_i|} + \sum_{i=1}^{2n} |z_i|^{k_2}\right], \tag{3.8}$$

$$W^\eta(t,z) = W(t_0,z_0) - \frac{1}{2\eta^2}(|t-t_0|^2 + |z-z_0|^2) \quad \text{for some } (t_0,z_0) \in B^\eta(t,z), \tag{3.9}$$

$$\frac{1}{2\eta^2}(|t-t_0|^2 + |z-z_0|^2) \to 0 \quad \text{uniformly on } Q \text{ as } \eta \to 0, \text{ and} \tag{3.10}$$

$$0 \le W^\eta(t,z) - W(t,z) \to 0 \quad \text{uniformly on } Q \text{ as } \eta \to 0, \tag{3.11}$$

where $C, k_1, k_2 > 0$ are constants independent of η. The estimates (3.8)–(3.10) *also hold when W^η is replaced by W_η, and*

$$0 \le W(t,z) - W_\eta(t,z) \to 0 \quad \text{uniformly on } Q \text{ as } \eta \to 0. \tag{3.12}$$

Proof. Inequality (3.8) follows from the definition of $\mathcal{G}$, and (3.9) is obvious. Moreover, by (3.9) we have

$$\frac{1}{2\eta^2}(|t-t_0|^2 + |z-z_0|^2) = W(t_0,z_0) - W^\eta(t,z) \le W(t_0,z_0) - W(t,z). \tag{3.13}$$

Since $|t-t_0| + |z-z_0| \to 0$ as $\eta \to 0$, by (3.13) and the uniform continuity of W on Q, (3.10) follows. The estimate (3.11) follows from (3.9) and (3.10). The equicontinuity of W^η (w.r.t. η) on Q can be established by (3.11) and the continuous dependence of W^η on $(\eta,t,z) \in [\varepsilon,1] \times Q$ for any $0 < \varepsilon \le 1$. The case of W_η can be treated similarly. □

We define

$$H^\eta(t,z,\xi,V) = \inf_{(s,w)\in B^\eta(t,z)} \sup_{u\in U} \widetilde{H}(s,w,u,\xi,V), \tag{3.14}$$

$$H_\eta(t,z,\xi,V) = \sup_{(s,w)\in B^\eta(t,z)} \sup_{u\in U} \widetilde{H}(s,w,u,\xi,V). \tag{3.15}$$

Then it can be shown that H^η and H_η converge to $H(t,z,\xi,V)$ uniformly on any compact subset of $[0,T] \times \mathbb{R}^{2n} \times \mathbb{R}^{2n} \times S^{2n}$ as $\eta \to 0$, where S^{2n} denotes the set of $2n \times 2n$ real symmetric matrices. The following lemma can be proved by a method similar to that in [4, 9, 10]; the proof is omitted here. Notice that the viscosity sub/supersolution properties hold on a domain smaller than $[0,T] \times \mathbb{R}^{2n}$.

LEMMA 3.6. *If $\underline{v}$ ($\overline{v}$, respectively) is a viscosity subsolution (supersolution, respectively) to* (3.2) *on $[0,T] \times \mathbb{R}^{2n}$, then $\underline{v}^\eta$ ($\overline{v}_\eta$, respectively) is a viscosity subsolution (supersolution, respectively) to HJB equation A (B, respectively) on $[0, T-\eta] \times \mathbb{R}^{2n}$, where the HJB equations A and B are given by*

$$A: \begin{cases} -v_t + H^\eta(t,z,v_z,v_{zz}) = 0, \\ v(T-\eta,z) = \underline{v}^\eta(T-\eta,z), \end{cases} \qquad B: \begin{cases} -v_t + H_\eta(t,z,v_z,v_{zz}) = 0, \\ v(T-\eta,z) = \overline{v}_\eta(T-\eta,z). \end{cases}$$

In the above, $\underline{v}^\eta$ and $\overline{v}_\eta$ are defined by (3.4)–(3.5).

4. Proof of Theorem 2.6. In this section we give a proof of Theorem 2.6. We note that certain technical but standard arguments are not included here for reasons of economy of exposition; complete references to the detailed versions of these parts of the proof are supplied at appropriate places in the text.

We follow the method in [12, 4], employing the particular structure of the system dynamics, and will make necessary modifications. For the viscosity subsolution and supersolution $\underline{v}, \overline{v} \in \mathcal{G}$ we prove that

$$\sup_{Q_1}(\underline{v}-\overline{v}) = \sup_{\partial^* Q_0}(\underline{v}-\overline{v}) \stackrel{\triangle}{=} c_0 \quad \text{for } Q_1 = [T_1, T]\times\mathbb{R}^{2n}, \tag{4.1}$$

where $T_1 = T - \frac{1}{4\Delta}$, $\Delta = 25n(C_g + C_\sigma) + 10C_{f^0}$, C_g is a finite constant such that for g given in (1.2), $|g_i(t,x,p,u)| \le C_g(1+\sum_{k=1}^n |p_k|)$ for $t\in[0,T]$, $x,p\in\mathbb{R}^n$, $u\in U$, $1\le i\le n$, and C_σ, C_{f^0} are given in assumptions (H1)–(H2) introduced in section 2. The maximum principle (2.12) follows by repeating the above procedure backward with time. Our proof by contradiction starts with the observation that if (4.1) is not true, there exists $(\hat{t},\hat{z})\in(T_1,T)\times\mathbb{R}^{2n}$ such that

$$\underline{v}(\hat{t},\hat{z}) - \overline{v}(\hat{t},\hat{z}) = c_0^+ > c_0. \tag{4.2}$$

We break the proof into several steps: (1) we construct a comparison function Λ depending on positive parameters α, β, ε, λ, and, based upon (4.2), Λ is used to induce a certain interior maximum; (2) using the viscosity sub/supersolution conditions, we get a set of inequalities at the interior maximum; and (3) we establish an inequality relation between α and β by taking appropriate vanishing subsequences of ε, λ, η, and this inequality relation is shown to lead to a contradiction. The weak coupling condition (H2) for x is used to obtain estimates used in Step 3 below.

Step 1 (constructing a comparison function and the interior maximum). To avoid introducing too many constants, we assume that $\underline{v}$ and $\overline{v}$ belong to the class $\mathcal{G}$ with associated constants $k_1 = k_2 = 4$. The more general case can be treated in exactly the same way. Now we define the comparison function

$$\begin{aligned}\Lambda(t,z,s,w) = {}& \frac{\alpha(2\mu T - t - s)}{2\mu T}\left\{\sum_{i=1}^n\left[e^{5\sqrt{z_i^2+1}} + e^{5\sqrt{w_i^2+1}}\right] + \sum_{i=1}^{2n}(z_i^6 + w_i^6)\right\}\\ & - \beta(t+s) + \frac{1}{2\varepsilon}|t-s|^2 + \frac{1}{2\varepsilon}|z-w|^2 + \frac{\lambda}{t-T_1} + \frac{\lambda}{s-T_1},\end{aligned}$$

where α, β, ε, λ are all taken from $(0,1]$; $\mu = 1 + \frac{1}{4T\Delta}$; $z,w\in\mathbb{R}^{2n}$; and $t,s\in(T_1,T]$. We write $\Phi(t,z,s,w) = \underline{v}^\eta(t,z) - \overline{v}_\eta(s,w) - \Lambda(t,z,s,w)$, where $\underline{v}^\eta$ and $\overline{v}_\eta$ are also in $\mathcal{G}$ by Lemma 3.5. Noticing that $\Phi\to-\infty$ as $t\wedge s\to T_1$ or $|z|+|w|\to\infty$, there exists (t_0,z_0,s_0,w_0) such that $\Phi(t_0,z_0,s_0,w_0) = \sup_{Q_1\times Q_1}\Phi(t,z,s,w)$. By $\Phi(t_0,z_0,s_0,w_0)\ge\Phi(T,0,T,0)$, one can find a constant C_α depending only on α such that (see Remark 2)

$$|z_0| + |w_0| + \frac{1}{2\varepsilon}|t_0-s_0|^2 + \frac{1}{2\varepsilon}|z_0-w_0|^2 \le C_\alpha \quad\text{and}\quad t_0, s_0 \in \left[T_1 + \frac{\lambda}{C_\alpha}, T\right]. \tag{4.3}$$

Combining $2\Phi(t_0,z_0,s_0,w_0)\ge\Phi(t_0,z_0,t_0,z_0)+\Phi(s_0,w_0,s_0,w_0)$, (4.3), and Lemma 3.5, we get for fixed $\alpha>0$ (see Remark 3)

$$\frac{1}{2\varepsilon}|t_0-s_0|^2 + \frac{1}{2\varepsilon}|z_0-w_0|^2 \to 0 \quad\text{uniformly as } \varepsilon\to 0. \tag{4.4}$$

In this section, we take $\beta \in (0, \frac{c_0^+ - c_0}{4T})$. We further show that there exists $\alpha_0 > 0$ such that for $\alpha < \alpha_0$ and for sufficiently small r_0 (which may depend upon α) and $\eta \le r_0$, $\varepsilon \le r_0$, $\lambda \le r_0$, the maximum of Φ on Q_1 is attained at an interior point (t_0, z_0, s_0, w_0) of the set

$$Q_\alpha = \left\{ (t,z,s,w) : T_1 + \frac{\lambda}{2C_\alpha} \le t,\ s \le T - \eta, \text{ and } |z|, |w| \le 2C_\alpha \right\}, \tag{4.5}$$

where C_α is determined in (4.3).

We begin by observing that $\Phi(t_0, z_0, s_0, w_0) \ge \Phi(\hat{t}, \hat{z}, \hat{t}, \hat{z})$ yields

$$\begin{aligned} \underline{v}^\eta(\hat{t},\hat{z}) - \overline{v}_\eta(\hat{t},\hat{z}) &\le \underline{v}^\eta(t_0,z_0) - \overline{v}_\eta(s_0,w_0) - \Lambda(t_0,z_0,s_0,w_0) + \Lambda(\hat{t},\hat{z},\hat{t},\hat{z}) \\ &\le \underline{v}^\eta(t_0,z_0) - \overline{v}_\eta(s_0,w_0) + 2\beta T + \frac{2\lambda}{\hat{t} - T_1} \\ &\quad + 2\alpha \left[\sum_{i=1}^n e^{5\sqrt{\hat{z}_i^2+1}} + \sum_{i=1}^{2n} \hat{z}_i^6 \right]. \end{aligned} \tag{4.6}$$

Let $\mathbb{H}^\beta$ stand for the assertion that there exists α_0 such that when $\alpha \le \alpha_0$ and $\max\{\eta, \varepsilon, \lambda\} \le r_0$ for sufficiently small r_0, (t_0, z_0, s_0, w_0) is an interior point of Q_α in (4.5).

If $\mathbb{H}^\beta$ is not true, then there exists an arbitrarily small $\alpha \in (0, 1]$ such that for this fixed α we can select $\eta^{(k)}, \varepsilon^{(k)}, \lambda^{(k)} \to 0$ for which the resulting $(t_0^{(k)}, z_0^{(k)}, s_0^{(k)}, w_0^{(k)}) \notin \mathrm{Int}(Q_\alpha)$. By (4.3) it necessarily follows that $t_0^{(k)} \vee s_0^{(k)} \ge T - \eta^{(k)} \to T$ and (4.4) gives $|t_0^{(k)} - s_0^{(k)}| + |s_0^{(k)} - w_0^{(k)}| \to 0$. It is also clear that $(t_0^{(k)}, z_0^{(k)}, s_0^{(k)}, w_0^{(k)})$ is contained in a compact set determined by α. Then by selecting an appropriate subsequence of $(t_0^{(k)}, z_0^{(k)}, s_0^{(k)}, w_0^{(k)})$ and taking the limit in (4.6) along this subsequence, we get

$$\begin{aligned} \underline{v}(\hat{t},\hat{z}) - \overline{v}(\hat{t},\hat{z}) &\le \underline{v}(T,z^\alpha) - \overline{v}(T,z^\alpha) + \frac{c_0^+ - c_0}{2} + 2\alpha \left[\sum_{i=1}^n e^{5\sqrt{\hat{z}_i^2+1}} + \sum_{i=1}^{2n} \hat{z}_i^6 \right] \\ &\le \frac{c_0^+ + c_0}{2} + 2\alpha \left[\sum_{i=1}^n e^{5\sqrt{\hat{z}_i^2+1}} + \sum_{i=1}^{2n} \hat{z}_i^6 \right], \end{aligned} \tag{4.7}$$

where z^α denotes the common limit of the selected subsequences of $z_0^{(k)}$ and $w_0^{(k)}$. Sending $\alpha \to 0$, we get $\underline{v}(\hat{t},\hat{z}) - \overline{v}(\hat{t},\hat{z}) < c_0^+$, which contradicts (4.2); hence $\mathbb{H}^\beta$ holds. From the argument leading to (4.7) it is seen that α_0 can be chosen independently of β.

Step 2 (applying Ishii's lemma). Hereafter, we assume that $\beta < \frac{c_0^+ - c_0}{4T}$, $\alpha < \alpha_0$, and $\max\{\eta, \varepsilon, \lambda\} \le r_0$ are always satisfied and thus $\mathbb{H}^\beta$ holds. We assume Φ attains a strict maximum at (t_0, z_0, s_0, w_0); otherwise we replace Λ by $\Lambda + |t - t_0|^2 + |s - s_0|^2 + |z - z_0|^4 + |w - w_0|^4$. Following the derivations in [12, 9, 4] and using the interior maximum obtained in Step 1, the semiconvexity of $\underline{v}^\eta$, and the semiconcavity of $\overline{v}_\eta$ for $\eta \le \eta_{Q_\alpha}$ by Lemma 3.4, and by Lemma 3.6, we obtain the so-called Ishii's lemma; i.e., there exist $2n \times 2n$ symmetric matrices M_k, $k = 1, 2$, such that

$$-\Lambda_t(t_0,z_0,s_0,w_0) + H^\eta(t_0,z_0,\Lambda_z(t_0,z_0,s_0,w_0),M_1) \le 0, \tag{4.8}$$

$$\Lambda_s(t_0,z_0,s_0,w_0) + H_\eta(s_0,w_0,-\Lambda_w(t_0,z_0,s_0,w_0),M_2) \ge 0, \tag{4.9}$$

$$\begin{pmatrix} M_1 & 0 \\ 0 & -M_2 \end{pmatrix} \le \begin{pmatrix} \Lambda_{zz} & \Lambda_{zw} \\ \Lambda_{zw}^\tau & \Lambda_{ww} \end{pmatrix} \Big|_{(t_0,z_0,s_0,w_0)}. \tag{4.10}$$

We note that it is important to have $t_0 \vee s_0 < T - \eta$ in order to establish (4.8)–(4.9) by Lemma 3.6 and an approximation procedure (see, e.g., [4] for the case of a bounded domain). Now (4.8) and (4.9) yield

$$\begin{aligned} &-\Lambda_t(t_0,z_0,s_0,w_0) - \Lambda_s(t_0,z_0,s_0,w_0) \\ &\quad \le H_\eta(s_0,w_0,-\Lambda_w(t_0,z_0,s_0,w_0),M_2) - H^\eta(t_0,z_0,\Lambda_z(t_0,z_0,s_0,w_0),M_1). \end{aligned} \tag{4.11}$$

Step 3 (estimates for LHS and RHS of (4.11)). The final stage in our deduction of a contradiction from (4.2) involves estimates of the LHS and RHS of (4.11). The estimates for both sides of (4.11) are taken at (t_0,z_0,s_0,w_0), but for brevity we omit the subscript 0 for each variable. We have

$$\begin{aligned} \text{LHS of (4.11)} &= \frac{\alpha}{\mu T}\left[\sum_{i=1}^n \left(e^{5\sqrt{z_i^2+1}} + e^{5\sqrt{w_i^2+1}}\right) + \sum_{i=1}^n (z_i^6 + w_i^6)\right] \\ &\quad + 2\beta + \frac{\lambda}{(t-T_1)^2} + \frac{\lambda}{(s-T_1)^2} \\ &\ge \frac{\alpha}{\mu T}\left[\sum_{i=1}^n \left(e^{5\sqrt{z_i^2+1}} + e^{5\sqrt{w_i^2+1}}\right) + \sum_{i=1}^n (z_i^6 + w_i^6)\right] + 2\beta \end{aligned} \tag{4.12}$$

and

$$\begin{aligned} \text{RHS of (4.11)} &= \sup_{u\in U}[\Lambda_w^\tau(t,z,s,w)\psi(\widehat{s},\widehat{w},u)] - \sup_{u\in U}[-\Lambda_z^\tau(t,z,s,w)\psi(\widehat{t},\widehat{z},u)] \\ &\quad + \frac{1}{2}\mathrm{tr}[G(\widehat{t},\widehat{z})G^\tau(\widehat{t},\widehat{z})M_1] - \frac{1}{2}\mathrm{tr}[G(\widehat{s},\widehat{w})G^\tau(\widehat{s},\widehat{w})M_2] + l(\widehat{z}) - l(\widehat{w}) \\ &\le \sup_{u\in U}[\Lambda_w^\tau(t,z,s,w)\psi(\widehat{s},\widehat{w},u) + \Lambda_z^\tau(t,z,s,w)\psi(\widehat{t},\widehat{z},u)] \\ &\quad + \frac{1}{2}\mathrm{tr}[G(\widehat{t},\widehat{z})G^\tau(\widehat{t},\widehat{z})M_1] - \frac{1}{2}\mathrm{tr}[G(\widehat{s},\widehat{w})G^\tau(\widehat{s},\widehat{w})M_2] - l(\widehat{z}) - l(\widehat{w}), \end{aligned}$$

which, together with (4.10) and (3.14)–(3.15), leads to

$$\begin{aligned} \text{RHS of (4.11)} &\le \sup_{u\in U}[\Lambda_w^\tau(t,z,s,w)\psi(\widehat{s},\widehat{w},u) + \Lambda_z^\tau(t,z,s,w)\psi(\widehat{t},\widehat{z},u)] \quad \left(\triangleq A_1\right) \\ &\quad + \frac{1}{2\varepsilon}\mathrm{tr}\{[G(\widehat{t},\widehat{z}) - G(\widehat{s},\widehat{w})]^\tau[G(\widehat{t},\widehat{z}) - G(\widehat{s},\widehat{w})]\} \quad \left(\triangleq A_2\right) \\ &\quad + \frac{\alpha(2\mu T - t - s)}{2\mu T} \\ &\quad \times \sum_{i,k=1}^n \frac{1}{2}\big[\sigma_{ik}^2(\widehat{t},\widehat{z})(\Gamma''(z_i) + 30z_i^4) \\ &\qquad\qquad + \sigma_{ik}^2(\widehat{s},\widehat{w})(\Gamma''(w_i) + 30w_i^4)\big] \quad \left(\triangleq A_3\right) \\ &\quad + [l(\widehat{z}) - l(\widehat{w})] \quad \left(\triangleq A_4\right) \\ &= A_1 + A_2 + A_3 + A_4, \end{aligned} \tag{4.13}$$

where $\Gamma(r) \triangleq e^{5\sqrt{r^2+1}}$, $\Gamma'' = \frac{d^2\Gamma}{dr^2}$ and $(\widehat{t}_0,\widehat{z}_0) \in B^\eta(t_0,z_0)$, $(\widehat{s}_0,\widehat{w}_0) \in B^\eta(s_0,w_0)$. Notice that the set $\mathcal{S}_{\eta,\varepsilon} = \{(t_0,z_0),(\widehat{t}_0,\widehat{z}_0),(s_0,w_0),(\widehat{s}_0,\widehat{w}_0)\}$ is contained in a compact

set Q_α^* determined by α. For $0 < \varepsilon \le 1$ appearing in $\Lambda(t,z,s,w)$, there exists $\eta_\varepsilon > 0$ such that, for all $0 < \eta \le \eta_\varepsilon$,

$$\text{RHS of (4.11)} \le A_1^0 + A_2^0 + A_3^0 + A_4^0 + \varepsilon, \tag{4.14}$$

where, again without writing the subscript 0 for (t_0, z_0, s_0, w_0), we define

$$\begin{aligned}
A_1^0 &= \sup_{u\in U}[\Lambda_w^\tau(t,z,s,w)\psi(s,w,u) + \Lambda_z^\tau(t,z,s,w)\psi(t,z,u)],\\
A_2^0 &= \frac{1}{2\varepsilon}\mathrm{tr}\{[G(t,z)-G(s,w)]^\tau[G(t,z)-G(s,w)]\},\\
A_3^0 &= \frac{\alpha(2\mu T - t - s)}{2\mu T}\sum_{i,k=1}^{n}\frac{1}{2}[\sigma_{ik}^2(t,z)(\Gamma''(z_i)+30z_i^4) + \sigma_{ik}^2(s,w)(\Gamma''(w_i)+30w_i^4)],\\
A_4^0 &= l(z) - l(w).
\end{aligned}$$

Since $\mathcal{S}_{\eta,\varepsilon}$ is contained in Q_α^* and the diameter of $\mathcal{S}_{\eta,\varepsilon}$ tends to 0 as $\eta, \varepsilon \to 0$, by taking an appropriate sequence $(\eta^{(k)}, \varepsilon^{(k)}, \lambda^{(k)}) \to 0$ satisfying $\eta^{(k)} \le \eta_{\varepsilon^{(k)}}$, we get convergent sequences $(t_0^{(k)}, z_0^{(k)})$, $(t_0^{(k)}, \widehat{z}_0^{(k)})$, $(s_0^{(k)}, w_0^{(k)})$, $(s_0^{(k)}, \widehat{w}_0^{(k)}) \to (\tilde{t}, \tilde{z})$ as $k \to \infty$. In the following we use the same C to denote different constants which are independent of α. Now we have the three relations

$$\limsup_{k\to\infty} \text{ LHS of (4.11)}(\eta^{(k)}, \varepsilon^{(k)}, \lambda^{(k)}) \ge \frac{2\alpha}{\mu T}\left[\sum_{i=1}^{n} e^{5\sqrt{\tilde{z}_i^2+1}} + \sum_{i=1}^{2n}|\tilde{z}_i|^6\right] + 2\beta, \tag{4.15}$$

$$\lim_{k\to\infty}(A_2^0 + A_4^0)\big(\eta^{(k)}, \varepsilon^{(k)}, \lambda^{(k)}\big) = 0, \tag{4.16}$$

$$\limsup_{k\to\infty} A_3^0\big(\eta^{(k)}, \varepsilon^{(k)}, \lambda^{(k)}\big) \le \frac{n\alpha C_\sigma(\mu T - \tilde{t})}{\mu T}\sum_{i=1}^{n}\left(25e^{5\sqrt{\tilde{z}_i^2+1}} + 30|\tilde{z}_i|^4\right), \tag{4.17}$$

where (4.15) follows from (4.12), and (4.16) follows from the continuity of $l(z)$, the Lipschitz continuity of $G(t,z)$ by assumption (H1), and (4.4). We proceed to analyze A_1^0:

$$\begin{aligned}
A_1^0 &\le \sup_{u\in U}\sum_{i=n+1}^{2n}[\Lambda_{z_i}(t,z,s,w)\psi_i(t,z,u) + \Lambda_{w_i}(t,z,s,w)\psi_i(s,w,u)]\\
&\quad + \sum_{i=1}^{n}[\Lambda_{z_i}(t,z,s,w)f_i(t,z) + \Lambda_{w_i}(t,z,s,w)f_i(s,w)] \stackrel{\triangle}{=} A_{11}^0 + A_{12}^0.
\end{aligned}$$

Then by (H1), (4.4), and recalling $|g_i(t,x,p,u)| \le C_g(1 + \sum_{k=1}^n |p_k|)$ for $t \in [0,T]$, $u \in U$, we have

$$\limsup_{k\to\infty} A_{11}^0\big(\eta^{(k)}, \varepsilon^{(k)}, \lambda^{(k)}\big) \le \frac{\alpha(\mu T - \tilde{t})}{\mu T}\sum_{i=n+1}^{2n} 12C_g\left[2n|\tilde{z}_i|^6 + |\tilde{z}_i|^5\right]. \tag{4.18}$$

Now we employ $a_i(t) \ge 0$ for $t \in [0,T]$ in the weak coupling condition (H2), and the Lipschitz property of $f_i(t,z) = a_i(t)z_i + f_i^0(t,z)$ by (H1) to obtain

$$
\begin{aligned}
A_{12}^0 &= \frac{\alpha(2\mu T - t - s)}{2\mu T} \\
&\quad \times \sum_{i=1}^n \left\{ \left[\frac{5z_i}{\sqrt{z_i^2+1}} e^{5\sqrt{z_i^2+1}} + 6z_i^5 + \frac{z_i - w_i}{\varepsilon} \right] [-a_i(t)z_i + f_i^0(t,z)] \right. \\
&\qquad \left. + \left[\frac{5w_i}{\sqrt{w_i^2+1}} e^{5\sqrt{w_i^2+1}} + 6w_i^5 + \frac{w_i - z_i}{\varepsilon} \right] [-a_i(s)w_i + f_i^0(s,w)] \right\} \\
&\le \frac{\alpha(2\mu T - t - s)}{2\mu T} \sum_{i=1}^n \left\{ \left[\frac{5z_i}{\sqrt{z_i^2+1}} e^{5\sqrt{z_i^2+1}} + 6z_i^5 \right] f_i^0(t,z) \right. \\
&\qquad \left. + \left[\frac{5w_i}{\sqrt{w_i^2+1}} e^{5\sqrt{w_i^2+1}} + 6w_i^5 \right] f_i^0(s,w) \right\} \\
&\quad + O\left(\frac{|t-s|^2}{\varepsilon} + \frac{|z-w|^2}{\varepsilon} \right).
\end{aligned} \tag{4.19}
$$

Hence, invoking (4.4), it follows that

$$
\limsup_{k\to\infty} A_{12}^0(\eta^{(k)}, \varepsilon^{(k)}, \lambda^{(k)}) \le \frac{\alpha C_{f^0}(\mu T - \tilde{t})}{\mu T} \sum_{i=1}^n \left[10 e^{5\sqrt{\tilde{z}_i^2+1}} + 12|\tilde{z}_i|^5 \right], \tag{4.20}
$$

which, together with (4.16)–(4.18), gives

$$
\begin{aligned}
&\limsup_{k\to\infty} \text{ RHS of (4.11) } (\eta^{(k)}, \varepsilon^{(k)}, \lambda^{(k)}) \\
&\quad \le \frac{[10C_{f^0} + 25n(C_\sigma + C_g)]\alpha(\mu T - \tilde{t})}{\mu T} \left[\sum_{i=1}^n e^{5\sqrt{\tilde{z}_i^2+1}} + \sum_{i=1}^{2n} |\tilde{z}_i|^6 + C \right] \\
&\quad \le \frac{\alpha}{2\mu T} \left[\sum_{i=1}^n e^{5\sqrt{\tilde{z}_i^2+1}} + \sum_{i=1}^{2n} |\tilde{z}_i|^6 + C \right],
\end{aligned} \tag{4.21}
$$

where C is independent of α. Hence it follows from (4.11), (4.15), and (4.21) that

$$
2\beta \le -\frac{3\alpha}{2\mu T} \left\{ \sum_{i=1}^n e^{5\sqrt{\tilde{z}_i^2+1}} + \sum_{i=1}^{2n} |\tilde{z}_i|^6 \right\} + \alpha C \le \alpha C. \tag{4.22}
$$

We recall from Step 1 that $\beta \le 1$ can take a strictly positive value from the interval $(0, \frac{c_0^+ - c_0}{4T})$ and $\alpha \in (0, \alpha_0)$. Letting $\alpha \to 0$ in (4.22) yields $\beta \le 0$, which is a contradiction to $\beta \in (0, \frac{c_0^+ - c_0}{4T})$, and this completes the proof.

Remark 2. By $\Phi(t_0, z_0, s_0, w_0) \ge \Phi(T, 0, T, 0)$ and $|\underline{v} - \overline{v}| = o([\sum_{i=1}^n (e^{5|z_i|} + e^{5|w_i|}) + \sum_{i=1}^{2n}(z_i^6 + w_i^6)])$, there exist $\delta_\alpha > 0$, $C > 0$ such that

$$
\begin{aligned}
&\frac{1}{2\varepsilon}|t_0 - s_0|^2 + \frac{1}{2\varepsilon}|z_0 - w_0|^2 + \frac{\lambda}{t_0 - T_1} + \frac{\lambda}{s_0 - T_1} \\
&\qquad + \delta_\alpha \left[\sum_{i=1}^n \left(e^{5\sqrt{1+z_{0,i}^2}} + e^{5\sqrt{1+w_{0,i}^2}} \right) + \sum_{i=1}^{2n} \left(z_{0,i}^6 + w_{0,i}^6 \right) \right] \le C.
\end{aligned}
$$

Then (4.3) follows readily.

Remark 3. By expanding $2\Phi(t_0, z_0, s_0, w_0) \geq \Phi(t_0, z_0, t_0, z_0) + \Phi(s_0, w_0, s_0, w_0)$ using all the individual terms, it can be shown that $\frac{1}{2\varepsilon}|t_0 - s_0|^2 + \frac{1}{2\varepsilon}|z_0 - w_0|^2$ is dominated by a continuous function $F(t_0, z_0, s_0, w_0)$, which goes to zero as $|t_0 - s_0| + |z_0 - w_0| \to 0$, which in turn follows from (4.3) when $\varepsilon \to 0$.

Remark 4. The proof of the theorem is based upon the methods in [12, 9, 10, 2]. Since here we deal with the function class $\mathcal{G}$ with a highly nonlinear growth condition on an unbounded domain, a localized semiconvex/semiconcave approximation technique is devised. The particular structure of the system dynamics also plays an important role in the proof of uniqueness, and in general it is more difficult to obtain uniqueness results under more general dynamics and the above fast growth condition. It is seen that the weak coupling feature of the dynamics of the state subprocess x is crucial for the above proof, and furthermore, when there exists an $a_i < 0$ (see assumption (H2)), the estimate (4.19) would not be valid.

5. Control with state constraints. In this section we consider the case when the state subprocess p is subject to constraints; i.e., the trajectory of each p_i must be maintained to be in a certain range. We term this situation as optimization under hard constraints. In [11] the author considered a deterministic model and obtained a constrained viscosity solution formulation for a first order HJB equation. Now due to the exogenous subprocess x, we come up with a second order HJB equation, and we will develop a similar formulation. Suppose that $u \in U$, where U is a compact convex set in $\mathbb{R}^n$, and that p satisfies $p_i \in [0, \overline{P}_i]$, where $\overline{P}_i$ is the upper bound. For simplicity we take $U = [-1, 1]^n$ and $\overline{P}_i = \infty$. For any fixed initial value $p_0 \geq 0$ (i.e., each $(p_0)_i \geq 0$), define the admissible control set

$$\begin{aligned}\mathcal{U}^{p_0} = \{u(\cdot) \mid u \text{ is adapted to } \sigma(z_s,\ s \leq t),\ u(t) \in U,\\ \text{and } P_\Omega(p_i(t) \geq 0 \ \text{ for all } 0 \leq t \leq T) = 1,\ 1 \leq i \leq n\}.\end{aligned}$$

In this section we consider the simple case of

$$g(t, x, p, u) = u.$$

Under the admissible control set $\mathcal{U}^{p_0}$, we will use the notation of section 2, for which the interpretation should be clear, and in the following we also use $\mathcal{U}^{p_0}$ with any initial time $s \leq T$. It is evident that $\mathcal{U}^{p_0}$ is a convex set. Under the norm $\|\cdot\|$ on $\mathcal{L}$ defined in section 2, $\mathcal{U}^{p_0}$ is also closed. Indeed, if $\|u^{(k)} - u\| \to 0$ as $k \to \infty$, where $u^{(k)} \in \mathcal{U}^{p_0}$, one can show that u will also generate positive p trajectories with probability 1 with initial value p_0. Thus $u \in \mathcal{U}^{p_0}$. As in the state unconstrained case, one can prove existence and uniqueness of the optimal control. Write

$$Q_T^0 = [0, T) \times \mathbb{R}^n \times (0, \infty)^n, \quad Q_T = [0, T) \times \mathbb{R}^n \times [0, \infty)^n,$$
$$\overline{Q}_T = [0, T] \times \mathbb{R}^n \times [0, \infty)^n.$$

We consider the HJB equation

$$\begin{aligned}0 &= -\frac{\partial v}{\partial t} + \sup_{u \in U}\left\{-\frac{\partial^\tau v}{\partial z}\psi\right\} - \frac{1}{2}\mathrm{tr}\left(\frac{\partial^2 v}{\partial z^2}GG^\tau\right) - l,\\ v|_{t=T} &= 0,\end{aligned} \tag{5.1}$$

where $(t, z) = (t, x, p) \in \overline{Q}_T$.

DEFINITION 5.1. *$v(t,z) \in C(\overline{Q}_T)$ is called a* constrained viscosity solution *to* (5.1) *if* (i) *$v|_{t=T} = 0$ and, for any $\varphi(t,z) \in C^{1,2}(\overline{Q}_T)$, whenever $v - \varphi$ takes a local maximum at $(t,z) \in Q_T^0$, we have*

$$-\frac{\partial \varphi}{\partial t} + \sup_{u \in U}\left\{-\frac{\partial^\tau \varphi}{\partial z}\psi\right\} - \frac{1}{2}\mathrm{tr}\left(\frac{\partial^2 \varphi}{\partial z^2} GG^\tau\right) - l \le 0, \quad z \in \mathbb{R}^{2n}, \tag{5.2}$$

at (t,z), and (ii) *for any $\varphi(t,z) \in C^{1,2}(\overline{Q}_T)$, whenever $v - \varphi$ takes a local minimum at $(t,z) \in Q_T$, in* (5.2) *we have an opposite inequality at (t,z). In short, we term the constrained viscosity solution $v(t,z) \in C(\overline{Q}_T)$ as a viscosity subsolution on Q_T^0 and a viscosity supersolution on Q_T.*

Remark 5. Conditions (i) and (ii) hold on Q_T^0 and Q_T, respectively. Here we give a heuristic interpretation on how the state constraints are captured by condition (ii). Suppose that $v - \varphi$ attains a minimum at $(\bar{t}, \overline{x}, \overline{p})$, where v is the value function and satisfies (5.1) at $(\bar{t}, \overline{x}, \overline{p})$ with classical derivatives, i.e.,

$$0 = -\frac{\partial v}{\partial t} + \left\{-\frac{\partial^\tau v}{\partial z}\psi\right\}\Big|_{u=\widehat{u}} - \frac{1}{2}\mathrm{tr}\left(\frac{\partial^2 v}{\partial z^2} GG^\tau\right) - l. \tag{5.3}$$

In addition, we assume that $\widehat{u}$ is admissible w.r.t. $(\overline{x}, \overline{p})$. Here $\bar{t} \in [0,T)$ and $\overline{p}$ lies on the boundary of $[0,\infty)^n$. By the necessary condition for a minimum, at $(\bar{t}, \overline{x}, \overline{p})$, we have

$$v_t - \varphi_t \ge 0, \quad v_{x_i} - \varphi_{x_i} = 0, \quad v_{x_i x_i} - \varphi_{x_i x_i} \ge 0, \quad 1 \le i \le n, \tag{5.4}$$

where the first inequality becomes equality when $\bar{t} \in (0,T)$. Since $\overline{p}$ is on the boundary of $[0,T)^n$, we can find an index set I such that $\overline{p}_i = 0$ when $i \in I$, and $\overline{p}_i > 0$ when $i \in \{1,\dots,n\}\backslash I$. Again, by the minimum property at $(\bar{t}, \overline{x}, \overline{p})$ we get

$$v_{p_i} - \varphi_{p_i} \ge 0 \quad \text{for } i \in I, \qquad v_{p_i} - \varphi_{p_i} = 0 \quad \text{for } i \in \{1,\dots,n\}\backslash I \tag{5.5}$$

at $(\bar{t}, \overline{x}, \overline{p})$. Since we assume that $\widehat{u}$ is admissible w.r.t. $(\overline{x}, \overline{p})$, then we have $\widehat{u}_i \ge 0$ for $i \in I$, and therefore by (5.5), at $(\bar{t}, \overline{x}, \overline{p})$

$$(v_p - \varphi_p)^\tau \widehat{u} \ge 0. \tag{5.6}$$

Now by (5.4) and (5.6) we see that

$$-\frac{\partial \varphi}{\partial t} + \left\{-\frac{\partial^\tau \varphi}{\partial z}\psi\right\}\Big|_{u=\widehat{u}} - \frac{1}{2}\mathrm{tr}\left(\frac{\partial^2 \varphi}{\partial z^2} GG^\tau\right) - l \ge 0,$$

and then condition (ii) holds at $(\bar{t}, \overline{x}, \overline{p})$.

As in section 2, we also define the set $\mathcal{U} = \{u(\cdot) | u \text{ is adapted to } \sigma(z_s,\ s \le t) \text{ and } u(t) \in U,\ t \le T\}$.

LEMMA 5.2. *For any initial pair (s_0, x_0, p_0) with each $(p_0)_i \ge 0$, and any $u \in \mathcal{U}$, there exists $\widetilde{u} \in \mathcal{U}^{p_0}$ such that*

$$P_\Omega\left\{\int_{s_0}^T |\widetilde{u} - u| ds \le 4\varepsilon\right\} = 1, \tag{5.7}$$

where with probability 1 *and for all* $1 \le i \le n$, *the constant* $\varepsilon > 0$ *satisfies*

$$\sup_{t\in[s_0,T]} \max\{-p_i(t, s_0, p_0, u), 0\} \le \varepsilon, \tag{5.8}$$

and $p(t, s_0, p_0, u)$ *denotes the value of* p *at* t *with initial condition* (s_0, p_0) *and control* u.

Proof. We need only to modify each component u_i of u in the following way. Define $\tau_i^0 = s_0$, and for $k \ge 1$,

$$\tau_i^k = \inf\{t > \tau_i^{k-1},\ p_i(t, s_0, p_0, \widetilde{u}) = 0\}, \tag{5.9}$$

$$\tau_i^k = T \quad \text{if } \ p_i(t, \tau_i^{k-1} + \varepsilon, p_i(\tau_i^{k-1} + \varepsilon), u) > 0 \quad \forall t \ge \tau_i^{k-1} + \varepsilon, \tag{5.10}$$

$$\widetilde{u}_i(t) = 1 \quad \text{on } [\tau_i^{k-1},\ \tau_i^{k-1} + \varepsilon), \tag{5.11}$$

$$\widetilde{u}_i(t) = u_i(t) \quad \text{on } [\tau_i^{k-1} + \varepsilon, \tau_i^k). \tag{5.12}$$

Then it is obvious that $\widetilde{u} \in \mathcal{U}^{p_0}$. Suppose that (5.7) is not true, and then there exist i and a set A^0 with $P_\Omega(A^0) > 0$ such that on A^0

$$\int_{s_0}^{T} |\widetilde{u}_i - u_i| ds > 4\varepsilon. \tag{5.13}$$

For any fixed $\omega \in A^0$, if $\tau_i^{k_0}$ is the last stopping time defined by (5.9) and (5.10), then by (5.13) we can easily show that $p_i(\tau_i^{k_0-1}, s_0, p_0, u) < -2\varepsilon$, which is a contradiction to (5.8). □

With Lemma 5.2, we can further show that the value function $v(t, z)$ is continuous on $\overline{Q}_T$ by a comparison method, as in the unconstrained case [3]. The details are omitted here. The growth condition of Proposition 2.2 also holds in the state constrained case.

PROPOSITION 5.3. *The value function* v *is a constrained viscosity solution to the HJB equation* (5.1).

Proof. We verify condition (i) first. For an initial condition pair (s, z) with $z \in Q_T^0$ and any $u \in U$ we construct control $\widetilde{u} = u$ on $[s, s+\varepsilon]$ and $\widetilde{u} = 0$ on $(s+\varepsilon, T]$. We see that when ε is sufficiently small, $\widetilde{u}$ is in the admissible control set w.r.t. (s, z) since each $p_i \in [0, \infty)$. All the remaining steps and the verification of condition (ii) can be done as in Theorem 2.4. □

REFERENCES

[1] C. D. CHARALAMBOUS AND N. MENEMENLIS, *Stochastic models for long-term multipath fading channels*, in Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, AZ, 1999, IEEE Press, Piscataway, NJ, pp. 4947–4952.

[2] M. G. CRANDALL, H. ISHII, AND P. L. LIONS, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27 (1992), pp. 1–67.

[3] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[4] W. H. FLEMING AND H. M. SONER, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[5] M. GUDMUNDSON, *Correlation model for shadow fading in mobile radio systems*, Electron. Lett., 27 (1991), pp. 2145–2146.

[6] M. HUANG, P. E. CAINES, C. D. CHARALAMBOUS, AND R. P. MALHAMÉ, *Power control in wireless systems: A stochastic control formulation*, in Proceedings of the American Control Conference, Arlington, VA, 2001, IEEE Press, Piscataway, NJ, pp. 750–755.

[7] M. Huang, P. E. Caines, C. D. Charalambous, and R. P. Malhamé, *Stochastic power control for wireless systems: Classical and viscosity solutions*, in Proceedings of the 40th IEEE Conference on Decision and Control, Orlando, FL, 2001, IEEE Press, Piscataway, NJ, pp. 1037–1042.

[8] M. Huang, P. E. Caines, and R. P. Malhamé, *Uplink power adjustment in wireless communication systems: A stochastic control analysis*, IEEE Trans. Automat. Control, 49 (2004), pp. 1693–1708.

[9] H. Ishii, *On uniqueness and existence of viscosity solutions of fully nonlinear second-order elliptic PDEs*, Comm. Pure Appl. Math., 42 (1989), pp. 15–45.

[10] R. Jensen, P. L. Lions, and P. E. Souganidis, *A uniqueness result for viscosity solutions of second order fully nonlinear partial differential equations*, Proc. Amer. Math. Soc., 102 (1988), pp. 975–978.

[11] H. M. Soner, *Optimal control with state-space constraint* I, SIAM J. Control Optim., 24 (1986), pp. 552–561.

[12] J. Yong and X. Y. Zhou, *Stochastic Controls: Hamiltonian Systems and HJB Equations*, Springer, New York, 1999.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 2, pp. 389–417

A BOUNDED VARIATION CONTROL PROBLEM FOR DIFFUSION PROCESSES*

ANANDA WEERASINGHE†

Abstract. We consider an infinite horizon discounted cost minimization problem for a one-dimensional stochastic differential equation model. The available control is an added bounded variation process. The cost structure involves a running cost function and a proportional cost for the use of the control process. The running cost function is not necessarily convex. We obtain sufficient conditions to guarantee the optimality of the zero control. Also, for unbounded cost functions, we provide sufficient conditions which make our optimal state process a reflecting diffusion on a compact interval. In both cases, the value function is a C^2 function. For bounded cost functions, under additional assumptions, we obtain a complex optimal strategy which turned out to be a mixture of jumps and local-time-type processes. In this case, we show that the value function is only a C^1 function and that it fails to be a C^2 function. We also discuss a related variance control problem.

Key words. stochastic optimal control, optimal stopping, local-time processes, diffusions with reflections

AMS subject classifications. 49A60, 93E20, 60H30

DOI. 10.1137/S0363012903436119

1. Introduction.

Consider a weak solution to the one-dimensional stochastic differential equation

$$X_x(t) = x + \int_0^t \mu(X_x(s-))\,ds + \int_0^t \sigma(X_x(s-))\,dW(s) + A(t), \tag{1.1}$$

where $x \in \mathbb{R}$, $\{W(t) : t \geq 0\}$ is a standard one-dimensional Brownian motion adapted to a right-continuous filtration $\{\mathfrak{F}_t : t \geq 0\}$ on a probability space $(\Omega, \mathfrak{F}, P)$. The σ-algebra $\mathfrak{F}_0$ contains all the P-null sets in $\mathfrak{F}$, and the Brownian increments $W(t+s) - W(t)$ are independent of $\mathfrak{F}_t$ for all t and $s \geq 0$. The control process $A(\cdot)$ is $\{\mathfrak{F}_t\}$-adapted, right continuous with left limits, and of bounded variation on finite intervals. Also $A(0) = 0$. Let $|A|(\cdot)$ be the total variation process of $A(\cdot)$, where $|A|(t)$ is the total variation of $A(\cdot)$ on $[0,t]$. We further assume that for each $X_x(\cdot)$, there is an increasing sequence of stopping times (τ_n) with respect to $\{\mathfrak{F}_t\}$ such that $\lim_{n\to\infty} \tau_n = +\infty$ and

$$\text{(i)} \qquad E_x \int_0^{T\wedge\tau_n} [|\mu(X_x(s-))| + \sigma^2(X_x(s-))]\,ds < \infty \quad \text{for each } T > 0 \text{ and} \tag{1.2}$$

$$\text{(ii)} \qquad \lim_{n\to\infty} E_x \left[|X_x(\tau_n)| e^{-\alpha\tau_n} I_{[\tau_n<\infty]}\right] = 0,$$

where $\alpha > 0$ is a constant which represents the discount rate in (1.3).

The first condition above helps us make sense of (1.1) and the second condition will be used in the proof of the verification lemma in section 2. Similar conditions are

*Received by the editors October 10, 2003; accepted for publication (in revised form) March 14, 2005; published electronically August 31, 2005.
http://www.siam.org/journals/sicon/44-2/43611.html
†Department of Mathematics, 400 Carver Hall, Iowa State University, Ames, IA 50011 (ananda@iastate.edu).

assumed in the verification theorems of [11, Theorem 4.1, p. 322], [21], and [35], all of which address similar singular control problems. If $E_x \int_0^T [|\mu(X_x(s-))| + \sigma^2(X_x(s-))]\, ds < \infty$ for each $T > 0$, then the choice $\tau_n \equiv +\infty$ a.s. for all n, satisfies (1.2).

The cost functional associated with (1.1) is given by

$$J(x, A) = E_x \int_0^\infty e^{-\alpha t}[C(X_x(t))\, dt + d|A|(t)], \tag{1.3}$$

where $C(\cdot)$ is a nonnegative running cost function with the additional properties described below, and $\alpha > 0$ is a constant which represents the discount factor.

We intend to minimize $J(x, A)$ over all available state processes satisfying (1.1), (1.2), and to find an optimal control $A(\cdot)$ to achieve the minimum value. We make the following assumptions on the functions $\mu(\cdot), \sigma(\cdot)$, and $C(\cdot)$ throughout this paper. Here μ', σ', and C' denote the first derivatives of μ, σ, and C, respectively.

(1.4) (i) The functions μ and σ are continuously differentiable on R,
$$\inf_R(\alpha - \mu'(y)) > 0, \inf_R \sigma(y) > 0, \text{ and } x\mu(x) < 0 \text{ for all } x \neq 0.$$

(1.5)

(ii)
$$\int_{-\infty}^0 \frac{\mu(x) - x}{\sigma^2(x)}\, dx = \int_0^\infty \frac{x - \mu(x)}{\sigma^2(x)}\, dx = +\infty.$$

(1.6) (iii) The cost function C is continuously differentiable on R, decreasing on $(-\infty, 0)$, increasing on $(0, +\infty)$, and it satisfies either $\liminf_{|x|\to\infty} \dfrac{C(x)}{|x|} > 0$ or $\limsup_{|x|\to\infty} \dfrac{C(x)}{|x|} < \infty$.

If $C(0) = r \neq 0$, we could introduce a new running cost function $\tilde{C}(x) = C(x) - r$ and define the functional $\tilde{J}(x, A)$ as similar to (1.3). Then $J(x, A) = \tilde{J}(x, A) + \frac{r}{\alpha}$ by (1.3), and hence the value functions V and $\tilde{V}$ defined as in (1.8) below are also related by $V(x) = \tilde{V}(x) + \frac{r}{\alpha}$. Therefore, without loss of generality, we assume $C(0) = 0$.

The assumption $x\mu(x) < 0$ for all $x \neq 0$ in (1.4) guarantees that the ordinary differential equation $\dot{x} = \mu(x)$ has a unique stable equilibrium point at the origin. This corresponds to (1.1) with the processes $A(\cdot)$ and $\sigma(\cdot)$ being identically zero. Hence, $X_x(\cdot)$ in (1.1) can be considered as a random perturbation of a stable dynamical system. We will provide two motivating examples from finance and operations research at the end of this section. In each example, the corresponding state process is a stochastic perturbation of a stable dynamical system.

Note that the diffusion coefficient $\sigma(\cdot)$ is allowed to be unbounded subject to the assumptions (1.4) and (1.5). Under assumptions (1.4)–(1.6), our main objectives here are to obtain sufficient conditions in terms of the functions μ, σ, and C for the following.

(i) Optimality of the zero control and the C^2 regularity of the corresponding value function.

(ii) Optimality of a reflected diffusion process in a compact interval and the C^2 regularity of the value function.

(iii) To illustrate the lack of C^2 regularity for the value function for a class of bounded cost functions regardless of the smoothness of μ, σ, and C, and to obtain an optimal policy.

(iv) To generalize the results in (i) and (ii) for a related variance control problem.

In cases (i), (ii), and (iv), the value function is a C^2 solution to the Hamilton–Jacobi–Bellman (HJB) equation (1.9), while in case (iii), it fails to be a C^2 function but it is a C^1 solution to the HJB equation (1.9). In (iv), we consider the problem of controlling the bounded variation process as well as the diffusion coefficient of (1.1). Our assumptions in (iv) guarantee the convexity of the value function, and hence it is optimal to choose the minimum diffusion coefficient. We generalize the results of [14], [21], and [35] to nonsymmetric diffusions and to a large class of nonsymmetric cost functions without the convexity assumption. In addition, we allow the cost functions to be of slow growth such as $C(x) \sim |x|^\alpha$ with $0 < \alpha < 1$, $C(x) \sim \log|x|$ for large $|x|$ as well as bounded cost functions. Here we employ the connection between stochastic control and optimal stopping, which is a known theme in optimal control theory. We refer to [3], [4], [5], [7], [15], [16], [27], and [35] for this approach. The articles [5], [15], and [16] consider the case of Brownian motion with an added bounded variation control process. The work of [5] considers the case of a diffusion process and their drift and diffusion coefficients are allowed to be time dependent, but the drift term is linear and the diffusion is independent of the space variable. The running cost function is a symmetric convex function in [5], [13], [15], and [16]. A control problem with a quite general cost function with a different cost structure was considered in [3].

For a given x in $\mathbb{R}$, we call $((\Omega, \mathfrak{F}, P), (\mathfrak{F}_t), W(\cdot), X_x(\cdot), A(\cdot))$ an admissible control system if (i) $X_x(\cdot)$ is a weak solution to (1.1), and (ii) $X_x(\cdot)$ satisfies (1.2) and $J(x, A) < +\infty$, where J is given in (1.3). To define the value function for the control problem, we first introduce

$$\mathcal{A}(x) = \{A(\cdot) : \ A(\cdot) \text{ is a bounded variation process in (1.1) with a corresponding admissible state process } X_x(\cdot)\}. \tag{1.7}$$

For each x in $\mathbb{R}$, using a reflecting diffusion on an interval containing the point x, one could obtain a control process $A(\cdot)$ so that the corresponding $J(x, A)$ is finite. Hence $\mathcal{A}(x)$ is nonempty and the value function is finite for all x. The value function is defined by

$$V(x) = \inf_{A \in \mathcal{A}(x)} J(x, A). \tag{1.8}$$

The formal HJB equation for the value function is given by

$$\min\left\{\frac{1}{2}\sigma^2(x)V''(x) + \mu(x)V'(x) - \alpha V(x) + C(x),\ 1 - |V'(x)|\right\} = 0 \text{ a.s. on } \mathbb{R}. \tag{1.9}$$

We show that under our assumptions, this value function is the minimal solution to (1.9).

The paper is organized as follows. In section 2, we establish several verification results. They will be used to sort out our optimal strategies. Some auxiliary results will be proved in section 3. In section 4, we show that when $\alpha - \mu'(x) \geq |C'(x)|$ for all x, then $A(\cdot)$ being identically zero is an optimal strategy and that the value function is a C^2 solution to (1.9). In section 5, when $(1+\varepsilon)(\alpha - \mu'(x)) < |C'(x)|$ for large $|x|$, for some $\varepsilon > 0$, we show that the value function is C^2 and the optimal state process

is a reflecting diffusion on a finite interval which contains the origin. Our approach in both sections 4 and 5 is to first observe that the value function can be completely determined by its first derivative and then to obtain this derivative function. In sections 4 and 5, we generalize the results of [35] to a large class of cost functions and to general diffusions, while section 6 is completely new. In section 5, we develop the idea of using a family of optimal stopping problems to obtain the derivative of the value function. In section 6, we consider the situation of bounded cost functions. If $\alpha - \mu'(x) \geq |C'(x)|$, then the results in section 4 remain valid and the zero control is optimal. Under additional symmetry assumptions, we obtain a complex optimal strategy which involves a possible jump at a stopping time followed by a local-time-type control when $|C'(x)|$ is larger than $\alpha - \mu'(x)$ in a large compact interval. In this case, the value function is a C^1 solution to (1.9) and it is C^2 everywhere except at two points. When μ is identically zero and σ is a constant, an example of a similar optimal process is given in Example 4.3, Chapter VIII of [11]. In section 7, we control the diffusion coefficient in addition to the bounded variation process. We observe that the monotonicity of the function $h(x) = \frac{C'(x)}{\alpha - \mu'(x)}$ implies the convexity of the value function related to the sections 4 and 5. Hence the choice of the minimal diffusion coefficient is optimal.

We intend to consider the ergodic control problem and Abelian limit relations between the value functions in a future work; see [34] for a related article. The articles [9] and [31] discuss issues related to the regularity of the value function in higher dimensions. The existence of an optimal Markovian control for related stochastic control problems was established in [20]. Next we describe two motivating examples from finance and operations research.

Example 1 (foreign exchange rates). We consider the currency exchange rate that governs the transactions between two countries. We assume that the economies of the two countries are stable and therefore the exchange rate fluctuates around a stable equilibrium point. In the presence of uncertainty, it is a common practice to model the currency exchange rate using stochastic differential equations (see Chapter 7 of [26] and also [8], [12], [18], and [25]). We consider the problem of a central bank in one of these countries which would like to keep the exchange rate as close as possible to a given target level through minimal intervention.

In several models (see [12], [18]), the exchange rate is assumed to take values in an exogenously given target interval which is called the "target band." To keep the exchange rates within the target band, the central bank may intervene while the exchange rate is still inside the band and there is empirical evidence to support this fact [6]. In [13], impulse control methods were used to find an optimal target zone. An exchange rate model with constant σ and discrete intervention times is considered in [25]. They derive an optimal target band in a specific example with a symmetric running cost function $C(\cdot)$. Also we refer the reader to [24] and [32] for related work. It is typically the responsibility of the central bank to have a monetary policy to guarantee the exchange rate to take values in the target band. The central bank may intervene in the foreign exchange market by adjusting the money supply, controlling the flow of foreign capitals via local interest rate adjustments, and by buying and selling large amounts of the foreign currency. These procedures imply a cost for the central bank and this cost increases with the level of intervention.

In our model, there is no exogenous target band for the exchange rate, but we consider a target value or a benchmark. There will be two types of costs involved: there is a running cost associated with the deviation of the exchange rate from the

target value and a cost for the intervention of the central bank which is proportional to the change in the interest rate. Our exchange rate model is described by (1.1). Here $X_x(t)$ represents the value of one unit of foreign currency in domestic currency units at time $t \geq 0$ and $X_x(0) = x$. The bounded variation control process $A(\cdot)$ in (1.1) represents the changes in the exchange rate due to the central bank interventions. We assume that the underlying cost of these interventions is proportional to the changes in the exchange rate and is represented by the total variation process. Without loss of generality, we consider the target level for the exchange rate to be at the origin, since if the target level were at $\rho > 0$ we could consider $X_x(t) - \rho$, which satisfies an equation similar to (1.1). The running cost function $C(\cdot)$ which satisfies (1.6) represents the cost due to the deviation of the exchange rate from the target. Notice that C has its absolute minimum at the target point. If there is no uncertainty, then the exchange rate should have a stable equilibrium point at the target level since we assume that μ and σ satisfy (1.4) and (1.5).

Our objective here is to find an optimal intervention policy $A(\cdot)$ of the central bank which minimizes the infinite horizon total cost $J(x, A)$ of (1.3), where $\alpha > 0$ is a constant discount factor. We would also like to derive conditions on μ, σ, and C which imply the existence of an optimal target band $[a^*, b^*]$ and also to obtain conditions which guarantee the nonexistence of an optimal target band. In [8], the authors address a similar problem for the geometric Brownian motion with $C(x) = x^2$ but allow the cost due to each intervention to have a fixed cost and a proportional cost. They also obtain an explicit expression for the value function. Our results are of more qualitative nature. In section 4, under assumptions (1.4)–(1.6) and (4.2), we guarantee the nonexistence of an optimal target band and show that it is optimal for the central bank not to intervene at all to minimize the cost functional $J(x, A)$. Our results in section 5 prove the existence of an optimal target band $[a^*, b^*]$ under assumptions (1.4)–(1.6) and (5.1)–(5.2). In Theorem 5.4, we derive an optimal intervention policy.

Example 2 (capacity change under uncertainty). Capacity change is the process of adding new facilities or deleting available facilities over time to satisfy a demand. Let $d(t)$ be the demand at time t and $e(t)$ be the available capacity at time t. We let $d(\cdot)$ and $e(\cdot)$ be random processes. At time t, if $e(t) < d(t)$, then a shortage cost will be paid. If $e(t) > d(t)$, then an overcapacity cost will be paid. Let $X_x(t) = d(t) - e(t)$ for $t \geq 0$ and $X_x(0) = x$. $X_x(\cdot)$ represents the "undercapacity" process and notice that $X_x(t)$ may take negative values. We introduce the penalty function C by letting $C(y)$ be the penalty paid per unit time if $X_x(t) = y$. If $|X_x(t)|$ is large, we expect that the corresponding penalty is also large and hence C satisfies the assumptions in (1.6) and $C(0) = 0$. Since the "overcapacity" and "undercapacity" penalty rates could be different, the function $C(\cdot)$ need not be symmetric. When a project is undertaken, the decision maker is provided with an initial fund and also an emergency fund. Initial funds are used by day-to-day control of the project to change the capacity $e(\cdot)$ continuously to match the demand process $d(\cdot)$, by using the future forecasts of the demand. Therefore, it is assumed that these changes are made so that the capacity process $e(\cdot)$ remain absolutely continuous with respect to t. However, uncertain future demands, significant forecasting errors or any other defects may result in critical shortages or unwanted gross surpluses of the capacity. Therefore, the $X_x(\cdot)$ process may deviate significantly from the origin. To correct these situations, the controller may use the emergency funds to provide or delete the additional capacity. These corrections are costly and the corresponding changes in capacity are represented by a bounded variation process $A(\cdot)$ which may allow jumps or nonabsolutely continuous changes with respect to t. It is assumed that the cost of

change of capacity is proportional to the amount changed. Thus, the total variation $|A|(t)$ represents the costs due to these emergency capacity changes during $[0,t]$. The decision maker would like to choose the $A(\cdot)$ process which minimizes the infinite horizon discounted cost $J(x,A) \equiv E_x \int_0^\infty e^{-\alpha t}[C(X_x(t))\,dt + d|A|(t)]$, where $\alpha > 0$ is a constant discount factor.

The condition $x\mu(x) < 0$ for $x \neq 0$ in (1.4) implies the stability of the system when there is no uncertainty or when perfect forecasting is available. With the above model, when μ, σ, and C are known to the decision maker, there are several issues of interest. The results in this paper will address the following issues.

1. Under what conditions should the controller rule out using emergency funds at all? We will provide an answer in section 4.
2. In general, what is the qualitative nature of the optimal strategies? Sections 4, 5, and 6 yield answers to this question.
3. When large losses are covered by an insurance, the cost function $C(\cdot)$ can be considered a bounded function satisfying (1.6) and thus $\lim_{|x|\to\infty} C(x)$ is finite. What are the available optimal strategies? We answer this question in section 6.

In a celebrated paper [22], Manne specified the demand as a Wiener process with a drift and used the regenerative behavior of $X_x(\cdot)$ to conclude that the optimal capacity can be found by considering an equivalent deterministic problem with a modified discount rate. In [10], Davis et al. considered the demand as a Poisson process with a similar infinite horizon cost functional related to a piecewise-deterministic Markov process. They developed algorithms to approximate the optimal strategy. In conclusion, they point out the importance as well as the difficulty of developing results of qualitative nature for optimal policies. Our results address this need when $X_x(\cdot)$ is modeled by (1.1). In [30], Ryan employs the geometric Brownian motion for the demand process and uses option pricing formulas to derive infinite horizon discounted cost. For a survey article on capacity expansion review, a list of references, and for new directions, we refer the reader to [33].

2. A verification lemma. The results in this section will enable us to obtain a lower bound for the value function. It is closely related to Theorem 2.1 of [21], Lemma 3.1 of [35], and also to the general verification theorems in Chapter 8 of [11]. We first establish the verification lemma under the assumption $\liminf_{|x|\to\infty} \frac{C(x)}{|x|} > 0$ and then we generalize it to the case $\limsup_{|x|\to\infty} \frac{C(x)}{|x|} < \infty$. We relabel these assumptions in (1.6). Let us assume

$$\text{either (i) } \liminf_{|x|\to\infty} \frac{C(x)}{|x|} > 0 \tag{2.1}$$

$$\text{or (ii) } \limsup_{|x|\to\infty} \frac{C(x)}{|x|} < \infty. \tag{2.2}$$

LEMMA 2.1. *Assume* (2.1). *Let $Q(x)$ be a twice continuously differentiable function satisfying the HJB equation* (1.9). *Then $V(x) \geq Q(x)$ for all x, where $V(x)$ is the value function given in* (1.8).

Proof. The proof is essentially the same as the proof of Lemma 3.1(a) of [35] with the following observations. First, by (2.1), there exist two constants C_0 and C_1 so that $C_1 > 0$ and $C(x) \geq C_0 + C_1|x|$ for all x. Second, we can use the sequence (τ_n) given in (1.2) in the proof of Lemma 3.1 in [35] and follow the proof there. ∎

Remark. The above lemma remains valid if Q is a C^2 function which satisfies

(2.3)
$$\operatorname{Min}\left\{\frac{1}{2}\sigma^2(x)Q''(x)+\mu(x)Q'(x)-\alpha Q(x)+C(x),1-|Q'(x)|\right\}\geq 0 \quad \text{for all } x.$$

COROLLARY 2.2. *Assume* (2.2). *Let $Q(x)$ be a twice continuously differentiable function which satisfies the HJB equation* (1.9). *Then $V(x)\geq Q(x)$ for all x, where V is the value function given in* (1.8).

Proof. Let $C(x)$ satisfy (2.2). Consider any process $X_x(\cdot)$ which satisfies (1.1) and (1.2) and has corresponding $J(x,A)<\infty$. Then $E_x\int_0^\infty e^{-\alpha t}\,d|A|(t)$ is finite. By (2.2) we have $C(x)<K_0+K_1|x|$ for all x, for some constants K_0 and $K_1>0$. We approximate $C(x)$ by a sequence of C^2 functions $\{C_N\}$ such that

(i) $C_N(0)=0$, C_N is increasing on $[0,\infty)$, and decreasing on $(-\infty,0)$;

(ii) $C_0(x)\geq C_N(x)\geq C(x)$ and $\lim_{N\to\infty}C_N(x)=C(x)$ for all x;

(iii) $C_N(x)\leq K_0+K_1|x|$ for all x and $C_N(x)=K_0+K_1|x|$ for $|x|\geq N+1$.

Next, we estimate $E_x[C_N(X_x(T))]$ by applying Itô's lemma to $C_N(X_x(T\wedge\tau_k))$, where (τ_k) is as in (1.2), and using the following facts:

(a) $\sup_R(|C_N'(x)|+|C_N''(x)|)<D_N$, where D_N is a positive constant;

(b) $\mu(x)C_N'(x)\leq 0$ for all x and $C_N''(x)=0$ when $|x|\geq N+1$; and

(c) $E_x\int_0^{T\wedge\tau_k}|C_N''(X_x(t))|\sigma^2(X_x(t))\,dt\leq D_N\,(\max_{|x|\leq N+1}\sigma^2(x))T$.

Thus, we obtain $E_x[C_N(X_x(T\wedge\tau_k)]\leq C_N(x)+B_N(T+E|A|(T))$, where B_N is a constant which depends only on N. Together with $E_x\int_0^\infty e^{-\alpha t}\,d|A|(t)<\infty$, we conclude

$$\begin{aligned}E_x\int_0^\infty e^{-\alpha t}C(X_x(t))\,dt &\leq E_x\left[\int_0^\infty e^{-\alpha t}C_N(X_x(t))\,dt\right]\\ &\leq E_x\left[\int_0^\infty e^{-\alpha t}C_0(X_x(t))\,dt\right]<+\infty.\end{aligned}$$

But $\lim_{N\to\infty}C_N(x)=C(x)$ for all x, and consequently $\lim_{N\to\infty}E_x\int_0^\infty e^{-\alpha t}C_N(X_x(t))\,dt=E_x\int_0^\infty e^{-\alpha t}C(X_x(t))\,dt$. We let $J_N(x,A)\equiv E_x\int_0^\infty e^{-\alpha t}(C_N(X_x(t))dt+d|A|(t))$. Then each $J_N(x,A)$ is finite and $\lim_{N\to\infty}J_N(x,A)=J(x,A)$. Since $C_N(x)\geq C(x)$, it follows that Q also satisfies $\operatorname{Min}\{\frac{1}{2}\sigma^2(x)Q''(x)+\mu(x)Q'(x)-\alpha Q(x)+C_N(x),1-|Q'(x)|\}\geq 0$ for all x. Thus, by Lemma 2.1 and the remark in (2.3), we obtain $J_N(x,A)\geq Q(x)$. Now letting N tend to infinity, we obtain $J(x,A)\geq Q(x)$ for all x. Hence the result. □

The following result will be used in section 6.

PROPOSITION 2.3. *Assume* (2.1) *or* (2.2). *Let Q be a bounded continuously differentiable function on R which is also twice continuously differentiable everywhere except on a finite set $\{b_1,b_2,\dots,b_N\}$. Also, assume that the limits, $\lim_{x\to b_j^-}Q''(x),\lim_{x\to b_j^+}Q''(x)$ exist and are finite for each $j=1,2,\dots,N$, and that Q satisfies the HJB equation* (1.9) *everywhere except on the set $\{b_1,b_2,\dots,b_N\}$. Then $V(x)\geq Q(x)$ for all x, where V is the value function in* (1.8).

Proof. Let $X_x(\cdot)$ be any admissible process which satisfies (1.1) and (1.2) and has corresponding $J(x,A)$ in (1.3) finite. Hence $E_x\int_0^\infty e^{-\alpha t}\,d|A|(t)$ and $E_x[|A|(T)]$ for any $T>0$ are finite. Next, consider the open set $G_n=\bigcup_{i=1}^N\left(b_i-\frac{1}{n},b_i+\frac{1}{n}\right)$. We can approximate Q by a sequence of C^2 functions $\{Q_n\}$ so that $Q(x)\equiv Q_n(x)$ for all x in $\mathbb{R}\setminus G_n$, $\lim_{n\to\infty}Q_n''(x)=Q''(x)$ for all $x\neq b_1,b_2,\dots,b_N$, and $|Q_n''(x)|\leq M_1$

for all x in G_1, where $M_1 > 0$ is a constant independent of n. By (1.9), we have $|Q'(x)| \le 1$ for all x. Thus $|Q_n'(x)| \le 1 + \rho_n$, where ρ_n is a positive constant and $\lim_{n\to\infty} \rho_n = 0$. Therefore, since Q is bounded, we can find a constant $M_2 > 0$ independent of n, so that $|Q_n(x)| < M_2$ for all n. Next, observe that $\frac{\sigma^2(x)}{2} Q_n''(x) + \mu(x)Q_n'(x) - \alpha Q_n(x) + C(x) \ge 0$ for all x in $\mathbb{R} \setminus G_n$, since $Q_n \equiv Q$ on $\mathbb{R} \setminus G_n$. On the other hand, $|\frac{\sigma^2(x)}{2} Q_n''(x) + \mu(x)Q_n'(x) - \alpha Q_n(x)| < M_3$ for all x in G_n, where $M_3 > 0$ is a constant independent of n. Using these estimates, Itô's lemma and following the computation in the proof of Lemma 3.1 in [35], we obtain

$$E_x\big[e^{-\alpha(T\wedge\tau_k)}|Q_n(X_x(T\wedge\tau_k))|\big] + J(x,A) \ge Q(x) + \epsilon_n(x,T,\tau_k), \tag{2.4}$$

where $\epsilon_n(x,T,\tau_k) = -\rho_n E_x \int_0^{T\wedge\tau_k} e^{-\alpha s}\, d|A|(s) - M_3 E_x \int_0^{T\wedge\tau_k} e^{-\alpha s} I_{G_n}(X_x(s-))\, ds$.

Therefore, by (2.4), we have

$$\begin{aligned} M_2 E_x\big(e^{-\alpha(T\wedge\tau_k)}\big) + J(x,A) &\ge Q(x) - \rho_n E_x \int_0^T e^{-\alpha s}\, d|A|(s) \\ &\quad - M_3 E_x \int_0^T e^{-\alpha s} I_{G_n}(X_x(s-))\, ds. \end{aligned} \tag{2.5}$$

Since $\int_0^t \sigma(X_x(s-)\, dW(s)$ is a continuous local Martingale and $\inf_{\mathbb{R}} \sigma(x) > 0$, it follows that $E_x \int_0^T I_{\{b_i i=1,\dots,N\}}(X_x(s-))\, ds = 0$ (see p. 225, Example 7.10 of [17]). Hence $\lim_{n\to\infty} E_x \int_0^T e^{-\alpha s} I_{G_n}(X_x(s-))\, ds = 0$. Now letting n tend to infinity in (2.5), we obtain $M_2 E_x(e^{-\alpha(T\wedge\tau_k)}) + J(x,A) \ge Q(x)$. Next, we let (τ_k) tend to infinity and then T tend to infinity to obtain $J(x,A) \ge Q(x)$ for all x. This completes the proof. □

3. An auxiliary result. Let $\sigma(\cdot)$ and $\mu(\cdot)$ satisfy assumptions (1.4) and (1.5), and $\sigma'(\cdot)$ be the derivative of $\sigma(\cdot)$. We consider a weak solution of

$$\begin{aligned} dY(t) &= [\sigma(Y(t))\sigma'(Y(t)) + \mu(Y(t))]\, dt + \sigma(Y(t))\, dW(t), \\ Y(0) &= x, \end{aligned} \tag{3.1}$$

where $\{W(t) : t \ge 0\}$ is a Brownian motion on some probability space $(\Omega, \mathfrak{F}, P)$ equipped with a Brownian filtration $\{\mathfrak{F}_t\}$. Let τ_∞ be the explosion time for $\{Y(t) : t \ge 0\}$. If $\sigma'(x)$ is a bounded function, then $\tau_\infty \equiv +\infty$ a.s. by Khasminski's criteria for nonexplosion [29, p. 297], and the following proposition is obvious. But we need it for the general case.

PROPOSITION 3.1. *Assume* (1.4) *and* (1.5). *Let* $\{Y(t) : t \ge 0\}$ *be a weak solution of* (3.1). *Then*

$$\int_0^{\tau_\infty} (\alpha - \mu'(Y(s)))\, ds = +\infty \text{ a.s. in } P. \tag{3.2}$$

To prove (3.2) we need the following technical lemma.

LEMMA 3.2. *Let* u *be the solution to the differential equation*

$$\frac{\sigma^2(x)}{2} u''(x) + (\sigma(x)\sigma'(x) + \mu(x))u'(x) - (\alpha - \mu'(x))u(x) = 0 \tag{3.3}$$

with the boundary conditions $u(0) = 1$ *and* $u'(0) = 0$. *Then* $u(x) \ge 1$ *for all* x, u *is increasing on* $(0, +\infty)$, *decreasing on* $(-\infty, 0)$, *and* $\lim_{x\to-\infty} u(x) = \lim_{x\to+\infty} u(x) = +\infty$.

Proof. This proof is elementary and we sketch the basic steps. First, observe that $u''(0) > 0$. Thus u has a strictly convex local minimum at the origin. Second, by (3.3), u cannot have any positive local maxima nor any negative local minima. By integrating (3.3), we obtain $u'(x) > 0$ for $x > 0$ and $u'(x) < 0$ for $x < 0$ and the estimate $u'(x) > \frac{2}{\sigma^2(x)}(\alpha x - \mu(x))$ for $x > 0$. Hence $u(x) > 1 + 2\int_0^x \frac{\alpha r - \mu(r)}{\sigma^2(r)}\, dr$. Using (1.5), it follows that $\lim_{x\to+\infty} u(x) = +\infty$. A similar proof yields $\lim_{x\to-\infty} u(x) = +\infty$. This completes the proof. □

Proof of Proposition 3.1. By Lemma 3.2, for $n > 1$, we can find $\alpha_n < 0 < \beta_n$ such that $u(\alpha_n) = u(\beta_n) = n$, α_n decreasing to $-\infty$ and β_n increasing to $+\infty$. For the process $\{Y(t)\}$ in (3.1) with $Y(0) = x$, let $n_0 > 1$ so that $\alpha_{n_0} < x < \beta_{n_0}$. For each $n \geq n_0$ we define the stopping time τ_n by

$$\begin{aligned} \tau_n &= \inf\{t \geq 0; Y(t) \notin (\alpha_n, \beta_n)\} \\ &= +\infty \text{ if the above set is empty.} \end{aligned} \tag{3.4}$$

Thus $\lim_{n\to\infty} \tau_n = \tau_\infty$ a.s., where τ_∞ is the explosion time of the process $\{Y(t)\}$ in (3.1).

For each x in $[\alpha_n, \beta_n]$, we introduce the function H_n by $H_n(x) = \frac{u(x)}{n}$. Then H_n satisfies (3.3) on $[\alpha_n, \beta_n]$ with $H_n(\alpha_n) = H_n(\beta_n) = 1$, $\frac{1}{n} \leq H_n(x) < 1$ for all x in (α_n, β_n) and H_n', H_n'' are bounded on $[\alpha_n, \beta_n]$. Next, we apply Itô's lemma to $H_n(Y(t))e^{-\int_0^t \gamma(Y(s))\, ds}$, where $\gamma(Y(t)) = (\alpha - \mu'(Y(t)))$, and we obtain

$$E_x\big[H_n(Y(T \wedge \tau_n))e^{-\int_0^{T\wedge\tau_n} \gamma(Y(t))\, dt}\big] = H_n(x).$$

Hence, by letting T tend to infinity,

$$E_x\big[e^{-\int_0^{\tau_n} \gamma(Y(t))\, dt} I_{[\tau_n<\infty]}\big] \leq H_n(x) = \frac{u(x)}{n}.$$

By letting τ_n increase to τ_∞ as n tends to infinity, we obtain

$$E_x\big[e^{-\int_0^{\tau_\infty} \gamma(Y(t))\, dt} I_{[\tau_\infty<\infty]}\big] = 0$$

and hence (3.2) follows. This completes the proof. □

4. Optimality of the zero control. Consider a weak solution of

$$Z_x(t) = x + \int_0^t \mu(Z_x(s))\, ds + \int_0^t \sigma(Z_x(s))\, dW(s) \tag{4.1}$$

corresponding to (1.1), where the process $A(\cdot)$ is identically zero. Our goal in this section is to furnish the conditions on $\mu(\cdot), \sigma(\cdot)$, and $C(\cdot)$ under which $\{Z_x(t) : t \geq 0\}$ is an optimal process for (1.8). Here we extend the results in section 4 of [35] for a larger class of cost functions and for general diffusions. Throughout this section, in addition to (1.4), (1.5), and (1.6), we also assume

$$\alpha - \mu'(x) \geq |C'(x)| \quad \text{for all } x. \tag{4.2}$$

Introduce a sequence of stopping times $\{\tau_n\}$ for each $n > |x|$ by

$$\begin{aligned} \tau_n &= \inf\{t \geq 0 : |Z_x(t)| \geq n\} \\ &= +\infty \text{ if the above set is empty.} \end{aligned} \tag{4.3}$$

PROPOSITION 4.1. *Assume* (1.4) *and* (1.5). *Then a weak solution* $\{Z_x(t) : t \geq 0\}$ *to* (4.1) *exists and satisfies*

$$\lim_{N\to\infty} E_x\left[|Z_x(\tau_N)|e^{-\alpha\tau_N} I_{[\tau_N<\infty]}\right] = 0. \tag{4.4}$$

Therefore, $Z_x(t)$ *is finite for each* $t > 0$ *and satisfies the admissibility condition* (1.2) *with respect to* $\{\tau_N\}$ *which is defined in* (4.3).

Proof. The existence of a weak solution to (4.1) follows from the assumptions (1.4) and (1.5). We need to establish (4.4). Let H_0 be the solution of

$$\frac{\sigma^2(x)}{2}H_0''(x)+\mu(x)H_0'(x) - \alpha H_0(x) = 0 \quad \text{for all } x, \text{ and} \tag{4.5}$$
$$H_0(0) = 1, \qquad H_0'(0) = 0.$$

Following the proof of the Proposition 4.1 of [35], we can establish $\lim_{x\to+\infty} H_0'(x) = +\infty$ and $\lim_{x\to-\infty} H_0'(x) = -\infty$. Using L'Hôpital rule, we conclude $\lim_{x\to-\infty} \frac{H_0(x)}{x} = -\infty$ and $\lim_{x\to+\infty} \frac{H_0(x)}{x} = +\infty$. Itô's lemma implies that $H_0(Z_x(t\wedge\tau_n))e^{-\alpha(t\wedge\tau_n)}$ is a positive martingale, and we obtain $E_x[|Z_x(\tau_n)|e^{-\alpha\tau_n}I_{[\tau_n<\infty]}] \leq \frac{nH_0(x)}{\text{Min}\{H_0(n),H_0(-n)\}}$ (see Proposition 4.1 of [35]). Consequently, $\lim_{N\to\infty} E_x[|Z_x(\tau_N)|e^{-\alpha\tau_N}I_{[\tau_N<\infty]}] = 0$. This also shows τ_∞, the explosion time for $\{Z_x(t) : t \geq 0\}$ is infinite a.s. To verify the admissibility condition (1.2), we observe $E_x[\int_0^{T\wedge\tau_N} (|\mu(Z_x(s))| + \sigma^2(Z_x(s)))\, ds] \leq T \cdot \max_{[-N,N]}(|\mu(x)| + \sigma^2(x)) < \infty$. Hence both conditions in (1.2) are satisfied and the proposition follows. □

Remarks.

1. Using (1.5), one can directly verify Khasminski's criteria for nonexplosion for $\{Z_x(t)\}$ (see [29, p. 297]).
2. Using Itô's lemma for $H_0(Z_x(T \wedge \tau_n))$, we can derive $E_x[H_0(Z_x(T))] \leq H_0(x)e^{\alpha T}$ and thus $E_x|Z_x(T)|$ is finite for each $T > 0$.

Our next step is to obtain a function W_∞ on $\mathbb{R}$ which satisfies the following conditions (4.6) and (4.7) everywhere on R:

$$\frac{\sigma^2(x)}{2}W_\infty''(x) + (\sigma(x)\sigma'(x) + \mu(x))W_\infty'(x) - (\alpha - \mu'(x))W_\infty(x) + C'(x) = 0 \tag{4.6}$$

and

$$|W_\infty(x)| \leq 1 \quad \text{for all } x. \tag{4.7}$$

Once we find a solution W_∞, we will be able to show that W_∞ is the derivative of the value function $V(x)$ and the process in (4.1) is optimal. To obtain a solution of (4.6) and (4.7), we employ the process $\{Y(t) : t \geq 0\}$ defined in (3.1). Let $\alpha_n < 0 < \beta_n$ be the constants introduced in the proof of Proposition 3.1. Introduce the function W_n which satisfies

$$\frac{\sigma^2(x)}{2}W_n''(x) + (\sigma(x)\sigma'(x) + \mu(x))W_n'(x) - (\alpha - \mu'(x))W_n(x) + C'(x) = 0 \tag{4.8}$$
$$\text{on } (\alpha_n, \beta_n),$$
$$W_n(\alpha_n) = -1 \quad \text{and} \quad W_n(\beta_n) = +1. \tag{4.9}$$

We extend W_n to $\mathbb{R}$ so that W_n satisfies (4.8) everywhere. The following lemma will lead to our main theorem.

LEMMA 4.2. *Assume* (1.4), (1.5), (1.6), *and* (4.2). *Let* W_n *be as defined above. Then the following conclusions hold.*

(i) $|W_n(x)| \le 1$ *on* (α_n, β_n) *and* $\lim_{n\to\infty} W_n(x)$ *exists for every* x.

(ii) *Let* $W_\infty(x) = \lim_{n\to\infty} W_n(x)$ *for each* x *in* $\mathbb{R}$. *Then* W_∞ *satisfies* (4.6) *and* (4.7). *Furthermore,* W_∞ *has the stochastic representation*

$$W_\infty(x) = E_x\left[\int_0^{\tau_\infty} e^{-\int_0^t \gamma(Y(s))\,ds} C'(Y(t))\,dt\right] \quad \text{for all } x, \tag{4.10}$$

where τ_∞ *is the explosion time of the process* $\{Y(t)) : t \ge 0\}$ *as in Proposition* 3.1, *and* $\gamma(y) = \alpha - \mu'(y)$ *for all* y.

Proof. If $W_n'(z) = 0$ for some z in (α_n, β_n), then

$$\frac{\sigma^2(z)}{2} W_n''(z) = (\alpha - \mu'(z))\left(W_n(z) - \frac{C'(z)}{\alpha - \mu'(z)}\right).$$

Hence, by (4.2), if $W_n(z) > 1$, then z is necessarily a local minimum and if $W_n(z) < -1$, then z is necessarily a local maximum. But $W_n(\alpha_n) = -1$ and $W_n(\beta_n) = 1$. Therefore, it follows that $|W_n(x)| \le 1$ on (α_n, β_n). Next we apply Itô's lemma to $W_n(Y(T \wedge \tau_n))e^{-\int_0^{T\wedge\tau_n} \gamma(Y(s))\,ds}$, where τ_n's are defined in (3.4) and obtain

$$\begin{aligned} &E_x\left[W_n(Y(T\wedge\tau_n))e^{-\int_0^{T\wedge\tau_n} \gamma(Y(s))\,ds}\right] \\ &\quad = W_n(x) - E_x\int_0^{T\wedge\tau_n} e^{-\int_0^t \gamma(Y(s))\,ds} C'(Y(t))\,dt. \end{aligned} \tag{4.11}$$

Using (4.2), we have

$$E_x\left[\int_0^{\tau_\infty} e^{-\int_0^t \gamma(Y(s))\,ds} |C'(Y(s))|\,ds\right] \le E_x\left[\int_0^{\tau_\infty} e^{-\int_0^t \gamma(Y(s))\,ds} \gamma(Y(t))\,dt\right] \le 1 \tag{4.12}$$

and $|W_n(x)| \le 1$ on $[\alpha_n, \beta_n]$. Thus by letting T tend to infinity in (4.11) we obtain

$$\left|W_n(x) - E_x\int_0^{\tau_n} e^{-\int_0^t \gamma(Y(s))\,ds} C'(Y(t))\,dt\right| \le E_x\left[e^{-\int_0^{\tau_n} \gamma(Y(s))\,ds}\right]. \tag{4.13}$$

By (4.12) we also have

$$\lim_{n\to\infty} E_x\int_0^{\tau_n} e^{-\int_0^t \gamma(Y(s))\,ds} C'(Y(t))\,dt = E_x\int_0^{\tau_\infty} e^{-\int_0^t \gamma(Y(s))\,ds} C'(Y(t))\,dt. \tag{4.14}$$

By letting n tend to infinity, the right-hand side of (4.13) becomes zero by Proposition 3.1. Hence by (4.13) and (4.14) we conclude that $W_\infty(x) \equiv \lim_{n\to\infty} W_n(x)$ exists, $W_\infty(x)$ has the representation (4.10) and $|W_\infty(x)| \le 1$ for all x by (4.12). It remains to verify that W_∞ also satisfies (4.6).

We fix a large interval $[-M, M]$ and consider $n > n_0$, where $[-M, M] \subseteq [\alpha_{n_0}, \beta_{n_0}]$. By integrating (4.8) twice, we obtain

$$\begin{aligned} W_n(x) = W_n(0) &- 2\int_0^x \frac{C(r) + \mu(r)W_n(r)}{\sigma^2(r)}\,dr + \sigma^2(0)W_n'(0)\int_0^x \frac{1}{\sigma^2(r)}\,dr \\ &+ 2\alpha\int_0^x \frac{1}{\sigma^2(y)}\int_0^y W_n(u)\,du\,dy \quad \text{for any } x \text{ in } [-M, M]. \end{aligned}$$

Since $|W_n(x)| \le 1, \lim_{n\to\infty} W_n(x) = W_\infty(x)$ and $|W_\infty(x)| \le 1$ for all x, it follows from the above equation that $\lim_{n\to\infty} W_n'(0) = \lambda_0$ exists and is finite. Also W_∞ satisfies

$$W_\infty(x) = W_\infty(0) - 2\int_0^x \frac{C(r)+\mu(r)W_\infty(r)}{\sigma^2(r)}\,dr + \lambda_0\sigma^2(0)\int_0^x \frac{1}{\sigma^2(r)}\,dr + 2\alpha\int_0^x \frac{1}{\sigma^2(y)}\int_0^y W_\infty(u)\,du\,dy \quad \text{for any } x \text{ in } [-M, M].$$

By differentiating this equation twice, we see that W_∞ satisfies (4.6) as desired and $W_\infty'(0) = \lambda_0$. This completes the lemma. □

Next we construct a function $F(x)$ so that $F'(x) \equiv W_\infty(x)$, $\alpha F(0) = \frac{\sigma^2(0)}{2} W_\infty'(0)$ and F will satisfy the HJB equation (1.9). Let

$$F(x) = \frac{\sigma^2(0)}{2\alpha} W_\infty'(0) + \int_0^x W_\infty(r)\,dr \quad \text{for all } x \text{ in } \mathbb{R}. \tag{4.15}$$

We claim that F satisfies

$$\frac{\sigma^2(x)}{2} F''(x) + \mu(x)F'(x) - \alpha F(x) + C(x) \equiv 0 \quad \text{for all } x \text{ in } \mathbb{R}. \tag{4.16}$$

Since $F'(x) \equiv W_\infty(x)$, clearly (4.16) holds at $x = 0$. Also W_∞ satisfies (4.6), thus the derivative of the left-hand side of (4.16) is equal to zero. Therefore, (4.16) holds for all x. Next we prove the main theorem in this section.

THEOREM 4.3. *Assume* (1.4), (1.5), (1.6), *and* (4.2). *Let F be defined by* (4.15). *Then $F(x) = V(x)$ for all x, where V is the value function given in* (1.8). *Moreover, the process $\{Z_x(t)\}$ defined by* (4.1) *is an optimal process for each x.*

Proof. By (4.16) and Lemma 4.2, F is a C^2 function which satisfies the HJB equation (1.9) and by the verification results in section 2, it follows that $F(x) \le V(x)$ for all x. But $\{Z_x(t) : t \ge 0\}$ in (4.1) is an admissible process as described in the Proposition 4.1. Applying Itô's lemma to $F(Z_x(T\wedge\tau_n))e^{-\alpha(T\wedge\tau_n)}$, where (τ_n) is given in (4.3), and using the Proposition 4.1, we obtain $F(x) = E_x\int_0^\infty e^{-\alpha t}C(Z_x(t))\,dt$. Thus $F(x) \ge V(x)$ and consequently, $F(x) \equiv V(x)$ for all x. Hence, the process $\{Z_x(t) : t \ge 0\}$ is optimal. □

5. A bounded optimal process.

5.1. Reflecting diffusion processes. In addition to (1.4), (1.5), and (1.6) we assume the following conditions in this section.

(5.1) (i) There exist two points $\theta_0 < \beta_0$ such that $|C'(x)| < (\alpha - \mu'(x))$ for all x in (θ_0, β_0) and $|C'(x)| > (\alpha - \mu'(x))$ outside the interval $[\theta_0, \beta_0]$.

(5.2) (ii) There is a $\varepsilon_0 > 0$ and a large $M > 0$ so that $|C'(x)| > (1+\varepsilon_0)(\alpha - \mu'(x))$ when $|x| > M$.

Remarks.

1. By (1.6), since $C'(0) = 0$, it follows that $\theta_0 < 0 < \beta_0$, $C'(x) + (\alpha - \mu'(x)) < 0$ for $x < \theta_0$ and $C'(x) - (\alpha - \mu'(x)) > 0$ for $x > \beta_0$.
2. Clearly $\lim_{|x|\to\infty} C(x) = +\infty$.

Under these assumptions, our candidate for an optimal process comes from a class of diffusion processes with reflection barriers at the points a and b, where $a < 0 < b$. Therefore, for each $a < 0 < b$, we consider a weak solution of

$$X_x(t) = x + \int_0^t \mu(X_x(s))\,ds + \int_0^t \sigma(X_x(s))\,dW(s) + K(t), \tag{5.3}$$

where $\{W(t) : t \geq 0\}$ is a Brownian motion and the bounded variation process $K(t)$ is given by

$$dK(t) = dL_a(t) - dL_b(t) \quad \text{for } t > 0 \text{ and} \tag{5.4}$$
$$d|K|(t) = dL_a(t) + dL_b(t) \quad \text{for } t > 0.$$

Here L_a and L_b are the local time processes of $\{X_x(t) : t \geq 0\}$ at the points a and b, respectively. If the initial point x is outside $[a, b]$, then there will be an initial jump from x to the nearest point of the set $\{a, b\}$. Hence, in this case we let $X_x(0-) = x$ and $|K|(0)$ is equal to the distance from x to $\{a, b\}$. Since $X_x(\cdot)$ satisfies assumption (1.2), it is an admissible process. The cost functional related to $X_x(\cdot)$ is given by

$$V_{ab}(x) = E_x \int_0^\infty e^{-\alpha t}(C(X_x(t))\, dt + d|K|(t)). \tag{5.5}$$

Using Itô's lemma, it is easy to verify that V_{ab} satisfies the differential equation

$$\frac{\sigma^2(x)}{2} V''_{ab}(x) + \mu(x)V'_{ab}(x) - \alpha V_{ab}(x) + C(x) = 0 \quad \text{for } a < x < b, \tag{5.6}$$

$$V'_{ab}(x) = -1 \text{ for } x \leq a \quad \text{and} \quad V'_{ab}(x) = 1 \quad \text{for } x \geq b. \tag{5.7}$$

Our aim here is to find a pair of points $a^* < b^*$ such that $V_{a^*b^*}$ is twice continuously differentiable and satisfies (5.6), (5.7), and, in addition to that, $V_{a^*b^*}$ satisfies $|V'_{a^*b^*}(x)| \leq 1$ for all x. Thus if we let $W(x) = V'_{a^*b^*}(x)$, it should satisfy

$$\frac{\sigma^2(x)}{2} W''(x) + (\sigma(x)\sigma'(x) + \mu(x))W'(x) - (\alpha - \mu'(x))W(x) + C'(x) = 0 \tag{5.8}$$

for $a^* < x < b^*$

and

$$W(a^*) = -1, \quad |W(x)| < 1 \text{ on } (a^*, b^*), \quad \text{and} \quad W(b^*) = 1. \tag{5.9}$$

For $V_{a^*b^*}$ to be twice continuously differentiable, W has to satisfy an additional requirement

$$W'(a^*) = W'(b^*) = 0. \tag{5.10}$$

Once we have the existence of a^* and b^*, it follows that $a^* < 0 < b^*$ by (5.8)–(5.10). Our approach here is first to derive such a function W together with an interval $[a^*, b^*]$ and then to construct $V_{a^*b^*}$ and to show that it is a C^2 solution to the HJB equation (1.9). As a first step, we need to obtain a C^1 solution to (5.8), (5.9), and (5.10). For this, we consider a class of optimal stopping problems.

5.2. A class of optimal stopping problems. Let $\{Y(t) : t \geq 0\}$ be a weak solution to (3.1) with $Y(0) = x$ as described in section 3, and let τ_∞ be the explosion time for $\{Y(t) : t \geq 0\}$. We introduce the function $\psi(x)$ by

$$\psi(x) = C'(x) - (\alpha - \mu'(x)) \quad \text{for all } x. \tag{5.11}$$

By assumption (5.1) (see also the remark after (5.1)), there is a point $\beta_0 > 0$ so that $\psi(\beta_0) = 0$ and

$$\psi(x) > 0 \quad \text{for all } x > \beta_0. \tag{5.12}$$

For each $p < \beta_0$ and $x \geq p$, introduce the stopping problem

$$R_p(x) = \inf_{0 \leq \tau \leq \tau_\infty} E_x \left[\int_0^{\tau \wedge \tau_p} e^{-\int_0^t \gamma(Y(s))\, ds} \psi(Y(t))\, dt \right], \tag{5.13}$$

where

$$\gamma(x) = \alpha - \mu'(x) \tag{5.14}$$

and

$$\begin{aligned} \tau_p &= \inf\{t \geq 0 : Y(t) \leq p\} \\ &= +\infty \text{ if the above set is empty.} \end{aligned} \tag{5.15}$$

For each $p < q$, introduce

$$R_{pq}(x) = E_x \int_0^{\tau_{pq}} e^{-\int_0^t \gamma(Y(s))\, ds} \psi(Y(t))\, dt \tag{5.16}$$

for each x in $[p, q]$, where $\tau_{pq} = \text{Min}\{\tau_p, \tau_q\}$. Using Itô's lemma, one can verify that R_{pq} satisfies

$$\begin{aligned} &\frac{1}{2}\sigma^2(x) R''_{pq}(x) + (\sigma(x)\sigma'(x) + \mu(x)) R'_{pq}(x) - (\alpha - \mu'(x)) R_{pq}(x) + \psi(x) = 0 \\ &\quad \text{for all } x \text{ in } (p, q) \end{aligned} \tag{5.17}$$

and

$$R_{pq}(p) = R_{pq}(q) = 0. \tag{5.18}$$

We extend R_{pq} to $\mathbb{R}$ so that it is a C^2 function which satisfies (5.17) everywhere. Using (5.12) and (5.17) we observe the following fact which will be used in the next few lemmas:

$$\text{for any } p < q, R_{pq} \text{ cannot have negative local minima on } (\beta_0, +\infty). \tag{5.19}$$

The next result contains the technical work necessary for finding a solution of (5.8)–(5.10).

LEMMA 5.1. *Assume* (5.1) *and* (5.2). *Then*

(i) *for each* $p < \beta_0$, *there exists a unique point* $\eta_p > \beta_0$ *such that* $R_{p\eta_p}$, *the solution of* (5.17) *and* (5.18) *with* $q = \eta_p$, *also satisfies* $R_{p\eta_p}(x) < 0$ *for all* x *in* (p, η_p), $R'_{p\eta_p}(\eta_p) = 0$, *and* $R''_{p\eta_p}(\eta_p) < 0$;

(ii) $R_{p\eta_p}$ *is increasing on* (β_0, η_p) *and decreasing on* (η_p, ∞), *and* $R_{p\eta_p}(x) < 0$ *on the set* $(p, \infty) \backslash \{\eta_p\}$;

(iii) *if* $p_1 < p_2 < \beta_0$, *then* $\beta_0 < \eta_{p_2} < \eta_{p_1}$ *and* $R_{p_1\eta_{p_1}}(x) < R_{p_2\eta_{p_2}}(x)$ *in* $[p_2, \eta_{p_2}]$;

(iv) *as a function of* p, η_p *is a continuous strictly decreasing function on the interval* $(-\infty, \beta_0)$;

(v) *if* $a < \beta_0$, *then* $\lim_{p \to a} R_{p\eta_p}(x) = R_{a\eta_a}(x)$ *for all* x *in* (a, η_a).

Proof. Let $p < \beta_0$ and $q = \beta_0$ in (5.17) and (5.18). Since $\psi < 0$ on (p, β_0), by Theorems 3 and 4 and the corollary in the pp. 6–7 of [28], we conclude

$$R_{p\beta_0}(x) < 0 \text{ for all } x \text{ in } (p, \beta_0), \quad R'_{p\beta_0}(p) < 0, \quad \text{and} \quad R'_{p\beta_0}(\beta_0) > 0. \tag{5.20}$$

Next, for each $p < \beta_0$, we pick a point $q_p > \beta_0$ so that $\int_p^{q_p} \psi(x)\,dx = 0$. We claim that for each $q > q_p$,

$$\sup_{[p,q]} R_{pq}(x) > 0. \tag{5.21}$$

Suppose (5.21) is false. Then there is a $q > q_p$ so that $\sup_{[p,q]} R_{pq} \le 0$. Hence $R'_{pq}(p) \le 0, R'_{pq}(q) \ge 0$ and by integrating (5.17), we have

$$\frac{\sigma^2(q)}{2} R'_{pq}(q) - \frac{\sigma^2(p)}{2} R'_{pq}(p) + \int_p^q \psi(x)\,dx = \alpha \int_p^q R_{pq}(x)\,dx. \tag{5.22}$$

Next, $\int_p^q \psi(x)\,dx > 0$ since $q > q_p > \beta_0$. Thus the left-hand side of (5.22) is strictly positive. But the right-hand side of (5.22) is negative, since $R_{pq}(x) \le 0$ on $[p, q]$. This is a contradiction and it proves (5.21).

For each $p < \beta_0$, we consider the set G defined by $G = \{q : q > p$ and $R_{pq}(x) < 0$ in $(p, q)\}$. By (5.20), β_0 is in G and thus G is nonempty. By (5.21), q_p is an upper bound for G. Thus $\sup G$ exists and is finite. We let $\eta_p = \sup G$. We intend to show that η_p is in G and $R'_{p\eta_p}(\eta_p) = 0$, similar to Lemma 4.3 in [27]. Clearly, $R_{p\eta_p}(x) \le 0$ for all x in (p, η_p). Suppose that $R_{p\eta_p}(x_0) = 0$ for some x_0 in (p, η_p). Then $R_{p\eta_p}(x_0)$ is a local maximum and by (5.17) we have $\psi(x_0) \ge 0$. Therefore, $x_0 \ge \beta_0$ and $R_{p\eta_p}$ will necessarily have a local minimum on (x_0, η_p). This contradicts (5.19). Hence $R_{p\eta_p}(x) < 0$ for all x in (p, η_p) and this implies that η_p is in G and $R'_{p\eta_p}(\eta_p) \ge 0$. Now suppose $R'_{p\eta_p}(\eta_p) > 0$. Then $R_{p\eta_p}$ is strictly increasing on an interval $[\eta_p - \epsilon, \eta_p + \epsilon]$. Let $y_0 = R_{p\eta_p}(\eta_p + \epsilon) > 0$ and $D(x)$ be the solution to the homogeneous equation $\frac{\sigma^2(x)}{2} D''(x) + (\sigma(x)\sigma'(x) + \mu(x))D'(x) - (\alpha - \mu'(x))D(x) = 0$ for all x in $(p, \eta_p + \epsilon)$ with the boundary conditions $D(p) = 0$ and $D(\eta_p + \epsilon) = y_0 > 0$. Using the corollary to Theorem 4 in page 7 of [28], we conclude that $D(x) > 0$ for all $x > p$. Now the uniqueness theorem for solutions to (5.17) (Theorem 7, page 13 of [28]) implies that $R_{p\eta_p+\epsilon}(x) = R_{p\eta_p}(x) - D(x)$, $R_{p\eta_p+\epsilon}(x) < 0$ for all x in $(p, \eta_p + \epsilon)$ and hence $\eta_p + \epsilon$ is in G. This is a contradiction and thus $R'_{p\eta_p}(\eta_p) = 0$. Using this together with $\psi(\eta_p) > 0$ yields $R''_{p\eta_p}(\eta_p) < 0$. Therefore, η_p satisfies all the conditions in part (i) of the lemma.

To obtain the uniqueness of η_p, suppose $\tilde{\eta}_p$ is another such point. Then $\tilde{\eta}_p$ is also in G and hence $\tilde{\eta}_p < \eta_p$. Consider the solution $R_{p\tilde{\eta}_p}$ of (5.17). Since $\tilde{\eta}_p$ is in G, we have $R''_{p\tilde{\eta}_p}(\tilde{\eta}_p) < 0$ and hence $\psi(\tilde{\eta}_p) \ge 0$. Consequently, $\tilde{\eta}_p \ge \beta_0$, but $\tilde{\eta}_p \ne \beta_0$ by (5.20). Hence $\tilde{\eta}_p > \beta_0$, $R_{p\tilde{\eta}_p}$ has a local maximum at $x = \tilde{\eta}_p$ and by (5.19), $R_{p\tilde{\eta}_p}$ is decreasing on $(\tilde{\eta}_p, \infty)$. Therefore, $R_{p\eta_p}$ and $R_{p\tilde{\eta}_p}$ meet at a point z in $(\tilde{\eta}_p, \eta_p)$. Now, by the uniqueness theorem for solutions to (5.17) on $[p, z]$, it follows that $R_{p\eta_p}(x) = R_{p\tilde{\eta}_p}(x)$ for all x in $[p, z]$. Hence $R_{p\eta_p}(\tilde{\eta}_p) = 0$ and this is a contradiction, since η_p is in G. Consequently, $\eta_p = \tilde{\eta}_p$ and part (i) follows.

Since $R_{p\eta_p}(\eta_p) = R'_{p\eta_p}(\eta_p) = 0$, $R_{p\eta_p}$ has a strict local maximum at $x = \eta_p$ by (5.17) and by applying (5.19), part (ii) follows.

To prove part (iii), let $p_1 < p_2 < \beta_0$ and suppose that $\eta_{p_1} \le \eta_{p_2}$. If $\eta_{p_1} = \eta_{p_2}$, since $R_{p_i\eta_{p_i}}(\eta_{p_i}) = R'_{p_i\eta_{p_i}}(\eta_{p_i}) = 0$ for $i = 1, 2$, by the uniqueness of solutions to (5.17) it follows that $p_1 = p_2$. Hence $\eta_{p_1} \ne \eta_{p_2}$ and we have $p_1 < p_2 < \beta_0 < \eta_{p_1} < \eta_{p_2}$. Since $R_{p_1\eta_{p_1}}(p_2) < 0$ and $R_{p_2\eta_{p_2}}(\eta_{p_1}) < 0$, there is a point c_1 such that $p_2 < c_1 < \eta_{p_1}$ and $R_{p_1\eta_{p_1}}(c_1) = R_{p_2\eta_{p_2}}(c_1)$. By part (ii), we can conclude $R_{p_1\eta_{p_1}}(\eta_{p_2}) < 0$ and $R_{p_2\eta_{p_2}}(\eta_{p_1}) < 0$ and this implies the existence of a point c_2 so that $\eta_{p_1} < c_2 < \eta_{p_2}$ and $R_{p_1\eta_{p_1}}(c_2) = R_{p_2\eta_{p_2}}(c_2)$. Hence, $R_{p_1\eta_{p_1}}$ and $R_{p_2\eta_{p_2}}$ are identical by the uniqueness of solutions to (5.17) (see Theorem 7, page 13 of [28]) and thus $p_1 = p_2$. This is

a contradiction. Hence $\eta_{p_2} < \eta_{p_1}$, thus by part (i) and again by the uniqueness of solutions to (5.17), we conclude $R_{p_1\eta_{p_1}}(x) < R_{p_2\eta_{p_2}}(x)$ for all x in $[p_2, \eta_{p_2}]$. This proves part (iii).

To prove part (iv), it remains to show η_p is continuous as a function of p for $p < \beta_0$. Let a be a point so that $a < \beta_0$, and (p_n) be any increasing sequence so that $\lim_n p_n = a$. Thus $p_n < a < \beta_0$. By part (iii), the sequence (η_{p_n}) is decreasing and $\eta_{p_n} > \eta_a$ for each n. Thus $\lim_n \eta_{p_n} \geq \eta_a$. Now suppose that $\lim_n \eta_{p_n} > \eta_a$. Then we can pick a point q so that $\lim_n \eta_{p_n} > q > \eta_a$ and consider a solution $R(x)$ to (5.17) with the initial condition $R(q) = R'(q) = 0$. Since $\psi(q) > 0$, by (5.17) we have $R''(q) < 0$, and R has a strict local maximum at $x = q$. Since R satisfies (5.17), we can also conclude that R has no local minima in (β_0, ∞) and it is increasing on $[\beta_0, q]$ and decreasing on $[q, +\infty)$. For each n, $R_{p_n\eta_{p_n}}(q) < 0$ and thus R meets $R_{p_n\eta_{p_n}}$ at some point in (q, η_{p_n}). But R cannot be identically equal to $R_{p_n\eta_{p_n}}$ and hence by applying the uniqueness theorem (Theorem 7, page 13 of [28]) we can conclude that $R(x) > R_{p_n\eta_{p_n}}(x)$ for each x in $[p_n, q]$ and thus $R(p_n) > 0$. Consequently, $R(a) \geq 0$. With a similar argument, R and $R_{a\eta_a}$ meet at a point in (η_a, q) and $R(\eta_a) < 0$. Consequently, $R(x) < R_{a\eta_a}(x)$ for all x in $[a, \eta_a]$. This yields $R(a) < 0$, which is a contradiction. Therefore, $\lim_n \eta_{p_n} = \eta_a$. A very similar proof can be given to show that if (p_n) is a decreasing sequence satisfying $p_n < \beta_0$ for all n and $\lim_n p_n = a$, then $\lim_n \eta_{p_n} = \eta_a$. Hence, part (iv) follows.

To prove part (v), we will show the following: if (p_n) is a monotone sequence with $\lim_{n\to\infty} p_n = a$, then $\lim_n R_{p_n\eta_{p_n}}(x) = R_{a\eta_a}(x)$ for each x in $[a, \eta_a]$. First, consider an increasing sequence (p_n) so that $\lim_{n\to\infty} p_n = a$. For each x in $[a, \eta_a]$, $R_{p_n\eta_{p_n}}(x)$ has the stochastic representation (5.16), since $[a, \eta_a] \subseteq [p_n, \eta_{p_n}]$. In (5.16), we relabel the stopping time $\tau_{p_n\eta_{p_n}}$ by τ_n for simplicity. Clearly, (τ_n) is decreasing and $\tau_n > \tau_{a\eta_a}$ for all n. $E_x[\tau_n]$ is given by formula (5.55) in page 343 of [17] with $a = p_n$ and $b = \eta_{p_n}$ (see also (5.59) in page 344 of [17]). By letting n tend to infinity there, we obtain $\lim_{n\to\infty} E_x[\tau_n] = E_x[\tau_{a\eta_a}]$. Thus $\lim_{n\to\infty} \tau_n = \tau_{a,\eta_a}$ a.s. Now again using (5.16) and the dominated convergence theorem, we obtain $\lim_{n\to\infty} R_{p_n\eta_{p_n}}(x) = R_{a\eta_a}(x)$ for each x in (a, η_a). Next, if (p_n) is a decreasing sequence so that $\lim_{n\to\infty} p_n = a$ and $p_n < \beta_0$ for all n, consider (5.16) to represent $R_{p_n\eta_{p_n}}(x)$ and $R_{a\eta_a}(x)$. Notice that $\tau_{p_n\eta_{p_n}}$ is increasing to $\tau_{a\eta_a}$ as n tends to infinity. Hence we apply the dominated convergence theorem to (5.16) and conclude that $\lim_{n\to\infty} R_{p_n\eta_{p_n}}(x) = R_{a\eta_a}(x)$ for all x in (a, η_a). This completes the proof. □

The next lemma derives an optimal stopping policy for the optimal stopping problem (5.13).

LEMMA 5.2. *Fix $p < \beta_0$. Consider $R_p(x)$ defined in* (5.13) *for each $x \geq p$. Let the function $R_{p\eta_p}$ and the point η_p be as in Lemma* 5.1. *Then*

(i) $R_p(x) = \begin{cases} R_{p\eta_p}(x) & \text{for } p \leq x \leq \eta_p, \\ 0 & \text{for } x > \eta_p. \end{cases}$

(ii) *The stopping rule τ^* given by* $\tau^* = \begin{cases} \min\{\tau_p, \tau_{\eta_p}\} & \text{when } p \leq x \leq \eta_p, \\ 0 & \text{when } x > \eta_p \end{cases}$

is optimal for the stopping problem (5.13).

Proof. Let

$$R(x) = \begin{cases} R_{p\eta_p}(x) & \text{for } x \leq \eta_p, \\ 0 & \text{for } x > \eta_p. \end{cases}$$

We observe that $R(x)$ is a C^1 function on $(p,+\infty)$, which is C^2 everywhere except at $x=\eta_p$. By Lemma 5.1, $R''(\eta_p^-)<0$. Clearly, $R''(\eta_p^+)=0$. Let $Y(0)=x>p$ and apply Itô's lemma (see [17, p. 219]) to $R(Y(t))e^{-\int_0^t \gamma(Y(s))\,ds}$ to obtain

$$E_x\left[R(Y(T\wedge\tau\wedge\tau_{pn}))e^{-\int_0^{T\wedge\tau\wedge\tau_{pn}}\gamma(Y(s))\,ds}\right]=R(x)$$

$$-E_x\left[\int_0^{T\wedge\tau\wedge\tau_{pn}} e^{-\int_0^t\gamma(Y(s))\,ds}\psi(Y(t))\,dt\right],$$

where $\tau_{pn}=\tau_p\wedge\tau_n$, $n>\eta_p$ is a large integer, and τ is any stopping time such that $\tau\le\tau_\infty$ a.s. Since $R(\cdot)$ is nonpositive and bounded below on $[p,+\infty)$, by letting T and n tend to infinity, we obtain $E_x[\int_0^{\tau\wedge\tau_p} e^{-\int_0^t\gamma(Y(s))\,ds}\psi(Y(t))\,dt]\ge R(x)$. Hence by (5.13), $R_p(x)\ge R(x)$ for all $x\ge p$. Let us define the stopping time τ^* as in the statement of Lemma 5.2(ii). Then, following the above calculations, we obtain $R(x)=E_x[\int_0^{\tau^*\wedge\tau_p} e^{-\int_0^t\gamma(Y(s))\,ds}\psi(Y(t))\,dt]$ for $x\ge p$. Consequently, $R_p(x)\equiv R(x)$ for all $x\ge p$. This completes the proof. □

LEMMA 5.3. *Consider the function $H(\cdot)$ defined by $H(p)=\min_{x\ge p}R_p(x)$ for each $p<\beta_0$. Then*

(i) *$H(\cdot)$ is continuous and increasing on $(-\infty,\beta_0)$, and*

(ii) $$\lim_{p\to\beta_0}H(p)=0 \quad \textit{and} \quad \lim_{p\to-\infty}H(p)<-2. \tag{5.23}$$

Proof. By the previous lemma, $H(p)=\min_{[p,\eta_p]}R_p(x)$ and $H(p)<0$ for each $p<\beta_0$. Moreover, by parts (ii) and (iii) of Lemma 5.1, it follows that $H(p)$ is increasing on $(-\infty,\beta_0)$. But $H(p)=\min_{[p,\beta_0]}R_p(x)$. Since $\psi>0$ on $(\beta_0,+\infty)$, $\inf_{[p,\eta_p]}\psi(x)=\inf_{[p,\beta_0]}\psi(x)$. Therefore, by (5.13) and (5.16), it easily follows that $\frac{1}{\alpha}\inf_{[p,\beta_0]}\psi(x)<H(p)<0$. Hence $\lim_{p\to\beta_0}H(p)=0$. Next, we show that $H(p)$ is continuous on $(-\infty,\beta_0)$. Let $a<\beta_0$, and we pick two points b and c so that $b<a<c<\beta_0$. For each p in $[b,c]$, consider the function $R_{p\eta_p}$ of Lemma 5.1 and we introduce the function $Q_p(x)$ defined on $[b,\eta_b]$ by

$$Q_p(x)=\begin{cases}R_{p\eta_p}(x) & \text{for } b\le x\le\eta_p,\\ 0 & \text{for } \eta_p\le x\le\eta_b.\end{cases}$$

Each Q_p is a C^1 function on $[b,\eta_b]$. If $b<p_1<p_2<c$, then $R_{p_1\eta_{p_1}}$ and $R_{p_2\eta_{p_2}}$ meet at some point on (η_{p_2},η_{p_1}) as observed in the proof of part (iii) of Lemma 5.1. Therefore, $R_{p_1\eta_{p_1}}(x)<R_{p_2\eta_{p_2}}(x)$ for all x in $[b,\eta_{p_2}]$. Consequently, we have $H(b)\le Q_p(x)\le\max_{[b,\eta_b]}Q_c(x)$ for all x in $[b,\eta_p]$ and for all p in $[b,c]$. Thus we can obtain a constant $M_1>0$ so that $\sup_{b\le p\le c}\sup_{b\le x\le\eta_b}|Q_p(x)|<M_1$. Next by integrating (5.17), we obtain

$$\frac{\sigma^2(x)}{2}Q_p'(x)=\int_x^{\eta_p}\psi(u)\,du-\mu(x)Q_p(x)-\alpha\int_x^{\eta_p}Q_p(u)\,du \quad \text{for } b\le x\le\eta_p,$$

and $Q_p'(x)=0$ for $x\ge\eta_p$. Thus using the estimate for $|Q_p(x)|$, we can find a constant $M_2>0$ so that $\sup_{b\le p\le c}\sup_{b\le x\le\eta_b}|Q_p'(x)|\le M_2$. Hence for each x,y in $[b,\eta_b]$, we have $|Q_p(x)-Q_p(y)|\le M_2|x-y|$ and in particular $|Q_p(x)|\le M_2|x-p|$. This in turn yields that $\lim_{p\to a}Q_p(a)=0$. Now let (p_n) be a sequence in (b,c)

so that $\lim_n p_n = a$. Then $H(p_n) = Q_{p_n}(x_n)$ for some x_n in $[p_n, \beta_0]$ by Lemma 5.1(ii). Thus $b \le p_n \le x_n \le \beta_0$. Let y be a limit point of (x_n). Then $a \le y \le \beta_0$ and $|Q_{p_n}(x_n) - Q_a(y)| \le |Q_{p_n}(x_n) - Q_{p_n}(y)| + |Q_{p_n}(y) - Q_a(y)|$. But $|Q_{p_n}(x_n) - Q_{p_n}(y)| \le M_2|x_n - y|$ and thus $\lim_{n\to\infty} |Q_{p_n}(x_n) - Q_{p_n}(y)| = 0$.

By part (v) of Lemma 5.1, and since $\lim_{p\to a} Q_p(a) = 0$ we have $\lim_{n\to\infty} |Q_{p_n}(y) - Q_a(y)| = 0$. Therefore, $\lim_{n\to\infty} Q_{p_n}(x_n) = Q_a(y)$ and, consequently, $\lim_{n\to\infty} H(p_n) = Q_a(y) \ge H(a)$. On the other hand, $H(a) = Q_a(x_0)$ for some x_0 in $[a, \beta_0]$ and thus $H(a) = \lim_{n\to\infty} Q_{p_n}(x_0) \ge \lim_{n\to\infty} H(p_n)$. Hence $H(\cdot)$ is continuous at $p = a$ for each $a < \beta_0$.

It remains to verify that $\lim_{p\to-\infty} H(p) < -2$. We pick $r_0 < -M$, where M is in (5.2). Then $\psi(x) < -(2+\varepsilon_0)\gamma(x)$ for all $x < r_0$. Let $p < r_0$ and $R_p(x)$ be given by (5.13). For any $p < r_0$, we obtain

$$R_p(x) \le E_x \int_0^{\tau_p \wedge \tau_{r_0}} e^{-\int_0^t \gamma(Y(s))\,ds} \psi(Y(t))\,dt \quad \text{for all } x \text{ in } (p, r_0).$$

Hence, we derive

$$R_p(x) < -(2+\varepsilon_0) + (2+\varepsilon_0) E_x\Big[e^{-\int_0^{\tau_p\wedge\tau_{r_0}} \gamma(Y(s))\,ds}\Big] \quad \text{for } p < x < r_0.$$

Since $H(p) \le R_p(x)$ and H is increasing on $(-\infty, \beta_0)$, we obtain

$$\lim_{p\to-\infty} H(p) \le -(2+\varepsilon_0) + (2+\varepsilon_0) \lim_{p\to-\infty} E_x\Big[e^{-\int_0^{\tau_p\wedge\tau_{r_0}} \gamma(Y(s))\,ds}\Big] \tag{5.24}$$

for any $x < r_0$, provided that the limit on right-hand side exists. Let τ_∞ be the explosion time for $\{Y(t) : t \ge 0\}$ as in Proposition 3.1, then it follows that

$$\begin{aligned} \lim_{p\to-\infty} E_x\Big[e^{-\int_0^{\tau_p\wedge\tau_{r_0}} \gamma(Y(s))\,ds}\Big] &= E_x\Big[e^{-\int_0^{\tau_\infty\wedge\tau_{r_0}} \gamma(Y(s))\,ds}\Big] \\ &= E_x\Big[e^{-\int_0^{\tau_{r_0}} \gamma(Y(s))\,ds} I_{[\tau_{r_0}<\tau_\infty]}\Big]. \end{aligned} \tag{5.25}$$

In the last equality we have used Proposition 3.1.

We let $P(x) = E_x[e^{-\int_0^{\tau_{r_0}} \gamma(Y(s))\,ds} I_{[\tau_{r_0}<\tau_\infty]}]$. We claim that $\lim_{x\to-\infty} P(x) = 0$. For each $N > |x|$, we define $P_N(x) = E_x[e^{-\int_0^{\tau_{r_0}} \gamma(Y(s))\,ds} I_{[\tau_{r_0}<\tau_{-N}]}]$, where τ_{-N} is defined by (5.15). By Itô's lemma, one can verify that

$$\begin{aligned} &\frac{\sigma^2(x)}{2} P_N''(x) + (\sigma(x)\sigma'(x) + \mu(x)) P_N'(x) - (\alpha - \mu'(x)) P_N(x) = 0 \\ &\text{on } (-N, r_0) \ \text{ and } \ P_N(-N) = 0,\ P_N(r_0) = 1. \end{aligned} \tag{5.26}$$

Observe that, $P_N(x) > 0$ on $(-N, r_0)$ and by (5.26), it cannot have positive local maxima. Hence P_N is increasing on $(-N, r_0)$. Since the solutions to (5.26) cannot intersect twice, it follows that $P_N(y) \ne P_{N'}(y)$ when $N \ne N'$ and $y < r_0$. Therefore, for a fixed $x < r_0$, as N tends to infinity, $\{P_N(x)\}$ is increasing and bounded above by 1. Let $P_\infty(x) = \lim_{N\to\infty} P_N(x)$. Then P_∞ also satisfies the same differential equation on $(-\infty, r_0)$ and $P_\infty(r_0) = 1$. It is clear that $P_\infty(x) \equiv P(x)$. Hence $P(x)$ satisfies (5.26) on $(-\infty, r_0)$ and we obtain

$$\frac{\sigma^2(r_0)}{2} P'(r_0) + \mu(r_0) = \frac{\sigma^2(x)}{2} P'(x) + \mu(x) P(x) + \alpha \int_x^{r_0} P(u)\,du. \tag{5.27}$$

Since P is increasing on $(-\infty, r_0)$, we let $\lim_{x\to-\infty} P(x) = \delta \geq 0$. If $\delta > 0$, then the right-hand side of (5.27) tends to infinity as x tend to $-\infty$. This leads to a contradiction.

Hence $\lim_{x\to-\infty} P(x) = 0$. Using this in (5.24), we obtain $\lim_{p\to-\infty} H(p) \leq -(2+\varepsilon_0)$. This completes the proof of Lemma 5.3. □

Our next proposition finds the two points a^*, b^* and a function W which satisfies (5.8)–(5.10).

PROPOSITION 5.4. *There exists two points a^*, b^* and a function W defined on $[a^*, b^*]$ satisfying* (5.8), (5.9), *and* (5.10). *Furthermore, $[\theta_0, \beta_0] \subseteq [a^*, b^*]$ and thus $[a^*, b^*]$ contains the origin. The function W can be extended to $\mathbb{R}$ so that $W(x) = -1$ for $x \leq a^*$ and $W(x) = 1$ for $x \geq b^*$. Then W is continuously differentiable on $\mathbb{R}$ and it is C^2 everywhere except at the points a^* and b^*.*

Proof. By Lemma 5.3, $H(p) = -2$ for some $p = p^*$ and we consider the corresponding function $R_{p^*}(x)$. Let a^* be a point where R_{p^*} achieves its absolute minimum. Then we have $p^* < a^* < \beta_0$ by Lemma 5.1 and $R_{p^*\eta_{p^*}}(a^*) = -2$. We let $b^* = \eta_{p^*}$. Then $p^* < a^* < \beta_0 < b^*$. Evaluating (5.17) at $x = a^*$ for $R_{p^*\eta_{p^*}}$, we obtain $C'(a^*) + (\alpha - \mu'(a^*)) \leq 0$ and thus $p^* < a^* \leq \theta_0 < \beta_0 < b^*$. Next we claim this absolute minimum point a^* is unique. For this, assume that $a_1 < a_2$ are two points of absolute minimum, and hence $R_{p^*\eta_{p^*}}(a_i) = -2$, and $p^* < a_1 < a_2 \leq \theta_0$. Then there will be a point of local maximum at $x = d$ in (a_1, a_2) and evaluating (5.17) at $x = d$ we obtain $C'(d) > -(\alpha - \mu'(d))$. Consequently, $C'(d) + (\alpha - \mu'(d)) > 0$ and hence $d > \theta_0$. This is a contradiction since $d < a_2 \leq \theta_0$. Therefore, the absolute minimum point $x = a^*$ is unique.

Next, we let $W(x) = R_{p^*\eta_{p^*}}(x) + 1$ on $[a^*, b^*]$. We extend W to $\mathbb{R}$ by $W(x) = -1$ for $x \leq a^*$ and $W(x) = +1$ for $x \geq b^*$. Also $W(a^*) = -1$, $W'(a^*) = 0$, and $W''(a^*+) \geq 0$. Using (5.17), it follows that W satisfies (5.8) and $W(b^*) = 1$. By the uniqueness of the absolute minimum point $x = a^*$ and by Lemma 5.1, it follows that $-1 < W(x) < 1$ on (a^*, b^*), $W(b^*) = 1$, $W'(b^*) = 0$, and $W''(b^*-) \leq 0$. Furthermore, W is continuously differentiable and C^2 everywhere except at the points a^* and b^*. This completes the proof. □

Next, we define the function $V^*(x)$ on $\mathbb{R}$ by

$$V^*(x) = \frac{C(a^*) - \mu(a^*)}{\alpha} + \int_{a^*}^{x} W(r)\, dr. \tag{5.28}$$

In the next theorem, we show that V^* is indeed the value function.

THEOREM 5.5. *Assume* (5.1) *and* (5.2). *Let a^* and b^* be the points described in Proposition* 5.4. *Consider the reflecting diffusion process $X_x^*(\cdot)$ on $[a^*, b^*]$ defined by* (5.3) *and* (5.4). *Then $\{X_x^*(t) : t \geq 0\}$ is an optimal process and the function V^* defined in* (5.28) *is the value function.*

Proof. Let $\{X_x^*(t) : t \geq 0\}$ be the reflecting diffusion on $[a^*, b^*]$ satisfying (5.3) and (5.4). Clearly, it is an admissible process which satisfies (1.2). The function V^* satisfies (5.6) and (5.7) and thus it is the pay-off function for $\{X_x^*(t) : t \geq 0\}$. Hence, we have $V^*(x) \geq V(x)$ for all x, where V is the value function given in (1.8).

To show that $V^*(x) \leq V(x)$ we will use the verification lemma. By Proposition 5.4, $|V^{*'}(x)| < 1$ on (a^*, b^*) and $|V^{*'}(x)| = 1$ outside (a^*, b^*). Also V^* satisfies (5.6) on (a^*, b^*). Let $I(x) = \frac{\sigma^2(x)}{2} V^{*''}(x) + \mu(x) V^{*'}(x) - \alpha V^*(x) + C(x)$ for each x. It remains to verify that $I(x) \geq 0$ outside (a^*, b^*). When $x < a^*$, by (5.28) we have $I(x) = \int_{a^*}^{x} [C'(u) + (\alpha - \mu'(u))]\, du$. But $C'(u) + (\alpha - \mu'(u)) < 0$ on $(-\infty, a^*)$ since $a^* \leq \theta_0$ as in Proposition 5.4. Thus, $I(x) > 0$ for $x < a^*$. Similarly, for $x > b^*$,

$I(x) = \int_{b^*}^{x} [C'(u) - (\alpha - \mu'(u))]\, du > 0$ since $b^* > \beta_0$ as described in the proof of Proposition 5.4. Consequently, V^* satisfies the HJB equation (1.9) and V^* is a C^2 function. Then by the verification lemma, $V^*(x) \leq V(x)$ for all x. Therefore, $V^*(x) = V(x)$ for all x and $\{X_x^*(t) : t \geq 0\}$ is an optimal process. This completes the proof. □

6. Bounded cost functions. The purpose of this section is to identify a new optimal strategy for a class of bounded cost functions and also to show that the value function is a C^1 function but it fails to be a C^2 function regardless of the smoothness of $\mu(\cdot)$, $\sigma(\cdot)$, and $C(\cdot)$.

Consider a bounded cost function $C(\cdot)$ which satisfies the basic assumption (1.6). Let $\|C\|_\infty = \sup_{\mathbb{R}} |C(x)|$ and using $A(t) \equiv 0$ for all t in (1.1) and in (1.3), we obtain

$$0 < V(x) \leq J(x,0) = E \int_0^\infty e^{-\alpha t} C(X_x(t))\, dt \leq \frac{\|C\|_\infty}{\alpha}, \tag{6.1}$$

where $V(\cdot)$ is the value function defined in (1.8). Therefore, if the value function is a C^1 solution of the HJB equation (1.9), then the set $\{x : |V'(x)| = 1\}$ should have finite Lebesgue measure and with the structure of our cost function, we expect the jump set $\{x : |V'(x)| = 1\}$ to be bounded. If the condition $|C'(x)| \leq \alpha - \mu'(x)$ holds for all x, then our results in section 4 remain valid, the set $\{x : |V'(x)| = 1\}$ is empty, and the zero control policy is an optimal strategy. But in the other cases, we cannot apply the results in section 5 since assumptions (5.1) and (5.2) imply that $C(\cdot)$ is unbounded.

In this section, we develop an interesting new optimal strategy when $|C'(x)| - (\alpha - \mu'(x))$ take large positive values on a large compact set. In addition to (1.4)–(1.6), we make the following assumptions.

(6.2) (i) Assume $\mu(\cdot)$ is an odd function and $\sigma(\cdot)$ and $C(\cdot)$ are even functions and $\lim_{x\to+\infty} C(x)$ is finite.

(6.3) (ii) Let $h(x) = \frac{C'(x)}{\alpha - \mu'(x)}$. There exists two points β_0 and δ_0 such that $0 < \beta_0 < \delta_0$, $0 < h(x) < 1$ on $(0, \beta_0) \cup (\delta_0, \infty)$ and $h(x) > 1$ on (β_0, δ_0). Also h is decreasing on the interval (δ_0, ∞).

(iii) There exist two points p_0 and q_0 which satisfy $\beta_0 < p_0 < q_0 < \delta_0$,

$$C(p_0) > \alpha p_0 - \mu(p_0) + \frac{\left(1 + 2\int_0^{\beta_0} \frac{C(r)}{\sigma^2(r)}\, dr\right)}{\int_0^{\beta_0} \frac{2}{\sigma^2(r)}\, dr} \tag{6.4}$$

and

$$\int_{p_0}^{q_0} \frac{2}{\sigma^2(x)} \int_{p_0}^{x} \psi(u)\, du\, dx > 1, \tag{6.5}$$

where $\psi(x) = C'(x) - (\alpha - \mu'(x))$ as in section 5.

With assumption (6.2), the value function will be an even function. By (6.3), we observe that $|C'(x)| > \alpha - \mu'(x)$ on $(-\delta_0, -\beta_0) \cup (\beta_0, \delta_0)$. The technical conditions (6.4) and (6.5) guarantee that the interval (β_0, δ_0) is quite large and also the function $\psi(x)$ takes large positive values on (β_0, δ_0). We intend to prove the existence of a feedback-type optimal strategy related to two points a^* and b^* so that $0 < \beta_0 < a^* < b^*$ as described below.

(a) If the initial point x is in $[-a^*, a^*]$, then our optimal state process is a reflecting diffusion on $[-a^*, a^*]$ with instantaneous reflections at $\pm a^*$ and satisfies (5.3) and (5.4) with $a = -a^*$ and $b = a^*$.

(b) If the initial point x is in $(a^*, b^*]$, our optimal process will be an initial jump to a^* and thereafter follows the reflecting diffusion described in (a). Similarly, if x is in $[-b^*, -a^*)$, then there will be an initial jump to $-a^*$.

(6.6) (c) If the initial position x is in $(b^*, +\infty)$, zero control will be used (i.e., $A(t) \equiv 0$) until the state process reaches b^* and then jumps to a^*. Thereafter, it follows the reflecting diffusion on $[-a^*, a^*]$. Similarly, if x is in $(-\infty, -b^*)$, use zero control until the state process reaches $-b^*$, then jump to $-a^*$ and follow the reflecting diffusion on $[-a^*, a^*]$.

Remarks.

1. If the initial point is in $(-\infty, -b^*) \cup (b^*, +\infty)$, then the state process satisfies (4.1) up to the entrance time to the interval $[-b^*, b^*]$ and has infinite explosion time as noticed in Proposition 4.1.
2. When μ is identically zero and σ is a constant, an example of an optimal control problem with a bounded cost function and an optimal policy similar to that above was given in Example 4.3, Chapter VIII of [11].

LEMMA 6.1. *For each $r > \beta_0$, there is a unique bounded solution $W_r(\cdot)$ to the boundary value problem*

$$\frac{\sigma^2(x)}{2} W_r''(x) + (\sigma(x)\sigma'(x) + \mu(x)) W_r'(x) - (\alpha - \mu'(x)) W_r(x) + C'(x) = 0 \tag{6.7}$$

for all $x \geq r$, and

$$W_r(r) = 1, \qquad \lim_{x \to \infty} W_r(x) = 0. \tag{6.8}$$

Furthermore, $W_r(\cdot)$ satisfies the following conditions.

(i) *$0 < W_r(x) \leq M_r$ for all $x \geq r$, where $M_r = \max\{1, \sup_{[r,\infty)} h(x)\}$ and h is as in* (6.3).

(ii) *There is a point $\eta_r > \delta_0$ such that $W_r(\cdot)$ is decreasing on $[\eta_r, \infty)$.*

(iii) *$W_r(x)$ and its derivative $W_r'(x)$ are jointly continuous with respect to (r, x) on the set $\{(r, x) : \beta_0 < r < x\}$.*

(iv) *Let p_0 and q_0 be as in* (6.5). *Then there exists a unique point b_0 such that $p_0 \leq b_0 < \delta_0$, and the corresponding function W_{b_0} satisfies $W_{b_0}'(b_0) = 0$ in addition to* (6.7) *and* (6.8) *above.*

Proof. We sketch the proof since it mostly depends on elementary calculus. We fix $r > \beta_0$. To obtain a solution W_r satisfying (6.7) and (6.8), we consider a sequence $\{W_N\}$ as follows. For each $N > r$, let W_N be the solution of the differential equation (6.7) on (r, N) with the boundary conditions $W_N(r) = 1$, $W_N(N) = 0$. Let $M_r = \max\{1, \sup_{[r,\infty)} h(x)\}$. Using (6.7) and by elementary calculus, $W_N(x) - M_r$ has no positive local maxima on (r, N). Similarly, W_N has no negative local minima on (r, N). Thus, $0 \leq W_N(x) \leq M_r$ on $[r, N]$. If $r < N_1 < N_2$, it follows that $0 \leq W_{N_1}(x) < W_{N_2}(x) \leq M_r$ for all x in $(r, N_1]$. Hence, for a fixed x, $\{W_N(x)\}$ is an increasing sequence in N and the limit $W_r(x) = \lim_{N \to \infty} W_N(x)$ exists. Thus, $0 < W_r(x) \leq M_r$ for all $x \geq r$, $W_r(r) = 1$. W_r also satisfies (6.7) and this can be verified as in the proof of Lemma 4.2(ii). This yields part (i).

Next we derive part (ii) and use it to obtain $\lim_{x \to \infty} W_r(x) = 0$. First, we show that W_r has no local minima on (δ_0, ∞). Suppose that W_r has a local minimum at z_0, where $z_0 > \delta_0$, then we can find a large enough N so that W_N has a local minimum

at y_1, a local maximum at y_2, $\delta_0 < y_1 < y_2 < N$ and $W_N(y_1) < W_N(y_2)$. Since W_N satisfies (6.7), we obtain that $h(y_1) \le W_N(y_1)$ and $W_N(y_2) \le h(y_2)$. This implies $h(y_1) < h(y_2)$ and it contradicts (6.3). Thus, W_r has no local minima on $(\delta_0, +\infty)$ and we deduce that there is a point $\eta_r \ge \delta_0$ so that W_r is monotone on $[\eta_r, +\infty)$. Hence, the limit $L \equiv \lim_{x\to+\infty} W_r(x)$ exists and $0 \le L \le M_r$. Our next step is to show that $L \le 1$. Suppose that $L > 1$. Then we can find a large N so that W_N has a local maximum at a point $z_1 > \eta_r$ and $W_N(z_1) > 1$. Using (6.7) for W_N at z_1 we obtain $C'(z_1) > (\alpha - \mu'(z_1))$ and this contradicts $z_1 > \eta_r > \delta_0$. Hence, $L \le 1$. Next, we show that W_r is decreasing on $[\eta_r, +\infty)$. We already know that it is monotone on $[\eta_r, +\infty)$. Suppose that W_r is increasing on $[\eta_r, +\infty)$. Since $L \le 1$ and $W_r(r) = 1$, there is a point $\xi > r$ such that $W_r(\xi) = \inf_{[r,+\infty)} W_r(x)$. Clearly $W_r(\xi) < 1$, and we let $r_1 = \inf\{a \ge r : W_r(x) < 1 \text{ on } (a, \xi)\}$. Thus $r_1 \ge r$, $W_{r_1}(x) \equiv W_r(x)$ for $x \ge r_1$ and $W_r'(r_1) \le 0$. Integrating (6.7), we obtain $\mu(\xi)W_r(\xi) + C(\xi) = \frac{\sigma^2(r_1)}{2} W_r'(r_1) + \mu(r_1) + C(r_1) + \alpha \int_{r_1}^{\xi} W_r(u)\,du$. This yields $\int_{r_1}^{\xi} \psi(u)\,du < 0$, where ψ is as in (6.5). Since $\beta_0 < r \le r_1$, this implies that $\xi > \delta_0$. This is a contradiction since W_r cannot have local minima on $(\delta_0, +\infty)$. Thus W_r is decreasing on $[\eta_r, +\infty)$ and part (ii) follows. To show that $\lim_{x\to+\infty} W_r(x) = 0$, we integrate (6.7) and use $W_r'(x) \le 0$ on $[\eta_r, +\infty)$ to obtain $\|C\|_\infty \ge C(x) \ge (\frac{\sigma^2(r)}{2} W_r'(r) + \mu(r) + C(r)) + \alpha \int_r^x W_r(u)\,du$. Thus $\int_r^{+\infty} W_r(u)\,du$ is convergent and $\lim_{x\to+\infty} W_r(x) = 0$. Hence, W_r satisfies (6.7) and (6.8).

To show the uniqueness of the solution to (6.7) and (6.8), let $r > \beta_0$ be fixed. Assume W_1 and W_2 are two solutions of (6.7) and (6.8). Introduce $U(x) = W_1(x) - W_2(x)$. Then $U(r) = 0$, $\lim_{x\to\infty} U(x) = 0$ and U is a bounded solution of the homogeneous, differential equation associated with (6.7). By elementary calculus, U is identically zero and $W_1(x) = W_2(x)$ for all $x \ge r$.

To prove (iii), take (r, x) so that $\beta_0 < r < x$. Pick r_1 and r_2 so that $\beta_0 < r_2 < r_1 < r$ and the solutions $W_{r_1}(x)$ and $W_{r_2}(x)$ are not identical to each other on $[r_1, \infty)$. Let $U(x) = W_{r_1}(x) - W_{r_2}(x)$. Then U is not identically zero on $[r_1, \infty)$ and is a solution of the homogeneous differential equation associated with (6.7). Furthermore, $\lim_{x\to\infty} U(x) = 0$. By an argument similar to the uniqueness proof above, $U(x) \ne 0$ for all $x \ge r_1$. Observe that for any $r > r_1$ fixed, the unique solution of (6.7) and (6.8) can be written as $W_r(x) = W_{r_1}(x) + (\frac{1 - W_{r_1}(r)}{U(r)})U(x)$ and thus $W_r'(x) = W_{r_1}'(x) + (\frac{1-W_{r_1}(r)}{U(r)})U'(x)$ for all $x \ge r$. Hence $W_r(x)$ and $W_r'(x)$ are jointly continuous in (r, x).

To obtain part (iv), let p_0 and q_0 be as in (6.5). If $W_{p_0}'(p_0) = 0$, then we can take $b_0^* = p_0$ and the proof is complete. Thus, let us consider $W_{p_0}'(p_0) \ne 0$. We claim $W_{p_0}'(p_0) > 0$. Suppose not; then $W_{p_0}'(p_0) < 0$. This yields $W_{p_0}(x) < 1$ for all $x > p_0$, otherwise there will be a local minima at $x = \xi$ so that $\xi > p_0$ and $W_{p_0}(\xi) < 1$. By evaluating (6.7) at $x = \xi$ we obtain $h(\xi) < 1$, where h is as in (6.3). Therefore, $\xi > \delta_0$, but by the proof of part (ii) above, W_{p_0} cannot have any local minima on (δ_0, ∞). Consequently, $W_{p_0}(x) < 1$ for all $x > p_0$. Next, we integrate (6.7) and use the above facts to obtain $\frac{\sigma^2(x)}{2} W_{p_0}'(x) + \mu(x) + C(x) < \frac{\sigma^2(p_0)}{2} W_{p_0}'(p_0) + \mu(p_0) + C(p_0) + \alpha \int_{p_0}^{x} W_{p_0}(u)\,du$ for all $x > p_0$ and consequently $\frac{\sigma^2(x)}{2} W_{p_0}'(x) + \int_{p_o}^{x} \psi(u)\,du < \frac{\sigma^2(p_0)}{2} W_{p_0}'(p_0) < 0$. Hence $W_{p_0}(q_0) + \int_{p_0}^{q_0} \frac{2}{\sigma^2(x)} \int_x^{p_0} \psi(u)\,du\,dx \le 1$ and by part (i), $W_{p_0}(q_0) > 0$. Thus, we obtain $\int_{p_0}^{q_0} \frac{2}{\sigma^2(x)} \int_x^{p_0} \psi(u)\,du\,dx < 1$ and this contradicts (6.5). Therefore, we conclude that $W_{p_0}'(p_0) > 0$. Since $\lim_{x\to\infty} W_{p_0}(x) = 0$, it follows that there is a point $r_0 > p_0$ such that $W_{p_0}(r_0) = 1$ and $W_{p_0}(x) < 1$ for all $x > r_0$. Hence, by the uniqueness of solutions to (6.7) and (6.8), $W_{p_0}(x) \equiv W_{r_0}(x)$ for all $x \ge r_0$. Now following an argument similar to the proof of Lemma 5.1 and using parts (i)–(iii), we

let $b_0 = \sup\{r : r \in [p_0, r_0] \text{ and } W_r'(r) > 0\}$, and obtain $W_{b_0}'(b_0) = 0$. Since $h(x) < 1$ for $x > \delta_0$, using the proofs of parts (i)–(iii), one can show that $W_r'(r) < 0$ for $r \geq \delta_0$ and thus $b_0 < \delta_0$ as claimed in part (iv).

To prove the uniqueness of b_0, assume that $\tilde{b}_0$ also satisfies part (iv). Without loss of generality, let $\tilde{b}_0 < b_0$. Then $\beta_0 < p_0 \leq \tilde{b}_0 < b_0 < \delta_0$. Next, we extend the corresponding functions W_{b_0} and $W_{\tilde{b}_0}$ to (β_0, ∞) as the solutions of (6.7) on (β_0, ∞). Then W_{b_0} has a local maxima at $x = b_0$. If $x = x_1$ is a local minimum for W_{b_0}, then by (6.7) we obtain $\psi(x_1) < 0$. Thus by (6.3), W_{b_0} cannot have local minima on $[\beta_0, \delta_0]$. Similarly, $W_{\tilde{b}_0}$ has a local maximum at $\tilde{b}_0$ and no local minima on $[\beta_0, \delta_0]$. Hence, W_{b_0} is increasing on $[\tilde{b}_0, b_0]$, $W_{\tilde{b}_0}$ is decreasing on $[\tilde{b}_0, b_0]$ and consequently W_{b_0} and $W_{\tilde{b}_0}$ meet at a point z in $(\tilde{b}_0, b_0)$. Now let $U(x) = W_{b_0}(x) - W_{\tilde{b}_0}(x)$ on (β_0, ∞). $U(x)$ is a solution of the homogeneous equation associated with (6.7), $U(z) = 0$ and $\lim_{x\to\infty} U(x) = 0$. Therefore, as in the proof of the uniqueness of solution of (6.7) and (6.8), $U(x)$ is identically zero on (β_0, ∞) and this leads to $\tilde{b}_0 = b_0$. This completes the proof. □

LEMMA 6.2. *There exist a point a^* such that $\beta_0 < a^* < \delta_0$ and a continuously differentiable function U_{a^*} defined on $[0, a^*]$ satisfying the following conditions.*

(i) *U_{a^*} satisfies the differential equation* (6.7) *on $(0, a^*)$.*

(ii) *$U_{a^*}(0) = 0$, $U_{a^*}(a^*) = 1$, $U_{a^*}'(a^*) = 0$, and $0 < U_{a^*}(x) < 1$ on $(0, a^*)$.*

Proof. For each $p > 0$, let U_p be the solution of (6.7) on $[0, p]$ with $U_p(0) = 0$ and $U_p(p) = 1$. Using (6.7), it follows that if U_p has a local minimum at a point z in $(0, p)$, then $U_p(z) > 0$. Hence, $U_p(x) > 0$ on $(0, p)$. Also, if U_p has a local maximum at a point z and $U_p(z) \geq 1$, then $C'(z) \geq (\alpha - \mu'(z))$ and thus $z \geq \beta_0$. Therefore, for each $p < \beta_0$, $0 < U_p(x) < 1$ on $(0, p)$ and $U_p'(p) \geq 0$. Suppose $U_p'(p) = 0$. Then, we extend U_p to $[0, \beta_0]$ as a solution of (6.7) and evaluate it at $x = p$. But $p < \beta_0$, therefore, $\psi(p) < 0$ and $U_p''(p) > 0$. Since $U_p(p) = 1$, this contradicts the fact $0 < U_p(x) < 1$ for $0 < x < p$. Hence, $U_p'(p) > 0$. Now consider $U_{\beta_0}(x)$. If $U_{\beta_0}'(\beta_0) = 0$, then we can take $a^* = \beta_0$, and we are done. Otherwise $U_{\beta_0}'(\beta_0) > 0$ and $0 < U_{\beta_0}(x) < 1$ on $(0, \beta_0)$. First, we estimate $U_{\beta_0}'(0)$. By integrating (6.7) and using $\mu(x)U_{\beta_0}(x) < 0$ on $(0, \beta_0)$, we obtain $U_{\beta_0}'(x) + \frac{2C(x)}{\sigma^2(x)} > \frac{\sigma^2(0)}{\sigma^2(x)} U_{\beta_0}'(0)$ for $0 < x < \beta_0$ and thus

$$1 + 2\int_0^{\beta_0} \frac{C(u)}{\sigma^2(u)}\, du > \sigma^2(0) U_{\beta_0}'(0) \int_0^{\beta_0} \frac{1}{\sigma^2(r)}\, dr. \tag{6.9}$$

We use (6.9) to show $U_{p_0}'(p_0) < 0$, where p_0 is given in (6.4) and $p_0 > \beta_0$. Consider $U_{p_0}(x)$ on $[0, p_0]$ and suppose that $\sup_{[0,p_0]} U_{p_0}(x) \leq 1$. Then $U_{p_0}(x) \leq 1 < U_{\beta_0}(x)$ for some x in (β_0, p_0), since $U_{\beta_0}'(\beta_0) > 0$ and $U_{\beta_0}(\beta_0) = 1$. Hence, U_{β_0} and U_{p_0} are two different solutions of (6.7). But $U_{\beta_0}(0) = U_{p_0}(0) = 0$, hence by the uniqueness of solutions to (6.7), $U_{\beta_0}(x) \neq U_{p_0}(x)$ for all $x > 0$ and $U_{\beta_0}'(0) \neq U_{p_0}'(0)$. Consequently, we have $0 < U_{p_0}(x) < U_{\beta_0}(x)$ for $0 < x < \beta_0$ and also $U_{p_0}'(0) < U_{\beta_0}'(0)$. By integrating (6.7) for U_{p_0}, we obtain $\frac{\sigma^2(p_0)}{2} U_{p_0}'(p_0) + \mu(p_0) + C(p_0) = \frac{\sigma^2(0)}{2} U_{p_0}'(0) + \alpha \int_0^{p_0} U_{p_0}(r)\, dr$ and hence $\frac{\sigma^2(p_0)}{2} U_{p_0}'(p_0) < \frac{\sigma^2(0)}{2} U_{\beta_0}'(0) + \alpha p_0 - \mu(p_0) - C(p_0)$. Using (6.4) and (6.9), we obtain

$$\frac{\sigma^2(p_0)}{2} U_{p_0}'(p_0) < \frac{\left[1 + 2\int_0^{\beta_0} \frac{C(u)}{\sigma^2(u)}\, du\right]}{\int_0^{\beta_0} \frac{2}{\sigma^2(r)}\, dr} + \alpha p_0 - \mu(p_0) - C(p_0) < 0.$$

This contradicts the fact $\sup_{[0,p_0]} U_{p_0}(x) \leq 1$. Thus $\sup_{[a,p_0]} U_{p_0}(x) > 1$. Now as similar to the proof of Lemma 5.1, we take $a^* = \sup\{p : 0 < p \leq p_0 \text{ and } U_p(x) < 1$

on $(0,p)\}$. Thus U_{a^*} satisfies (6.7) on $(0,a^*)$, $U_{a^*}(0)=0$, $U_{a^*}(a^*)=1$, $U'_{a^*}(a^*)=0$, and $\beta_0 \le a^* < p_0 < \delta_0$. This completes the proof. □

LEMMA 6.3. *Let W_r be the solution of* (6.7) *and* (6.8) *as described in Lemma* 6.1. *Introduce the function $G(\cdot)$ by $G(r)=\frac{\sigma^2(r)}{2}W'_r(r)+\int_{a^*}^{r}\psi(u)\,du$ for all $r\ge a^*$, where a^* is given in Lemma* 6.2.

Then there is a point b^* such that $G(b^*)=0$ and $b^*>b_0$, where b_0 is given in Lemma 6.1(iv). *Furthermore, W_{b^*} satisfies parts* (i)–(iii) *of Lemma* 6.1, *$0<W_{b^*}(x)<1$ on $(b^*,+\infty)$, and W_{b^*} is decreasing on $[b^*,+\infty)$.*

Proof. The function G is continuous on $[a^*,\infty)$ since $W'_r(x)$ is jointly continuous in (r,x) as shown in Lemma 6.1. By the same lemma, $G(b_0)>0$ since $\beta_0<a^*<b_0<\delta_0$. Furthermore, by (6.3), $\int_{\delta_0}^{\infty}\psi(u)\,du=-\infty$ and thus there is a point $b_1>\delta_0$ such that $\int_{a^*}^{b_1}\psi(u)\,du=0$ and $\int_{a^*}^{x}\psi(u)\,du>0$ for $a^*<x<b_1$. By the proof of part (iv) of Lemma 6.1, $W'_r(r)<0$ for all $r>\delta_0$ and therefore $G(b_1)<0$. Since G is continuous, there is a point b^* such that $b_1>b^*>b_0$ and $G(b^*)=0$.

We consider W_{b^*} on $[b^*,+\infty)$. Since $W'_{b^*}(b^*)+\int_{a^*}^{b^*}\psi(u)\,du=0$ and $\int_{a^*}^{b^*}\psi(u)\,du>0$, we obtain $W'_{b^*}(b^*)<0$. But $W_{b^*}(b^*)=1$ and $\lim_{x\to\infty}W_{b^*}(x)=0$. Let us show that W_{b^*} is decreasing on $[b^*,+\infty)$. Suppose not; then there exist a local minimum at a point $x=\xi_1$ and a local maximum at $x=\xi_2$ so that $\xi_2>\xi_1$ and $W_{b^*}(\xi_2)>W_{b^*}(\xi_1)$. Using (6.7), we obtain $\frac{C'(\xi_1)}{(\alpha-\mu'(\xi_1))}\le W_{b^*}(\xi_1)<W_{b^*}(\xi_2)\le\frac{C'(\xi_2)}{(\alpha-\mu'(\xi_2)}$. Hence, $h(\xi_1)<h(\xi_2)$. Also by integrating (6.7), we obtain $C(\xi_1)+\mu(\xi_1)<\mu(\xi_1)W_{b^*}(\xi_1)+C(\xi_1)=\frac{\sigma^2(b^*)}{2}W'_{b^*}(b^*)+\mu(b^*)+C(b^*)+\int_{b^*}^{\xi_1}\alpha W_{b^*}(u)\,du<C(b^*)+\mu(b^*)+\alpha(\xi_1-b^*)$. Therefore, $\int_{b^*}^{\xi_1}\psi(u)\,du<0$ and hence $\xi_1>\delta_0$, where δ_0 is given in (6.3). Thus $\delta_0<\xi_1<\xi_2$ and we have $h(\xi_1)<h(\xi_2)$. By (6.3), this is a contradiction. Hence W_{b^*} is decreasing on $(b^*,+\infty)$, $\lim_{x\to\infty}W_{b^*}(x)=0$ and $W'_{b^*}(b^*)<0$. This completes the proof. □

LEMMA 6.4. *There exists two points $0<a^*<b^*$ and a continuous function W on $[0,\infty)$ such that*

(i) *$\beta_0<a^*<p_0<b^*$, $W(0)=0$, $W(x)=1$ on $[a^*,b^*]$;*

(ii) *W is C^1 everywhere except at $x=b^*$, W satisfies* (6.7) *on $(0,a^*)\cup(b^*,+\infty)$, $W'(a^*)=0$, and $0<W(x)<1$ on $(0,a^*)\cup(b^*,+\infty)$; also the limits $W'(b^*-)$ and $W'(b^*+)$ exist and are finite;*

(iii) *let*

$$P(x)=\frac{1}{2}(\sigma^2(x)W'(x)-\sigma^2(0)W'(0))+\mu(x)W(x)+C(x) \qquad (6.10)$$
$$-\alpha\int_0^x W(u)\,du \quad \text{for } x\ne b^*;$$

then $P(x)\ge 0$ for all $x>0$, and $x\ne b^*$; furthermore, $P(b^*-)$ and $P(b^*+)$ are finite, $P(b^*-)>0$ and $P(b^*+)=0$.

Proof. Let the points a^* and b^* and the corresponding functions U_{a^*} and W_{b^*} be as described in Lemmas 6.2 and 6.3. Introduce the function W by

$$W(x)=U_{a^*}(x)I_{[0,a^*]}(x)+I_{(a^*,b^*)}(x)+W_{b^*}(x)I_{[b^*,\infty)}(x),$$

where I_A represents the indicator function of the set A. Then parts (i) and (ii) follows from Lemmas 6.2 and 6.3. Notice that $W'(b^*-)=0$ and $W'(b^*+)=W'_{b^*}(b^*+)<0$. It remains to verify part (iii).

Observe that $P'(x)=0$ on $(0,a^*)\cup(b^*,+\infty)$. Since $P(0)=0$, and using part (ii), we obtain $P(x)=0$ on $[0,a^*]$. Also $P(a^*)=0$ and this implies

that $C(a^*) + \mu(a^*) = \frac{\sigma^2(0)}{2} W'(0) + \alpha \int_0^{a^*} W(u)\,du$. Using this together with the fact $W(x) = 1$ on $[a^*, b^*]$ and (6.10), we obtain $P(x) = C(x) + \mu(x) - \alpha(x - a^*) - (C(a^*) + \mu(a^*)) = \int_{a^*}^{x} \psi(u)\,du$ for $a^* \le x < b^*$. But $\int_{a^*}^{x} \psi(u)\,du > 0$ for $a^* < x < b^*$ from the proof of Lemma 6.3. Hence $P(x) > 0$ on $[a^*, b^*)$ and $P(b^*-) > 0$. On the set $(b^*, +\infty)$, $P'(x) = 0$, since W_{b^*} satisfies (6.7). By a computation similar to above, we obtain

$$\begin{aligned} P(b^*+) &= \frac{\sigma^2(b^*)}{2} W'_{b^*}(b^*) + C(b^*) + \mu(b^*) \\ &\quad - \left(\frac{\sigma^2(0)}{2} W'(0) + \alpha \int_0^{a^*} W(u)\,du \right) - \alpha(b^* - a^*) \\ &= \frac{\sigma^2(b^*)}{2} W'_{b^*}(b^*) + C(b^*) + \mu(b^*) - (C(a^*) + \mu(a^*)) - \alpha(b^* - a^*) \\ &= \frac{\sigma^2(b^*)}{2} W'_{b^*}(b^*) + \int_{a^*}^{b^*} \psi(u)\,du. \end{aligned}$$

Thus $P(b^*+) = 0$ by using Lemma 6.3 and consequently $P(x) = 0$ for all $x > b^*$. This completes the proof. □

We construct an even function $F(\cdot)$ on $\mathbb{R}$ by

$$F(x) = \begin{cases} \dfrac{\sigma^2(0)}{2\alpha} W'(0) + \displaystyle\int_0^x W(u)\,du & \text{for } x \ge 0, \\ F(-x) & \text{for } x < 0, \end{cases} \tag{6.11}$$

where W is given in Lemma 6.4. Next, we describe the main theorem of this section.

THEOREM 6.5. *Assume* (1.4)–(1.6) *and* (6.2)–(6.5). *Let F be the function defined in* (6.11). *Then*

(i) *F is a bounded C^1-function which is C^2 everywhere except at $x = \pm b^*$ and F satisfies the HJB equation* (1.9) *everywhere except at the points $x = \pm b^*$;*

(ii) *$F(x) = V(x)$ for all x, where V is the value function given in* (1.8)*; let $0 < a^* < b^*$ be as in Lemma* 6.4*, then the strategy described in* (6.6) *is optimal.*

Proof. For part (i), we use Lemma 6.4. For $x \ge 0$, $F'(x) = W(x)$ by (6.11) and, therefore, $0 < F'(x) \le 1$, F is C^1 on $[0, \infty)$ and C^2 everywhere except at $x = b^*$. Furthermore, $\frac{1}{2}\sigma^2(x)F''(x) + \mu(x)F'(x) - \alpha F(x) + C(x) = P(x)$, where $P(\cdot)$ is given in Lemma 6.4 and $P(x) \ge 0$, $P(x) = 0$ on $[0, a^*] \cup (b^*, \infty)$, and $F'(x) = 1$ on $[a^*, b^*]$. Therefore, F satisfies (1.9) on $[0, \infty)$. By the proof of Lemma 6.1, $\int_0^\infty W(u)\,du$ is convergent, and hence F is bounded on $[0, \infty)$. Since F is an even function, part (i) follows.

For part (ii), we can apply Proposition 2.3 in section 2 to conclude that $F(x) \le V(x)$ for all x, where V is the value function in (1.8). To prove $F \ge V$, we begin with the strategy (6.6) with a^* and b^* as in Lemma 6.4. If the initial point x is in $[-a^*, a^*]$, then the strategy in (6.6) yields a reflecting diffusion on $[-a^*, a^*]$ which satisfies (5.3) and (5.4). Since F satisfies (5.6) and (5.7) on $[-a^*, a^*]$, F can be represented by (5.5) with respect to the reflecting diffusion on $[-a^*, a^*]$ and $F(x) \ge V(x)$ for all x in $[-a^*, a^*]$. If x is in $[a^*, b^*]$, then there is an initial jump to a^* and thereafter it follows a reflecting diffusion on $[-a^*, a^*]$. Since $W(x) = 1$ on $[a^*, b^*]$, we have $F(x) = F(a^*) + (x - a^*)$ and $F(x) \ge V(x)$ on $[a^*, b^*]$. Similarly, $F(x) \ge V(x)$ on $[-b^*, -a^*]$. If $x > b^*$, our candidate $\{X_x^*(t) : t \ge 0\}$ for an optimal process satisfies

(4.1) up to the first entrance time τ_{b^*} of $[-b^*, b^*]$ and at the time τ_{b^*} it will jump to a^* using the control process $A(\cdot)$ in (1.1). Thereafter, it remains in $[-a^*, a^*]$ as a reflecting diffusion satisfying (5.3) and (5.4). The process $X_x^*(\cdot)$ is admissible with respect to the stopping times $\{\tau_n\}$ defined is (4.3) and the explosion time of $X_x^*(\cdot)$ is infinity, as verified in Proposition 4.1. Also $P_x[\tau_{b^*} < \infty] = 1$, and this can be established by standard methods (see [17, p. 345]). To verify that the pay-off from $X_x^*(\cdot)$ is indeed $F(x)$, we apply Itô's lemma to $F(X_x^*(t))e^{-\alpha t}$ and obtain

$$F(x) = F(b^*)E_x(e^{-\alpha\tau_{b^*}}) + E_x\left[\int_0^{\tau_{b^*}} e^{-\alpha t}C(X_x^*(t))\,dt\right] \tag{6.12}$$
$$= F(a^*)E_x(e^{-\alpha\tau_{b^*}}) + E_x\left[\int_0^{\tau_{b^*}} e^{-\alpha t}(C(X_x^*(t))\,dt + d|A^*|(t))\right],$$

where $\{A^*(t)\}$ is the bounded variation control process associated with $X_x^*(\cdot)$ as in (1.1), and A^* satisfies $A^*(t) = 0, \quad 0 \le t < \tau_{b^*}, A^*(\tau_{b^*}) = -(b^* - a^*)$, and for $t > \tau_{b^*}$, $A^*(t) = A^*(\tau_{b^*}) + L_{-a^*}(t) - L_{a^*}(t)$, where L_{-a^*} and L_{a^*} are the local time processes at $-a^*$, and a^* is as described in (5.4). Since $F(a^*)$ can be represented by (5.5), (6.12) leads to $F(x) = E_x \int_0^\infty e^{-\alpha t}(C(X_x^*(t))\,dt + d|A^*|(t))$. Hence, $F(x) \ge V(x)$ for all $x \ge b^*$. Similar argument works when $x < -b^*$. Consequently, $F(x) = V(x)$ for all x and the strategy described in (6.6) is optimal. This completes the proof. □

Next, we give an explicit example which satisfies all our assumptions.

Example 6.6. Let $\mu(x) = -\theta x$ and $\sigma(x) = \sigma_0$, where θ and σ_0 are positive constants. Hence the zero-control in (1.1) yields on Ornstein–Uhlenbeck process. To construct a twice continuously differentiable even cost function $C(\cdot)$, we introduce two points $p_0 > 0$ and $q_0 > 0$ by $p_0 = \frac{3}{2}(\alpha+\theta) + \frac{\sigma_0}{\sqrt{\alpha+\theta}}$ and $q_0 = p_0 + \frac{\sigma_0}{\sqrt{\alpha+\theta}}$.

We take $C(x) = x^2$ on $[0, q_0]$. On $[q_0, +\infty]$, we construct $C(x)$ so that $C(\cdot)$ is C^2 on $[0,\infty), C'(x)$ is increasing on $[0, q_0 + 1]$, $C'(x)$ is nonnegative and decreasing on $[q_0 + 1, \infty)$, and $\int_{q_0}^\infty C'(u)\,du$ is finite. We extend $C(x)$ to $\mathbb{R}$ as an even function and consequently $C(\cdot)$ is a bounded cost function. Assumption (6.2) is obvious. In (6.3), $h(x) = \frac{C'(x)}{(\alpha+\theta)}$ for $x > 0$, $\beta_0 = \frac{(\alpha+\theta)}{2}$, and $\delta_0 > q_0 + 1$. Thus (6.3) is also satisfied. Let p_0 and q_0 be as introduced above. Condition (6.4) can be reduced to $(p_0 - \frac{(\alpha+\theta)}{2})^2 > \frac{\sigma_0^2}{(\alpha+\theta)} + \frac{(\alpha+\theta)^2}{3}$ and it is satisfied by our choice of p_0. Condition (6.5) can be reduced to $\frac{1}{3}(q_0^3 + 2p_0^3) > \frac{\sigma_0^2}{2} + p_0^2 q_0 + \frac{(\alpha+\theta)}{2}(q_0 - p_0)^2$ and this inequality is also satisfied by our choice of p_0 and q_0.

Hence, there exist two points $a^* > 0$ and $b^* > 0$ and an optimal strategy of the type described in (6.6). Moreover, the value function fails to be C^2 at the points $\pm b^*$.

7. Variance control. Here we allow the controller to control the diffusion coefficient in addition to the bounded variation control process. We assume that the controlled state process is a weak solution of the stochastic differential equation

$$X_x(t) = x + \int_0^t \mu(X_x(s-))\,ds + \int_0^t u(s)\,dW(s) + dA(x), \tag{7.1}$$

where x is in $\mathbb{R}, \{W(t) : t \ge 0\}$ is the Brownian motion adapted to a right continuous filtration $\{\mathfrak{F}_t : t \ge 0\}$ on a probability space $\{\mathfrak{F}_t\}$ and the bounded variation control process $A(\cdot)$ is $\{\mathfrak{F}_t\}$-adapted and satisfies all the conditions described below (1.1) in section 1. In addition, the process $\{u(t) : t \ge 0\}$ is $\{\mathfrak{F}_t\}$-adapted, and for each $t > 0, u(t)$ belongs to a control set $D(X_x(t-))$, where $D(y) \subseteq (0,\infty)$ for each y, and the collection of control sets $\mathcal{D} = \{D(y) : y \text{ is in } R\}$ is a priori known to the controller.

A process $X_x(\cdot)$ is *admissible* if it satisfies (7.1), $u(t)$ belongs to $D(X_x(t-))$ for each $t > 0$, and if there exists an increasing sequence of stopping times $\{\tau_n\}$ such that $\lim_{n\to\infty} \tau_n = +\infty$ and

$$\text{(7.2)} \quad \text{(i)} \qquad E_x \int_0^{T\wedge\tau_n} [|\mu(X_x(s-))| + (u(s))^2]\, ds < \infty \quad \text{for each } T > 0 \text{ and}$$

$$\text{(ii)} \qquad \lim_{n\to\infty} E_x\big[|X_x(\tau_n)| e^{-\alpha\tau_n} I_{[\tau_n<\infty]}\big] = 0.$$

The control problem addressed in this section is to minimize the function

$$J(x,u,A) = E_x \int_0^\infty e^{-\alpha t}[C(X_x(t))\, dt + d|A|(t)] \tag{7.3}$$

over all available state processes $X_x(\cdot)$ satisfying (7.1) and (7.2). The cost function $C(\cdot)$ satisfies assumption (1.6).

Analogous to section 1, we call $((\Omega, \mathfrak{F}, P), \{\mathfrak{F}_t\}, W(\cdot), X_x(\cdot), u(\cdot), A(\cdot))$ an admissible control system if (i) $X_x(\cdot)$ is a weak solution to (7.1) which also satisfies (7.2) with respect to control processes $u(\cdot)$, $A(\cdot)$, and (ii) $J(x,u,A)$ described in (7.3) is finite. We let

$$\Sigma(x) = \{(u(\cdot), A(\cdot)) : \text{ there exists an admissible state process } X_x(\cdot) \text{ which satisfies (7.1) corresponding to control processes } u(\cdot) \text{ and } A(\cdot)\}.$$

The value function $V(\cdot)$ is defined by

$$V(x) = \inf_{\Sigma(x)} J(x,u,A). \tag{7.4}$$

The key to the derivation of the optimal strategies in this section is the function $\sigma(\cdot)$ defined by

$$\sigma(y) = \inf\{u : u \in D(y)\}. \tag{7.5}$$

We make the following assumptions on the collection of control sets $\mathcal{D}$:

(7.6) (i) for each y, $\sigma(y)$ belongs to $D(y)$;

(7.7) (ii) the function $\sigma(\cdot)$ satisfies conditions (1.4) and (1.5).

In contrast to many articles in stochastic control, the control sets $D(y)$ need not be bounded. We assume condition (1.6) related to the cost function $C(\cdot)$. We need the following additional assumption in this section:

$$\text{the function } h(x) = \frac{C'(x)}{\alpha - \mu'(x)} \text{ is monotone increasing on } \mathbb{R}. \tag{7.8}$$

Condition (7.8) guarantees the convexity of the value function, and hence the use of minimal variance is optimal. The HJB equation related to this problem is given by

$$\min\left\{ \inf_{u\in\mathcal{D}(x)} \frac{u^2}{2} V''(x) + \mu(x)V'(x) - \alpha V(x) + C(x),\ 1 - |V'(x)| \right\} = 0. \tag{7.9}$$

Our next lemma will extend the verification results of section 2.

LEMMA 7.1. *Let $\sigma(\cdot)$ be defined by* (7.5) *and let the functions μ, σ, and C satisfy* (1.4)–(1.6) *and* (7.5)–(7.7). *Let Q be a twice continuously differentiable function which*

satisfies the HJB equation (7.9) *with* $Q''(x) \geq 0$ *for all* x. *Then* $Q(x) \leq V(x)$ *for all* x, *where* V *is the value function given by* (7.4).

Proof. Since $Q'' \geq 0$, observe that $\inf_{u \in \mathcal{D}(x)} \frac{u^2}{2} Q''(x) = \frac{\sigma^2(x)}{2} Q''(x)$, where $\sigma(\cdot)$ is given in (7.5). Since Q satisfies (7.9), it also satisfies (1.9) with the diffusion coefficient $\sigma(\cdot)$. For any admissible process $X_x(\cdot)$ with controls $u(\cdot)$ and $A(\cdot)$, we observe that

$$\frac{u(t)^2}{2} Q''(X_x(t-)) \geq \frac{\sigma^2(X_x(t-))}{2} Q''(X_x(t-)) \quad \text{for all } t \geq 0. \tag{7.10}$$

Using (7.10) in the proofs of Lemma 2.1 and Corollary 2.2, the conclusion follows. □

Remark. If $\sigma(\cdot)$ is as in (7.5) and Q is a C^2 function which satisfies (1.9) and $Q''(x) \geq 0$ for all x, then Q also satisfies (7.9).

Next, we describe the main theorem of this section.

THEOREM 7.2. *Assume the same conditions as in Lemma* 7.1. *Then the following holds.*

(i) *If* (4.2) *holds, then the process* $Z_x(\cdot)$ *of* (4.1) *is optimal for the variance control problem* (7.4).

(ii) *If* (5.1) *and* (5.2) *holds, then the reflecting diffusion process* $X_x^*(\cdot)$ *described in the Theorem* 5.5 *is optimal for* (7.4).

Proof. To prove part (i) we use the proof of Theorem 4.3. With the aid of Lemma 7.1, it suffices to show that $F''(x) \geq 0$ for all x, where F is given in (4.15). This will follow if we prove W_∞ in (4.15) is increasing. Suppose not, then there exist $z_1 < z_2$ so that $W_\infty(z_1) > W_\infty(z_2)$. Hence, we can find large n, so that W_n has a local maximum at ξ_1 and a local minimum at ξ_2, $\xi_1 < \xi_2$, $W_n(\xi_1) > W_n(\xi_2)$, and W_n satisfies (4.8). Using (4.8), it follows that $h(\xi_1) \geq W_n(\xi_1) > W_n(\xi_2) > h(\xi_2)$. This contradicts (7.8). Hence part (i) follows.

For part (ii), we follow the proof of Theorem 5.5. Again it suffices to show that the function W described in Proposition 5.4 is monotone increasing on $[a^*, b^*]$. This can be proved similarly to part (i). Thus V^* given in (5.28) satisfies the assumptions of Lemma 7.1, and the process $X_x^*(\cdot)$ described in Theorem 5.5 is optimal for the variance control problem. This completes the proof. □

Acknowledgment. The author would like to thank an anonymous referee for a careful reading of the manuscript and for many suggestions to improve the presentation of the material.

REFERENCES

[1] A.B. ABEL AND J.C. EBERLY, *A unified model of investment under uncertainty*, Amer. Econ. Rev., 84 (1994), pp. 1369–1384.

[2] A.B. ABEL AND J.C. EBERLY, *Optimal investment with cost reversibility*, Rev. Econom. Stud., 63 (1996), pp. 581–593.

[3] L.R. ALVAREZ, *A class of solvable singular stochastic control problems*, Stoch. Stoch. Rep., 67 (1999), pp. 83–122.

[4] L.R. ALVAREZ, *Singular stochastic control, linear diffusions and optimal stopping: A class of solvable problems*, SIAM J. Control Optim., 39 (2001), pp. 1697–1710.

[5] F.M. BALDURSSON, *Singular stochastic control and optimal stopping*, Stoch. Stoch. Rep., 21 (1987), pp. 1–40.

[6] G. BERTOLA AND R. CABALLERO, *Target zones and realignments*, Amer. Econ. Rev., 27 (1992), pp. 520–536.

[7] F. BOETIUS AND M. KOHLMANN, *Connections between optimal stopping and singular stochastic control*, Stochastic Process. Appl., 77 (1998), pp. 253–281.

[8] A. Cadenillas and F. Zapatero, *Optimal central bank intervention in the foreign exchange market*, J. Econom. Theory, 87 (1999), pp. 218–242.
[9] M.B. Chiarolla and U.G. Haussmann, *The free boundary of the monotone follower*, SIAM J. Control Optim., 32 (1994), pp. 690–727.
[10] M.H.A. Davis, M.A.H. Dempster, S.P. Sethi, and D. Vermes, *Optimal capacity expansion under uncertainty*, Adv. in Appl. Probab., 19 (1987), pp. 156–176.
[11] W.H. Fleming and H.M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.
[12] P.M. Garber and L.E.O. Svensson, *The operation and collapse of fixed exchange rate regimes*, in Handbook of International Economics, K. Rogoff and G. Grossman, eds., North-Holland, Amsterdam, 1996.
[13] M. Jeanblanc-Picqué, *Impulse control method and exchange rate*, Math. Finance, 3 (1993), pp. 161–177.
[14] I. Karatzas, *A class of singular stochastic problems*, Adv. in Appl. Probab., 15 (1983), pp. 225–254.
[15] I. Karatzas and S.E. Shreve, *Connections between optimal stopping and singular control* I. *Monotone follower problems*, SIAM J. Control Optim., 22 (1984), pp. 856–877.
[16] I. Karatzas and S.E. Shreve, *Connections between optimal stopping and singular control* II. *Reflected follower problems*, SIAM J. Control Optim., 23 (1985), pp. 433–451.
[17] I. Karatzas and S.E. Shreve, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, 1988.
[18] P.R. Krugman, *Target zones and exchange rate dynamics*, Quart. J. Econom., 106 (1991), pp. 669–682.
[19] L. Kruk, *Optimal policies for n-dimensional singular stochastic control problems, Part* I*: The Skrokhod problem*, SIAM J. Control Optim., 38 (2000), pp. 1603–1622.
[20] T.G. Kurtz and R.H. Stockbridge, *Existence of Markov controls and characterization of optimal Markov controls*, SIAM J. Control Optim., 36 (1998), pp. 609–653.
[21] J. Ma, *On the principle of smooth-fit for a class of singular stochastic control problems for diffusions*, SIAM J. Control Optim., 30 (1992), pp. 975–999.
[22] A.S. Manne, *Capacity expansion and probabilistic growth*, Econometrica, 29 (1961), pp. 632–649.
[23] P.A. Meyer, *Un cours sur les integrales stochastiques*, in Seminaire de Probabilities X, Lecture Notes in Math. 511, Springer-Verlag, New York, 1974.
[24] M. Miller and L. Zhang, *Optimal target zones: How an exchange rate mechanism can improve upon discretion*, J. Econom. Dynam. Control, 20 (1996), pp. 1641–1660.
[25] G. Mundaca and B. Øksendal, *Optimal stochastic intervention control with application to the exchange rate*, J. Math. Econom., 29 (1998), pp. 225–243.
[26] M. Musiela and M. Rutkowski, *Martingale Methods in Financial Modelling*, Springer-Verlag, New York, 1997.
[27] D. Ocone and A. Weerasinghe, *Degenerate variance control of a one-dimensional diffusion*, SIAM J. Control Optim., 38 (2000), pp. 1–24.
[28] M.H. Protter and H.F. Weinberger, *Maximum Principles in Differential Equations*, Springer-Verlag, New York, 1984.
[29] L.C.G. Rogers and D. Williams, *Diffusions, Markov Processes and Martingales, Vol.* 2*, Itô Calculus*, Wiley, New York, 1987.
[30] S.M. Ryan, *Capacity expansion for random exponential growth with lead times*, Manage. Sci., 50 (2004), pp. 740–748.
[31] H.M. Soner and S.E. Shreve, *Regularity of the value function for a two dimensional singular stochastic control problem*, SIAM J. Control Optim., 27 (1989), pp. 876–907.
[32] L. Svensson, *The term structure of interest rate differentials in a target zone*, J. Monet. Econom., 28 (1991), pp. 87–116.
[33] J.A. van Mieghem, *Capacity portfolio investment and hedging: Review and new directions*, Manuf. Service Oper. Manag., 5 (2003), pp. 269–302.
[34] A. Weerasinghe, *Stationary stochastic control for Itô processes*, Adv. in Appl. Probab., 34 (2002), pp. 128–140.
[35] A. Weerasinghe, *Minimizing infinite time horizon discounted cost with mean, variance and bounded variation controls*, SIAM J. Control Optim., 42 (2003), pp. 1395–1415.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 2, pp. 418–443

GRAPH TOPOLOGIES, GAP METRICS, AND ROBUST STABILITY FOR NONLINEAR SYSTEMS*

W. BIAN† AND M. FRENCH†

Abstract. Graph topologies for nonlinear operators which admit coprime factorizations are defined w.r.t. a gain function notion of stability in a general normed signal space setting. Several metrics are also defined and their relationship to the graph topologies are examined. In particular, relationships between nonlinear generalizations of the gap and graph metrics, Georgiou-type formulae, and the graph topologies are established. Closed loop robustness results are given w.r.t. the graph topology, where the role of a coercivity condition on the nominal plant is emphasized.

Key words. gap metric, graph metric, graph topology, robust stability, nonlinear systems

AMS subject classifications. 93D09, 93D25, 93C10

DOI. 10.1137/S0363012903421200

1. Introduction.

The theory of coprime factorizations of linear signal operators is well known to be a significant tool in the study of robustness of stability for linear feedback systems and has been extensively studied (see [5, 16, 20]). Perturbations to normalized coprime factors form a good description of physically realistic deviations from nominal models, since they allow a unified treatment of both low and high frequency uncertainties [8]. In the linear theory, it is well known that the graph topology is the appropriate topological description for studying robustness of stability and that coprime factor perturbations can be used to induce the graph topology. Furthermore, the graph topology is metrizable, and both the gap metric [3, 21] and the graph metric [20] provide suitable metrizations, the former being more suitable for calculations by standard H^∞ optimizations, (although both metrics are topologically equivalent) [5, 16, 21]. There is thus a rich set of equivalences between the notions of coprime factorizations, gap/graph metrics and topologies and their attendant robust stability theorems. Moreover, this framework is a cornerstone of modern robust linear control theory.

Given the richness and importance of this framework in the linear setting, it is natural to seek extensions to the nonlinear case, and to alternative signal spaces. Indeed, by adopting a notion of stability corresponding to the existence of a linear gain (typically either in an L^2 or L^∞ setting), a number of authors have previously considered a nonlinear theory of coprime factorization. Here we highlight three contributions of particular relevance to the context of this paper. In [18], Verma defined a notion of coprime factorization for nonlinear mappings and presented a stability result for a nonlinear system. In [2], Anderson, James, and Limebeer generalized the linear theory of normalized coprime factor robustness optimization to the case of affine input nonlinear systems and presented an optimal robustness margin. In [10], a new definition of "normalized" was introduced for left representation for the graph of a nonlinear system and different gap metrics were studied. Many further pointers

*Received by the editors January 15, 2003; accepted for publication (in revised form) November 26, 2004; published electronically August 31, 2005.
http://www.siam.org/journals/sicon/44-2/42120.html
†School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, United Kingdom (wb@ecs.soton.ac.uk, mcf@ecs.soton.ac.uk).

to a growing literature on nonlinear coprime factorization can be found in the monograph [14] and the references therein.

On the other hand, the gap metric has also been generalized into a nonlinear setting in a fundamental contribution by Georgiou and Smith, [7]. In further recent papers [1, 22, 10], generalizations of Vinnicombe's ν-gap metric [21] to nonlinear operators have been considered (for linear systems the ν gap is always smaller than the gap and has sharper properties; however, the nonlinear theory does not as yet reflect these extra properties).

The purpose of this paper is three-fold.

1. To further extend the existing robust stability theory by replacing the restrictive requirement of the existence of an induced gain by the weaker requirement of the existence of a gain function; see also [7].

2. To provide the topological descriptions underpinning the convergence and robust stability notions in both the case of linear gain and gain function stability; in particular to provide a topological characterization of nonlinear gap topologies in terms of coprime factor perturbations.

3. To establish links between the nonlinear graph topologies, the recent results on the nonlinear gap metric [7], and other metrizations (e.g., graph metrics [20], and Georgiou-type formulae [4]).

In the context of the first and second items, directly related work of which the authors are aware can be found in [17], where a sufficient condition for the existence of coprime factorization of nonlinear mappings was given in the sense of IOS; we will show much more of the theory for linear gains can be extended to this more general setting. In [7], robustness of stability results was given in a gain function setting using a generalization of the gap metric. Interestingly, whilst the results in the gain function setting given in [7] implicitly define a notion of plant convergence, the underlying topology is not explicit. In particular, in contrast to the case of the linear gain, a metric was not defined, hence a topology cannot be automatically induced. One contribution of this paper is to provide the underlying topology, and to provide explicit metrizations. In the case of stability of nonlinear operators defined via a linear gain, we show that the graph metric naturally generalizes and induces the graph topology. In the more general case of gain function stability, we only show that the gap topology is stronger than the (weighted) graph topology. The converse relationship remains open. However, we do establish many other relationships and equivalences between a variety of gap and graph metrics and topologies.

An outline of this paper is as follows. Section 2 is devoted to the preliminaries, in particular known results on coprime factorization for nonlinear systems are briefly reviewed. The main results are arranged in three sections. In section 3, we define *pointwise* and *weighted* graph topologies and study the associated convergence over a general subset of signal operators admitting coprime factorizations. In section 4, we study the metrizability of the weighted graph topology. Seven gap metrics are considered. Equivalences and other relationships between the metrics and their associated topologies (including equivalence to the weighted graph topology) are presented. Finally in section 5, we apply the graph topologies to study the robust stability of nonlinear feedback systems. A summary and discussion of future work is given in section 6.

2. Background on coprime factorization. The material in this section is mostly directly based on (and straightforward generalizations of) work of previous authors [7, 17, 18, 19]. However, we need to present this material within the language of this paper and for completeness.

We let $\mathcal{U}$, $\mathcal{Y}$ be two signal spaces, respectively, representing the input and output signal spaces. These could be the spaces $L_n^\infty := L^\infty(\mathbb{R}_+, \mathbb{R}^n)$, $L_n^{\infty,e}$, L_n^p, $L_n^{p,e}$, l^p, or even a general set on which a truncation can be defined and for which any truncated domain is a normed linear space and $\sup_{\tau>0} \|T_\tau x\| < \infty$ implies $x \in \mathcal{U}_s$. In particular, for one-dimensional continuous domains, we define the truncation operator and the truncated norm for a signal u, say $u \in L_n^{\infty,e}$, by

$$(T_\tau u)(t) = \begin{cases} u(t), & t \le \tau \\ 0, & t > \tau \end{cases}, \qquad \|u\|_\tau = \|T_\tau u\|,$$

where norm of a normed space X is denoted by $\|\cdot\|_X$ or $\|\cdot\|$ if the usage is unambiguous. Note, however, that the notion of truncation and all the material in this paper equally apply to signal spaces with discrete domains, e.g., $L^\infty(\mathbb{Z}_+, \mathbb{R}^n)$, and to multidimensional domains, e.g., $L^\infty(\mathbb{R}_+^m, \mathbb{R}^n)$, under a suitably modified notion of truncation. Let $\mathcal{U}_s$, $\mathcal{Y}_s$ be the auxiliary normed subspaces which consist of all bounded signals in $\mathcal{U}$, $\mathcal{Y}$, respectively. In the case where $\mathcal{U}$ (resp., $\mathcal{Y}$) is a normed space, $\mathcal{U}_s = \mathcal{U}$ (resp., $\mathcal{Y}_s = \mathcal{Y}$). Typically, $\mathcal{U}$, $\mathcal{Y}$ are taken to be extended spaces (e.g., $L_n^{\infty,e}$), and $\mathcal{U}_s$, $\mathcal{Y}_s$ are their nonextended subspaces (e.g., L_n^∞).

The identity operator on any space Y is denoted by I_Y or I if the usage is clear. Given a matrix operator (A, B), let $(A, B)^\top$ be its transpose, that is $(A, B)^\top = \binom{A}{B}$. We also let $\mathcal{K}_\infty$ denote the set of functions $\omega : [0, \infty) \to [0, \infty)$ which are continuous, strictly increasing, and $\omega(0) = 0$, $\omega(\infty) = \infty$.

Any signal operator $P : \mathrm{Dom}(P) \to \mathcal{Y}$ is assumed to be causal and its domain is denoted by

$$\mathrm{Dom}(P) = \{u \in \mathcal{U}_s : Pu \in \mathcal{Y}_s\}.$$

It is worthwhile to observe that unstable plant operators $\hat{P}$ are often thought of as operators $\mathcal{U} \to \mathcal{Y}$ for suitably large signal spaces $\mathcal{U}$, $\mathcal{Y}$. We will only have to be interested in the relation between elements in $\mathrm{Dom}(P)$ and $\mathcal{Y}_s$ so do not consider the definition of P on the wider signal spaces. However, it should be noted that under extra assumptions such as causal extendibility [6] and for appropriate choices of signal space, the operator $P : \mathrm{Dom}(P) \to \mathcal{Y}_s$ uniquely extends to an operator $\hat{P} : \mathcal{U} \to \mathcal{Y}$, hence the topologies we will define on sets of operators $P : \mathrm{Dom}(P) \to \mathcal{Y}_s$ can be thought of as topologies on sets of operators $\hat{P} : \mathcal{U} \to \mathcal{Y}$.

Linear gains of operators $P : \mathrm{Dom}(P) \to \mathcal{Y}$ are defined by

$$\|P\| := \sup \left\{ \frac{\|Pu\|}{\|u\|} : u \in \mathrm{Dom}(P) \text{ with } \|u\| \neq 0 \right\}.$$

If P is causal, one can prove that

$$\|P\| = \sup \left\{ \frac{\|Pu\|_\tau}{\|u\|_\tau} : \tau > 0,\ u \in \mathrm{Dom}(P) \text{ with } \|u\|_\tau \neq 0 \right\}$$

which is used in [7] as the definition of linear gain. When P is a linear operator, $\|P\|$ is the induced operator norm of P. In the nonlinear setting, in contrast to linear systems, it is often the case that $\mathrm{Dom}(P) = \mathcal{U}_s$ and yet no linear gain exists. Therefore a weaker notion of stability is adopted, namely that of the existence of a gain function. The gain function of an operator P is defined by

$$\gamma(P)(r) := \sup\{\|Pu\|_\tau : \tau > 0,\ u \in \mathrm{Dom}(P) \text{ with } \|u\|_\tau \le r\} \quad \text{for } r \ge 0.$$

In the case where P is causal, we also have

$$\gamma(P)(r) = \sup\{\|Pu\| : u \in \mathrm{Dom}(P),\ \|u\| \le r\} \quad \text{for } r \ge 0.$$

We summarize elementary properties of the linear gain $\|P\|$ and the gain function $\gamma(P)$ in the following lemma.

LEMMA 2.1. *The linear gain and gain function have the following properties:*

1. $\gamma(P)(0) = 0$, *if* $P(0) = 0$,
2. $\gamma(P)(r_1) \le \gamma(P)(r_2)$, *if* $r_1 \le r_2$,
3. *For any two well-defined operators* P_1, P_2 *and any* $r > 0$, $\lambda \in \mathbb{R}$, *we have*

$$\begin{aligned} \|\lambda P_1\| = 0 &\iff P_1 = 0, & \gamma(P_1) = 0 &\iff P_1 = 0,\\ \|\lambda P_1\| &\le |\lambda|\|P_1\|, & \gamma(\lambda P_1)(r) &\le |\lambda|\gamma(P_1)(r),\\ \|P_1 + P_2\| &\le \|P_1\| + \|P_2\|, & \gamma(P_1 + P_2)(r) &\le \gamma(P_1)(r) + \gamma(P_2)(r),\\ \|P_1P_2\| &\le \|P_1\|\|P_2\|, & \gamma(P_1P_2)(r) &\le \gamma(P_1)(\gamma(P_2)(r)). \end{aligned}$$

4. $\gamma(P)(r) \le r\|P\|$ *for all* $r > 0$. *In particular, if* P *is linear and bounded, then* $\gamma(P)(r) = r\|P\|$.

DEFINITION 2.2. *A signal operator* P *is said to be*

(i) *gain stable if* $\|P\| < \infty$,

(ii) (gf)*-stable if* $\gamma(P)(r) < \infty$ *for each* $r \ge 0$.

We remark that gain stability implies (gf)-stability and both imply $P(\mathrm{Dom}(P)) \subset \mathcal{Y}_s$. In fact, a stable operator P maps bounded subsets of $\mathcal{U}_s$ into bounded subsets of $\mathcal{Y}_s$ (compare to [19]). As a shorthand, in the rest of this paper, a stable operator is taken to mean that the operator is stable in the sense of (gf)-stability unless specified otherwise.

DEFINITION 2.3. *A causal operator* $P : \mathrm{Dom}(P) \subset \mathcal{U}_s \to \mathcal{Y}$ *is said to admit a (right) coprime factorization if and only if there exist causal stable operators* $N : \mathcal{U}_s \to \mathcal{Y}_s$ *and* $D : \mathcal{U}_s \to \mathcal{U}_s$ *such that*

(i) D *is causally invertible with* $\mathrm{Dom}(D^{-1}) = \mathrm{Dom}(P)$,

(ii) $P = ND^{-1}$,

(iii) *there exists a causal stable mapping* $L : \mathcal{U}_s \times \mathcal{Y}_s \to \mathcal{U}_s$ *such that* $L(D, N)^\top = I$.

In that case, we also say that P *admits the coprime factorization* (N, D) *and we write* $P = ND^{-1}$. *For convenience, we call* L *the associated operator to this coprime factorization. The set of all coprime factorizations of* P *is denoted by* $\mathrm{rcf}(P)$.

In this definition, and henceforth, an operator $D : \mathcal{U}_s \to \mathcal{U}_s$ is said to be invertible with inverse D^{-1} if $D^{-1} : \mathrm{Dom}(D^{-1}) \subset \mathcal{U}_s \to \mathcal{U}_s$ is a well-defined operator and $DD^{-1}|_{\mathrm{Dom}(D^{-1})} = I$, $D^{-1}D|_{\mathcal{U}_s} = I$. Equivalently, D is required to be both left and right invertible.

DEFINITION 2.4. *Suppose* (N, D) *is a coprime factorization of* P. *If*

$$\|(D, N)^\top u\| = \|u\| \quad \text{for all } u \in \mathcal{U},$$

we say that (N, D) *is a normalized right coprime factorization of* P. *The set of all normalized right coprime factorizations is denoted by* $\mathrm{nrcf}(P)$.

Definitions 2.3 and 2.4 are generalizations of the coprime factorization and normalized coprime factorization for linear operators (see [20]) to the nonlinear case, as considered previously by various authors. Definition 2.3 is given by Verma and Hunt in [19] (see also [12, 18]) where the stability is in the sense of "bounded input implies bounded output" (resp., linear gain) between normed spaces. Sontag [17] also

defined the concept in which L is required to be of the form (B, A) with $A : \mathcal{Y} \to \mathcal{U}$, $B : \mathcal{U} \to \mathcal{U}$. Others using this Bezout identity to define coprime factorizations for nonlinear systems include Hammer [9], James, Smith, and Vinnicombe [10], Moore and Irlichet [11], etc. Whilst the Bezout identity $BN + AD = I$ always appears in the linear case, the more general form of L is less restrictive in the nonlinear setting. Generalizations of normalized coprime factorization, including those for specific signal operators, can be found in [2, 10, 13, 14, 15] and the references therein.

Existence and construction of (normalized) coprime factorizations for certain classes of nonlinear systems have been considered previously. For example, in [2, 10, 13, 15], normalized coprime factorizations for stabilizable nonlinear affine systems

$$x' = f(x) + g(x)u, \qquad y = C(x)$$

were constructed; Sontag [17] proved that, in the sense of IOS, if the above system with $C = I$ is smoothly input to stay stable by a controller of the form $u = k(x) + v$, then its input to state mapping $P : u \mapsto x$ admits a coprime factorization with $L = (I, A)$, where $-A$ is the memoryless operator induced by the smooth state feedback controller $u = k(x)$, N is the input to state mapping $v \mapsto x$ of the closed-loop system, and $D = I - AN$. Similar existence results were obtained by Verma and Hunt [19], in the sense of (gf)-stability, for the causal I/O mapping of the system

$$x'(t) = f(x(t), u(t), t), \qquad x(0) = x_0, \qquad y(t) = h(x(t), u(t), t)$$

in the case when $\mathcal{U}$, $\mathcal{Y}$ are L^p spaces. Other references to the state space construction of coprime factors can be found in [14] and the references therein.

The following two results can be found in [18] where the notion of stability is in the sense of a finite linear gain. However, the proofs remain valid in the context of (gf)-stability, hence we omit the proofs.

PROPOSITION 2.5. *Suppose P admits coprime right factorization (N, D). Then*

$$\mathrm{Graph}(P) := \{(u, Pu)^\top : u \in \mathrm{Dom}(P)\} = \{(Du, Nu)^\top : u \in \mathcal{U}_s\}.$$

Proof. See [18]. □

PROPOSITION 2.6. *$(N, D), (N_1, D_1) \in \mathrm{rcf}(P)$ if and only if there exists a causally stable operator U on $\mathcal{U}_s$, where U^{-1} exists, is stable, and is such that $N = N_1 U$, $D = D_1 U$.*

Proof. See [18]. □

If the coprime factorizations in Proposition 2.6 are also normalized, then we also have the following proposition.

PROPOSITION 2.7. *If $(N, D), (N_1, D_1) \in \mathrm{nrcf}$, then the operator U in Proposition 2.6 is such that $\|Uu\| = \|U^{-1}u\| = \|u\|$ for all $u \in \mathcal{U}_s$.*

Proof. Let $u \in \mathcal{U}_s$. By Proposition 2.6 and the definition of normalized coprime factorization, we see $\|U^{-1}u\| = \|(D, N)^\top U^{-1}u\| = \|(D_1, N_1)^\top U U^{-1}u\| = \|(D_1, N_1)^\top u\| = \|u\|$ and, similarly, $\|Uu\| = \|(D_1, N_1)^\top Uu\| = \|(D, N)^\top U^{-1}Uu\| = \|(D, N)^\top u\| = \|u\|$. This proves the proposition. □

3. Graph topologies. In this section, we will study graph topologies on the set of certain signal operators having coprime factorizations. As in the linear case, we will show that the graph topologies play a natural role in the theory of closed loop robust stability.

In practice, the signal spaces, operators, and associated coprime factorizations concerned are constrained to lie within certain classes for different control problems.

For example, one may only be interested in the case when $\mathcal{U} = L_n^{\infty,e}$, $\mathcal{Y} = L_m^{\infty,e}$, and where all operators considered lie in the subset of I/O operators of all affine (nonlinear) systems.

Here we list some particular categories that will be considered in this paper.

- Category nor : $\mathcal{U}, \mathcal{Y}$ are general signal spaces as assumed and all operators considered are those that admit normalized coprime factorizations in the sense defined in the last section.
- Category ω with $\omega \in \mathcal{K}_\infty$: $\mathcal{U}, \mathcal{Y}$ are general signal spaces and all operators F considered are such that $\mathrm{rcf}(F) \neq \emptyset$ and $\sup_{r>0} \frac{\gamma((N,D)^\top)(\omega(r))}{r} < \infty$ for all $(N, D) \in \mathrm{rcf}(F)$.
- Category $\mathbf{L}$: $\mathcal{U}$, $\mathcal{Y}$ are both the frequency domain Hardy spaces $\mathcal{H}_2$ (see [20]), operators are the real rational $p \times q$ transfer function matrices, and the associated coprime factorizations are those linear factorizations over $\mathcal{RH}_\infty$[1] as in [20] or [21].

Since $F \equiv 0$ has normalized coprime factorization $(0, I)$, we see that each category is nonempty.

The graph topologies will be defined on a general category although in the next section we mainly consider $\mathbf{N}_{\mathrm{nor}}(\mathcal{U}, \mathcal{Y})$ and $\mathbf{N}_\omega(\mathcal{U}, \mathcal{Y})$. So we use Γ to represent the category concerned and write

$$\mathbf{N}_\Gamma(\mathcal{U}, \mathcal{Y}) := \left\{ P : \mathrm{Dom}(P) \subset \mathcal{U} \to \mathcal{Y} : \begin{array}{l} P \text{ and the associated rcfs} \\ N, D \text{ are within category } \Gamma \end{array} \right\}.$$

Correspondingly, we have notations $\mathbf{N}_{\mathrm{nor}}(\mathcal{U}, \mathcal{Y})$, $\mathbf{N}_\omega(\mathcal{U}, \mathcal{Y})$, and $\mathbf{N}_{\mathbf{L}}(\mathcal{H}_2, \mathcal{H}_2) =: \mathbf{N}_{\mathbf{L}}$. For example

$$\mathbf{N}_{\mathrm{nor}}(\mathcal{U}, \mathcal{Y}) := \{P : \mathrm{Dom}(P) \subset \mathcal{U} \to \mathcal{Y} \text{ and } \mathrm{nrcf}(P) \neq \emptyset\}.$$

$$\mathbf{N}_\omega(\mathcal{U}, \mathcal{Y}) := \left\{ P : \mathrm{Dom}(P) \subset \mathcal{U} \to \mathcal{Y} : \begin{array}{l} \mathrm{rcf}(P) \neq \emptyset \text{ and for all } (N, D) \in \mathrm{rcf}(P), \\ \sup\limits_{r>0} \dfrac{\gamma((N,D)^\top)(\omega(r))}{r} < \infty \end{array} \right\}.$$

A graph topology for $\mathbf{N}_{\mathbf{L}}$, denoted by $\mathcal{T}_{\mathbf{L}}$, has been defined in [20] by the following local base for $P \in \mathcal{RH}_\infty$:

$$\mathcal{N}(N, D; \varepsilon) = \{N_1 D_1^{-1} : \|(N_1 - N, D_1 - D)^\top\|_\infty < \varepsilon,\ N, D, N_1, D_1 \in \mathcal{RH}_\infty\} \tag{3.1}$$

with $\varepsilon > 0, (N, D) \in \mathrm{rcf}(P)$, and $(N_1, D_1) \in \mathrm{rcf}(N_1 D_1^{-1})$, respectively. We will show that our L^2 topologies for $\mathbf{N}_{\mathbf{L}}$ are the same as $\mathcal{T}_{\mathbf{L}}$.

For notational ease in what follows, any pairs N, D or N_k, D_k are always assumed to be coprime factorizations of $P = ND^{-1}$ and $P_k = N_k D_k^{-1}$, respectively, and P and P_k are taken to be well-defined operators from $\mathrm{Dom}(P) \to \mathcal{Y}_s$, $\mathrm{Dom}(P_k) \to \mathcal{Y}_s$, respectively.

[1] $\mathcal{H}_2$ is the space of Fourier transforms of signals in $L^2(\mathbb{R}_+, \mathbb{R}^n)$ endowed with the norm $\|x\|_2^2 := \frac{1}{2\pi} \int_{-\infty}^{\infty} x^*(j\omega) x(j\omega) d\omega$. By Parseval's theorem, it is isometrically isomorphic to $L^2(\mathbb{R}_+, \mathbb{R}^n)$ and therefore the two notations are not distinguished. $\mathcal{RH}_\infty$ is the space of rational transfer functions of stable linear, time-invariant, continuous time systems endowed with the norm $\|P\|_\infty := \sup_{\omega \in \mathbb{R}} \bar{\sigma}[P(j\omega)]$, where $\bar{\sigma}$ denotes the maximum singular value. Equivalently, $\|P\|_\infty := \sup\{\|Pu\|_{\mathcal{H}_2} / \|u\|_{\mathcal{H}_2} : u \in \mathcal{H}_2, u \neq 0\}$. So by Paserval's theorem, the H_∞-norm in the frequency domain corresponds to the induced L^2 norm in the time domain.

3.1. Pointwise graph topology. Let $\Re$ be the vector space of all functions from $\mathbb{R}^+$ to $\mathbb{R}$. For any open subset $\Omega \subset \mathbb{R}$ and a finite subset $\{t_1, \dots, t_n\} \subset \mathbb{R}^+$, let

$$\mathcal{V}(t_1, \dots, t_n; \Omega) = \{f \in \Re : f(t_i) \in \Omega\}.$$

It can be proved that $\{\mathcal{V}(t_1, \dots, t_n; \Omega) : t_i \in \mathbb{R}^+,\ n > 0,\ \Omega \subset \mathbb{R},\ \Omega \text{ open}\}$ forms a subbase for a topology on $\Re$. Moreover, the family of subsets

$$\Re_0 = \{\mathcal{V}(t_1, \dots, t_n; \varepsilon) := \mathcal{V}(t_1, \dots, t_n; (-\varepsilon, \varepsilon)) : \varepsilon > 0,\ t_i \in \mathbb{R}^+,\ n > 0\}$$

is a base for the neighborhood of $f(t) \equiv 0$ in $\Re$ under such a topology.

For each $P \in \mathbf{N}_\Gamma(\mathcal{U}, \mathcal{Y})$ with coprime factorization (N, D) and each $V \in \Re_0$, we define

$$O(N, D; V) = \{P_1 = N_1 D_1^{-1} \in \mathbf{N}_\Gamma(\mathcal{U}, \mathcal{Y}) : \gamma((D - D_1, N - N_1)^\top) \in V\}.$$

Obviously, $P = ND^{-1} \in O(N, D; V)$ for each $V \in \Re_0$. Moreover, we have the following proposition.

PROPOSITION 3.1. *If $ND^{-1} \in O(N_1, D_1; V_1) \cap O(N_2, D_2; V_2)$, then there exist $V \in \Re_0$ such that $O(N, D; V) \subset O(N_1, V_1) \cap O(N_2, D_2; V_2)$.*

Proof. We may suppose $V_1 = \mathcal{V}(t_1, \dots, t_n; \varepsilon_1), V_2 = \mathcal{V}(s_1, \dots, s_m; \varepsilon_2)$ with $t_i > 0$, $s_j > 0$, $\varepsilon_k > 0$ $(i = 1, \dots, n,\ j = 1, \dots, m,\ k = 1, 2)$. Then $\varepsilon_1 - \gamma((D - D_1, N - N_1)^\top)(t_i)$ and $\varepsilon_2 - \gamma((D - D_2, N - N_2)^\top)(s_j)$ are all positive numbers. Let $\varepsilon > 0$ such that

$$\varepsilon < \min \begin{Bmatrix} \varepsilon_1 - \gamma((D - D_1, N - N_1)^\top)(t_i), & i = 1, \dots, n, \\ \varepsilon_2 - \gamma((D - D_2, N - N_2)^\top)(s_j), & j = 1, \dots, m \end{Bmatrix}.$$

If $\tilde{N}\tilde{D}^{-1} \in O(N, D; V)$ with $V = \mathcal{V}(t_1, \dots, t_n, s_1, \dots, s_m; \varepsilon)$, then

$$\begin{aligned} \gamma((\tilde{D} - D_1, \tilde{N} - N_1)^\top)(t_i) &\le \gamma((\tilde{D} - D, \tilde{N} - N)^\top)(t_i) + \gamma((D - D_1, N - N_1)^\top)(t_i) \\ &< \varepsilon_1 - \gamma((D - D_1, N - N_1)^\top)(t_i) \\ &+ \gamma((D - D_1, N - N_1)^\top)(t_i) = \varepsilon_1 \end{aligned}$$

for all $i = 1, \dots, n$. This gives $\tilde{N}\tilde{D}^{-1} \in O(N_1, D_1; V_1)$. Similarly, we can show $\tilde{N}\tilde{D}^{-1} \in O(N_2, D_2; V_2)$. Therefore, $\tilde{N}\tilde{D}^{-1} \in O(N_1, D_1; V_1) \cap O(N_2, D_2; V_2)$ which means $O(N, D; V) \subset O(N_1, D_1; V_1) \cap O(N_2, D_2; V_2)$. ∎

From the above result, it follows that a topology on $\mathbf{N}_\Gamma(\mathcal{U}, \mathcal{Y})$ can be uniquely determined by the base $\mathbb{B}$, where

$$\mathbb{B} = \{O(N, D; V) : ND^{-1} \in \mathbf{N}_\Gamma(\mathcal{U}, \mathcal{Y}),\ V \in \Re_0\},$$

and $\{O(N, D; V) : ND^{-1} = P,\ V \in \Re_0\}$ a local base of P. We denote this topology by $\mathcal{T}$ and call it the *pointwise* (*graph*) *topology* (see the preceding footnote). The following proposition provides an alternative base for this topology.

PROPOSITION 3.2. *Let Q_+ be the set of all positive rational numbers and*

$$O'(N, D; r, \varepsilon) = \{N_1 D_1^{-1} \in \mathbf{N}_\Gamma(\mathcal{U}, \mathcal{Y}) : \gamma((D - D_1, N - N_1)^\top)(r) < \varepsilon\}.$$

Then a base for the pointwise graph topology $\mathcal{T}$ is the family of subsets

$$\mathbb{B}' = \{O'(N, D; r, \varepsilon) : ND^{-1} \in \mathbf{N}_\Gamma(\mathcal{U}, \mathcal{Y}),\ r, \varepsilon \in Q_+\}.$$

Proof. Obviously $\mathbb{B}' \subset \mathbb{B}$. Suppose $O(N, D; V) \in \mathbb{B}$ with $V = \mathcal{V}(t_1, \dots, t_n; \varepsilon)$. Let $r, \varepsilon_1 \in Q$ such that $r > \max\{t_1, \dots, t_n\}$ and $\varepsilon_1 < \varepsilon$. Then for each $N_1 D_1^{-1} \in O'(N, D; r, \varepsilon_1)$, from Lemma 2.1 (2.), it follows that

$$\gamma((D_1 - D, N_1 - N)^\top)(t_i) \le \gamma((D_1 - D, N_1 - N)^\top)(t) < \varepsilon_1 < \varepsilon$$

for all $i = 1, \dots, n$. This means $N_1 D_1^{-1} \in O(N, D; V)$ and therefore $O'(N, D; r, \varepsilon) \subset O(N, D; V)$. Hence $\mathbb{B}$ and $\mathbb{B}'$ are equivalent. □

If we restrict our consideration to $\mathbf{N_L}$, from Lemma 2.1 (4.) and (3.1), we see

$$\begin{aligned} O'(N, D; r, \varepsilon) &= \{N_1 D_1^{-1} : \gamma((D - D_1, N - N_1)^\top)(r) < \varepsilon,\ N, D, N_1, D_1 \in \mathcal{RH}_\infty\} \\ &= \left\{N_1 D_1^{-1} : \|(D - D_1, N - N_1)^\top\| < \frac{\varepsilon}{r} =: \varepsilon_1,\ N, D, N_1, D_1 \in \mathcal{RH}_\infty\right\} \\ &= \mathcal{N}(N, D; \varepsilon_1), \end{aligned}$$

hence we have the following corollary.

COROLLARY 3.3. *The pointwise graph topology $\mathcal{T}$ in the category $\mathbf{L}$ is the same as the graph topology $\mathcal{T}_{\mathbf{L}}$.*

Now we begin to consider the convergence of sequences under the pointwise topology. The following result shows that any convergent sequence has only one limit.

PROPOSITION 3.4. *The pointwise graph topology $\mathcal{T}$ is Hausdorff. Therefore, the limit point of a convergent sequence is unique.*

Proof. Let $P_1 \neq P_2$ be two distinct plants. Then there exist $(N_1, D_1) \in \mathrm{rcf}(P_1)$, $(N_2, D_2) \in \mathrm{rcf}(P_2)$ with $(D_1, N_1)^\top \neq (D_2, N_2)^\top$. This shows

$$\varepsilon := \gamma((D_1 - D_2, N_1 - N_2)^\top)(r) > 0 \quad \text{for some } r > 0.$$

Consider the neighborhoods $O(N_1, D_1; r, \varepsilon/3)$ of P_1 and $O(N_2, D_2; r, \varepsilon/3)$ of P_2. If there exists $P = ND^{-1} \in O(N_1, D_1; r, \varepsilon/3) \cap O(N_2, D_2; r, \varepsilon/3)$, since

$$\begin{aligned} \gamma((N_1 - N_2, D_1 - D_2)^\top)(r) \le{}& \gamma((N_1 - N, D_1 - D)^\top)(r) \\ &+ \gamma((N_2 - N, D_2 - D)^\top)(r), \end{aligned}$$

we see $\gamma((N_1 - N_2, D_1 - D_2)^\top)(r) < \varepsilon$. This is a contradiction. Hence we have that $O(N_1, D_1; r, \varepsilon/3) \cap O(N_2, D_2; r, \varepsilon/3) = \emptyset$, which proves the proposition. □

Suppose $\{P_n\} \subset \mathbf{N}_\Gamma(\mathcal{U}, \mathcal{Y})$ is a sequence. We let $P_n \xrightarrow{\mathcal{T}} P$ denote the convergence of the sequence $\{P_n\}_{n \ge 1}$ to P under the graph topology $\mathcal{T}$. From Proposition 3.2, we see that $P_n \xrightarrow{\mathcal{T}} P$ means that, for any $r > 0$, $\varepsilon > 0$ and each coprime factorization ND^{-1} of P, there exist $n_0 > 0$ and coprime factorization $N_n D_n^{-1}$ of P_n such that $N_n D_n^{-1} \in O(N, D; r, \varepsilon)$ for all $n \ge n_0$. Necessary and sufficient conditions for this convergence are given below.

THEOREM 3.5. *The following statements are equivalent.*

(i) $P_n \xrightarrow{\mathcal{T}} P$.

(ii) *For each $(N, D) \in \mathrm{rcf}(P)$, there exists $(N_n, D_n) \in \mathrm{rcf}(P_n)$ such that*

$$\gamma((D_n - D, N_n - N)^\top)(r) \to 0 \quad \textit{for each } r > 0.$$

(iii) *There exists $(N, D) \in \mathrm{rcf}(P)$ and, for each n, there exists $(N_n, D_n) \in \mathrm{rcf}(P_n)$ such that*

$$\gamma((D_n - D, N_n - N)^\top)(r) \to 0 \quad \textit{for all } r > 0.$$

Proof. (ii)⇒(i) and (ii)⇒(iii) are immmediate; we need only to prove (i)⇒(ii) and (iii)⇒(ii).

(i)⇒(ii). Let $r > 0$ and $P = ND^{-1}$ be given. According to the assumptions, for each $\varepsilon > 0$ and $n > 0$, there exists coprime factorization $N_{n,\varepsilon}D_{n,\varepsilon}^{-1}$ of P_n and $n_\varepsilon > 0$ such that $N_{n,\varepsilon}D_{n,\varepsilon}^{-1} \in O(N, D; r, \varepsilon)$ for all $n \geq n_\varepsilon$. Let $\varepsilon = 1, 1/2, \ldots, 1/2^k, \cdot$, respectively, to obtain the corresponding integers $n_k := n_{1/2^k}$. Define

$$N_n = N_{n,1/2^k} \quad \text{and} \quad D_n = D_{n,1/2^k} \quad \text{for } n_k \leq n < n_{k+1}.$$

Then $N_nD_n^{-1}$ is a coprime factorization of P_n and $N_nD_n^{-1} \in O(N, D; r, 1/2^k)$ for $n \geq n_k$. Hence $\gamma((D_n - D, N_n - N)^\top)(r) \to 0$.

(iii)⇒(ii). Suppose $(\tilde{N}, \tilde{D})$ is an arbitrary coprime factorization of P. Then by Proposition 2.6, there exists stable operator U with $\tilde{N} = NU$, $\tilde{D} = DU$. Moreover, $(\tilde{N}_n, \tilde{D}_n) := (N_nU, D_nU)$ is a coprime factorization of P_n due to the same proposition. Using Lemma 2.1, we have

$$\gamma((\tilde{D}_n - \tilde{D}, \tilde{N}_n - \tilde{N})^\top)(r) \leq \gamma((D_n - D, N_n - N)^\top)(\gamma(U)(r)).$$

The stability of U and the assumption ensure $\gamma((\tilde{D}_n - \tilde{D}, \tilde{N}_n - \tilde{N})^\top)(r) \to 0$ as $n \to \infty$ and, therefore, (ii) has been established. This completes the proof. □

Because the continuity of a mapping from a first-countable topological space to another topological space can be described by the convergence of sequences, we have the following corollary.

COROLLARY 3.6. *Let Λ be a first-countable topological space, $P_\lambda : \Lambda \to \mathbf{N}_\Gamma(\mathcal{U}, \mathcal{Y})$. Then $\lambda \mapsto P_\lambda$ is continuous at $\lambda = \lambda_0$ under the pointwise graph topology $\mathcal{T}$ if and only if there exist coprime factorizations $P_{\lambda_0} = N_0D_0^{-1}$ and $P_\lambda = N_\lambda D_\lambda^{-1}$ for each $\lambda \in \Lambda$ such that*

$$\gamma((D_0 - D_\lambda, N_0 - N_\lambda)^\top)(r) \to 0 \quad \textit{for all } r \geq 0 \textit{ as } \lambda \to \lambda_0.$$

3.2. Weighted graph topology. In this section, we consider another topology on the set $\mathbf{N}_\Gamma(\mathcal{U}, \mathcal{Y})$, which will be related to a given function $\omega \in \mathcal{K}_\infty$ and a weighted gain $\|\cdot\|_\omega$ defined by

$$\|\mathbf{P}\|_\omega = \sup_{r>0} \frac{\gamma(\mathbf{P})(\omega(r))}{r} \quad \text{for any signal operator } \mathbf{P}.$$

It is straightforward to prove that $\|\cdot\|_\omega$ is a norm. Moreover, if $\omega(r) \geq c_1 r$ with $c_1 > 0$ for all $r > 0$, then $\|\mathbf{P}\|_\omega \geq c_1\|\mathbf{P}\|$. If $\mathbf{P}0 = 0$, $c_2 > 0$, and $\omega(r) \leq c_2 r$ for all $r > 0$, then $\|\mathbf{P}\|_\omega \leq c_2\|\mathbf{P}\|$.

Let

$$\Sigma = \{\mathbf{P} : \mathcal{U} \to \mathcal{U} \times \mathcal{Y} \text{ with } \|\mathbf{P}\|_\omega < \infty\},$$

It can be seen from the basic properties of γ that Σ is a linear space and therefore $(\Sigma, \|\cdot\|_\omega)$ is a normed space. The norm induces a corresponding topology on Σ, of which a local base of open ball neighborhoods of $\mathbf{P} \equiv 0$ is denoted by $\mathcal{B}$.

For each $P \in \mathbf{N}_\Gamma(\mathcal{U}, \mathcal{Y})$ with coprime factorization $P = ND^{-1}$ and each $V \in \mathcal{B}$, we denote by

$$O_\omega(N, D; V) = \{N_1D_1^{-1} \in \mathbf{N}_\Gamma(\mathcal{U}, \mathcal{Y}) : (D - D_1, N - N_1)^\top \in V\}.$$

Obviously, $P = ND^{-1} \in O_\omega(N, D; V)$ for each $V \in \mathcal{B}$. Moreover, we have the following proposition.

PROPOSITION 3.7. *If $ND^{-1} \in O_\omega(N_1, D_1; V_1) \cap O_\omega(N_2, D_2; V_2)$, then there exist $V \in \mathcal{B}$ such that $O_\omega(N, D; V) \subset O_\omega(N_1, V_1) \cap O_\omega(N_2, D_2; V_2)$.*

Proof. We may suppose that $V_i = \{\mathbf{P} \in \Sigma : \|\mathbf{P}\|_\omega < \varepsilon_i\}$ with $\varepsilon_i > 0$, $i = 1, 2$. Let

$$\alpha_i = \sup_{r>0} \frac{\gamma((D - D_i, N - N_i)^\top)(\omega(r))}{r}, \quad i = 1, 2$$

and let ε be a positive number such that $\varepsilon < \min\{\varepsilon_1 - \alpha_1, \varepsilon_2 - \alpha_2\}$. Then for each $\tilde{N}\tilde{D}^{-1} \in O_\omega(N, D; \varepsilon)$ and each $r > 0$, from the third property of γ, it follows that

$$\begin{aligned}\frac{\gamma((\tilde{D} - D_i, \tilde{N} - N_i)^\top)(\omega(r))}{r} &\le \frac{\gamma((\tilde{D} - D, \tilde{N} - N)^\top)(\omega(r))}{r} \\ &\quad + \frac{\gamma((D - D_i, N - N_i)^\top)(\omega(r))}{r} \\ &< \varepsilon + \alpha_i, \quad i = 1, 2.\end{aligned}$$

Hence

$$\sup_{r>0} \frac{\gamma((\tilde{D} - D_i, \tilde{N} - N_i)^\top)(\omega(r))}{r} \le \varepsilon + \alpha_i < \varepsilon_i - \alpha_i + \alpha_i = \varepsilon_i, \quad i = 1, 2.$$

This implies $\tilde{N}\tilde{D}^{-1} \in O_\omega(N_1, D_1; \varepsilon_1) \cap O_\omega(N_2, D_2; \varepsilon_2)$ and, therefore, $O_\omega(N, D; \varepsilon) \subset O_\omega(N_1, D_1; \varepsilon_1) \cap O_\omega(N_2, D_2; \varepsilon_2)$. □

Let

$$\mathbb{B}_\omega = \{O_\omega(N, D; V) : ND^{-1} \in \mathbf{N}_\Gamma(\mathcal{U}, \mathcal{Y}),\ V \in \mathcal{B}\}.$$

From the above result, it follows that a topology on $\mathbf{N}_\Gamma(\mathcal{U}, \mathcal{Y})$ can be uniquely determined with $\mathbb{B}_\omega$ as its base. We denote this topology by $\mathcal{T}_\omega$ and call it the *weighted (graph) topology* related to function ω. Obviously, $\mathcal{T}_\omega$ has a countable local base.

If $\mathbf{P} \in \Sigma$ is linear and $\omega(t) \equiv t$, then from Lemma 2.1 (4.), we see that $\|\mathbf{P}\|_\omega = \|\mathbf{P}\|$. Therefore, if we restrict attention to $\mathbf{N}_\mathbf{L}$, then for each $P = ND^{-1} \in \mathcal{RH}_\infty$ and $V = \{\mathbf{P} : \mathcal{H}_2 \to \mathcal{H}_2 \times \mathcal{H}_2,\ \|\mathbf{P}\|_\omega < \varepsilon\}$, we have

$$O_\omega(N, D; V) = \{N_1 D_1^{-1} : \|(N_1 - N, D_1 - D)^\top\| < \varepsilon,\ N, D, N_1, D_1 \in \mathcal{RH}_\infty\}$$

and $O_\omega(N, D; V) = \mathcal{N}(N, D; \varepsilon)$. This fact yields the following corollary.

COROLLARY 3.8. *For $\omega(t) \equiv t$ and $\mathbf{N}_\Gamma(\mathcal{U}, \mathcal{Y}) = \mathbf{N}_\mathbf{L}$, the weighted graph topology $\mathcal{T}_\omega$ is the same as the graph topology $\mathcal{T}_\mathbf{L}$ defined for $\mathbf{N}_\mathbf{L}(\mathcal{U}, \mathcal{Y})$.*

From Proposition 3.7, we see that a sequence of operators $\{P_n\}_{n\ge1}$ converging to P under this graph topology, denoted by $P_n \xrightarrow{\mathcal{T}_\omega} P$, means that, for any $\varepsilon > 0$ and each coprime factorization ND^{-1} of P, there exist $n_0 > 0$ and coprime factorization $N_n D_n^{-1}$ of P_n such that $\|(D_n - D, N_n - N)^\top\|_\omega < \varepsilon$ for all $n \ge n_0$.

Using a method similar to the one used in Proposition 3.4, we can also prove that the weighted topology is a Hausdorff topology. So a convergent sequence has a unique limit.

THEOREM 3.9. *$P_n \xrightarrow{\mathcal{T}_\omega} P$ if and only if for each coprime factorization ND^{-1} of P, there exists coprime factorization $N_n D_n^{-1}$ of P_n such that*

$$\sup_{r>0} \frac{\gamma((D_n - D, N_n - N)^\top)(\omega(r))}{r} \to 0.$$

Proof. The proof is omitted for brevity as it follows the same reasoning as used in the first part proof for Theorem 3.5. □

Two immediate corollaries are as follows.

COROLLARY 3.10. *If* $P_n \xrightarrow{\mathcal{T}_\omega} P$, *then* $P_n \to P$ *under the pointwise graph topology* $\mathcal{T}$.

COROLLARY 3.11. *Let* Λ *be a first-countable topological space,* $P_\lambda : \Lambda \to \mathbf{N}_\Gamma(\mathcal{U}, \mathcal{Y})$. *Then* $\lambda \mapsto P_\lambda$ *is continuous at* $\lambda = \lambda_0$ *under a weighted graph topology* $\mathcal{T}_\omega$ *if and only if for each coprime factorization* $P_{\lambda_0} = N_0 D_0^{-1}$, *there exist coprime factorization* $P_\lambda = N_\lambda D_\lambda^{-1}$ *for each* $\lambda \in \Lambda$ *such that*

$$\gamma((D_0 - D_\lambda, N_0 - N_\lambda)^\top)(r) \to 0 \quad \text{for all } r \geq 0 \text{ as } \lambda \to \lambda_0.$$

To conclude this section, we observe that given two functions $\omega_1, \omega_2 \in \mathcal{K}_\infty$, each generates a weighted graph topology $\mathcal{T}_{\omega_1}, \mathcal{T}_{\omega_2}$. If $\omega_1(r) \leq \omega_2(r)$ for all $r \in \mathbb{R}^+$, then $\|\mathbf{P}\|_{\omega_1} \leq \|\mathbf{P}\|_{\omega_2}$ for all $\mathbf{P} : \mathcal{U} \to \mathcal{U} \times \mathcal{Y}$. Therefore

$$O_{\omega_2}(N, D; V) \subset O_{\omega_1}(N, D; V) \quad \text{for all } ND^{-1} \in \mathbf{N}_\Gamma(\mathcal{U}, \mathcal{Y}),\ V \in \mathcal{B}.$$

This implies the following comparison theorem.

THEOREM 3.12. *Suppose* $\omega_1, \omega_2 \in \mathcal{K}_\infty$ *satisfying* $\omega_1(r) \leq \omega_2(r)$ *for all* $r > 0$. *Then* $\mathcal{T}_{\omega_2}$ *is stronger than* $\mathcal{T}_{\omega_1}$ *(i.e., any sequence converging under* $\mathcal{T}_{\omega_2}$ *will converge under* $\mathcal{T}_{\omega_1}$*). Additionally,* $\mathcal{T}_{c\omega_1}$ *and* $\mathcal{T}_{\omega_1}$ *are equivalent for any* $c > 0$ *(i.e., they induce the same convergence).*

In particular we have the following corollary.

COROLLARY 3.13. *The linear gain induces a graph topology (denoted by* $\mathcal{T}_{lg}$*) on* $\mathbf{N}_\Gamma(\mathcal{U}, \mathcal{Y})$. *If* $c_1 r \leq \omega(r) \leq c_2 r$ *for all* $r \geq 0$, *then* $\mathcal{T}_\omega$ *and* $\mathcal{T}_{lg}$ *are equivalent.*

Hence it can be seen that the weighted graph topologies inherit the partial order given by the natural partial order on the weights.

4. Metrizability. The question addressed in this section is simply whether the nonlinear graph topologies introduced earlier can be sensibly metrized. In the linear case it is well known that the answer is affirmative. We will show that useful metrics can also be given for the weighted nonlinear graph topology. We will introduce a number of metrics on specific subsets of $\mathbf{N}(\mathcal{U}, \mathcal{Y})$ and prove that some of them induce the weighted graph topology.

Throughout this section, we suppose $\omega \in \mathcal{K}_\infty$ is a given function, $\|\cdot\|_\omega$ is the weighted gain, and $\mathcal{U}$, $\mathcal{Y}$, $\mathcal{U}_s$, $\mathcal{Y}_s$ are defined as before. Every signal operator P (say) is assumed to be causal and $P(0) = 0$.

4.1. The metric formulas. We define

$$\begin{aligned}
\mathcal{Q} &= \{Q : \mathcal{U}_s \to \mathcal{U}_s \text{ is stable and } Q^{-1} \text{ exists and is also stable}\},\\
\mathcal{Q}^* &= \{Q : \mathcal{U}_s \to \mathcal{U}_s \text{ is stable and } Q^{-1} \text{ exists (bijective)}\},\\
\mathcal{Q}^s &= \{Q : \mathcal{U}_s \to \mathcal{U}_s \text{ is stable and surjective}\}.
\end{aligned}$$

The subsets of signal operators we will consider are $\mathbf{N}_\omega(\mathcal{U}, \mathcal{Y})$ and $\mathbf{N}_{\text{nor}}(\mathcal{U}, \mathcal{Y})$ as defined in the last section. Recall that

$$\begin{aligned}
\mathbf{N}_\omega(\mathcal{U}, \mathcal{Y}) &= \{P \in \mathbf{N}(\mathcal{U}, \mathcal{Y}) : \|(D, N)^\top\|_\omega < \infty \text{ for all } (N, D) \in \text{rcf}(P)\},\\
\mathbf{N}_{\text{nor}}(\mathcal{U}, \mathcal{Y}) &= \{P \in \mathbf{N}_\omega(\mathcal{U}, \mathcal{Y}) : \text{nrcf}(P) \neq \emptyset\}.
\end{aligned}$$

We now define seven functionals over the above sets:

$$d_1(P_1,P_2) = \max\{\vec{d}_1(P_1,P_2),\vec{d}_1(P_2,P_1)\}, \quad \text{for } P_1,P_2 \in \mathbf{N}_\omega(\mathcal{U},\mathcal{Y}),$$

$$\text{where } \vec{d}_1(P_1,P_2) := \sup_{(N_1,D_1)\in \mathrm{rcf}(P_1)} \inf_{(N_2,D_2)\in \mathrm{rcf}(P_2)} \|(D_1-D_2, N_1-N_2)^\top\|_\omega;$$

$$d_2(P_1,P_2) = \max\{\vec{d}_2(P_1,P_2),\vec{d}_2(P_2,P_1)\} \quad \text{for } P_1,P_2 \in \mathbf{N}_{\mathrm{nor}}(\mathcal{U},\mathcal{Y}),$$

$$\text{where } \vec{d}_2(P_1,P_2) = \inf_{\substack{Q\in\mathcal{Q}\\ \|Q\|\le 1}} \|(D_1-D_2Q, N_1-N_2Q)^\top\|_\omega,\ (N_i,D_i)\in \mathrm{nrcf}(P_i),\ i=1,2;$$

$$d_3(P_1,P_2) = \max\{\vec{d}_3(P_1,P_2),\vec{d}_3(P_2,P_1)\} \quad \text{for } P_1,P_2 \in \mathbf{N}_{\mathrm{nor}}(\mathcal{U},\mathcal{Y}),$$

$$\text{where } \vec{d}_3(P_1,P_2) = \inf_{\substack{Q\in\mathcal{Q}^*\\ \|Q\|\le 1}} \|(D_1-D_2Q, N_1-N_2Q)^\top\|_\omega,\ (N_i,D_i)\in \mathrm{nrcf}(P_i),\ i=1,2;$$

$$d_4(P_1,P_2) = \log(1+\max\{\vec{d}_4(P_1,P_2),\vec{d}_4(P_2,P_1)\}) \quad \text{for any } P_1,P_2 : \mathcal{U}\to\mathcal{Y},$$

$$\text{where } \vec{d}_4(P_1,P_2) = \begin{cases} \inf\left\{\|I-\Phi\|_\omega : \begin{array}{l}\Phi \text{ is a surjective mapping from}\\ \mathrm{Graph}(P_1) \text{ to } \mathrm{Graph}(P_2),\ \Phi(0)=0\end{array}\right\},\\ \infty \text{ if no such operator } \Phi \text{ exists;}\end{cases}$$

$$d_5(P_1,P_2) = \log(1+\max\{\vec{d}_5(P_1,P_2),\vec{d}_5(P_2,P_1)\}) \quad \text{for any } P_1,P_2 : \mathcal{U}\to\mathcal{Y},$$

$$\text{where } \vec{d}_5(P_1,P_2) = \begin{cases} \inf\left\{\|I-\Phi\|_\omega : \begin{array}{l}\Phi \text{ is a bijective mapping from}\\ \mathrm{Graph}(P_1) \text{ to } \mathrm{Graph}(P_2),\ \Phi(0)=0\end{array}\right\},\\ \infty \text{ if no such operator } \Phi \text{ exists;}\end{cases}$$

$$d_6(P_1,P_2) = \log(1+\max\{\vec{d}_6(P_1,P_2),\vec{d}_6(P_2,P_1)\}) \quad \text{for } P_1,P_2 \in \mathbf{N}_{\mathrm{nor}}(\mathcal{U},\mathcal{Y}),$$

$$\text{where } \vec{d}_6(P_1,P_2) = \inf_{Q\in\mathcal{Q}^*} \|(D_1-D_2Q, N_1-N_2Q)^\top\|_\omega,\ (N_i,D_i)\in \mathrm{nrcf}(P_i),\ i=1,2;$$

$$d_7(P_1,P_2) = \log(1+\max\{\vec{d}_7(P_1,P_2),\vec{d}_7(P_2,P_1)\}) \quad \text{for } P_1,P_2 \in \mathbf{N}_{\mathrm{nor}}(\mathcal{U},\mathcal{Y}),$$

$$\text{where } \vec{d}_7(P_1,P_2) = \inf_{Q\in\mathcal{Q}^s} \|(D_1-D_2Q, N_1-N_2Q)^\top\|_\omega,\ (N_i,D_i)\in \mathrm{nrcf}(P_i),\ i=1,2.$$

Notice, when ω is the identity d_3 is closely related to the graph metric studied in [20] for finite dimensional linear systems, while d_5 is the gap metric defined in [7] where $\vec{d}_5$ is extensively exploited for the robustness of stability of nonlinear systems. In most cases, $\vec{d}_5$ can be replaced by $\vec{d}_4$ as later shown in Lemma 5.1 . The functionals d_6 and d_7 are closely related to the Georgiou formula for the gap metric [4].

We will prove that $d_1,\dots,d_7$ are metrics on suitable sets of signal operators and show relations between all seven functionals $d_1,\dots,d_7$ and their induced topologies.

The results developed in this section are as follows:

1. The weighted topology $\mathcal{T}_\omega$ on $\mathbf{N}_{\mathrm{nor}}(\mathcal{U},\mathcal{Y})$ can be metrized by graph metrics d_2, d_3 provided $cr \le \omega(r)$;

2. The weighted graph topology can also be metrized by Georgiou and Smith's gap metrics d_4, d_5 provided $r \le \omega(r)$;

3. The graph metrics $\vec{d}_2$, $\vec{d}_3$ and gap metric $\vec{d}_5$ are equivalent to each other. Therefore, the graph metrics give rise to the same robust stability margin as the gap metric [7];

4. The gap metrics d_4 and d_5 can be equivalently expressed by the Georgiou-type formulae, d_7 and d_6, respectively.

The following diagrams show the relations among the discussed topologies and (gap) metrics that will be established.

$$d_1 \geq d_6, \qquad d_2 \geq d_3 \geq d_6 = d_5 \geq d_4 = d_7$$

Diagram 1: Metric relations.

$$\mathcal{T}_{d_1} \overset{\text{P}4.1}{\geq} \mathcal{T}_\omega \overset{\text{T}4.4}{=} \mathcal{T}_{d_2} \overset{\text{T}4.7}{=} \mathcal{T}_{d_3} \overset{\text{T}4.4}{=} \mathcal{T}_{d_5} \overset{\text{P}4.6}{=} \mathcal{T}_{d_6} \overset{\text{P}4.6}{\geq} \mathcal{T}_{d_4} \overset{\text{P}4.6}{=} \mathcal{T}_{d_7}$$
$$\mathcal{T}_\omega \overset{\text{C}3.10}{\geq} \mathcal{T}$$

Diagram 2: Topological relations.

Here, the letters "T," "P," and "C" represent "Theorem," "Proposition," and "Corollary," respectively, and $\mathcal{T}_{d_i}$ means the topology induced by d_i.

4.2. The gap metric d_1. The first result is the following proposition.

PROPOSITION 4.1. *$d_1(\cdot,\cdot)$ is a metric on $\mathbf{N}_\omega(\mathcal{U},\mathcal{Y})$, whose topology, $\mathcal{T}_{d_1}$, is stronger than the weighted graph topology $\mathcal{T}_\omega$.*

Proof. From Lemma 2.1 and the definition of d_1, it follows that to prove d_1 is a metric we need only to verify $d_1(P_1,P_2)=0$ implies $P_1=P_2$.

In fact $d_1(P_1,P_2)=0$ implies that for each $(N_1,D_1)\in \mathrm{rcf}(P_1)$,

$$\inf_{(N_2,D_2)\in \mathrm{rcf}(P_2)} \|(D_1-D_2,N_1-N_2)^\top\|_\omega = 0.$$

So there exists a sequence $\{(N_{2,n},D_{2,n})\}\subset \mathrm{rcf}(P_2)$ with $\|(D_1-D_{2,n},N_1-N_{2,n})^\top\|_\omega \to 0$ as $n\to\infty$, from which it follows that, for each $r>0$, $\gamma((D_1-D_{2,n},N_1-N_{2,n})^\top)(r)\to 0$ as $n\to\infty$. By Theorem 3.5, $P_n \xrightarrow{\mathcal{T}} P$ and therefore $P_2=P_1$.

Now we suppose $P_n\in\mathbf{N}_\omega(\mathcal{U},\mathcal{Y})$ with $d_1(P_n,P)\to 0$ as $n\to\infty$. Then

$$\sup_{(N,D)\in\mathrm{rcf}(P)} \inf_{(N_n,D_n)\in\mathrm{rcf}(P_n)} \|(D-D_n,N-N_n)^\top\|_\omega \to 0.$$

So, for each $ND^{-1}\in\mathrm{rcf}(P)$, $\inf_{(N_n,D_n)\in\mathrm{rcf}(P_n)}\|(D-D_n,N-N_n)^\top\|_\omega\to 0$ and therefore there exists $N_nD_n^{-1}\in\mathrm{rcf}(P_n)$ such that $\|(D_n-D,N_n-N)^\top\|_\omega\to 0$ as $n\to\infty$. This proves $P_n\xrightarrow{\mathcal{T}_\omega}P$ and hence $\mathcal{T}_{d_1}$ is stronger than $\mathcal{T}_\omega$. ∎

4.3. The graph metrics d_2 and d_3. In this subsection, we will show that both d_2 and d_3 are well-defined metrics on $\mathbf{N}_{\mathrm{nor}}(\mathcal{U},\mathcal{Y})$ and the topologies induced are equivalent to the weighted graph topology $\mathcal{T}_\omega$ provided $\omega(r)\geq cr$ with $c>0$.

PROPOSITION 4.2. *$d_2(\cdot,\cdot)$ is a metric defined on $\mathbf{N}_{\mathrm{nor}}(\mathcal{U},\mathcal{Y})$.*

Proof. First, we need to prove d_2 is well-defined, that is, $d_2(P_1,P_2)$ is independent of the choice of normalized coprime factorizations and is finite for all $P_1,P_2\in\mathbf{N}_{\mathrm{nor}}(\mathcal{U},\mathcal{Y})$. So let $P_i\in\mathbf{N}_{\mathrm{nor}}(\mathcal{U},\mathcal{Y})$, $(N_i,D_i),(\hat N_i,\hat D_i)\in\mathrm{nrcf}(P_i)$, $i=1,2$. By Propositions 2.6 and 2.7, there exist $Q_i\in\mathcal{Q}$ $(i=1,2)$ with $\|Q_iu\|=\|Q_i^{-1}u\|=\|u\|$ (for all $u\in\mathcal{U}_s$) such that $\hat D_i=D_iQ_i$, $\hat N_i=N_iQ_i$. Notice, for every stable operator A

$$\|AQ_1\|_\omega \leq \sup_{r>0}\frac{\gamma(A)(\gamma(Q_1)(\omega(r)))}{r} = \sup_{r>0}\frac{\gamma(A)(\omega(r))}{r} = \|A\|_\omega, \tag{4.1}$$

so we have

$$\begin{aligned}\inf_{\substack{Q\in\mathcal{Q}\\ \|Q\|\le 1}} \|(\hat{D}_1-\hat{D}_2Q,\hat{N}_1-\hat{N}_2Q)^\top\|_\omega &= \inf_{\substack{Q\in\mathcal{Q}\\ \|Q\|\le 1}} \|(D_1Q_1-D_2Q_2Q,N_1Q_1-N_2Q_2Q)^\top\|_\omega\\ &\le \inf_{\substack{\hat{Q}\in\hat{\mathcal{Q}}\\ \|\hat{Q}\|\le 1}} \|(D_1-D_2\hat{Q},N_1-N_2\hat{Q})^\top\|_\omega.\end{aligned}$$

Replacing Q_i by Q_i^{-1}, we see that the opposite inequality is also true and, therefore

$$\inf_{\substack{\hat{Q}\in\mathcal{Q}\\ \|\hat{Q}\|\le 1}} \|(D_1-D_2\hat{Q},N_1-N_2\hat{Q})^\top\|_\omega = \inf_{\substack{Q\in\mathcal{Q}\\ \|Q\|\le 1}} \|(\hat{D}_1-\hat{D}_2Q,\hat{N}_1-\hat{N}_2Q)^\top\|_\omega.$$

This shows the value of $\vec{d}_2(P_1,P_2)$ is independent of the choice of normalized coprime factorizations. Similarly, we can prove $\vec{d}_2(P_2,P_1)$ is independent of the choice of normalized coprime factorizations and hence so is d_2. Also, for any $Q\in\mathcal{Q}$ with $\|Q\|\le 1$, we have

$$\|(D_1-D_2Q,N_1-N_2Q)^\top\|_\omega \le \|(D_1,N_1)^\top\|_\omega + \|(D_2,N_2)^\top\|_\omega \le 2 < \infty.$$

Hence d_2 is well-defined.

Next we prove d_2 is a metric.

Obviously, d_2 is symmetric and $d_2(P,P)=0$ for any $P\in\mathbf{N}_{\mathrm{nor}}(\mathcal{U},\mathcal{Y})$. Conversely, suppose $d_2(P_1,P_2)=0$ with $P=N_1D_1^{-1}$, $P_2=N_2D_2^{-1}$. Then for all $n>0$, there exist $Q_n\in\mathcal{Q}$ such that $\|(D_1-D_2Q_n,N_1-N_2Q_n)^\top\|_\omega\to 0$ as $n\to\infty$, from which we have

$$\gamma((D_1-D_2Q_n,N_1-N_2Q_n)^\top)(r)\to 0 \quad \text{for all } r>0 \text{ as } n\to\infty.$$

By Proposition 2.6, (N_2Q_n,D_2Q_n) is also a coprime factorization of $P_n\equiv P_2$ for each n. From Theorem 3.5, it follows that $P_2=P_n\xrightarrow{\mathcal{T}}P_1$ in $\mathbf{N}(\mathcal{U},\mathcal{Y})$, which implies $P_1=P_2$ since the pointwise graph topology is Hausdorff.

To prove the triangle inequality, we suppose $N_iD_i^{-1}$ are normalized coprime factorizations for P_i with $i=1,2,3$. Then for each $\varepsilon>0$, there exists $Q_1,Q_2\in\mathcal{Q}$ with $\|Q_1\|\le 1$, $\|Q_2\|\le 1$ such that

$$\begin{aligned}\|(D_1-D_3Q_1,N_1-N_3Q_1)^\top\|_\omega &\le \vec{d}_2(P_1,P_3)+\varepsilon,\\ \|(D_3-D_2Q_2,N_3-N_2Q_2)^\top\|_\omega &\le \vec{d}_2(P_3,P_2)+\varepsilon.\end{aligned}$$

Since $Q_2Q_1\in\mathcal{Q}$, $\|Q_2Q_1\|\le 1$ and by using (4.1), we have

$$\begin{aligned}\vec{d}_2(P_1,P_2) &\le \|(D_1-D_2Q_2Q_1,N_1-N_2Q_2Q_1)^\top\|_\omega\\ &\le \|(D_1-D_3Q_1,N_1-N_3Q_1)^\top+(D_3Q_1-D_2Q_2Q_1,N_3Q_1-N_2Q_2Q_1)^\top\|_\omega\\ &\le \vec{d}_2(P_1,P_3)+\varepsilon+\|(D_3-D_2Q_2,N_3-N_2Q_2)^\top Q_1\|_\omega\\ &\le \vec{d}_2(P_1,P_3)+\vec{d}_2(P_3,P_2)+2\varepsilon \le d_2(P_1,P_3)+d_2(P_3,P_2)+2\varepsilon.\end{aligned}$$

Since ε is arbitrary, we see that $\vec{d}_2(P_1,P_2)\le d_2(P_1,P_3)+d_2(P_3,P_2)$. By changing the order of P_1, P_2 on the left-hand side (they are arbitrary) and noticing that the right-hand side is symmetric, we have $\vec{d}_2(P_2,P_1)\le d_2(P_1,P_3)+d_2(P_3,P_2)$. This proves the triangle inequality and completes the proof. ∎

PROPOSITION 4.3. *Suppose $c > 0$ and $cr \le w(r)$ for all $r \ge 0$. Then d_3 is a metric which is topologically equivalent to d_2 on $\mathbf{N}_{\rm nor}(\mathcal{U}, \mathcal{Y})$.*

Proof. The proof for the well-definedness and the triangle inequality for d_3 is exactly the same as in Proposition 4.2.

Suppose $d_3(P_1, P_2) = 0$. Hence there exists a sequence $\{Q_n\} \subset \mathcal{Q}^*$ satisfying

$$\|(D_1 - D_2Q_n, N_1 - N_2Q_n)^\top\|_\omega \to 0 \quad \text{as } n \to \infty, \tag{4.2}$$

and therefore there exists $n_0 > 0$ such that

$$\sup_{\|u\| \le \omega(r)} \|(D_1 - D_2Q_n, N_1 - N_2Q_n)^\top u\| < \frac{c}{2} r \quad \text{for all } r > 0,\ n \ge n_0.$$

For any $u \in \mathcal{U}_s$, let $r = \|u\|/c$. Then $\|u\| \le cr \le \omega(r)$ and therefore,

$$\|(D_1 - D_2Q_n, N_1 - N_2Q_n)^\top u\| < \frac{c}{2}\frac{\|u\|}{c} = \frac{1}{2}\|u\| \quad \text{for all } u \in \mathcal{U}_s,\ n \ge n_0.$$

Since (N_1, D_1), (N_2, D_2) are normalized coprime factorizations, we see that

$$\begin{aligned} \|Q_n u\| &= \|(D_2, N_2)^\top Q_n u\| \\ &\ge \|(D_1, N_1)^\top u\| - \|(D_1 - D_2Q_n, N_1 - N_2Q_n)^\top u\| \\ &\ge \|u\| - \frac{1}{2}\|u\| = \frac{1}{2}\|u\| \quad \text{for all } u \in \mathcal{U}_s,\ n \ge n_0. \end{aligned}$$

This means that $\|Q_n^{-1}\| \le 2$ and Q_n^{-1} is stable for all $n > n_0$. By Proposition 2.6, for all $n > n_0$, (N_2Q_n, D_2Q_n) is a coprime factorization of P_2. Also from (4.2), we see that

$$\gamma((D_1 - D_2Q_n, N_1 - N_2Q_n)^\top)(r) \to 0 \quad \text{for all } r > 0, \text{ as } n \to \infty.$$

From Theorem 3.5, it follows that $P_2 \equiv P_n \xrightarrow{\mathcal{T}} P_1$ in $\mathbf{N}(\mathcal{U}, \mathcal{Y})$, which implies $P_1 = P_2$. Hence d_3 is a well-defined metric on $\mathbf{N}_{\rm nor}(\mathcal{U}, \mathcal{Y})$.

To prove the equivalence between d_2 and d_3, we first notice that $d_3 \le d_2$. This yields that convergence under d_2 implies convergence under d_3. On the other hand, let $P_n, P \in \mathbf{N}_{\rm nor}(\mathcal{U}, \mathcal{Y})$ with $(N, D) \in \text{nrcf}(P)$, $(N_n, D_n) \in \text{nrcf}(P_n)$, and $d_3(P_n, P) \to 0$ as $n \to \infty$. Then $\vec{d}_3(P, P_n) \to 0$ which means

$$\inf_{Q \in \mathcal{Q}^*} \|(D - D_nQ_n, N - N_nQ_n)^\top\|_\omega \to 0 \quad \text{as } n \to \infty.$$

This shows that for each $\varepsilon \in (0, c/2]$, there exists $n_\varepsilon > 0$ such that

$$\|(D - D_nQ_n, N - N_nQ_n)^\top\|_\omega < \varepsilon \quad \text{for all } n \ge n_\varepsilon. \tag{4.3}$$

Without loss of generality, we may suppose that $n_{\varepsilon_1} \le n_{\varepsilon_2}$ if $\varepsilon_1 > \varepsilon_2$. By letting $\varepsilon = c/2$, we see that there exists $0 < n_0 \le n_\varepsilon$ such that for each $n \ge n_0$ there is $Q_n \in \mathcal{Q}^*$ satisfying

$$\sup_{\|u\| \le \omega(t)} \|(D - D_nQ_n, N - N_nQ_n)^\top u\| \le \frac{c}{2} r \quad \text{for all } r > 0,\ n \ge n_0.$$

Using the same method as used in the first part (just replace (N_1, D_1) by (N, D) and (N_2, D_2) by (N_n, D_n), respectively), we can prove that Q_n^{-1} is stable for $n \ge n_0$. So from (4.3), it follows that

$$\vec{d}_2(P, P_n) \le \|(D - D_nQ_n, N - N_nQ_n)^\top\|_\omega < \varepsilon \quad \text{for all } n \ge n_\varepsilon$$

and therefore $\vec{d}_2(P, P_n) \to 0$ as $n \to \infty$. Similarly, $\vec{d}_2(P_n, P) \to 0$ and therefore $d_2(P_n, P) \to 0$ as $n \to \infty$. This completes the proof. □

THEOREM 4.4. *Suppose $c > 0$ and $cr \le w(r)$ for all $r \ge 0$. Then the topology induced by either d_2 or d_3 is equivalent to the weighted graph topology $\mathcal{T}_\omega$ on $\mathbf{N}_{\rm nor}(\mathcal{U}, \mathcal{Y})$.*

Proof. By Proposition 4.3, we only need to show $d_3(P_n, P) \to 0$ if and only if $P_n \xrightarrow{\mathcal{T}_\omega} P$, for any $P_n, P \in \mathbf{N}_{\rm nor}(\mathcal{U}, \mathcal{Y})$.

First, suppose $d_3(P_n, P) \to 0$. Then for every normalized coprime factorization $P = ND^{-1}$, $P_n = N_n D_n^{-1}$, there exist $Q_n \in \mathcal{Q}^*$ such that

$$\|(D - D_nQ_n, N - N_nQ_n)^\top\|_\omega \to 0 \quad \text{as } n \to \infty. \tag{4.4}$$

Let $(\hat{N}, \hat{D})$ be an arbitrary coprime factorization of P. By Proposition 2.6, there exists stable operator Q on $\mathcal{U}_s$, with Q^{-1} also stable, such that $\hat{D} = DQ$, $\hat{N} = NQ$, from which we see that

$$\|Q\|_\omega = \|(D, N)^\top Q\|_\omega = \|(\hat{D}, \hat{N})^\top\|_\omega < \infty.$$

Write $\hat{D}_n = D_nQ_nQ$, $\hat{N}_n = N_nQ_nQ$. Then from (4.4), it follows that

$$\begin{aligned}\|(\hat{D} - \hat{D}_n, \hat{N} - \hat{N}_n)^\top\|_\omega &= \|(D - D_nQ_n, N - N_nQ_n)^\top Q\|_\omega \\ &\le \|(D - D_nQ_n, N - N_nQ_n)^\top\|\|Q\|_\omega \\ &= \frac{1}{c}\|(D - D_nQ_n, N - N_nQ_n)^\top\|_\omega\|Q\|_\omega \to 0 \quad \text{as } n \to \infty.\end{aligned}$$

Using the same method as used in Proposition 4.3, we can prove that $(\hat{N}_n, \hat{D}_n)$ is a coprime factorization of P_n for all large n. Hence from Theorem 3.9, we see that $P_n \xrightarrow{\mathcal{T}_\omega} P$.

Secondly, suppose $P_n, P \in \mathbf{N}_{\rm nor}(\mathcal{U}, \mathcal{Y})$ with $P_n \xrightarrow{\mathcal{T}_\omega} P$. Let (N, D) be a normalized coprime factorization of P. Then there exist coprime factorizations $N_nD_n^{-1}$ of P_n with $\|(D - D_n, N - N_n)^\top\|_\omega \to 0$ as $n \to \infty$. Therefore, $\{\|(D_n, N_n)^\top\|\}$ is bounded and $\|(D - D_n, N - N_n)^\top\| \to 0$. Hence

$$\|(D_n, N_n)^\top\| \to \|(D, N)^\top\| = 1 \quad \text{as } n \to \infty \tag{4.5}$$

and for each $\varepsilon > 0$, there exists $n_\varepsilon > 0$ such that

$$\|(D - D_n, N - N_n)^\top u\| \le \varepsilon\|u\| \quad \text{for all } u \in \mathcal{U}_s,\ n > n_\varepsilon. \tag{4.6}$$

Let $\hat{N}_n\hat{D}_n^{-1}$ be a normalized coprime factorization of P_n. Then there exists stable operator U_n on $\mathcal{U}_s$, where U_n^{-1} exists and is stable, such that $D_n = \hat{D}_nU_n$, $N_n = \hat{N}_nU_n$. Since $\|U_nu\| = \|(\hat{D}_n, \hat{N}_n)^\top U_nu\| = \|(D_n, N_n)^\top u\|$ for any $u \in \mathcal{U}_s$, we see $\{\|U_n\|_\omega\}$ is bounded and from (4.5) it follows that

$$\|U_n\| = \|(\hat{D}_n, \hat{N}_n)^\top U_n\| = \|(D_n, N_n)^\top\| \to 1 \quad \text{as } n \to \infty.$$

From (4.6), it follows that for each $u \in \mathcal{U}_s$ and each $n > n_\varepsilon$ that

$$\|U_nu\| = \|(D_n, N_n)^\top u\| \ge \|(D, N)^\top u\| - \|(D_n - D, N_n - N)^\top u\| \ge (1 - \varepsilon)\|u\|,$$

which implies that $\|U_n^{-1}u\| \le \frac{1}{1-\varepsilon}\|u\|$ and therefore $\|U_n^{-1}\| \le \frac{1}{1-\varepsilon}$. Since $\|U_n^{-1}\| \ge 1/\|U_n\|$, we see $\|U_n^{-1}\| \to 1$ as $n \to \infty$.

Let $Q_n u = U_n u/\|U_n\|$ for each $u \in \mathcal{U}_s$. Then $\|Q_n\| \le 1$ and since $Q_n^{-1} = U_n^{-1} \cdot \|U_n\|$ exists and is stable, we have $Q_n \in \mathcal{Q}^*$. Also

$$\begin{aligned}\|(\hat{D}_n U_n - \hat{D}_n Q_n, \hat{N}_n U_n - \hat{N}_n Q_n)^\top\|_\omega &= \|(\hat{D}_n, \hat{N}_n)^\top (U_n - Q_n)\|_\omega \\ &= \|(U_n - Q_n)\|_\omega = \frac{|\|U_n\| - 1|}{\|U_n\|}\|U_n\|_\omega \to 0,\end{aligned}$$

which implies

$$\begin{aligned}\vec{d}_3(P, P_n) &\le \|(D - \hat{D}_n Q_n, N - \hat{N}_n Q_n)^\top\|_\omega \\ &\le \|(D - \hat{D}_n U_n, N - \hat{N}_n U_n)^\top\|_\omega \\ &\quad + \|(\hat{D}_n U_n - \hat{D}_n Q_n, \hat{N}_n U_n - \hat{N}_n Q_n)^\top\|_\omega \to 0 \quad \text{as } n \to \infty.\end{aligned}$$

Similarly, for $\tilde{Q}_n u = U_n^{-1} u/\|U_n^{-1}\|$, we can prove

$$\vec{d}_3(P_n, P) \le \|(\hat{D}_n - D\tilde{Q}_n, \hat{N}_n - N\tilde{Q}_n)^\top\|_\omega \to 0 \quad \text{as } n \to \infty.$$

This shows $d_3(P_n, P) \to 0$ and completes the proof. □

We remark that the first part of the proof shows, in the case $cr \le \omega(r)$, that $P_n \xrightarrow{\mathcal{T}_\omega} P$ in $\mathbf{N}_{\text{nor}}(\mathcal{U}, \mathcal{Y})$ if and only if there exist normalized coprime factorization (N, D) of P and coprime factorizations (N_n, D_n) of P_n such that $\|(D - D_n, N - N_n)^\top\|_\omega \to 0$ as $n \to \infty$.

In the case of $\omega(r) = r$ and stability is taken to be in the sense of linear gain, the above theorem shows that the graph topology induced by the linear gain is metrizable.

4.4. The gap metrics d_4, d_5, d_6, and d_7. In this subsection, we present the metric properties of $d_4, \dots, d_7$ over the subset $\mathbf{N}_{\text{nor}}(\mathcal{U}, \mathcal{Y})$. In particular, the equivalence between the weighted graph topology and the topologies induced by either d_5 or d_6 will be established.

Using the same method as used in [7], we can prove that d_4, d_5 are pseudometrics on the set of signal operators from $\mathcal{U}$ to $\mathcal{Y}$ provided $\omega(r) \ge r$ for all $r > 0$. Here pseudometric means that $d_4(P_1, P_2) = 0$ (resp., $d_5(P_1, P_2) = 0$) does not necessarily imply $P_1 = P_2$ unless extra conditions are imposed. Moreover, as in [7], they are only "generalized" pseudometrics, which means that possibly (say) $d_5(P_1, P_2) = \infty$ for some P_1, P_2. The following comparison results show that they both become well-defined metrics if restricted to $\mathbf{N}_{\text{nor}}(\mathcal{U}, \mathcal{Y})$ (no extra condition required).

We first give a key lemma.

LEMMA 4.5. *Suppose $P_i \in \mathbf{N}_{\text{nor}}(\mathcal{U}, \mathcal{Y})$ with $(D_i, N_i) \in \text{nrcf}(P_i)$, $i = 1, 2$. Then there exists a mapping $\Phi : \text{Graph}(P_1) \to \text{Graph}(P_2)$ if and only if there exists a mapping $Q : \mathcal{U}_s \to \mathcal{U}_s$ such that*

$$\Phi \begin{pmatrix} D_1 \\ N_1 \end{pmatrix} u = \begin{pmatrix} D_2 \\ N_2 \end{pmatrix} Qu \quad \text{for all } u \in \mathcal{U}_s. \tag{4.7}$$

Moreover,

(i) *Φ is surjective if and only if Q is surjective;*

(ii) *$\|Q\| = \|\Phi\|$ and $\gamma(\Phi)(r) = \gamma(Q)(r)$ for any $r > 0$ (so Φ is stable if and only if Q is stable);*

(iii) *Φ is injective if and only if Q is injective;*

(iv) *$\|Q^{-1}\| = \|\Phi^{-1}\| =: \|\Phi^{-1}|_{\mathcal{M}_2}\|$ and $\gamma(\Phi^{-1})(r) = \gamma(Q^{-1})(r)$ for any $r > 0$;*

(v) *Φ is causal if and only if Q is causal, and $\Phi 0 = 0$ if and only if $Q0 = 0$.*

Proof. Write $\mathcal{M}_i = \mathrm{Graph}(P_i)$ for $i = 1, 2$.

Let $\Phi : \mathcal{M}_1 \to \mathcal{M}_2$ be a given mapping. Then, for any $u \in \mathcal{U}_s$, by Proposition 2.5, $\Phi(D_1, N_1)^\top u \in \mathcal{M}_2$ and therefore there exists $v_u \in \mathcal{U}_s$ such that $\Phi(D_1, N_1)^\top u = (D_2, N_2)^\top v_u$. Since $(D_2, N_2)^\top$ is left invertible, such a point v_u is unique. This yields that the mapping $Qu = v_u$ is well defined on $\mathcal{U}_s$ and satisfies (4.7).

Conversely, let Q be a given mapping on $\mathcal{U}_s$. For any $w \in \mathcal{M}_1$, let $\Phi w = (D_2, N_2)^\top Q L_1 w$, where L_1 is the left inverse of $(D_1, N_1)^\top$ which is stable by the definition of the coprime factorization. Then, obviously, Φ is a well-defined mapping from $\mathcal{M}_1$ to $\mathcal{M}_2$ satisfying (4.7).

Now we prove the other claims.

(i) First, we suppose $\Phi : \mathcal{M}_1 \to \mathcal{M}_2$ is given and surjective. Since $(D_1, N_1)^\top : \mathcal{U}_s \to \mathcal{M}_1$ is surjective, for any $v \in \mathcal{U}_s$ there exists $u \in \mathcal{U}_s$ with $\Phi(D_1, N_1)^\top u = (D_2, N_2)^\top v$. The left invertibility of $(D_2, N_2)^\top$ and (4.7) show $Qu = v$. Therefore, Q is surjective.

If Q is surjective on $\mathcal{U}_s$, then for any $w \in \mathcal{M}_2$, the surjectivity of $(D_2, N_2)^\top$ implies the existence of $u \in \mathcal{U}_s$ such that $(D_2, N_2)^\top Qu = w$. Hence $\Phi(D_1, N_1)^\top u = w$ which shows that Φ is surjective.

(ii) From (4.7), we see that

$$\|Qu\| = \left\| \begin{pmatrix} D_2 \\ N_2 \end{pmatrix} Qu \right\| = \left\| \Phi \begin{pmatrix} D_1 \\ N_1 \end{pmatrix} u \right\| \quad \text{for all } u \in \mathcal{U}_s. \tag{4.8}$$

Since $\|u\| = \|(D_1, N_1)^\top u\|$, $\mathcal{M}_1 = (D_1, N_1)^\top \mathcal{U}_s$, the conclusions follow.

(iii) From (4.7) and the left invertibility of $(D_i, N_i)^\top$ $(i = 1, 2)$, it follows that

$$\begin{aligned}
\Phi \text{ is injective } &\Leftrightarrow \Phi w_1 = \Phi w_2 \text{ implies } w_1 = w_2 \text{ for any } w_1, w_2 \in \mathcal{M}_1 \\
&\Leftrightarrow \Phi \begin{pmatrix} D_1 \\ N_1 \end{pmatrix} u_1 = \Phi \begin{pmatrix} D_1 \\ N_1 \end{pmatrix} u_2 \text{ implies } u_1 = u_2 \text{ for any } u_1, u_2 \in \mathcal{U}_s \\
&\Leftrightarrow \begin{pmatrix} D_2 \\ N_2 \end{pmatrix} Qu_1 = \begin{pmatrix} D_2 \\ N_2 \end{pmatrix} Qu_2 \text{ implies } u_1 = u_2 \text{ for any } u_1, u_2 \in \mathcal{U}_s \\
&\Leftrightarrow Qu_1 = Qu_2 \text{ implies } u_1 = u_2 \text{ for any } u_1, u_2 \in \mathcal{U}_s \\
&\Leftrightarrow Q \text{ is injective.}
\end{aligned}$$

(iv) Since $\|w\| \le \|\Phi^{-1}\| \|\Phi w\|$ for any $w \in \mathcal{M}_1$, we have

$$\|u\| = \left\| \begin{pmatrix} D_1 \\ N_1 \end{pmatrix} u \right\| \le \|\Phi^{-1}\| \left\| \Phi \begin{pmatrix} D_1 \\ N_1 \end{pmatrix} u \right\| = \|\Phi^{-1}\| \left\| \begin{pmatrix} D_2 \\ N_2 \end{pmatrix} Qu \right\| = \|\Phi^{-1}\| \|Qu\|$$

for any $u \in \mathcal{U}_s$. So $\|Q^{-1}\| \le \|\Phi^{-1}\|$. Similarly, for any $w = (D_1, N_1)^\top u \in \mathcal{M}_1$,

$$\left\| \begin{pmatrix} D_1 \\ N_1 \end{pmatrix} u \right\| = \|u\| \le \|Q^{-1}\| \|Qu\| = \|Q^{-1}\| \left\| \begin{pmatrix} D_2 \\ N_2 \end{pmatrix} Qu \right\| = \|Q^{-1}\| \left\| \Phi \begin{pmatrix} D_1 \\ N_1 \end{pmatrix} u \right\|$$

which gives the reverse inequality. Hence $\|Q^{-1}\| = \|\Phi^{-1}\|$.

For any $r > 0$, (4.8) and the surjectivity of $\Phi, (D_i, N_i)^\top$, and Q yield

$$\gamma(Q^{-1})(r) = \sup_{\|Qv\| \le r} \|v\| = \sup_{\|\Phi(D_1, N_1)^\top v\| \le r} \left\| \begin{pmatrix} D_1 \\ N_1 \end{pmatrix} v \right\| = \sup_{\|w\| \le r} \|\Phi^{-1} w\| = \gamma(\Phi^{-1})(r).$$

(v) Let L_1, L_2 be the associated operators to $(D_1, N_1)^\top$, $(D_2, N_2)^\top$, respectively. By applying L_2 to (4.7), we have $Qu = L_2\Phi(D_1, N_1)^\top$. By the definition of Φ, $\Phi w = (D_2, N_2)^\top Q L_1$. So, the conclusions follow from the preassumptions on signal operators. This completes the proof. □

PROPOSITION 4.6. *$\vec{d}_5(P_1, P_2) = \vec{d}_6(P_1, P_2)$, $\vec{d}_4(P_1, P_2) = \vec{d}_7(P_1, P_2)$ for $P_i \in \mathbf{N}_{\text{nor}}(\mathcal{U}, \mathcal{Y})$, $i = 1, 2$.*

Proof. Let Q be a given stable bijective mapping on $\mathcal{U}_s$. Then there exists a stable and bijective map $\Phi : \mathcal{M}_1 \to \mathcal{M}_2$ satisfying (4.7), for which

$$\|\Phi - I\|_\omega = \|(D_1, N_1)^\top - (D_2, N_2)^\top Q\|_\omega. \tag{4.9}$$

Therefore, $\vec{d}_5(P_1, P_2) \le \|(D_1, N_1)^\top - (D_2, N_2)^\top Q\|_\omega$ and $\vec{d}_5(P_1, P_2) \le \vec{d}_6(P_1, P_2)$ as Q is arbitrary.

Since $\Phi_1(D_1, N_1)^\top u = (D_2, N_2)^\top u$ is a bijective operator from $\mathcal{M}_1$ to $\mathcal{M}_2$ and $\|\Phi_1 - I\|_\omega < \infty$, we have

$$\vec{d}_5(P_1, P_2) = \inf\{\|\Phi - I\|_\omega : \Phi : \mathcal{M}_1 \to \mathcal{M}_2 \text{ bijective } \|\Phi - I\|_\omega < \infty\}.$$

Notice that $\|\Phi - I\|_\omega < \infty$ implies the stability of Φ. So, given any bijective map $\Phi : \mathcal{M}_1 \to \mathcal{M}_2$ with $\|\Phi - I\|_\omega < \infty$, by Lemma 4.5, there exists a stable, bijective mapping Q on $\mathcal{U}_s$ satisfying (4.7) and, therefore, (4.9). Hence $\vec{d}_6(P_1, P_2) \le \|\Phi - I\|_\omega$ which indicates that $\vec{d}_6(P_1, P_2) \le \vec{d}_5(P_1, P_2)$. This proves that $\vec{d}_6(P_1, P_2) = \vec{d}_5(P_1, P_2)$. The equality $\vec{d}_4(P_1, P_2) = \vec{d}_7(P_1, P_2)$ can be proved similarly. □

THEOREM 4.7. *d_5, d_6 are well-defined metrics on $\mathbf{N}_{\text{nor}}(\mathcal{U}, \mathcal{Y})$ and the graph topology $\mathcal{T}_\omega$ is equivalent to the topology induced by either d_5 or d_6, provided $r \le \omega(r)$ for all $r \ge 0$.*

Proof. Using the same methods as in Propositions 4.2 and 4.3, we see that d_6 is well-defined and $d_6(P_1, P_2) = 0$ if and only if $P_1 = P_2$ on $\mathbf{N}_{\text{nor}}(\mathcal{U}, \mathcal{Y})$. By Proposition 4.6, d_5 satisfies the same property.

To prove the triangle inequality for d_5, we suppose $P_1, P_2, P_3 \in \mathbf{N}_{\text{nor}}(\mathcal{U}, \mathcal{Y})$ and $\Phi_1 : \text{Graph}(P_1) \to \text{Graph}(P_2), \Phi_i : \text{Graph}(P_2) \to \text{Graph}(P_3)$ are bijective mappings. Then $\Phi := \Phi_2\Phi_1$ is a bijective mapping from $\text{Graph}(P_1)$ to $\text{Graph}(P_3)$ and $\Phi - I = (\Phi_2 - I)\Phi_1 - I$. So

$$\begin{aligned}\|\Phi - I\|_\omega &\le \|\Phi_2 - I\|\|\Phi_1\|_\omega + \|\Phi_1 - I\|_\omega \\ &\le \|\Phi_2 - I\|_\omega\|\Phi_1\|_\omega + \|\Phi_1 - I\|_\omega \\ &\le \|\Phi_2 - I\|_\omega(\|\Phi_1 - I\|_\omega) + 1 + \|\Phi_1 - I\|_\omega\end{aligned}$$

and therefore

$$\hat{d}_5(P_1, P_3) \le \hat{d}_5(P_1, P_2) + \hat{d}_5(P_2, P_3).$$

This means that d_5 satisfies the triangle inequality. Hence d_5 is a well-defined metric on $\mathbf{N}_{\text{nor}}(\mathcal{U}, \mathcal{Y})$ and so is d_6 because of Proposition 4.6.

Since $d_6 \le d_3$ and by Theorem 4.4, the convergence of the sequence under $\mathcal{T}_\omega$ implies the convergence under d_6. Conversely, if $d_6(P, P_n) \to 0$ as $n \to \infty$, then by using the same method as in Theorem 4.4 (see the Theorem's remark), we can prove that $P_n \xrightarrow{\mathcal{T}_\omega} P$. This shows the equivalence between $\mathcal{T}_\omega$ and the topology induced by either d_6 or d_5. □

Proposition 4.6 and Theorem 4.7 suggest that the two metrics d_3 and d_6 might be equivalent (we already know that $d_6 \le d_3$). In fact, Georgiou [4] has proved

$d_3(P_1,P_2) \le 2d_6(P_1,P_2)$ in the linear setting. In the nonlinear setting and in the case where $(D_2,N_2)^\top$ is incrementally stable, are where

$$\|(D_2,N_2)^\top\|_\triangle := \sup\left\{\frac{\|(D_2,N_2)^\top u_1 - (D_2,N_2)^\top u_2\|_\tau}{\|u_1-u_2\|_\tau} : \tau>0,\ u_1,u_2\in\mathcal{U}_s\right\} < \infty;$$

this claim can be proved by exactly the same technique as in [4].

Finally, we consider the relationship between d_1 and d_6. For $(N_i,D_i)\in \mathrm{rcf}(P_i)$, $i=1,2$, by Proposition 2.6, we have that

$$\inf_{Q\in\mathcal{Q}}\left\|\begin{pmatrix}D_1\\N_1\end{pmatrix}-\begin{pmatrix}D_2\\N_2\end{pmatrix}Q\right\|_\omega = \inf_{(\tilde N_2,\tilde D_2)\in\mathrm{rcf}(P_2)}\left\|\begin{pmatrix}D_1\\N_1\end{pmatrix}-\begin{pmatrix}\tilde D_2\\\tilde N_2\end{pmatrix}\right\|_\omega \le \vec d_1(P_1,P_2).$$

This gives a direct relation between d_1 and d_6 as below.

PROPOSITION 4.8. *For $P_i\in\mathbf{N}_{\mathrm{nor}}(\mathcal{U},\mathcal{Y})$ with $(D_i,N_i)\in\mathrm{nrcf}(P_i)$, $i=1,2$, $\vec d_1(P_1,P_2)\ge\vec d_6(P_1,P_2)$ and therefore $d_1(P_1,P_2)\ge d_6(P_1,P_2)$.*

5. Robustness of stability of nonlinear feedback systems. The importance of graph topology in the linear case is well known. In this section, we will show that it may also play a significant role in the nonlinear case by considering the system described by the configuration of Figure 5.1.

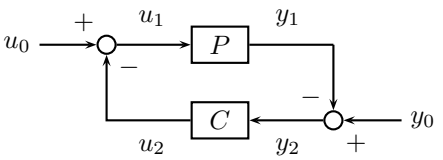


FIG. 5.1. *Standard feedback configuration.*

In this configuration, $u_i\in\mathcal{U}$, $y_i\in\mathcal{Y}$ for $i=0,1,2$, and both the plant P and compensator C are, in general, causal and nonlinear. We suppose all systems in this section are well-posed, that is, for each $(u_0,y_0)^\top\in\mathcal{U}_s\times\mathcal{Y}_s$, there exist unique signals $u_1,u_2\in\mathcal{U}$ and $y_1,y_2\in\mathcal{Y}$ such that

$$u_0=u_1+u_2,\qquad y_0=y_1+y_2,\qquad y_1=Pu_1,\qquad u_2=Cy_2,$$

and the feedback operator

$$H_{P,C}:\mathcal{W}_s\to\mathcal{W}\times\mathcal{W}:\begin{pmatrix}u_0\\y_0\end{pmatrix}\mapsto\left(\begin{pmatrix}u_1\\y_1\end{pmatrix},\begin{pmatrix}u_2\\y_2\end{pmatrix}\right)$$

is causal. Here, $\mathcal{W}_s=\mathcal{U}_s\times\mathcal{Y}_s,\mathcal{W}=\mathcal{U}\times\mathcal{Y}$. The feedback stability of this system is the requirement that $H_{P,C}$ is stable in a suitable sense. We are concerned with the following robustness problem: when is $H_{P_\lambda,C}$ stable given that $H_{P,C}$ is stable and P_λ is a perturbation to P?

In [7], this problem has been studied using a gap metric. Particularly, in the case where the linear gain is considered, it is proved that if $H_{P,C}$ is gain stable and P_λ is close enough to P in the sense of gap metric, then $H_{P_\lambda,C}$ is gain stable. Similar results are also given when $H_{P,C}$ is (gf)-stable with superlinear growth. However, in the (gf)-stability case, the notion of convergence was not made explicit as no topology was indicated. In this paper, we consider the robustness of (gf)-stability when the

convergence of P_λ to P is in the sense of any of the two graph topologies defined in the previous sections.

We suppose Λ is a topological space and for each $\lambda \in \Lambda$, P_λ is a perturbation to the nominal plant $P = P_{\lambda_0}$. Define $\mathcal{M} = \mathrm{Graph}(P)$, $\mathcal{M}_\lambda = \mathrm{Graph}(P_\lambda)$, $\mathcal{N} = \mathrm{Graph}(C)$, and let $\Pi_{\mathcal{M}//\mathcal{N}}$ be the parallel projection which maps $(u_0, y_0)^\top$ to $(u_1, y_1)^\top$ and $\Pi_{\mathcal{N}//\mathcal{M}} = I - \Pi_{\mathcal{M}//\mathcal{N}}$. It is known that $H_{P,C}$ is (gf)-stable (resp., gain stable) if and only if $\Pi_{\mathcal{M}//\mathcal{N}}$ is (gf)-stable (resp., gain stable); see [7].

A signal operator $F : \mathcal{U} \to \mathcal{Y}$ is said to be causally extendable if, for each $u \in \mathcal{U}$, $y = Fu$ and for each $\tau > 0$, there exists $u_\tau \in \mathrm{Dom}(F)$ such that $T_\tau(u, y)^\top = T_\tau(u_\tau, y_\tau)^\top$ with $y_\tau = Fu_\tau$. Henceforth, we suppose that P, C and each P_λ are causally extendable.

LEMMA 5.1. *Suppose Φ is a surjective mapping from $\mathcal{M}$ to $\mathcal{M}_\lambda$. Then, for any $z \in \mathcal{W}_s$ and any $\tau > 0$, there exists $x_\tau \in \mathcal{W}_s$ such that*

$$T_\tau z = T_\tau x_\tau + T_\tau(\Phi - I)\Pi_{\mathcal{M}//\mathcal{N}} T_\tau x_\tau \quad \textit{and} \quad T_\tau \Pi_{\mathcal{M}_\lambda//\mathcal{N}} z = T_\tau \Phi \Pi_{\mathcal{M}//\mathcal{N}} T_\tau x_\tau.$$

Proof. Let $H_{P_\lambda,C} z = (z_1, z_2)$ with $z_1 = (u_1, P_\lambda u_1)^\top$, $z_2 = (Cy_2, y_2)^\top$ for some $u_1 \in \mathcal{U}$, $y_2 \in \mathcal{Y}$. Then $z = z_1 + z_2$ and $\Pi_{\mathcal{M}_\lambda//\mathcal{N}} z = z_1$. By the causal extendability, for each $\tau > 0$, there exist $z_1^\tau \in \mathcal{M}_\lambda$, $z_2^\tau \in \mathcal{N}$ such that $T_\tau z_1 = T_\tau z_1^\tau$, $T_\tau z_2 = T_\tau z_2^\tau$. Since Φ is surjective from $\mathcal{M}$ to $\mathcal{M}_\lambda$, there exists $z_3^\tau \in \mathcal{M}$ with $\Phi z_3^\tau = z_1^\tau$. Write $x_\tau = z_3^\tau + z_2^\tau$. Then $x_\tau \in \mathcal{W}_s$ and $\Pi_{\mathcal{M}//\mathcal{N}} x_\tau = z_3^\tau$, $\Pi_{\mathcal{N}//\mathcal{M}} x_\tau = z_2^\tau$. Hence

$$\begin{aligned} T_\tau z &= T_\tau z_1 + T_\tau z_2 = T_\tau z_1^\tau + T_\tau z_2^\tau = T_\tau \Phi z_3^\tau + T_\tau z_2^\tau \\ &= T_\tau \Phi \Pi_{\mathcal{M}//\mathcal{N}} x_\tau + T_\tau \Pi_{\mathcal{N}//\mathcal{M}} x_\tau = T_\tau \Phi \Pi_{\mathcal{M}//\mathcal{N}} T_\tau x_\tau + T_\tau \Pi_{\mathcal{N}//\mathcal{M}} T_\tau x_\tau \\ &= T_\tau(\Phi - I)\Pi_{\mathcal{M}//\mathcal{N}} T_\tau x_\tau + T_\tau x_\tau \end{aligned}$$

and

$$T_\tau \Pi_{\mathcal{M}_\lambda//\mathcal{N}} z = T_\tau z_1 = T_\tau \Phi z_3^\tau = T_\tau \Phi \Pi_{\mathcal{M}//\mathcal{N}} x_\tau = T_\tau \Phi \Pi_{\mathcal{M}//\mathcal{N}} T_\tau x_\tau. \qquad \square$$

For our main results, we will always require that the nominal plant satisfies a k-coercive condition as stated below; note that this assumption will not be imposed on the perturbed plant P_λ.

DEFINITION 5.2. *A signal operator $P : \mathcal{U} \to \mathcal{Y}$ is said to be k-coercive, with $k \in \mathcal{K}_\infty$, if P has a coprime factorization (N, D) such that*

$$\|(D, N)^\top u\| \geq k(\|u\|) \quad \textit{for all } u \in \mathcal{U}_s. \tag{5.1}$$

Notice that P is k-coercive if and only if

$$\|Lw\| \leq k^{-1}(\|w\|) \quad \text{for all } w \in \mathrm{Graph}(P), \tag{5.2}$$

where L is the associated operator of (N, D). Hence any operator P with coprime factors is $\gamma(L)^{-1}$-coercive, where L is the associated operator of a coprime factorization, since (5.2) always holds with $k^{-1}(r) = \gamma(L)(r)$. It is of interest to observe that a linear operator with coprime factors is always k-coercive with $k(r) = cr$, $c > 0$. Also note that if P has a normalized coprime factorization, then P is 1-coercive and therefore c-coercive for any $c > 0$.

In the case when $k(r) = cr$ is linear, (5.2) is required by James, Smith, and Vinnicombe [10] in their definition of (right) coprime factorization, while (5.1) is

required by Verma [18] in one of his definitions and exploited for the stability of another system in the sense of linear gain.

PROPOSITION 5.3. *Suppose that the nominal plant P is k-coercive and $\lambda \mapsto P_\lambda$ is continuous at λ_0 under a weighted topology $\mathcal{T}_\omega$ with $w \in \mathcal{K}_\infty$. Then, for each λ, there exists a surjective mapping $\Phi_\lambda : \mathcal{M} \to \mathcal{M}_\lambda$ such that*

$$\sup_{r>0} \frac{\gamma(I-\Phi_\lambda)(k(\omega(r)))}{r} \to 0, \quad \text{as } \lambda \to \lambda_0. \tag{5.3}$$

Proof. Let ND^{-1} be the coprime factorization of $P = P_{\lambda_0}$ satisfying the coercive condition (5.1). From Corollary 3.11, it follows that P_λ has coprime factorization $N_\lambda D_\lambda^{-1}$ such that

$$\sup_{r>0} \frac{\gamma((D-D_\lambda, N-N_\lambda)^\top)(\omega(r))}{r} \to 0, \quad \text{as } \lambda \to \lambda_0. \tag{5.4}$$

For each $\lambda > 0$ and each $u \in \mathcal{U}_s$, let

$$\Phi_\lambda((Du, Nu)^\top) = ((D_\lambda u, N_\lambda u)^\top). \tag{5.5}$$

Φ_λ is a well-defined, causal, and surjective mapping from $\mathcal{M}$ to $\mathcal{M}_\lambda$ since $\Phi_\lambda w = ((D_\lambda, N_\lambda)^\top Lw)$ with L the left inverse of $(D,N)^\top$.

Now let $r > 0$, $(Du, Nu)^\top \in \mathcal{M}$ with $u \in \mathcal{U}_s$, and $\|(Du, Nu)^\top\| \le k(w(r))$. From (5.1), it follows that $\|u\| \le w(r)$, which implies

$$\|((D-D_\lambda)u, (N-N_\lambda)u)^\top\| \le \sup_{\|v\| \le w(r)} \|((D-D_\lambda)v, (N-N_\lambda)v)^\top\|.$$

Therefore

$$\begin{aligned} \gamma(I-\Phi_\lambda)(k(w(r))) &= \sup_{\substack{(Du,Nu)^\top \in \mathrm{Dom}(I-\Phi_\lambda) \\ \|(Du,Nu)^\top\| \le k(w(r))}} \|((D-D_\lambda)u, (N-N_\lambda)u)^\top\| \\ &\le \sup_{\|v\| \le w(r)} \|((D-D_\lambda)v, (N-N_\lambda)v)^\top\| \\ &= \gamma((D-D_\lambda, N-N_\lambda)^\top)(w(r)). \end{aligned} \tag{5.6}$$

By (5.4), we see that (5.3) holds. ∎

Remark. From (5.3), we see that $\|I-\Phi_\lambda\|_{k\circ\omega} \to 0$ which implies $\vec{\delta}(P, P_\lambda) \to 0$ under a new weighted function $\omega_1 = k \circ \omega$. However, we cannot show whether $\vec{\delta}(P_\lambda, P) \to 0$ unless each P_λ is also k-coercive. Also notice here the l.a.c. assumption was not imposed. So (5.3) does not imply $P_\lambda \to P$ under d_4.

Similarly, for the pointwise continuity, we have the following proposition.

PROPOSITION 5.4. *Suppose that P is k-coercive and $\lambda \mapsto P_\lambda$ is continuous at λ_0 under the pointwise topology $\mathcal{T}$. Then, the mapping $\Phi_\lambda : \mathcal{M} \to \mathcal{M}_\lambda$ defined in* (5.5) *is surjective and $\gamma(I-\Phi_\lambda)(r) \to 0$ for each $r > 0$, as $\lambda \to \lambda_0$.*

Henceforth, we define the map Φ_λ to be as in Proposition 5.3 or 5.4 and give robustness results under each graph topology. The following results follow as consequences of Proposition 5.3 or 5.4 and the results of [7]; however, we give the entire proofs for completeness. First, we consider the case when weighted topology is involved.

THEOREM 5.5. *Suppose P is k-coercive and $H_{P,C}$ is (gf)-stable. If $\lambda \mapsto P_\lambda$ is continuous at λ_0 under a weighted topology $\mathcal{T}_\omega$ with $\omega \in \mathcal{K}_\infty$ and for all $r > 0$*

$$\gamma(\Pi_{\mathcal{M}//\mathcal{N}})(r) \le k(w(r)), \tag{5.7}$$

then, for any $n > 0$, there exists a neighborhood V_n of λ_0 such that $H_{P_\lambda,C}$ is (gf)-stable for $\lambda \in V_n$ and

$$\gamma(\Pi_{\mathcal{M}_\lambda//\mathcal{N}})(r) \le \gamma(\Pi_{\mathcal{M}//\mathcal{N}})\left(\frac{n+1}{n}r\right) + \frac{1}{n}r.$$

Proof. Let $\tau > 0$, $r > 0$, and $z \in \mathcal{W}$ be given with $\|z\|_\tau \le r$. By Proposition 5.3, for each λ, there exists a surjective mapping $\Phi_\lambda : \mathcal{M} \to \mathcal{M}_\lambda$ satisfying (5.3). From Lemma 5.1, it follows that for each λ, there exists $x_\lambda^\tau \in \mathcal{W}_s$ such that

$$T_\tau x_\lambda^\tau = T_\tau z - T_\tau(\Phi_\lambda - I)\Pi_{\mathcal{M}//\mathcal{N}} T_\tau x_\lambda^\tau, \qquad T_\tau \Pi_{\mathcal{M}_\lambda//\mathcal{N}} z = T_\tau \Phi_\lambda \Pi_{\mathcal{M}//\mathcal{N}} T_\tau x_\lambda^\tau. \tag{5.8}$$

By (5.3) and the properties of γ, there exists a neighborhood V_n of λ_0 such that

$$\frac{\gamma(\Phi_\lambda - I)(\gamma(\Pi_{\mathcal{M}//\mathcal{N}})(\|x_\lambda^\tau\|_\tau))}{\|x_\lambda^\tau\|_\tau} \le \frac{\gamma(\Phi_\lambda - I)(k(w(\|x_\lambda^\tau\|_\tau)))}{\|x_\lambda^\tau\|_\tau} < \frac{1}{n+1}$$

for all $n > 0$ and $\lambda \in V_n$. So, from (5.8), it follows that

$$\begin{aligned}\|x_\lambda^\tau\|_\tau &\le \|z\|_\tau + \|(\Phi_\lambda - I)\Pi_{\mathcal{M}//\mathcal{N}} T_\tau x_\lambda^\tau\|_\tau \\ &\le \|z\|_\tau + \gamma(\Phi_\lambda - I)(\gamma(\Pi_{\mathcal{M}//\mathcal{N}})(\|x_\lambda^\tau\|_\tau)) \\ &\le \|z\|_\tau + \frac{1}{n+1}\|x_\lambda^\tau\|_\tau,\end{aligned}$$

which implies that $\|x_\lambda^\tau\|_\tau \le (n+1)\|z\|_\tau/n \le (n+1)r/n$ for all $\lambda \in V_n$. By (5.8), we have

$$T_\tau \Pi_{\mathcal{M}_\lambda//\mathcal{N}} z = T_\tau \Phi_\lambda \Pi_{\mathcal{M}//\mathcal{N}} T_\tau x_\lambda^\tau = T_\tau \Pi_{\mathcal{M}//\mathcal{N}} T_\tau x_\lambda^\tau + T_\tau(\Phi_\lambda - I)\Pi_{\mathcal{M}//\mathcal{N}} T_\tau x_\lambda^\tau$$

and therefore

$$\begin{aligned}\|\Pi_{\mathcal{M}_\lambda//\mathcal{N}} z\|_\tau &\le \|\Pi_{\mathcal{M}//\mathcal{N}} T_\tau x_\lambda^\tau\|_\tau + \|(\Phi_\lambda - I)\Pi_{\mathcal{M}//\mathcal{N}} T_\tau x_\lambda^\tau\|_\tau \\ &\le \gamma(\Pi_{\mathcal{M}//\mathcal{N}})\left(\frac{n+1}{n}r\right) + \frac{1}{n+1}\|x_\lambda\|_\tau \le \gamma(\Pi_{\mathcal{M}//\mathcal{N}})\left(\frac{n+1}{n}r\right) + \frac{1}{n}r.\end{aligned}$$

Since τ is arbitrary, $\gamma(\Pi_{\mathcal{M}_\lambda//\mathcal{N}})(r) \le \gamma(\Pi_{\mathcal{M}//\mathcal{N}})(\frac{n+1}{n}r)r + \frac{1}{n}r < \infty$ for $\lambda \in V_0$. ∎

We remark that condition (5.7) can be replaced by the weaker condition

$$\gamma(\Pi_{\mathcal{M}//\mathcal{N}})(r) \le k(cw(r)) \quad \text{with } c > 0$$

since P is also kc-coercive due to the remark made after Definition 5.2. This claim is also supported by Theorem 3.12 from which we see that $\lambda \mapsto P_\lambda$ is also continuous at λ_0 under a weighted topology $\mathcal{T}_{c\omega}$, so the ω in (5.7) can be replaced by $c\omega$. This replacement gives a weaker bound for $\gamma(\Pi_{\mathcal{M}//\mathcal{N}})(r)$.

In the case of the pointwise topology, we have the following theorem.

THEOREM 5.6. *Suppose that P is k-coercive, $H_{P,C}$ is (gf)-stable and $\lambda \mapsto P_\lambda$ is continuous at λ_0 under the pointwise topology $\mathcal{T}$. If, for each λ, $T_\tau(\Phi_\lambda - I)\Pi_{\mathcal{M}//\mathcal{N}}$ is continuous and compact as a mapping from any subset $S_r = \{w \in \mathcal{W} : \sup_{\tau>0} \|w\|_\tau \leq r\}$ to $\mathcal{W}$, then for each $r > 0$, there exists a neighborhood V_r of λ_0 in Λ such that $\gamma(H_{P_\lambda,C})(r) < \infty$ for all $\lambda \in V_r$. Here Φ_λ is defined as in* (5.5).

Proof. Let $r > 0$ and $w \in \mathcal{W}$ be given with $\|w\|_\tau \leq r$. Consider the operator

$$\mathbf{A}_\lambda : \mathbf{A}_\lambda x = w + (\Phi_\lambda - I)\Pi_{\mathcal{M}//\mathcal{N}} x, \quad x \in \mathcal{W}.$$

Since $H_{P,C}$ is stable, $\gamma(\Pi_{\mathcal{M}//\mathcal{N}})(k) < \infty$ for all $k > 0$. Using Proposition 5.3, we see there exists a neighborhood V_r of λ_0 such that

$$\|(\Phi_\lambda - I)\Pi_{\mathcal{M}//\mathcal{N}} x\|_\tau \leq \gamma(\Phi_\lambda - I)(\gamma(\Pi_{\mathcal{M}//\mathcal{N}})(2r)) < r \tag{5.9}$$

for all $x \in S_{2r}$ and $\lambda \in V_r$. This implies that $\|\mathbf{A}_\lambda x\|_\tau < \|w\|_\tau + r < 2r$ for all $x \in S_{2r}$. Due to our assumption, we may suppose that $\mathbf{A}_\lambda$ is continuous and compact on S_{2r}. From Schauder's fixed point theorem, it follows that there exists $x_\lambda \in S_{2r}$ such that

$$x_\lambda = w + (\Phi_\lambda - I)\Pi_{\mathcal{M}//\mathcal{N}} x_n \quad \text{for } \lambda \in V_r,$$

i.e.,

$$w = \Pi_{\mathcal{N}//\mathcal{M}} x_\lambda + \Phi_\lambda \Pi_{\mathcal{M}//\mathcal{N}} x_\lambda.$$

Since $\Pi_{\mathcal{N}//\mathcal{M}} x_\lambda \in \mathcal{N}$, $\Phi_\lambda \Pi_{\mathcal{M}//\mathcal{N}} x_\lambda \in \mathcal{M}_\lambda$, and the perturbed system is well-posed, we have

$$\Pi_{\mathcal{M}_\lambda//\mathcal{N}} w = \Phi_\lambda \Pi_{\mathcal{M}//\mathcal{N}} x_\lambda = \Pi_{\mathcal{M}//\mathcal{N}} x_\lambda + (\Phi_\lambda - I)\Pi_{\mathcal{M}//\mathcal{N}} x_\lambda$$

and therefore

$$\begin{aligned} \|\Pi_{\mathcal{M}_\lambda//\mathcal{N}} w\|_\tau &= \|\Pi_{\mathcal{M}//\mathcal{N}} x_\lambda\|_\tau + \|(\Phi_\lambda - I)\Pi_{\mathcal{M}//\mathcal{N}} x_\lambda\|_\tau \\ &\leq \gamma(\Pi_{\mathcal{M}//\mathcal{N}})(2r) + \gamma(\Phi_\lambda - I)(\gamma(\Pi_{\mathcal{M}//\mathcal{N}})(2r)) \\ &\leq \gamma(\Pi_{\mathcal{M}//\mathcal{N}})(2r) + r. \end{aligned}$$

Hence $\gamma(\Pi_{\mathcal{M}_\lambda//\mathcal{N}})(r) \leq \gamma(\Pi_{\mathcal{M}//\mathcal{N}})(2r) + r < \infty$ for $\lambda \in V_r$. ∎

With the technical assumption of compactness of $T_\tau(\Phi_\lambda - I)\Pi_{\mathcal{M}//\mathcal{N}}$, this result states the boundedness of $H_{P_\lambda,C}(u_0, y_0)^\top$ for $\|(u_0, y_0)^\top\| \leq r$ and λ sufficiently close to λ_0. Obviously, if the neighborhood V_r for (5.9) is independent of r, that is, if

$$\gamma(\Phi_\lambda - I)(\gamma(\Pi_{\mathcal{M}//\mathcal{N}})(2r)) < r \quad \text{for all } \lambda \text{ sufficiently close to } \lambda_0$$

hold for all (large) r, then $H_{P_\lambda,C}$ would be (gf)-stable.

6. Conclusions. The main contributions of this paper are as follows. Natural generalizations of the graph topology w.r.t. a gain function notion of stability for nonlinear systems in a general normed signal space setting were defined. Convergence in the graph topology was shown to have a natural application in robust stability results. Various metrizations of the graph topologies were given; in particular it was shown that the generalizations of the gap metric given by [7] and the natural generalization of the graph metric both induce the graph topology when the stability notion is that of an (unweighted) induced gain, subject to certain assumptions on local asymptotic completeness and the existence of normalized coprime factorizations.

Weaker results have been derived for the more general cases (including the weighted case). Georgiou-type formulae [4] have been derived and are shown to be equivalent to other alternative formulations of the gap metric.

There are many directions for future work. An important topic is the extension of the above results to the ν-gap setting; in particular, the investigation of a coprime factor characterization of the underlying induced topology of the nonlinear generalizations of the ν-gap. A more fundamental area for future research concerns the investigation of the continuity of the closed loop response w.r.t. gap perturbations to the loop, probably involving greater regularity assumptions [7]. A final area of worthy future study concerns the explicit study of the numerical computation of the gap, possibly based on the Georgiou-type formulae, but with additional regularity assumptions on the minimizer Q, perhaps allowed by greater regularity assumptions on P and C. In this regard, nonlinear generalizations of the commutant lifting theory may be the appropriate tool.

Acknowledgments. The second author would like to acknowledge the hospitality of M. C. Smith and the Cambridge University Engineering Department for a visit during which this paper was completed. Informative discussions with M. C. Smith are gratefully acknowledged. The authors also thank the referees for valuable suggestions concerning the definition of normalized coprime factorizations and the presentation of the paper.

REFERENCES

[1] B. D. O. Anderson, T. S. Brinsmead, and F. D. Bruyne, *The Vinnicombe metric for nonlinear operators*, IEEE Trans. Automat. Control, 47 (2002), pp. 1450–1465.

[2] B. D. O. Anderson, M. R. James, and D. J. N. Limebeer, *Robust stabilization of nonlinear systems via normalized coprime factor representations*, Automatica J. IFAC, 34 (1998), pp. 1593–1599.

[3] M. Cantoni and G. Vinnicombe, *Linear feedback systems and the graph topology*, IEEE Trans. Automat. Control, 47 (2002), pp. 710–719.

[4] T. T. Georgiou, *On the computation of the gap metric*, Systems Control Lett., 11 (1988), pp. 253–257.

[5] T. T. Georgiou and M. C. Smith, *Optimal robustness in the gap metric*, IEEE Trans. Automat. Control, 35 (1990), pp. 673–686.

[6] T. T. Georgiou and M. C. Smith, *Graphs, causality, and stabilizability: Linear, shift invariant systems on* $L^2[0,\infty)$, Math. Control Signals Systems, 6 (1993), pp. 195–223.

[7] T. T. Georgiou and M. C. Smith, *Robustness analysis of nonlinear feedback systems: An input-output approach*, IEEE Trans. Automat. Control, 42 (1997), pp. 1200–1221.

[8] K. Glover and D. McFarlane, *Robust stabilization of normalized coprime factor plant descriptions with* H_∞*-bounded uncertainty*, IEEE Trans. Automat. Control, 34 (1989), pp. 821–830.

[9] J. Hammer, *Fraction representations of nonlinear systems: A simplified approach*, Internat. J. Control, 46 (1987), pp. 455–472.

[10] M. R. James, M. C. Smith, and G. Vinnicombe, *Gap metrics, representations and nonlinear robust stability*, SIAM J. Contr. Optim., 43 (2005), pp. 1535–1582.

[11] J. B. Moore and L. Irlicht, *Coprime factorization over a class of nonlinear systems*, Internat. J. Robust Nonlinear Control, 2 (1992), pp. 261–290.

[12] A. D. B. Paice and A. J. van der Schaft, *The class of stabilizing nonlinear plant controller pairs*, IEEE Trans. Automat. Control, 41 (1996), pp. 634–645.

[13] A. J. van der Schaft, *Robust stabilization of nonlinear systems via stable kernel representations with* L_2 *gain bounded uncertainly*, Systems Control Lett., 24 (1995), pp. 75–81.

[14] A. J. van der Schaft, L^2*-Gain and Passivity Techniques in Nonlinear Control*, 2nd ed., Springer-Verlag, London, 2000.

[15] J. M. A. Scherpen and A. J. Van der Schaft, *Normalized coprime factorizations and balancing for unstable nonlinear systems*, Internat. J. Control, 60 (1994), pp. 1193–1222.

[16] J. A. Sefton and R. J. Ober, *On the gap metric and coprime factor perturbations*, Automatica J. IFAC, 29 (1993), pp. 723–734.
[17] E. D. Sontag, *Smooth stabilization implies coprime factorization*, IEEE Trans. Automat. Control, 34 (1989), pp. 435–443.
[18] M. S. Verma, *Coprime fractional representations and stability of nonlinear feedback systems*, Internat. J. Control, 48 (1988), pp. 897–918.
[19] M. S. Verma and L. R. Hunt, *Right coprime fractorizations and stabilization for nonlinear systems*, IEEE Trans. Automat. Control, 38 (1993), pp. 222–231.
[20] M. Vidyasagar, *Control System Synthesis*, MIT Press, Cambridge, MA, 1985.
[21] G. Vinnicombe, *Uncertainty and Feedback: $\mathcal{H}_\infty$-shaping Control System Synthesis*, Imperial College Press, London, 2001.
[22] G. Vinnicombe, *A ν-gap distance for uncertain and nonlinear systems*, in Proceedings of 38th IEEE CDC, Phoeniz, AZ, 1999, pp. 2557–2562.
[23] G. Zames and A. K. El-Sakkary, *Unstable systems and feedback: The gap metric*, in Proceedings of the Allerton Conference, 1980, pp. 380–385.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 2, pp. 444–466

CONSTRAINED STOCHASTIC LQ CONTROL WITH RANDOM COEFFICIENTS, AND APPLICATION TO PORTFOLIO SELECTION*

YING HU† AND XUN YU ZHOU‡

Abstract. This paper is devoted to the study of a stochastic linear-quadratic (LQ) optimal control problem where the control variable is constrained in a cone, and all the coefficients of the problem are random processes. Employing Tanaka's formula, optimal control and optimal cost are explicitly obtained via solutions to two extended stochastic Riccati equations (ESREs). The ESREs, introduced for the first time in this paper, are highly nonlinear backward stochastic differential equations (BSDEs), whose solvability is proved based on a truncation function technique and Kobylanski's results. The general results obtained are then applied to a mean-variance portfolio selection problem for a financial market with random appreciation and volatility rates, and with short-selling prohibited. Feasibility of the problem is characterized, and efficient portfolios and efficient frontier are presented in closed forms.

Key words. stochastic LQ control, extended stochastic Riccati equation, backward stochastic differential equation, mean-variance portfolio selection, efficient portfolio, efficient frontier

AMS subject classifications. 93E20, 60H10, 91B28

DOI. 10.1137/S0363012904441969

1. Introduction.

Linear-quadratic (LQ) optimal control is a problem where the system dynamics are linear in state and control variables and the cost functional is quadratic in the two variables. It is a classical yet fundamental problem in control theory, pioneered by Kalman [11] (for deterministic control). Extension to stochastic LQ control was first carried out by Wonham [25]. Bismut [3] performed a detailed analysis for stochastic LQ control with random coefficients. With the joint effort of many researchers in the last 40 years, there has been an enormously rich theory on LQ control, deterministic and stochastic alike. Recently, starting with Chen, Li, and Zhou [6], there has been emerging interest in the so-called *indefinite* stochastic LQ control, where, quite contrary to the conventional belief, the cost weighting matrices are allowed to be indefinite; see [7, 8, 1, 26]. This new theory turns out to be useful in solving the continuous-time version of Markowitz's Nobel-winning mean-variance portfolio selection model; see [28, 14, 16, 18, 13, 17]. For systematic accounts of the deterministic and stochastic LQ theory, refer to [2] and [27], respectively.

One of the elegant features of the LQ theory is that it is able to give in explicit forms the optimal state feedback control and the optimal cost value through the celebrated Riccati equation; hence the LQ control problem is completely solved. What essentially enables this closed-form solution, besides the special LQ structure, is that the control is *not* constrained. Specifically, since the control is unconstrained, the feedback control constructed via the Riccati equation is *automatically* admissible.

*Received by the editors March 9, 2004; accepted for publication (in revised form) December 8, 2004; published electronically August 31, 2005. This work was supported by the RGC Earmarked Grants CUHK 4435/99E, CUHK 4175/00E, and CUHK 4234/01E.

http://www.siam.org/journals/sicon/44-2/44196.html

†Institut de Recherche Mathématique de Rennes, Université de Rennes 1, Campus de Beaulieu, 35042 Rennes Cedex, France (Ying.Hu@univ-rennes1.fr).

‡Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong (xyzhou@se.cuhk.edu.hk). Part of this work was completed when this author was visiting the Université de Rennes 1, whose hospitality is greatly appreciated.

If, on the other hand, there are pointwise control constraints, then the whole LQ approach would collapse.

One should acknowledge that LQ control with control constraints is a well-posed problem which is important in both theory and applications. For example, in many real applications the control variable is required to take only nonnegative values. The mean-variance portfolio selection problem with no-shorting constraint, which is to be tackled in this paper, is exactly one of such problems. There were some attempts in attacking the deterministic LQ problems with positive controls; see for example [23, 5, 9]. In these works, however, only some implicit necessary and sufficient conditions for optimality were derived and some numerical schemes suggested, and the special LQ structure was not fully taken advantage of and no explicit result was obtained. On the other hand, to our best knowledge research on pointwise constrained stochastic LQ control has been completely absent in the literature.

The main purpose of this paper is to tackle a stochastic LQ control problem where the control variable is constrained in a cone (which, certainly, includes the nonnegative orthant in a Euclidean space as a special case), and all the coefficient matrices of the model are random. Moreover, the problem is allowed to be "singular" in the sense that the control weighting matrix in the cost functional is possibly singular. However, we are able to treat only the case when the state variable is scalar-valued, although it is sufficient to cover many meaningful practical applications, in particular in financial area where the one-dimensional wealth process is typically taken as the state. We aim to obtain *explicit* solutions comparable to the classical unconstrained-control counterpart. To this end, we introduce two equations termed extended stochastic Riccati equations (ESREs). These two equations are highly nonlinear backward stochastic differential equations (BSDEs), the solvability of which is interesting in its own right. Based on a truncation function technique and a delicate result of Kobylanski [12], we are able to prove the existence of solutions to the introduced ESREs. Then, applying Tanaka's formula and going through some detailed analysis, we obtain explicit optimal feedback control as well as the optimal cost value in terms of the solutions to the two ESREs.

The other purpose of the paper is to solve the continuous-time mean-variance portfolio selection model with short selling prohibition and random market parameters. Indeed, this is the very problem that motivated us to tackle the general constrained stochastic LQ problem. Notice that a mean-portfolio selection model with no-shorting was solved in [16] using Hamilton–Jacobi–Bellman equation and viscosity solution theory; however, among other additional assumptions all the market parameters are assumed to be deterministic in [16]. The partial differential equation approach there does not extend to the current case with random parameters. To overcome this difficulty, we reformulate the problem so that it falls exactly into the general constrained LQ problem that has been solved. Hence, by applying the general results obtained we are able to solve the portfolio selection problem, once again, explicitly and completely.

There are a large number of papers devoted to applying stochastic control theory for mean-variance efficient/hedging portfolio selection models; see, to name a few recent ones, [14, 21, 13, 18, 17, 22, 4]. While some of these works are on more general asset price models (such as semimartingale ones) and/or markets (including incomplete markets), none of them deal with directly constrained portfolios. As a result, the value functions there remain quadratic, which is no longer the case, as will be demonstrated in this paper, with the conic constrained portfolios.

The rest of the paper is organized as follows. In section 2 we formulate the constrained stochastic LQ model, and in section 3 we present some mathematical preliminaries including Tanaka's formula and the introduction of the two ESREs. Section 4 is devoted to the solvability of the ESREs in two common cases. Section 5 gives the solution to the LQ problem. Application of the general results to a mean-variance portfolio selection problem is presented in section 6. Finally, section 7 concludes the paper with some remarks.

2. Problem formulation. We assume throughout that $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, P)$ is a given complete, filtered probability space and that $W(\cdot)$ is a k-dimensional standard Brownian motion on this space with $W(0) = 0$. In addition, we assume that $\mathcal{F}_t$ is the augmentation of $\sigma\{W(s) \,|\, 0 \leq s \leq t\}$ by all the P-null sets of $\mathcal{F}$. When no confusion would occur, we leave out P-a.s. for a statement that holds almost surely with respect to P.

Throughout this paper, we denote by $\mathbb{R}^m$ the set of m-dimensional *column* vectors, by $\mathbb{R}^m_+$ the set of m-dimensional column vectors whose components are nonnegative, by $\mathbb{R}^{m\times n}$ the set of $m \times n$ real matrices, and by $\mathbb{S}^n$ the set of symmetric $n \times n$ real matrices. Therefore, $\mathbb{R}^m \equiv \mathbb{R}^{m\times 1}$. If $M = (m_{ij}) \in \mathbb{R}^{m\times n}$, we denote its transpose by M', and its norm by $|M| = \sqrt{\sum_{i,j} m_{ij}^2}$. If $M \in \mathbb{S}^n$ is positive (positive semi-) definite, we write $M > (\geq)\ 0$. Suppose $\eta : \Omega \to \mathbb{R}^n$ is a $\mathcal{G}$-measurable random variable. We write $\eta \in L^2_{\mathcal{G}}(\Omega; \mathbb{R}^n)$ if η is square integrable (i.e., $E|\eta|^2 < \infty$), and $\eta \in L^\infty_{\mathcal{G}}(\Omega; \mathbb{R}^n)$ if η is uniformly bounded. Consider now the case when $f : [0, T] \times \Omega \to \mathbb{R}^n$ is an $\{\mathcal{F}_t\}_{t\geq 0}$-adapted process. If $f(\cdot)$ is square integrable (i.e., $E\int_0^T |f(t)|^2\, dt < \infty$) we shall write $f(\cdot) \in L^2_{\mathcal{F}}(0, T; \mathbb{R}^n)$; if $f(\cdot)$ is uniformly bounded (i.e., ess $\sup_{(t,\omega)\in[0,T]\times\Omega} |f(t)| < \infty$), then $f(\cdot) \in L^\infty_{\mathcal{F}}(0, T; \mathbb{R}^n)$. If $f(\cdot)$ has (P-a.s.) continuous sample paths and $E \sup_{t\in[0,\,T]} |f(t)|^2 < \infty$ we write $f(\cdot) \in L^2_{\mathcal{F}}(\Omega; C(0, T; \mathbb{R}^n))$. These definitions generalize in the obvious way to the case when $f(\cdot)$ is $\mathbb{R}^{n\times m}$- or $\mathbb{S}^n$-valued. In addition, we say that $N \in L^2_{\mathcal{F}}(0, T; \mathbb{S}^n)$ is positive (positive semi-) definite, which is sometimes denoted simply by $N > (\geq)\ 0$, if $N(t, \omega) > (\geq)\ 0$ for a.e. $t \in [0, T]$ and P-a.s., and say that N is uniformly positive definite if $N \geq cI_n$ for a.e. $t \in [0, T]$ and P-a.s. with some given deterministic constant $c > 0$, where I_n is the n-dimensional identity matrix.

Finally, for any real number we define $x^+ := \max\{x, 0\}$ and $x^- := \max\{-x, 0\}$.

Consider the following linear stochastic differential equation (SDE):

$$\text{(2.1)}\qquad \begin{cases} dx(t) &= [A(t)\,x(t) + B(t)\,u(t)]\,dt + [x(t)C(t)' + u(t)'D(t)']\,dW(t), \quad t \in [0, T], \\ x(0) &= x_0, \end{cases}$$

where A, B, C, and D are $\{\mathcal{F}_t\}_{t\geq 0}$-adapted processes (possibly matrix-valued), and $x_0 \in \mathbb{R}$ is a nonrandom scalar. Precise assumptions on these data will be specified below. Let $\Gamma \subset \mathbb{R}^m$ be a given closed cone; i.e., Γ is closed, and if $u \in \Gamma$, then $\alpha u \in \Gamma$ $\forall \alpha \geq 0$. Typical examples of such a cone are $\Gamma = \mathbb{R}^m_+$, $\Gamma = \{u \in \mathbb{R}^m \,|\, Mu \leq 0\}$, and $\Gamma = \{u \in \mathbb{R}^m \,|\, Mu = 0\}$, where $M \in \mathbb{R}^{n\times m}$. The class of *admissible controls* is the set

$$\mathcal{U} := \Big\{u(\cdot) \in L^2_{\mathcal{F}}(0, T; \mathbb{R}^m) \,\Big|\, \begin{array}{l} u(t) \in \Gamma,\ P-\text{a.s., a.e. } t \in [0, T], \text{ and (2.1) has a} \\ \text{unique solution under } u(\cdot) \end{array}\Big\}.$$

If $u(\cdot) \in \mathcal{U}$ and $x(\cdot)$ is the associated solution of (2.1), then we refer to $(x(\cdot), u(\cdot))$ as an *admissible pair*.

Suppose that the cost functional is given by

$$J(x_0, u(\cdot)) := E\Big\{\int_0^T \Big[Q(t)x(t)^2 + u(t)'R(t)u(t)\Big]\, dt + Gx(T)^2\Big\}. \tag{2.2}$$

Throughout this paper, we shall assume the following:

Assumption (A1):

$$\begin{cases} A,\, Q & \in & L^\infty_{\mathcal{F}}(0,\, T; \mathbb{R}), \\ B & \in & L^\infty_{\mathcal{F}}(0,\, T; \mathbb{R}^{1\times m}), \\ C & \in & L^\infty_{\mathcal{F}}(0,\, T; \mathbb{R}^{k}), \\ D & \in & L^\infty_{\mathcal{F}}(0,\, T; \mathbb{R}^{k\times m}), \\ R & \in & L^\infty_{\mathcal{F}}(0,\, T; \mathbb{S}^{m}), \\ G & \in & L^\infty_{\mathcal{F}_T}(\Omega;\, \mathbb{R}). \end{cases}$$

Note that all the parameters involved may be random. Also, by standard SDE theory, (2.1) admits a unique solution $x(\cdot) \in L^2_{\mathcal{F}}(\Omega;\, C(0,\, T;\, \mathbb{R}))$ for any $u(\cdot) \in L^2_{\mathcal{F}}(0,\, T;\, \mathbb{R}^m)$ under Assumption (A1). The *stochastic LQ problem* associated with (2.1)–(2.2) is as follows:

$$\begin{cases} \text{Minimize} & J(x_0, u(\cdot)), \\ \text{subject to} & (x(\cdot),\, u(\cdot)) \text{ admissible for (2.1).} \end{cases} \tag{2.3}$$

The problem (2.3) is said to be *finite* (w.r.t. x_0) if there exists some finite constant $K \in \mathbb{R}$ such that

$$J(x_0, u(\cdot)) \geq K \quad \forall\, u(\cdot) \in \mathcal{U},$$

and *solvable* (w.r.t. x_0) if there exists a control $u^*(\cdot) \in \mathcal{U}$ such that

$$J(x_0, u^*(\cdot)) \leq J(x_0, u(\cdot)) \quad \forall\, u(\cdot) \in \mathcal{U}.$$

In this case, the control $u^*(\cdot)$ is referred to as the *optimal control* (w.r.t. x_0). We say that (2.3) is *uniquely solvable* if it is solvable and the optimal control is unique. Note that a finite LQ problem is not necessarily solvable.

3. Preliminaries. In this section we present some mathematical preliminaries required in what follows, including in particular Tanaka's formula which plays a critical technical role in the subsequent analysis.

LEMMA 3.1 (Tanaka's formula). *Let $X(t)$ be a continuous semimartingale. Then*

$$\begin{aligned} dX^+(t) &= 1_{(X(t)>0)}dX(t) + \tfrac{1}{2}dL(t), \\ dX^-(t) &= -1_{(X(t)\leq 0)}dX(t) + \tfrac{1}{2}dL(t), \end{aligned} \tag{3.1}$$

where $L(\cdot)$ is an increasing continuous process, called the local time of $X(\cdot)$ at 0, satisfying

$$\int_0^t |X(s)|dL(s) = 0, \quad P - a.s.. \tag{3.2}$$

Proof. See, for example, [24, Chapter VI, Theorem 1.2, and Proposition 1.3]. □

Next, define the following mappings:

$$
\begin{aligned}
H_+(t,\omega,v,P,\Lambda) := \;& v'[R(t,\omega)+PD(t,\omega)'D(t,\omega)]v \\
& +2v'[B(t,\omega)'P+D(t,\omega)'PC(t,\omega)+D(t,\omega)'\Lambda], \\
H_-(t,\omega,v,P,\Lambda) := \;& v'[R(t,\omega)+PD(t,\omega)'D(t,\omega)]v \\
& -2v'[B(t,\omega)'P+D(t,\omega)'PC(t,\omega)+D(t,\omega)'\Lambda], \\
& (t,\omega,v,P,\Lambda)\in[0,T]\times\Omega\times\mathbb{R}^m\times\mathbb{R}\times\mathbb{R}^k,
\end{aligned}
\tag{3.3}
$$

and

$$
\begin{aligned}
&H_+^*(t,\omega,P,\Lambda) := \inf_{v\in\Gamma} H_+(t,\omega,v,P,\Lambda), \\
&H_-^*(t,\omega,P,\Lambda) := \inf_{v\in\Gamma} H_-(t,\omega,v,P,\Lambda), \quad (t,\omega,P,\Lambda)\in[0,T]\times\Omega\times\mathbb{R}\times\mathbb{R}^k.
\end{aligned}
\tag{3.4}
$$

Remark 3.1. $H_+^*(t,\omega,P,\Lambda)$ and $H_-^*(t,\omega,P,\Lambda)$ have finite values if $R(t,\omega)+PD(t,\omega)'D(t,\omega)>0$. Indeed, in this case, there exist $C_1(t,\omega,P,\Lambda)>0$, $C_2(t,\omega,P,\Lambda)>0$ such that

$$
H_+(t,\omega,v,P,\Lambda)\geq C_1|v|^2-C_2|v|=C_1|v|\left(|v|-\frac{C_2}{C_1}\right).
$$

If $|v|>\frac{C_2}{C_1}$, then $H_+>0$. Recall that $0\in\Gamma$, hence $\inf_{v\in\Gamma}H_+(t,\omega,v,P,\Lambda)\leq 0$. These facts imply that

$$
H_+^*(t,\omega,P,\Lambda)=\inf_{v\in\Gamma,|v|\leq\frac{C_2}{C_1}} H_+(t,\omega,v,P,\Lambda)\geq \min_{|v|\leq\frac{C_2}{C_1}} H_+(t,\omega,v,P,\Lambda)>-\infty.
$$

Hence, $\inf_{v\in\Gamma}H_+(t,\omega,v,P,\Lambda)$ is finite. The same is true for H_-.

Now we introduce the following two nonlinear backward stochastic differential equations—BSDEs (the arguments t and ω are suppressed):

$$
\begin{cases}
dP_+ = -\Big\{(2A+C'C)P_+ + 2C'\Lambda_+ + Q + H_+^*(P_+,\Lambda_+)\Big\}dt + \Lambda_+'dW, \quad t\in[0,\,T], \\
P_+(T)=G, \\
R+P_+D'D>0,
\end{cases}
\tag{3.5}
$$

$$
\begin{cases}
dP_- = -\Big\{(2A+C'C)P_- + 2C'\Lambda_- + Q + H_-^*(P_-,\Lambda_-)\Big\}dt + \Lambda_-'dW, \quad t\in[0,\,T], \\
P_-(T)=G, \\
R+P_-D'D>0.
\end{cases}
\tag{3.6}
$$

The equations (3.5) and (3.6) are referred to as the ESREs. The following gives a precise definition of their solutions.

Definition 3.1. *A stochastic process* $(P_+,\Lambda_+)\in L^2_{\mathcal{F}}(\Omega;C(0,\,T;\,\mathbb{R}))\times L^2_{\mathcal{F}}(0,\,T;\,\mathbb{R}^k)$ *is called a* solution *to the ESRE* (3.5) *if it satisfies the first equation of* (3.5) *in the Itô sense as well as the second (the terminal condition) and third (the positive definiteness) constraints of* (3.5). *A solution* (P_+,Λ_+) *of* (3.5) *is called* bounded *if* $P_+\in L^\infty_{\mathcal{F}}(0,\,T;\,\mathbb{R})$, *called* positive *(respectively,* nonnegative*) if* $P_+(t)>0$ *(respectively,* $P_+(t)\geq 0$*)* $\forall\ t\in[0,\,T]$ *and* P*-a.s., and called* uniformly positive *if* $P_+(t)\geq c$

$\forall\ t \in [0, T]$ *and* P*-a.s. with some deterministic constant* $c > 0$. *Similar terms can be defined for the other ESRE* (3.6).

4. Existence of solution to the ESREs. As the existence of a solution to the ESREs (3.5) and (3.6) is essential to solving the underlying stochastic LQ problem, we devote this section to this issue. Note that the existence problem is interesting in its own right from BSDE point of view, for both (3.5) and (3.6) are nonlinear BSDEs that do not satisfy the standard assumptions for existence.

We will deal with the following two cases.

Standard case. $Q \geq 0$, $R > 0$ with $R^{-1} \in L^{\infty}_{\mathcal{F}}(0, T; \mathbb{R}^{m\times m})$, and $G \geq 0$.

Singular case. $Q \geq 0$, $R \geq 0$, $G > 0$ with $G^{-1} \in L^{\infty}_{\mathcal{F}}(0, T; \mathbb{R})$, and $D'D > 0$ with $(D'D)^{-1} \in L^{\infty}_{\mathcal{F}}(0, T; \mathbb{R}^{m\times m})$.

4.1. Standard case. In this subsection we solve the standard case.

THEOREM 4.1. *For the standard case, there exists a bounded, nonnegative solution* (P_+, Λ_+) *(respectively,* (P_-, Λ_-)*) to the ESRE* (3.5) *(respectively,* (3.6)*).*

Proof. Let us first consider the following BSDE:

$$\begin{cases} dP_1 = -\Big\{(2A + C'C)P_1 + 2C'\Lambda_1 + Q\Big\}dt + \Lambda_1' dW, \quad t \in [0, T], \\ P_1(T) = G. \end{cases} \tag{4.1}$$

This is a linear BSDE with bounded coefficients (by virtue of Assumption (A1)), and with $Q \geq 0$ and $G \geq 0$. Hence there exists a unique, nonnegative bounded solution (P_1, Λ_1). Denote by $c_1 > 0$ an upper bound of P_1. Now, consider the following BSDE:

$$\begin{cases} dP = -F_1(t, P, \Lambda)dt + \Lambda' dW, \quad t \in [0, T], \\ P(T) = G, \end{cases} \tag{4.2}$$

where the function F_1 is defined by

$$\begin{aligned} F_1(t, \omega, P, \Lambda) := \quad & [2A(t, \omega) + C(t, \omega)'C(t, \omega)]P + 2C(t, \omega)'\Lambda + Q(t, \omega) \\ & + H^*_+(t, \omega, P^+, \Lambda)g_1(P^+), \ (t, \omega, P, \Lambda) \in [0, T] \times \Omega \times \mathbb{R} \times \mathbb{R}^k \end{aligned}$$

(recall that P^+ denotes the positive part of P), whereas $g_1 : \mathbb{R}^+ \to [0, 1]$ is a smooth truncation function satisfying $g_1(x) = 1$ for $x \in [0, c_1]$, and $g_1(x) = 0$ for $x \in [2c_1, +\infty)$.

The function F_1 is continuous in (P, Λ). To see this, for a given $n \in \mathbb{N}$ (the set of positive integers), if $|P| \leq n$, $|\Lambda| \leq n$, then by the assumption on R there exist two constants $C_1 > 0$ and $C_2(n) > 0$, such that

$$H_+(t, v, P^+, \Lambda) \geq C_1|v|^2 - C_2(n)|v|.$$

Thus, for $|P| \leq n, |\Lambda| \leq n$, the argument in Remark 3.1 results in

$$H^*_+(t, P^+, \Lambda) = \min_{v\in\Gamma, |v|\leq \frac{C_2(n)}{C_1}} H_+(t, v, P^+, \Lambda).$$

This implies that F_1 is continuous in (P, Λ).

On the other hand, there exists a $C_2 > 0$ such that

$$H_+(t, v, P^+, \Lambda) \geq C_1|v|^2 - C_2(P^+ + |\Lambda|)|v|,$$

which yields

$$0 \geq H_+^*(t, P^+, \Lambda) \geq -\frac{C_2^2(P^+ + |\Lambda|)^2}{4C_1}.$$

Hence, F_1 satisfies the hypothesis (H1) of Kobylanski [12] noting the role of the truncation function g_1. According to [12, Theorem 2.3], there is a bounded, maximal solution (see [12, p. 565] for its definition) (P_+, Λ_+) to the BSDE (4.2). Now, as $H_+^*(t, P, \Lambda) \leq 0$ and (P_1, Λ_1) is the only, hence maximal, bounded solution to (4.1), we get $P_+ \leq P_1 \leq c_1$. Furthermore, $G \geq 0$ and $Q \geq 0$ and $H_+^*(t, P^+, \Lambda) \geq -\frac{C_2^2(P^+ + |\Lambda|)^2}{4C_1}$ implies $P_+ \geq 0$, using the facts that (P_+, Λ_+) is the bounded, maximal solution to (4.2) and that (0,0) is an obvious solution to (4.2) with $G = 0$, $Q = 0$, and $H_+^*(t, P^+, \Lambda)g_1(P^+)$ replaced by $-\frac{C_2^2(P^+ + |\Lambda|)^2}{4C_1} g_1(P^+)$. This proves that (P_+, Λ_+) is a bounded nonnegative solution of the ESRE (3.5). The same argument concludes also the existence of a solution to the ESRE (3.6). □

4.2. Singular case. The singular case is the one that will be used in the financial application in the second half of the paper.

THEOREM 4.2. *For the singular case, there exists a bounded, uniformly positive solution* (P_+, Λ_+) *(respectively,* (P_-, Λ_-)*) to the ESRE* (3.5) *(respectively,* (3.6)*).*

Proof. Let us first consider the following BSDE (the argument t is suppressed):

$$\text{(4.3)} \quad \begin{cases} dP_2 = -[(2A + C'C)P_2 + 2C'\Lambda_2 + H^*(P_2, \Lambda_2)]dt + \Lambda_2' dW, \quad t \in [0, T], \\ P_2(T) = G, \\ P_2 > 0, \end{cases}$$

where

$$H^*(t, P, \Lambda) := -[PB + (C'P + \Lambda')D]P^{-1}(D'D)^{-1}[B'P + D'(PC + \Lambda)].$$

This is a BSDE studied independently in [13] and [10], using different approaches. The existence of its solution was proved in [13] with a rather involved proof, and was proved in [10] with a short proof, nonetheless, for the case $k = m$. For completeness, we will prove the existence of a solution to this BSDE using a simpler method in the lemma following the end of the proof of this theorem. By this lemma, there exists a unique bounded, uniformly positive solution (P_2, Λ_2). In particular, there exists a constant $c_2 > 0$ such that $P_2(t) \geq c_2$ $\forall t \in [0, T]$, P-a.s..

Now, let us consider the following BSDE:

$$\text{(4.4)} \quad \begin{cases} dP = -F_2(t, P, \Lambda)dt + \Lambda' dW, \quad t \in [0, T], \\ P(T) = G, \end{cases}$$

where

$$\begin{aligned} F_2(t, \omega, P, \Lambda) := \quad & [2A(t, \omega) + C(t, \omega)'C(t, \omega)]P + 2C(t, \omega)'\Lambda + Q(t, \omega) \\ & + H_+^*(t, \omega, P, \Lambda)g_2(P^+), \ (t, \omega, P, \Lambda) \in [0, T] \times \Omega \times \mathbb{R} \times \mathbb{R}^k \end{aligned}$$

with $g_2 : \mathbb{R}^+ \to [0, 1]$ being another smooth truncation function satisfying $g_2(x) = 0$ for $x \in [0, \frac{1}{2}c_2]$, and $g_2(x) = 1$ for $x \in [c_2, +\infty)$. As with the proof for the standard case we can show that the function F_2 is continuous in (P, Λ) and satisfies the assumption of [12]. Thus, there exists a bounded, maximal solution of the BSDE (4.4), denoted as (P_+, Λ_+).

Notice the following inequality:

$$\begin{aligned} H_+^*(t,P,\Lambda) &\geq \inf_{v\in\mathbb{R}^m} H_+(t,v,P,\Lambda) \\ &\geq \inf_{v\in\mathbb{R}^m} \{v'PD(t)'D(t)v + 2v'[B(t)'P + D(t)'PC(t) + D(t)'\Lambda]\} \\ &= H^*(t,P,\Lambda). \end{aligned}$$

Hence, noting $Q \geq 0$, the maximal solution argument gives

$$P_+(t) \geq P_2(t) \geq c_2, \quad \forall t \in [0,T].$$

This implies that (P_+, Λ_+) is actually a bounded, uniformly positive solution of the ESRE (3.5). The same argument also leads to the existence of a solution to the ESRE (3.6). □

LEMMA 4.1. *Equation* (4.3) *has a bounded, uniformly positive solution* (P_2, Λ_2).

Proof. Set

$$\alpha := 2A + C'C - (B + C'D)(D'D)^{-1}(B' + D'C), \ \beta := 2C - 2D(D'D)^{-1}(B' + D'C).$$

Then, (4.3) can be rewritten as

$$\text{(4.5)} \quad \begin{cases} dP_2 = -[\alpha P_2 + \beta'\Lambda_2 - \frac{1}{P_2}\Lambda_2'D(D'D)^{-1}D'\Lambda_2]dt + \Lambda_2'dW, \quad t \in [0,T], \\ P_2(T) = G, \\ P_2 > 0. \end{cases}$$

Let $c_3 > 0$ and $c_4 > 0$ be two constants satisfying $G \geq c_3$ and $|\alpha| \leq c_4$, and set

$$c_2 := c_3 e^{-c_4 T}.$$

Now, consider the following BSDE:

$$\text{(4.6)} \quad \begin{cases} dP = -[\alpha P^+ + \beta'\Lambda - \frac{1}{P}\Lambda'D(D'D)^{-1}D'\Lambda g_2(P)]dt + \Lambda'dW, \quad t \in [0,T], \\ P(T) = G, \end{cases}$$

where g_2 is the truncation function defined in the proof of Theorem 4.2 corresponding to the constant c_2. According again to [12, Theorem 2.3], there exists a bounded, maximal solution to this BSDE denoted as (P_2, Λ_2).

Finally, the following BSDE:

$$\text{(4.7)} \quad \begin{cases} dP = -[-c_4 P^+ + \beta'\Lambda - \frac{1}{P}\Lambda'D(D'D)^{-1}D'\Lambda g_2(P)]dt + \Lambda'dW, \quad t \in [0,T], \\ P(T) = c_3, \end{cases}$$

has an obvious solution $(c_3 e^{-c_4(T-t)}, 0)$. As (P_2, Λ_2) is a maximal solution to (4.6), we deduce that $P_2(t) \geq c_3 e^{-c_4(T-t)} \geq c_3 e^{-c_4 T} = c_2$. This implies that (P_2, Λ_2) is also a bounded, uniformly positive solution to (4.5), hence to (4.3). □

Remark 4.1. When there is no control constraint, i.e., when $\Gamma = \mathbb{R}^m$, then

$$H_+^*(t,P,\Lambda) = H_-^*(t,P,\Lambda) = -[PB + (C'P + \Lambda')D](R + PD'D)^{-1}[B'P + D'(PC + \Lambda)],$$

provided $R + PD'D > 0$. Hence both ESREs (3.5) and (3.6) reduce to the normal stochastic Riccati equation

$$
(4.8)\quad \begin{cases} dP = -\Big\{(2A + C'C)P + 2C'\Lambda + Q \\ \qquad\qquad - [PB + (C'P + \Lambda')D](R + PD'D)^{-1}[B'P + D'(PC + \Lambda)]\Big\}dt \\ \qquad\qquad + \Lambda' dW, \qquad t \in [0, T], \\ P(T) = G, \\ R + PD'D > 0. \end{cases}
$$

The above equation has been studied in great detail in [13, 10]. Moreover, in [10] the case when R is possibly indefinite was investigated.

Remark 4.2. The uniqueness of solutions to (3.5) and (3.6) will be proved in the next section, interestingly, as a direct consequence of the solution to the LQ problem. It should be noted that in [12], the uniqueness of solutions is proved under the additional assumption that the generator (i.e., the drift coefficient) is differentiable which is not satisfied here.

5. Solution to the LQ problem. In this section we give explicit solution to the LQ problem (2.1)–(2.2) in terms of the solutions to the two ESREs, for both the standard and singular cases defined in the previous section.

First, when $R + PD'D > 0$, define

$$
(5.1)\quad \begin{aligned} &\xi_+(t, \omega, P, \Lambda) := \operatorname{argmin}_{v\in\Gamma} H_+(t, \omega, v, P, \Lambda), \\ &\xi_-(t, \omega, P, \Lambda) := \operatorname{argmin}_{v\in\Gamma} H_-(t, \omega, v, P, \Lambda), \quad (t, \omega, P, \Lambda) \in [0, T] \times \Omega \times \mathbb{R} \times \mathbb{R}^k. \end{aligned}
$$

Note that the minimizers above are achievable due to a similar argument in Remark 3.1 and the assumption that Γ is closed.

THEOREM 5.1. *In both the standard and singular cases, let* $(P_+, \Lambda_+) \in L^2_{\mathcal{F}}(\Omega; C(0, T; \mathbb{R})) \times L^2_{\mathcal{F}}(0, T; \mathbb{R}^k)$ *and* $(P_-, \Lambda_-) \in L^2_{\mathcal{F}}(\Omega; C(0, T; \mathbb{R})) \times L^2_{\mathcal{F}}(0, T; \mathbb{R}^k)$ *be bounded, nonnegative (in the standard case) or bounded, uniformly positive (in the singular case) solutions to the ESREs* (3.5) *and* (3.6)*, respectively. Then the state feedback control,*

$$
(5.2)\qquad u^*(t) = \xi_+(t, P_+(t), \Lambda_+(t))x^+(t) + \xi_-(t, P_-(t), \Lambda_-(t))x^-(t),
$$

is optimal for the problem (2.1)–(2.2)*. Moreover, in this case the optimal cost is*

$$
(5.3)\qquad J^*(x_0) := \inf_{u(\cdot)\in\mathcal{U}} J(x_0, u(\cdot)) = P_+(0)(x_0^+)^2 + P_-(0)(x_0^-)^2.
$$

Proof. First note that Theorems 4.1 and 4.2 ensure that (3.5) and (3.6) admit bounded, nonnegative (in the standard case) or bounded, uniformly positive (in the singular case) solutions $(P_+, \Lambda_+) \in L^2_{\mathcal{F}}(\Omega; C(0, T; \mathbb{R})) \times L^2_{\mathcal{F}}(0, T; \mathbb{R}^k)$ and $(P_-, \Lambda_-) \in L^2_{\mathcal{F}}(\Omega; C(0, T; \mathbb{R})) \times L^2_{\mathcal{F}}(0, T; \mathbb{R}^k)$, respectively. Let $x(\cdot)$ be the solution of (2.1) under an arbitrary given admissible control $u(\cdot)$. By Tanaka's formula (Lemma 3.1), we obtain

$$
(5.4)\quad \begin{aligned} dx^+(t) = {} & 1_{(x(t)>0)}[A(t)x(t) + B(t)u(t)]dt + 1_{(x(t)>0)}[x(t)C(t)' + u(t)'D(t)']dW(t) \\ & + \frac{1}{2}dL(t), \end{aligned}
$$

where $L(\cdot)$ is the local time of $x(\cdot)$ at 0 as specified in Lemma 3.1. Applying Ito's formula to the above, we get

(5.5)
$$\begin{aligned}
& dx^+(t)^2 \\
= \;& 2x^+(t)\Big\{1_{(x(t)>0)}[A(t)x(t)+B(t)u(t)]dt + 1_{(x(t)>0)}[x(t)C(t)'+u(t)'D(t)']dW(t) \\
& \quad +\tfrac{1}{2}dL(t)\Big\} + 1_{(x(t)>0)}[x(t)C(t)'+u(t)'D(t)'][C(t)x(t)+D(t)u(t)]dt \\
= \;& \Big\{2A(t)x^+(t)^2 + 2u(t)'B(t)'x^+(t) + 1_{(x(t)>0)}[x(t)C(t)'+u(t)'D(t)'][C(t)x(t) \\
& \quad +D(t)u(t)]\Big\}dt + 2x^+(t)[x(t)C(t)'+u(t)'D(t)']dW(t),
\end{aligned}$$

where we have used the fact that $x^+(t)dL(t) = 0$ by virtue of (3.2). Using Ito's formula again to (3.5) and (5.5), and writing

(5.6)
$$\Theta_+(t) := -\left\{[2A(t)+C(t)'C(t)]P_+(t) + 2C(t)'\Lambda_+(t) + Q(t) + H_+^*(t, P_+(t), \Lambda_+(t))\right\},$$

we have (after some reorganization)

(5.7)
$$\begin{aligned}
& d[P_+(t)x^+(t)^2] \\
= \;& \Big\{u(t)'[1_{(x(t)>0)}P_+(t)D(t)'D(t)]u(t) + 2u(t)'[P_+(t)B(t)' + P_+(t)D(t)'C(t) \\
& \quad +D(t)'\Lambda_+(t)]x^+(t) \\
& \quad +[\Theta_+(t) + (2A(t)+C(t)'C(t))P_+(t) + 2C(t)'\Lambda_+(t)]x^+(t)^2\Big\}dt \\
& \quad +\Big\{2P_+(t)x^+(t)[x(t)C(t)'+u(t)'D(t)'] + x^+(t)^2\Lambda_+(t)'\Big\}dW(t).
\end{aligned}$$

Similarly, we can derive

(5.8)
$$\begin{aligned}
& d[P_-(t)x^-(t)^2] \\
= \;& \Big\{u(t)'[1_{(x(t)\le 0)}P_-(t)D(t)'D(t)]u(t) - 2u(t)'[P_-(t)B(t)' + P_-(t)D(t)'C(t) \\
& \quad +D(t)'\Lambda_-(t)]x^-(t) \\
& \quad +[\Theta_-(t) + (2A(t)+C(t)'C(t))P_-(t) + 2C(t)'\Lambda_-(t)]x^-(t)^2\Big\}dt \\
& \quad +\Big\{-2P_-(t)x^-(t)[x(t)C(t)'+u(t)'D(t)'] + x^-(t)^2\Lambda_-(t)'\Big\}dW(t),
\end{aligned}$$

where

(5.9)
$$\Theta_-(t) := -\left\{[2A(t)+C(t)'C(t)]P_-(t) + 2C(t)'\Lambda_-(t) + Q(t) + H_-^*(t, P_-(t), \Lambda_-(t))\right\}.$$

Next, we define, for $n \ge 1$, the following stopping time τ_n:

(5.10)
$$\begin{aligned}
\tau_n := \inf\Big\{t \ge 0 \;\Big|\; & \int_0^t \big\{|2P_+(s)x^+(s)[x(s)C(s)'+u(s)'D(s)'] + x^+(s)^2\Lambda_+(s)'|^2\big\}ds \\
& + \int_0^t \big\{|-2P_-(s)x^-(s)[x(s)C(s)'+u(s)'D(s)'] + x^-(s)^2\Lambda_-(s)'|^2\big\}ds \ge n\Big\} \wedge T,
\end{aligned}$$

where $\inf\emptyset := T$. Obviously, τ_n, $n \ge 1$, is an increasing sequence of stopping times converging to T almost surely.

Summing (5.7) and (5.8), taking integration from 0 to τ_n, and then taking expectation, we have (t is suppressed)

$$\begin{aligned}
&E\Big\{P_+(\tau_n)x^+(\tau_n)^2+P_-(\tau_n)x^-(\tau_n)^2\Big\}+E\Big\{\int_0^{\tau_n}[Q(t)x(t)^2+u(t)'R(t)u(t)]dt\Big\}\\
=\;& P_+(0)(x_0^+)^2+P_-(0)(x_0^-)^2\\
&+E\int_0^{\tau_n}\Big\{u'(R+1_{(x(t)>0)}P_+D'D+1_{(x(t)\le 0)}P_-D'D)u\\
&+2u'(P_+B'+P_+D'C+D'\Lambda_+)x^+-2u'(P_-B'+P_-D'C+D'\Lambda_-)x^-\\
&+[\Theta_++(2A+C'C)P_++2C'\Lambda_++Q](x^+)^2\\
&+[\Theta_-+(2A+C'C)P_-+2C'\Lambda_-+Q](x^-)^2\Big\}dt\\
=\;& P_+(0)(x_0^+)^2+P_-(0)(x_0^-)^2\\
&+E\int_0^{\tau_n}\Big\{u'(R+1_{(x(t)>0)}P_+D'D+1_{(x(t)\le 0)}P_-D'D)u\\
&+2u'(P_+B'+P_+D'C+D'\Lambda_+)x^+-2u'(P_-B'+P_-D'C+D'\Lambda_-)x^-\\
&-H_+^*(P_+,\Lambda_+)(x^+)^2-H_-^*(P_-,\Lambda_-)(x^-)^2\Big\}dt. \qquad (5.11)
\end{aligned}$$

Let us now send $n\to\infty$. Then by noting that $x(\cdot)\in L^2_{\mathcal{F}}(\Omega;C(0,T;\mathbb{R}))$ we get, from the dominated convergence theorem, that,

$$J(x_0,u(\cdot))=P_+(0)(x_0^+)^2+P_-(0)(x_0^-)^2+E\int_0^T\varphi(x(t),u(t))dt, \qquad (5.12)$$

where $\varphi(x(t),u(t))$ denotes the integrand on the right-hand side of (5.11).

Now, we are to show that $\varphi(x(t),u(t))\ge 0$ for any $t\in[0,T]$. Indeed, if $x(t)>0$ for some t, then set $u(t)=x(t)v(t)$. Notice $u(t)\in\Gamma$ if and only if $v(t)\in\Gamma$ since Γ is a cone. Then (again t is suppressed)

$$\begin{aligned}
\varphi(x,u)=\;& u'(R+P_+D'D)u+2u'(P_+B'+P_+D'C+D'\Lambda_+)x-H_+^*(P_+,\Lambda_+)x^2\\
=\;& [v'(R+P_+D'D)v+2v'(P_+B'+P_+D'C+D'\Lambda_+)]x^2-H_+^*(P_+,\Lambda_+)x^2\\
\ge\;& H_+^*(P_+,\Lambda_+)x^2-H_+^*(P_+,\Lambda_+)x^2=0. \qquad (5.13)
\end{aligned}$$

Moreover, the inequality becomes an equality when $u^*(t)=x(t)v^*(t)=x^+(t)\xi_+(t,P_+(t),\Lambda_+(t))\in\Gamma$. Next, if $x(t)<0$ for some t, then put $u(t)=-x(t)v(t)$. In this case,

$$\begin{aligned}
\varphi(x,u)=\;& u'(R+P_-D'D)u+2u'(P_-B'+P_-D'C+D'\Lambda_-)x-H_-^*(P_-,\Lambda_-)x^2\\
=\;& [v'(R+P_+D'D)v-2v'(P_-B'+P_-D'C+D'\Lambda_-)]x^2-H_-^*(P_-,\Lambda_-)x^2\\
\ge\;& H_-^*(P_-,\Lambda_-)x^2-H_-^*(P_-,\Lambda_-)x^2=0, \qquad (5.14)
\end{aligned}$$

where the equality holds at $u^*(t)=-x(t)v^*(t)=x^-(t)\xi_-(t,P_-(t),\Lambda_-(t))\in\Gamma$. Finally, when $x(t)=0$, then $\varphi(x,u)=u'(R+P_-D'D)u\ge 0$; here, the equality holds at $u^*(t)=0$.

The above analysis together with (5.12) shows that

$$J(x_0,u(\cdot))\ge P_+(0)(x_0^+)^2+P_-(0)(x_0^-)^2\ \ \forall u(\cdot)\in\mathcal{U}, \qquad (5.15)$$

whereas the equality is achieved when $u^*(\cdot)$ is defined by (5.2).

If we can prove that the control $u^*(\cdot)$ defined by (5.2) is in $L^2_{\mathcal{F}}(0,T;\mathbb{R}^m)$, then the proof of the theorem will be finished. To this end, first note that, as shown before, there exist two constants $C_1 > 0$ and $C_2 > 0$, such that

$$H_+(t,v,P,\Lambda) \geq C_1|v|^2 - C_2(|P|+|\Lambda|)v \geq C_1|v|[|v| - \frac{C_2}{C_1}(|P|+|\Lambda|)]. \tag{5.16}$$

Hence $H_+(t,v,P,\Lambda) > 0$ if $|v| > \frac{C_2}{C_1}(|P|+|\Lambda|)$. From the definition of $\xi_+(t,P,\Lambda)$, it follows that

$$|\xi_+(t,P,\Lambda)| \leq \frac{C_2}{C_1}(|P|+|\Lambda|). \tag{5.17}$$

The same is true for ξ_-.

Now, under the feedback control (5.2), the system dynamics (2.1) read

(5.18)
$$\begin{cases} dx(t) = & [A(t)x(t) + B(t)\xi_+(t,P_+(t),\Lambda_+(t))x^+(t) + B(t)\xi_-(t,P_-(t),\Lambda_-(t))x^-(t)]dt \\ & +[x(t)C(t)' + x^+(t)\xi_+(t,P_+(t),\Lambda_+(t))'D(t)' \\ & \quad +x^-(t)\xi_-(t,P_-(t),\Lambda_-(t))'D(t)']dW(t), \quad t\in[0,T], \\ x(0) = & x_0. \end{cases}$$

This equation has a unique continuous $\{\mathcal{F}_t\}_{t\geq 0}$-adapted solution; see the lemma following the end of the proof of this theorem. We denote this solution by $x^*(\cdot)$, and hence $u^*(t) = \xi_+(t,P_+(t),\Lambda_+(t))x^{*+}(t) + \xi_-(t,P_-(t),\Lambda_-(t))x^{*-}(t)$. The continuity of $x^*(\cdot)$ along with (5.17) leads to

$$\int_0^T (|x^*(t)|^2 + |u^*(t)|^2)dt < +\infty, \quad P-a.s.. \tag{5.19}$$

Denote by τ_n^*, $n \geq 1$, the sequence of stopping times defined by (5.10) where the state–control pair is taken as $(x^*(\cdot), u^*(\cdot))$. It follows from (5.19) that $\tau_n^* \to T$ as $n \to +\infty$, P-a.s.. On the other hand, (5.11) yields

(5.20)
$$\begin{aligned} &E\Big\{P_+(\tau_n^*)x^{*+}(\tau_n^*)^2 + P_-(\tau_n^*)x^{*-}(\tau_n^*)^2\Big\} + E\int_0^{\tau_n^*}[Q(t)x^*(t)^2 + u^*(t)'R(t)u^*(t)]dt \\ &\quad = P_+(0)(x_0^+)^2 + P_-(0)(x_0^-)^2. \end{aligned}$$

We are now in position to prove $u^*(\cdot) \in L^2_{\mathcal{F}}(0,T;\mathbb{R}^m)$. To do so, we will treat the standard case and the singular case separately.

For the standard case, denote by $c > 0$ such that $R \geq cI_m$; then it follows from (5.20) that

$$cE\int_0^{\tau_n^*} |u^*(t)|^2 dt \leq P_+(0)(x_0^+)^2 + P_-(0)(x_0^-)^2. \tag{5.21}$$

This implies that $u^*(\cdot) \in L^2_{\mathcal{F}}(0,T;\mathbb{R}^m)$.

For the singular case, construct a sequence of stopping times as follows:

$$\theta_n := \inf\left\{t \geq 0 \,\middle|\, \int_0^t (|x^*(s)|^2 + |C(s)x^*(s) + D(s)u^*(s)|^2)ds \geq n\right\} \wedge T.$$

Again θ_n increasingly converges to T a.s. due to (5.19). Rewrite (2.1) under $u^*(\cdot)$ as a kind of BSDE with a random terminal time,

(5.22)
$$\begin{cases} dx^*(t) = [(A - B(D'D)^{-1}D'C)x^*(t) + B(D'D)^{-1}D'z(t)]dt + z(t)'dW(t), \\ t \in [0, \tau_n^* \wedge \theta_n], \\ x^*(\tau_n^* \wedge \theta_n) = x^*(\tau_n^* \wedge \theta_n), \end{cases}$$

where $z(t) := C(t)x^*(t) + D(t)u^*(t)$. Theorem 4.2 provides that there is a constant $c > 0$ such that for any $t \in [0, T]$, $P_+(t) \geq c$ and $P_-(t) \geq c$. Thus, (5.20) with τ_n^* replaced by $\tau_n^* \wedge \theta_n$ leads to

$$cE[x^*(\tau_n^* \wedge \theta_n)^2] \leq P_+(0)(x_0^+)^2 + P_-(0)(x_0^-)^2.$$

On the other hand, the standard estimate for the BSDE (5.22) states that, for a constant $\tilde{c} > 0$,

$$\begin{aligned} E\int_0^{\tau_n^* \wedge \theta_n} (|x^*(s)|^2 + |z(s)|^2)ds &\leq \tilde{c}E[x^*(\tau_n^* \wedge \theta_n)^2] \\ &\leq \frac{\tilde{c}}{c}[P_+(0)(x_0^+)^2 + P_-(0)(x_0^-)^2]. \end{aligned}$$

Appealing to Fatou's lemma, we conclude that $x^*(\cdot) \in L^2_{\mathcal{F}}(0, T; \mathbb{R})$ and $z(\cdot) \times L^2_{\mathcal{F}}(0, T; \mathbb{R}^k)$. This in turn implies $u^*(\cdot) \in L^2_{\mathcal{F}}(0, T; \mathbb{R}^m)$ as $u^*(t) = [D(t)'D(t)]^{-1} D(t)'[z(t) - C(t)x^*(t)]$. □

LEMMA 5.1. *Equation* (5.18) *has a unique continuous* $\{\mathcal{F}_t\}_{t\geq 0}$*-adapted solution.*

Proof. Setting

$$\begin{aligned} &B_+(t) := B(t)\xi_+(t, P_+(t), \Lambda_+(t)), \ B_-(t) := B(t)\xi_-(t, P_-(t), \Lambda_-(t)), \\ &D_+(t) := D(t)\xi_+(t, P_+(t), \Lambda_+(t)), \ D_-(t) := D(t)\xi_-(t, P_-(t), \Lambda_-(t)), \end{aligned}$$

then (5.18) becomes

$$\begin{cases} dx(t) = & [A(t)x(t) + B_+(t)x^+(t) + B_-(t)x^-(t)]dt \\ & +[x(t)C(t)' + x^+(t)D_+(t)' + x^-(t)D_-(t)']dW(t), \quad t \in [0, T], \\ x(0) = & x_0. \end{cases} \tag{5.23}$$

Consider the following two linear SDEs:

(5.24)
$$\begin{cases} dx_+(t) = & [A(t) + B_+(t)]x_+(t)dt + x_+(t)[C(t) + D_+(t)]'dW(t), \quad t \in [0, T], \\ x_+(0) = & x_0^+, \end{cases}$$

and

(5.25)
$$\begin{cases} dx_-(t) = & [A(t) - B_-(t)]x_-(t)dt + x_-(t)[C(t) - D_-(t)]'dW(t), \quad t \in [0, T], \\ x_-(0) = & x_0^-. \end{cases}$$

It is well known that both (5.24) and (5.25) have unique continuous $\{\mathcal{F}_t\}_{t\geq 0}$-adapted solutions which can be represented explicitly as

$$\begin{aligned} x_+(t) = x_0^+ \exp\Big\{ &\int_0^t [A(s) + B_+(s)]ds + \int_0^t [C(s) + D_+(s)]'dW(s) \\ &- \frac{1}{2}\int_0^t |C(s) + D_+(s)|^2 ds\Big\} \end{aligned}$$

and

$$x_-(t) = x_0^- \exp\left\{\int_0^t [A(s) - B_-(s)]ds + \int_0^t [C(s) - D_-(s)]'dW(s) - \frac{1}{2}\int_0^t |C(s) - D_-(s)|^2 ds\right\}.$$

Define

$$x(t) := x_+(t) - x_-(t).$$

Since $x_+(t) \geq 0$, $x_-(t) \geq 0$, and $x_+(t)x_-(t) = 0$, we conclude that

$$x^+(t) = x_+(t), \quad x^-(t) = x_-(t) \ \ \forall t \in [0, T].$$

Subtracting (5.25) from (5.24) we get that $x(\cdot)$ is a continuous adapted solution of (5.23).

Let us turn to the uniqueness of the solution. Suppose that $x_1(\cdot)$ and $x_2(\cdot)$ are two continuous adapted solutions of (5.23). Put $\hat{x}(\cdot) := x_1(\cdot) - x_2(\cdot)$. We apply a linearization procedure as follows. Set

$$\alpha_+(t) := \frac{x_1^+(t) - x_2^+(t)}{x_1(t) - x_2(t)} 1_{\{x_1(t) \neq x_2(t)\}}, \ \alpha_-(t) := \frac{x_1^-(t) - x_2^-(t)}{x_1(t) - x_2(t)} 1_{\{x_1(t) \neq x_2(t)\}}.$$

Then $\hat{x}(\cdot)$ is a continuous adapted solution of the following linear SDE:

$$\text{(5.26)} \quad \begin{cases} d\hat{x}(t) = & [A(t) + B_+(t)\alpha_+(t) + B_-(t)\alpha_-(t)]\hat{x}(t)dt \\ & + \hat{x}(t)[C(t) + D_+(t)\alpha_+(t) + D_-(t)\alpha_-(t)]'dW(t), \quad t \in [0, T], \\ \hat{x}(0) = & 0. \end{cases}$$

Hence $\hat{x}(t) \equiv 0$ via a similar representation as (5.24) or (5.25), and the uniqueness of the solution is proved. □

Let us conclude this section by noting that a byproduct of Theorem 5.1 is the uniqueness of the solution to the ESREs (3.5) and (3.6). Indeed, consider an LQ control problem in an interval $[s, T]$, with $s \in [0, T)$, where the system dynamics is (2.1) with initial time s and initial state $x(s) = x_s \in L^2_{\mathcal{F}_s}(\Omega; \mathbb{R})$, and the cost functional is

$$J_s(x_s, u(\cdot)) := E\left\{\int_s^T \big[Q(t)x(t)^2 + u(t)'R(t)u(t)\big]dt + Gx(T)^2 \Big| \mathcal{F}_s\right\}.$$

Let $(P_+, \Lambda_+) \in L^2_{\mathcal{F}}(\Omega; C(0, T; \mathbb{R})) \times L^2_{\mathcal{F}}(0, T; \mathbb{R}^k)$ and $(P_-, \Lambda_-) \in L^2_{\mathcal{F}}(\Omega; C(0, T; \mathbb{R})) \times L^2_{\mathcal{F}}(0, T; \mathbb{R}^k)$ be *any* bounded, nonnegative (in the standard case) or bounded, uniformly positive (in the singular case) solutions to (3.5) and (3.6), respectively. Then, going through the same analysis as in the proof of Theorem 5.1 we deduce that the optimal cost is

$$\text{(5.27)} \qquad J_s^*(x_s) := \inf_{u(\cdot) \text{ admissible}} J_s(x_s, u(\cdot)) = P_+(s)(x_s^+)^2 + P_-(s)(x_s^-)^2.$$

This proves the following uniqueness result.

THEOREM 5.2. *Each of the ESREs* (3.5) *and* (3.6) *admits at most one bounded, nonnegative (in the standard case) or bounded, uniformly positive (in the singular case) solution.*

6. Application to a mean-variance portfolio selection problem. Consider a financial market with $m+1$ securities, consisting of a bank account and m stocks. The value of the bank account, $S_0(t)$, satisfies an ordinary differential equation,

$$\begin{cases} dS_0(t) = r(t)\, S_0(t)\, dt, \quad t \in [0,\, T], \\ S_0(0) = s_0 > 0, \end{cases} \tag{6.1}$$

where the interest rate $r(t) > 0$ is a deterministic, uniformly bounded, scalar-valued function. The price of each of the stocks, $S_1(t), \ldots, S_m(t)$, satisfies the SDE,

$$\begin{cases} dS_i(t) = S_i(t)\Big\{\mu_i(t)\, dt + \sum_{j=1}^m \sigma_{ij}(t)\, dW^j(t)\Big\}, \quad t \in [0,\, T], \\ S_i(0) = s_i > 0, \end{cases} \tag{6.2}$$

where $\mu_i(t) > 0$ and $\sigma_i(t) = [\sigma_{i1}, \ldots, \sigma_{im}(t)]$ are the appreciation rate and dispersion (or volatility) rate of the ith stock. Here, $\mu_i(t)$ and $\sigma_{ij}(t)$ are scalar-valued, $\{\mathcal{F}_t\}_{t\geq 0}$-adapted, uniformly bounded *stochastic processes*. Denoting

$$\sigma(t) := \begin{bmatrix} \sigma_1(t) \\ \vdots \\ \sigma_m(t) \end{bmatrix} \in \mathbb{R}^{m\times m}, \tag{6.3}$$

we assume throughout that $\sigma(t)$ is uniformly nondegenerate: that is, there exists a deterministic $\delta > 0$ such that

$$\sigma(t)\, \sigma(t)' \geq \delta\, I_m, \quad \forall\, t \in [0,\, T],\ P - a.s.. \tag{6.4}$$

In particular, $\sigma(t)$ must be nonsingular a.e. $t \in [0,\, T]$, P-a.s..

Suppose that the total wealth of an agent at time $t \geq 0$ is denoted by $x(t)$. If transaction costs and consumption are ignored and share trading takes place in continuous time, then we have

$$\begin{cases} dx(t) &= \Big\{r(t)\, x(t) + \sum_{i=1}^m [\mu_i(t) - r(t)]\, u_i(t)\Big\}\, dt \\ &\quad + \sum_{j=1}^m \sum_{i=1}^m \sigma_{ij}(t)\, u_i(t)\, dW^j(t), \quad t \in [0,\, T], \\ x(0) &= x_0 > 0, \end{cases} \tag{6.5}$$

where $u_i(t)$ is the total market value of the agent's wealth in the ith asset. We refer to $u(\cdot) := (u_1(\cdot), \ldots, u_m(\cdot))'$ as the *portfolio* of the agent. In our model, short-selling of the stocks is not allowed; hence we have the following constraints on a portfolio $u(\cdot) = (u_1(\cdot), \ldots, u_m(\cdot))'$:

$$u_i(t) \geq 0 \ \ \forall t \in [0,T], \ \ i = 1, \ldots, m. \tag{6.6}$$

Note that $u_0(\cdot)$ has been excluded from a portfolio since it is completely determined by the allocation of stocks and the total wealth $x(\cdot)$. Moreover, we do allow $u_0(t) < 0$, meaning that the agent is borrowing the amount $|u_0(t)|$ from the bank at rate $r(t)$.

DEFINITION 6.1. *A portfolio $u(\cdot)$ is said to be* admissible *if it is $\mathbb{R}^m$-valued, square-integrable (i.e., $E\int_0^T |u(t)|^2 dt < +\infty$), $\{\mathcal{F}_t\}_{t\geq 0}$-adapted, and satisfies* (6.6). *In this case, we refer to $(x(\cdot),\, u(\cdot))$ as an* admissible (wealth–portfolio) pair.

In a mean-variance portfolio selection problem, an agent's objective is to find an admissible portfolio $u(\cdot)$ such that the expected terminal wealth satisfies $Ex(T) = z$,

for some $z \geq x_0 e^{\int_0^T r(s)ds}$, while the risk measured by the variance of the terminal wealth

$$\text{Var } x(T) := E[x(T) - Ex(T)]^2 = E[x(T)]^2 - z^2 \tag{6.7}$$

is minimized. The restriction of the targeted payoff $z \geq x_0 e^{\int_0^T r(s)ds}$ is natural as the latter can always be achieved by putting all the money in the bank. Mathematically, it can be formulated as the following problem parameterized by $z \geq x_0 e^{\int_0^T r(s)ds}$:

$$\begin{cases} \text{Minimize} & J_{\text{MV}}(x_0, u(\cdot)) := E[x(T)]^2 - z^2, \\ \text{subject to:} & Ex(T) = z, \\ & (x(\cdot), u(\cdot)) \text{ is admissible for (6.5).} \end{cases} \tag{6.8}$$

The above problem is called *feasible* if there is at least one portfolio satisfying the constraints of (6.8). Finally, an optimal portfolio to (6.8) is called an *efficient portfolio* corresponding to z, the corresponding $(\text{Var } x(T), z)$ is called an *efficient point*, whereas the set of all the efficient points, with $z \geq x_0 e^{\int_0^T r(s)ds}$, is called an *efficient frontier*.

Equation (6.5) can be rewritten as

$$\begin{cases} dx(t) = [r(t)x(t) + B(t)u(t)]dt + u(t)'\sigma(t)dW(t), \\ x(0) = x_0, \end{cases} \tag{6.9}$$

where

$$B(t) := (\mu_1(t) - r(t), \ldots, \mu_m(t) - r(t)). \tag{6.10}$$

Since the problem (6.8) involves a terminal constraint $Ex(T) = z$, we first investigate conditions under which the problem is feasible for *any* $z \in [x_0 e^{\int_0^T r(s)ds}, +\infty)$.

THEOREM 6.1. *The mean-variance problem* (6.8) *is feasible for every* $z \in [x_0 e^{\int_0^T r(s)ds}, +\infty)$ *if and only if*

$$\sum_{i=1}^{m} E \int_0^T [\mu_i(t) - r(t)]^+ dt > 0. \tag{6.11}$$

Proof. We first prove the "if" part. Define

$$M_i := \{(t, \omega) : \mu_i(t, \omega) > r(t)\}, \quad i = 1, 2, \ldots, m.$$

Condition (6.11) implies that at least one of the sets M_i has a nonzero measure (in terms of the product of the Lebesgue measure and P). Suppose M_{i_0} has a nonzero measure. Construct a family of admissible portfolios $u^\beta(\cdot) := \beta u(\cdot)$, where $\beta \geq 0$ and the components of $u(\cdot)$ are all zero except its i_0th component which is defined to be

$$u_{i_0}(t, \omega) := \begin{cases} \mu_{i_0}(t, \omega) - r(t), & \text{if } (t, \omega) \in M_{i_0}, \\ 0, & \text{if } (t, \omega) \notin M_{i_0}. \end{cases} \tag{6.12}$$

Let $x^\beta(\cdot)$ be the wealth process corresponding to $u^\beta(\cdot)$. By linearity of the wealth equation, we have $x^\beta(t) = x^0(t) + \beta x^1(t)$, where $x^0(t) := x_0 e^{\int_0^t r(s)ds}$ and $x^1(\cdot)$ is the solution to the following equation:

$$\begin{cases} dx^1(t) = [r(t)x^1(t) + B(t)u(t)]dt + u(t)'\sigma(t)dW(t), \\ x^1(0) = 0. \end{cases} \tag{6.13}$$

Therefore, problem (6.8) is feasible for every $z \in [x_0 e^{\int_0^T r(s)ds}, +\infty)$ if there exists $\beta \geq 0$ such that $z = Ex^\beta(T) \equiv x^0(T) + \beta Ex^1(T)$. Equivalently, (6.8) is feasible for every $z \in [x_0 e^{\int_0^T r(s)ds}, +\infty)$ if $Ex^1(T) > 0$. However, applying Itô's formula we get

$$d[e^{\int_t^T r(s)ds} x^1(t)] = e^{\int_t^T r(s)ds} B(t)u(t)dt + \{\cdots\}dW(t).$$

Integrating from 0 to T and taking expectation we obtain

$$Ex^1(T) = E\int_0^T e^{\int_t^T r(s)ds} B(t)u(t)dt > 0, \tag{6.14}$$

due to the way $u(\cdot)$ was constructed. Consequently, (6.8) is feasible if (6.11) holds.

Conversely, suppose that problem (6.8) is feasible for every $z \in [x_0 e^{\int_0^T r(s)ds}, +\infty)$. Then for each z, there is an admissible portfolio $u(\cdot)$ so that $Ex(T) = z$. However, we can always decompose $x(t) = x^0(t) + x^1(t)$, where $x^1(\cdot)$ satisfies (6.13). This leads to $Ex^0(T) + Ex^1(T) = z$. Now, $Ex^0(T) \equiv z_0$ is independent of $u(\cdot)$; thus it is necessary that there is a $u(\cdot)$ with $Ex^1(T) > 0$. It follows then from (6.14) that (6.11) must be valid. □

Now we are going to solve the optimization problem (6.8) under the feasibility assumption (6.11). To handle the constraint $Ex(T) = z$ we apply the Lagrange multiplier technique. Define

$$\begin{aligned} J(x_0, u(\cdot), \lambda) :=\ & E\big\{x(T)^2 - z^2 - 2\lambda[x(T) - z]\big\} \\ =\ & E\big[|x(T) - \lambda|^2\big] - (\lambda - z)^2, \quad \lambda \in \mathbb{R}. \end{aligned} \tag{6.15}$$

Since $J_{\rm MV}(x_0, u(\cdot))$ is strictly convex in $u(\cdot)$ and the constraint function $Ex(T) - z$ is affine in $u(\cdot)$, we can apply the well-known duality theorem (see, e.g., [19]). Based on this theorem, we may first solve the following unconstrained problem parameterized by the Lagrange multiplier $\lambda \in \mathbb{R}$:

$$\begin{cases} \text{Minimize} & J(x_0, u(\cdot), \lambda) := E\big[|x(T) - \lambda|^2\big] - (\lambda - z)^2, \\ \text{subject to:} & (x(\cdot), u(\cdot)) \text{ is admissible for (6.9).} \end{cases} \tag{6.16}$$

This problem is exactly a singular case of the general LQ model solved in section 5, with $Q(t) = R(t) = 0$ and $\Gamma = \mathbb{R}^m_+$. Thus we will apply the general result to the problem. Let us first write down the specialization of the ESREs (3.5) and (3.6) as

$$\begin{cases} dP_+(t) = -[2r(t)P_+(t) + H^*_+(t, P_+(t), \Lambda_+(t))]dt + \Lambda_+(t)'dW(t), \quad t \in [0, T], \\ P_+(T) = 1, \\ P_+(t) > 0, \end{cases} \tag{6.17}$$

$$\begin{cases} dP_-(t) = -[2r(t)P_-(t) + H^*_-(t, P_-(t), \Lambda_-(t))]dt + \Lambda_-(t)'dW(t), \quad t \in [0, T], \\ P_-(T) = 1, \\ P_-(t) > 0, \end{cases} \tag{6.18}$$

where

$$\begin{aligned} H^*_+(t, P, \Lambda) &:= \min_{v \in \mathbb{R}^m_+} \{v'P\sigma(t)\sigma(t)'v + 2v'[B(t)'P + \sigma(t)\Lambda]\}, \\ H^*_-(t, P, \Lambda) &:= \min_{v \in \mathbb{R}^m_+} \{v'P\sigma(t)\sigma(t)'v - 2v'[B(t)'P + \sigma(t)\Lambda]\}. \end{aligned} \tag{6.19}$$

Also, define

$$\begin{aligned}&\xi_+(t,P,\Lambda) := \operatorname{argmin}_{v\in\mathbb{R}^m_+} H_+(t,v,P,\Lambda),\\ &\xi_-(t,P,\Lambda) := \operatorname{argmin}_{v\in\mathbb{R}^m_+} H_-(t,v,P,\Lambda),\quad (t,P,\Lambda)\in[0,T]\times\mathbb{R}\times\mathbb{R}^k.\end{aligned}\tag{6.20}$$

Clearly, Theorems 4.2 and 5.2 apply to (6.17) and (6.18) ensuring that they admit unique bounded, uniformly positive solutions.

LEMMA 6.1. *Assume that* (6.11) *holds, and let* $(P_+, \Lambda_+) \in L^2_{\mathcal{F}}(\Omega; C(0, T; \mathbb{R})) \times L^2_{\mathcal{F}}(0, T; \mathbb{R}^m)$ *and* $(P_-, \Lambda_-) \in L^2_{\mathcal{F}}(\Omega; C(0, T; \mathbb{R}))\times L^2_{\mathcal{F}}(0, T; \mathbb{R}^m)$ *be the unique bounded, uniformly positive solutions to the ESREs* (6.17) *and* (6.18)*, respectively. Then it must hold that*

$$P_+(0)e^{-2\int_0^T r(s)ds} - 1 \le 0, \quad \textit{and} \quad P_-(0)e^{-2\int_0^T r(s)ds} - 1 < 0.\tag{6.21}$$

Proof. Define $g(t) := P_-(t)e^{-2\int_t^T r(s)ds}$. Then it is straightforward that

$$dg(t) = -e^{-2\int_t^T r(s)ds}H^*_-(t, P_-(t), \Lambda_-(t))dt + e^{-2\int_t^T r(s)ds}\Lambda_-(t)'dW(t).$$

Integrating from 0 to T and taking expectation we have

$$1 - g(0) = -E\int_0^T e^{-2\int_t^T r(s)ds}H^*_-(t, P_-(t), \Lambda_-(t))dt \ge 0,\tag{6.22}$$

since $H^*_-(t, P_-(t), \Lambda_-(t)) \le 0$ by its very definition. Hence $P_-(0)e^{-2\int_0^T r(s)ds} \equiv g(0) \le 1$. Similarly, we can prove that $P_+(0)e^{-2\int_0^T r(s)ds} - 1 \le 0$.

It remains to prove the strict inequality $P_-(0)e^{-2\int_0^T r(s)ds} - 1 < 0$. In fact, if $P_-(0)e^{-2\int_0^T r(s)ds} - 1 = 0$, then it follows from (6.22) that $H^*_-(t, P_-(t), \Lambda_-(t)) = 0$, a.e. $t \in [0, T]$, P-a.s.. Thus we deduce, from the uniqueness of solution to the BSDE (6.18), that $P_-(t) = e^{2\int_t^T r(s)ds}$, and $\Lambda_-(t) = 0$. Consequently, $H^*_-(t, P_-(t), 0) = 0$.

On the other hand,

$$\begin{aligned}H^*_-(t, P_-(t), 0) &= \min_{v\in\mathbb{R}^m_+} P_-(t)[v'\sigma(t)\sigma(t)'v - 2v'B(t)']\\ &\le \min_{v\in\mathbb{R}^m_+} K_3[K_4|v|^2 - 2B(t)v],\end{aligned}$$

for some constants $K_3 > 0$ and $K_4 > 0$. Notice that the minimum value on the right-hand side of the above equation *strictly* negative whenever $B(t,\omega)^+ := (B_1(t,\omega)^+, \ldots, B_m(t,\omega)^+) \ne 0$. In view of the assumption (6.11), the set of (t,ω) on which $B(t,\omega)^+$ is nonzero has a nonzero measure. Thus we get a contradiction. ∎

Remark 6.1. One does not have the strict inequality $P_+(0)e^{-2\int_0^T r(s)ds} - 1 < 0$ since no information about $B(t,\omega)^- := (B_1(t,\omega)^-, \ldots, B_m(t,\omega)^-)$ is available. On the other hand, the inequality $P_-(0)e^{-2\int_0^T r(s)ds} - 1 < 0$ is exactly what is required in what follows.

THEOREM 6.2. *Let* $(P_+, \Lambda_+) \in L^2_{\mathcal{F}}(\Omega; C(0, T; \mathbb{R}))\times L^2_{\mathcal{F}}(0, T; \mathbb{R}^m)$ *and* $(P_-, \Lambda_-) \in L^2_{\mathcal{F}}(\Omega; C(0, T; \mathbb{R})) \times L^2_{\mathcal{F}}(0, T; \mathbb{R}^m)$ *be the unique bounded, uniformly positive solutions to the ESREs* (6.17) *and* (6.18)*, respectively. Then the state feedback control*

$$\begin{aligned}u^*(t) = {}&\xi_+(t, P_+(t), \Lambda_+(t))\left(x(t) - \lambda e^{-\int_t^T r(s)ds}\right)^+\\ &+ \xi_-(t, P_-(t), \Lambda_-(t))\left(x(t) - \lambda e^{-\int_t^T r(s)ds}\right)^-\end{aligned}\tag{6.23}$$

is optimal for the problem (6.16). *Moreover, in this case the optimal cost is*

(6.24)
$$J^*(x_0,\lambda) := \inf_{u(\cdot)\in L^2_{\mathcal{F}}(0,T;\mathbb{R}^m_+)} J(x_0,u(\cdot),\lambda)$$
$$= \begin{cases} [P_+(0)e^{-2\int_0^T r(s)ds}-1]\lambda^2 - 2[x_0P_+(0)e^{-\int_0^T r(s)ds}-z]\lambda + P_+(0)x_0^2 - z^2, \\ \qquad \text{if } x_0 > \lambda e^{-\int_0^T r(s)ds}, \\ [P_-(0)e^{-2\int_0^T r(s)ds}-1]\lambda^2 - 2[x_0P_-(0)e^{-\int_0^T r(s)ds}-z]\lambda + P_-(0)x_0^2 - z^2, \\ \qquad \text{if } x_0 \le \lambda e^{-\int_0^T r(s)ds}. \end{cases}$$

Proof. Set

$$y(t) := x(t) - \lambda e^{-\int_t^T r(s)ds}. \tag{6.25}$$

It turns out the wealth equation (6.9) in terms of $y(\cdot)$ has exactly the same form except for the initial condition,

$$\begin{cases} dy(t) = [r(t)y(t) + B(t)u(t)]dt + u(t)'\sigma(t)dW(t), \\ y(0) = x_0 - \lambda e^{-\int_0^T r(s)ds}, \end{cases} \tag{6.26}$$

whereas the cost function (6.15) can be written as

$$J(y_0,u(\cdot),\lambda) = Ey(T)^2 - (\lambda - z)^2. \tag{6.27}$$

The above problem (6.26)–(6.27) is exactly a special case of the general problem we have solved in section 5 (ignoring the constant term $-(\lambda-z)^2$ in (6.27)). Hence the optimal feedback control (6.23) follows from (5.2). Finally, the optimal cost is

$$J^*(x_0,\lambda) = P_+(0)[(x_0 - \lambda e^{-\int_0^T r(s)ds})^+]^2 + P_-(0)[(x_0 - \lambda e^{-\int_0^T r(s)ds})^-]^2 - (\lambda - z)^2$$

which equals the right-hand side of (6.24) after some simple manipulations. □

THEOREM 6.3 (efficient portfolios and efficient frontier). *Assume that* (6.11) *holds. Then the efficient portfolio corresponding to* $z \ge x_0 e^{\int_0^T r(s)ds}$, *as a feedback of the wealth process, is*

$$\begin{aligned} u^*(t) = {} & \xi_+(t,P_+(t),\Lambda_+(t))\left(x^*(t) - \lambda^* e^{-\int_t^T r(s)ds}\right)^+ \\ & + \xi_-(t,P_-(t),\Lambda_-(t))\left(x^*(t) - \lambda^* e^{-\int_t^T r(s)ds}\right)^-, \end{aligned} \tag{6.28}$$

where

$$\lambda^* := \frac{z - x_0P_-(0)e^{-\int_0^T r(s)ds}}{1 - P_-(0)e^{-2\int_0^T r(s)ds}}. \tag{6.29}$$

Moreover, the efficient frontier is

(6.30)
$$\operatorname{Var} x^*(T) = \frac{P_-(0)e^{-2\int_0^T r(s)ds}}{1 - P_-(0)e^{-2\int_0^T r(s)ds}}\left[Ex^*(T) - x_0e^{\int_0^T r(s)ds}\right]^2, \quad Ex^*(T) \ge x_0e^{\int_0^T r(s)ds}.$$

Proof. First note that λ^* in (6.29) is well defined thanks to Lemma 6.1. Now, if $z = x_0e^{\int_0^T r(s)ds}$, then it is straightforward that the corresponding efficient portfolio

is $u^*(t) \equiv 0$, meaning that all the wealth is to be put in the bank account. The resulting wealth process is $x^*(t) = x_0 e^{\int_0^t r(s)ds}$. On the other hand, in this case the associated $\lambda^* = x_0 e^{\int_0^T r(s)ds}$. Thus the portfolio given by (6.28) reduces to $u^*(t) \equiv 0$ with $x^*(t) = x_0 e^{\int_0^t r(s)ds}$. This implies that (6.28) is indeed the efficient portfolio when $z = x_0 e^{\int_0^T r(s)ds}$.

So now we need only to prove the theorem for any fixed $z > x_0 e^{\int_0^T r(s)ds}$. Applying the duality theorem (see, e.g., [19, p. 224, Theorem 1][1]) we have

(6.31)
$$J^*_{\mathrm{MV}}(x_0) := \inf_{u(\cdot)\in L^2_{\mathcal{F}}(0,T;\mathbb{R}^m_+)} J_{\mathrm{MV}}(x_0, u(\cdot)) = \sup_{\lambda\in\mathbb{R}} \inf_{u(\cdot)\in L^2_{\mathcal{F}}(0,T;\mathbb{R}^m_+)} J(x_0, u(\cdot), \lambda) > -\infty,$$

and the optimal feedback control for (6.8) is (6.28), due to Theorem 6.2, with λ replaced by λ^* which maximizes $J^*(x_0, \lambda)$ $(= \inf_{u(\cdot)\in L^2_{\mathcal{F}}(0,T;\mathbb{R}^m_+)} J(x_0, u(\cdot), \lambda))$ over $\lambda \in \mathbb{R}$.

If $\lambda \in (-\infty, x_0 e^{\int_0^T r(s)ds})$, then the expression (6.24) gives, taking into consideration (6.21) and the fact that $z \geq x_0 e^{\int_0^T r(s)ds}$,

$$\begin{aligned}
\frac{\partial}{\partial\lambda} J^*(x_0, \lambda) &= 2[P^+(0)e^{-2\int_0^T r(s)ds} - 1]\lambda - 2[x_0 P_+(0)e^{-\int_0^T r(s)ds} - z] \\
&\geq 2[P^+(0)e^{-2\int_0^T r(s)ds} - 1]x_0 e^{\int_0^T r(s)ds} \\
&\quad -2[x_0 P_+(0)e^{-\int_0^T r(s)ds} - x_0 e^{\int_0^T r(s)ds}] \\
&= 0.
\end{aligned}$$

Hence,

$$\sup_{\lambda\in\mathbb{R}} J^*(x_0, \lambda) = \sup_{\lambda\in[x_0 e^{\int_0^T r(s)ds}, +\infty)} J^*(x_0, \lambda).$$

But for $\lambda \in [x_0 e^{\int_0^T r(s)ds}, +\infty)$, it follows from (6.24) that $J^*(x_0, \lambda)$ is a quadratic function in λ whose maximizer is given by (6.29) (noticing Lemma 6.1), whereas

(6.32)
$$\begin{aligned}
J^*_{\mathrm{MV}}(x_0) &= \sup_{\lambda\in\mathbb{R}} \inf_{u(\cdot)\in L^2_{\mathcal{F}}(0,T;\mathbb{R}^m_+)} J^*(x_0, \lambda) \\
&= \sup_{\lambda\in\mathbb{R}} \Big\{ [P_-(0)e^{-2\int_0^T r(s)ds} - 1]\lambda^2 - 2[x_0 P_-(0)e^{-\int_0^T r(s)ds} - z]\,\lambda \\
&\qquad + P_-(0)x_0^2 - z^2 \Big\} \\
&= \frac{P_-(0)e^{-2\int_t^T r(s)ds}}{1 - P_-(0)e^{-2\int_0^T r(s)ds}} \left[z - x_0 e^{\int_0^T r(s)ds} \right]^2, \quad z \geq x_0 e^{\int_0^T r(s)ds}.
\end{aligned}$$

This proves (6.30), noting that $Ex^*(T) = z$. □

Corollary 6.1. *Assume that* (6.11) *holds. Then the efficient portfolio* (6.28) *can be rewritten as*

$$u^*(t) = \xi_-(t, P_-(t), \Lambda_-(t)) \left(\lambda^* e^{-\int_t^T r(s)ds} - x^*(t) \right). \tag{6.33}$$

[1] To be precise, one should apply [19, p. 236, Problem 7] together with the proof of [19, p. 224, Theorem 1] in our case, as there is an *equality* constraint in (6.8). To be able to use the result there, one needs to check a condition posed in [19, p. 236, Problem 7], namely, 0 is an interior point of the set $\mathcal{T} := \{Ex(T) - z \,\big|\, x(\cdot) \text{ is the wealth process of an admissible portfolio } u(\cdot) \text{ with } x(0) = x_0\}$. In view of Theorem 6.1 and the assumption (6.11), we have $[x_0 e^{\int_0^T r(s)ds} - z, +\infty) \subset \mathcal{T}$. Hence 0 is an interior point of $\mathcal{T}$ because $x_0 e^{\int_0^T r(s)ds} - z < 0$.

Proof. It suffices to prove that under the feedback policy (6.28) the corresponding wealth trajectory $x^*(\cdot)$ satisfies

$$x^*(t) - \lambda^* e^{-\int_t^T r(s)ds} \le 0. \tag{6.34}$$

To this end, write $y(t) := x^*(t) - \lambda^* e^{-\int_t^T r(s)ds}$. Then it is immediate from (6.9) that $y(\cdot)$ follows

$$\begin{cases} dy(t) = [r(t)y(t) + B_+(t)y^+(t) + B_-(t)y^-(t)]dt + [D_+(t)y^+(t) + D_-(t)y^-(t)]'dW(t), \\ y(0) = x_0 - \lambda^* e^{-\int_0^T r(s)ds}, \end{cases} \tag{6.35}$$

where

$$\begin{aligned} &B_+(t) := B(t)\xi_+(t, P_+(t), \Lambda_+(t)),\ B_-(t) := B(t)\xi_-(t, P_-(t), \Lambda_-(t)), \\ &D_+(t) := \sigma(t)'\xi_+(t, P_+(t), \Lambda_+(t)),\ D_-(t) := \sigma(t)'\xi_-(t, P_-(t), \Lambda_-(t)). \end{aligned}$$

Note that $y(0) = x_0 - \lambda^* e^{-\int_0^T r(s)ds} \le 0$ by virtue of (6.29) and the fact that $z \ge x_0 e^{\int_0^T r(s)ds}$. Hence the proof of Lemma 5.1 yields that $y^+(t) = y(0)^+ \exp\{\cdots\} = 0$, which proves that $y(t) \le 0$. □

It is interesting to note that, as indicated by the preceding theorem and corollary, after all only one of the two Riccati equations, (6.18) is necessary in the final solution to the portfolio selection problem. This is essentially due to the fact that one is only interested in the "nonsatiation portion" (i.e., that which corresponds to $z \ge x_0 e^{\int_0^T r(s)ds}$) of the entire variance-minimizing boundary. Because of this, the form of the efficient portfolio (6.33) turns out to be strikingly similar to its shorting-permitted counterpart [28]. In particular, if shorting is allowed, then the definition (6.20) should be modified so that $\xi_-(t, P, \Lambda)$ is the minimum point of $H_-(t, v, P, \Lambda)$ over $v \in \mathbb{R}^m$, or $\xi_-(t, P, \Lambda) = (\sigma(t)\sigma(t)')^{-1}\left[B(t)' + \sigma(t)\frac{\Lambda}{P}\right]$. Furthermore, if all the market coefficients are deterministic, then $\Lambda(t) \equiv 0$ and the result of [28] is recovered.

Also, it follows from (6.34) that the wealth trajectory under the efficient portfolio is capped almost surely at any time by the present value of a *deterministic* constant λ^*.

Finally, we remark that in the portfolio selection application the control (portfolio) constraint is taken to be $\Gamma = \mathbb{R}^m_+$, for it has significant financial interpretation (no-shorting). We can easily cope with other forms of constrained portfolio thanks to the general results established in sections 4–5. An example is, in the case of two stocks, $\Gamma = \{(u_1, u_2) \in \mathbb{R}^2 | u_1 \le 2u_2\}$. Such a constraint can be interpreted as maintaining certain weights on different stocks. Note that if we deal with a general portfolio constraint, then explicit characterization of the feasibility such as (6.11) may no longer be possible. However, it is still possible to obtain a certain implicit feasibility condition based on the dual of the constraint cone, Γ. Details are left to the interested reader.

7. Concluding remarks. In this paper, we have solved explicitly a stochastic LQ control problem where the control is constrained by a cone and all the coefficients are random. The solution is heavily dependent on the two nonlinear BSDEs which are introduced in this paper for the first time. The study on these two equations is interesting from the point view of BSDE theory. A continuous-time mean-variance portfolio selection problem has been then solved as a special case of the general constrained stochastic LQ model.

A major assumption of the paper is that the state variable be one-dimensional. Although this assumption is valid in many interesting applications including the financial one, it is a very challenging open problem to obtain an explicit solution to a multidimensional problem and study the possible associated BSDEs.

REFERENCES

[1] M. Ait Rami and X. Y. Zhou, *Linear matrix inequalities, Riccati equations, and indefinite stochastic linear quadratic controls*, IEEE Trans. Automat. Control, 45 (2000), pp. 1131–1143.

[2] B. D. O. Anderson and J. B. Moore, *Optimal Control—Linear Quadratic Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[3] J. M. Bismut, *Linear quadratic optimal stochastic control with random coefficients*, SIAM J. Control Optim., 14 (1976), pp. 419–444.

[4] O. Bobrovnytska and M. Schweizer, *Mean–variance hedging and stochastic control: Beyond the Brownian setting*, IEEE Trans. Automat. Control, 49 (2004), pp. 396–408.

[5] S. L. Campbell, *On positive controllers and linear quadratic optimal control problems*, Internat. J. Control, 36 (1982), pp. 885–888.

[6] S. Chen, X. J. Li, and X. Y. Zhou, *Stochastic linear quadratic regulators with indefinite control weight costs*, SIAM J. Control Optim., 36 (1998), pp. 1685–1702.

[7] S. Chen and J. Yong, *Stochastic linear quadratic optimal control problems*, Appl. Math. Optim., 43 (2001), pp. 21–45.

[8] S. Chen and X. Y. Zhou, *Stochastic linear quadratic regulators with indefinite control weight costs.* II, SIAM J. Control Optim., 39 (2000), pp. 1065–1081.

[9] W. P. Heemels, S. van Eijndhoven, and A. A. Stoorvogel, *Linear quadratic regulator problem with positive controls*, Internat. J. Control, 70 (1998), pp. 551–578.

[10] Y. Hu and X. Y. Zhou, *Indefinite stochastic Riccati equations*, SIAM J. Control Optim., 42 (2003), pp. 123–137.

[11] R. E. Kalman, *Contribution to the theory of optimal control*, Bol. Soc. Mat. Mexicana, 5 (1960), pp. 102–119.

[12] M. Kobylanski, *Backward stochastic differential equations and partial differential equations with quadratic growth*, Ann. Probab., 28 (2000), pp. 558–602.

[13] M. Kohlmann and S. Tang, *Global adapted solution of one-dimensional backward stochastic Riccati equations, with application to the mean–variance hedging*, Stochastic Process. Appl., 97 (2002), pp. 255–288.

[14] M. Kohlmann and X. Y. Zhou, *Relationship between backward stochastic differential equations and stochastic controls: A linear-quadratic approach*, SIAM J. Control Optim., 38 (2000), pp. 1392–1407.

[15] J.-P. Lepeltier and J. San Martin, *Existence for BSDE with superlinear–quadratic coefficient*, Stochastics Stochastics. Rep., 63 (1998), pp. 227–240.

[16] X. Li, X. Y. Zhou, and A. E. B. Lim, *Dynamic mean–variance portfolio selection with no-shorting constraints*, SIAM J. Control Optim., 40 (2002), pp. 1540–1555.

[17] A. E. B. Lim, *Quadratic hedging and mean-variance portfolio selection with random parameters in an incomplete market*, Math. Oper. Res., 29 (2004), pp. 132–161.

[18] A. E. B. Lim and X. Y. Zhou, *Mean-variance portfolio selection with random parameters in a complete market*, Math. Oper. Res., 27 (2002), pp. 101–120.

[19] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley and Sons, New York, 1969.

[20] J. Ma and J. Yong, *Forward-Backward Stochastic Differential Equations and Their Applications*, Lecture Notes in Mathematics 1702, Springer-Verlag, Berlin, 1999.

[21] M. Mania, *A general problem of an optimal equivalent change of measure and contingent claim pricing in an incomplete market*, Stochastic Process. Appl., 90 (2000), pp. 19–42.

[22] M. Mania and R. Tevzadze, *Backward stochastic PDE and imperfect hedging*, Int. J. Theor. Appl. Finance, 6 (2003), pp. 663–692.

[23] M. Pachter, *The linear–quadratic optimal control problem with positive controllers*, Internat. J. Control, 32 (1980), pp. 589–608.

[24] D. Revuz and M. Yor, *Continuous Martingales and Brownian Motion*, 3rd ed., Springer-Verlag, Berlin, 1999.

[25] W. M. Wonham, *On a matrix Riccati equation of stochastic control*, SIAM J. Control, 6 (1968), pp. 681–697.

[26] D. Yao, S. Zhang, and X. Zhou, *Stochastic linear-quadratic control via semidefinite program-*

ming, SIAM J. Control Optim., 40 (2001), pp. 801–823.
[27] J. Yong and X. Y. Zhou, *Stochastic Controls. Hamiltonian Systems and HJB Equations*, Springer-Verlag, New York, 1999.
[28] X. Y. Zhou and D. Li, *Continuous-time mean-variance portfolio selection: A stochastic LQ framework*, Appl. Math. Optim., 42 (2000), pp. 19–33.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 2, pp. 467–494

A PRIMAL-DUAL ACTIVE SET STRATEGY FOR OPTIMAL BOUNDARY CONTROL OF A NONLINEAR REACTION-DIFFUSION SYSTEM*

R. GRIESSE† AND S. VOLKWEIN†

Abstract. This paper is concerned with optimal boundary control of an instationary reaction-diffusion system in three spatial dimensions. This problem involves a coupled nonlinear system of parabolic differential equations with bilateral as well as integral control constraints. We include the integral constraint in the cost by a penalty term whereas the bilateral control constraints are handled explicitly. First- and second-order conditions for the optimization problem are analyzed. A primal-dual active set strategy is utilized to compute optimal solutions numerically. The algorithm is compared to a semismooth Newton method.

Key words. reaction-diffusion equations, optimal boundary control, bilateral and integral control constraints, primal-dual active set strategy, semismooth Newton methods

AMS subject classifications. 35Kxx, 49Lxx, 65Kxx

DOI. 10.1137/S0363012903438696

1. Introduction.

The subject matter of this paper is an optimal control problem for a coupled system of semilinear parabolic reaction-diffusion equations. The equations model a chemical or biological process where the species involved are subject to diffusion and reaction among each other. As an example, we consider the reaction $A+B \to C$ which obeys the law of mass action. To simplify the discussion, we assume that the backward reaction $C \to A+B$ is negligible and that the forward reaction proceeds with a constant (e.g., not temperature-dependent) rate. This leads to a coupled semilinear parabolic system for the respective concentrations; see (2.3). We consider the state equations in three spatial dimensions and prove existence and uniqueness of solutions in $W(0,T)$. This is a nontrivial result, and from the analysis in [9] it follows that higher order polynomial reaction terms do not admit solutions in $W(0,T)$, in three space dimensions. Simplifications to the two- or even one-dimensional situation are of course possible in a straightforward way.

The control function acts through the Neumann boundary values for one of the reaction components on some subset of the two-dimensional boundary manifold. It is natural to impose bilateral pointwise bounds on the control function: on one hand, the substance can never be extracted through the boundary, i.e., the lower control bound should be nonnegative. On the other hand, only a limited amount may be added *at any given time*. In addition, we impose a constraint on the *total amount* of control action. This scalar integral constraint (see (2.11)) is very much in contrast with the usual pointwise bounds.

Optimality conditions for optimal control problems governed by semilinear parabolic equations together with mixed inequality constraints were considered theoret-

*Received by the editors December 17, 2003; accepted for publication (in revised form) July 1, 2004; published electronically August 31, 2005. This work was supported in part by the Fonds zur Förderung der wissenschaftlichen Forschung under the Special Research Center F003 "Optimization and Control."

http://www.siam.org/journals/sicon/44-2/43869.html

†Karl-Franzens-Universität Graz, Institut für Mathematik, Heinrichstrasse 36, A-8010 Graz, Austria and Johann Radon Institute for Computational and Applied Mathematics (RICAM), Altenbergerstrasse 69, A–4040 Linz, Austria (roland.griesse@oeaw.ac.at, stefan.volkwein@uni-graz.at).

ically in the literature. We refer, for instance, the reader to [5, 17, 30]. Note that the theory presented in [27] cannot be applied directly to system (2.3) since the assumptions for the semilinear part are not satisfied. Less attention is paid to the numerical realization of the infinite-dimensional optimality systems. Let us mention, for example, [22].

The integral constraint is included into the cost functional by a penalty term, whereas the bilateral control constraints are treated explicitly by a primal-dual active set strategy for nonlinear problems. The primal-dual active set method has proved to be an efficient numerical tool in the context of diverse applications; see, for instance, [1, 2, 12]. So far it was mainly investigated for linear-quadratic problems, which arise, for example, as a subproblem within SQP or Newton methods (compare, e.g., [13, 14, 21, 34]). Ito and Kunisch studied the primal-dual active set algorithm for nonlinear problems and bilateral control constraints in [19]. Utilizing the close relationship between the primal-dual active set strategy and semismooth Newton methods, local superlinear convergence was shown as well. Let us mention that the primal-dual active set strategy for nonlinear problems was already applied numerically, combined with SQP and nonlinear conjugate gradient methods in [8, 20] and [37], respectively. Semismooth Newton methods for general purpose nonlinear finite-dimensional optimal control problems are well studied; see, for instance, [24] and [7, section 7.5]. Much less is known about such methods in infinite dimensions, and specifically in the context of optimal control problems. Let us refer here, e.g., to [15, 16, 35]. We will compare the primal-dual active set strategy with a semismooth Newton method in the numerical test examples. Both utilize an inexact Newton method as an inner iteration. The penalty term leads to a nonquadratic objective, which gives rise to an interesting update step for the control constraint multiplier.

This paper is organized in the following manner. In section 2, the state equations are analyzed and the optimal control problem is investigated. The integral control constraint is treated using a penalization approach. Section 3 is devoted to the optimality conditions for the penalized optimization problem. The primal-dual active set algorithm and its relationship to a semismooth Newton method are discussed in section 4. Numerical examples are presented in section 5 and we draw some conclusions in the last section.

2. The problem formulation. The goal of this section is to introduce the infinite-dimensional optimal control problem. The cost functional is of tracking type, the equality constraints are given by a coupled nonlinear parabolic system, and the inequality constraints are bilateral control constraints as well as an integral constraint for the control. We study the state equations, propose the optimal control problem, and prove existence of optimal controls. Finally we introduce the optimization problem with the penalized objective.

2.1. The state equations. Let Ω denote an open and bounded subset of $\mathbb{R}^3$ with a Lipschitz-continuous boundary $\Gamma = \partial\Omega$ such that Γ is decomposed into two parts $\Gamma = \Gamma_n \cup \Gamma_c$ with $\Gamma_n \cap \Gamma_c = \emptyset$. For terminal time $T > 0$ let $Q = (0,T) \times \Omega$, let $\Sigma = (0,T) \times \Gamma$, and let $\Sigma_c = (0,T) \times \Gamma_c$.

By $L^2(0,T;H^1(\Omega))$ we denote the space of all measurable functions $\varphi : [0,T] \to H^1(\Omega)$, which are square integrable, i.e.,

$$\int_0^T \|\varphi(t)\|^2_{H^1(\Omega)} \,\mathrm{d}t < \infty,$$

where $\varphi(t)$ stands for the function $\varphi(t,\cdot)$ considered as a function in Ω only. The space $W(0,T)$ is defined by

$$W(0,T) = \{\varphi \in L^2(0,T;H^1(\Omega)) : \varphi_t \in L^2(0,T;H^1(\Omega)')\}. \tag{2.1}$$

Here $H^1(\Omega)'$ denotes the dual space of $H^1(\Omega)$. Recall that $W(0,T)$ is a Hilbert space endowed with the common inner product and the induced norm; see, e.g., [6, p. 286]. Since $W(0,T)$ is continuously embedded into $C([0,T];L^2(\Omega))$, the space of all continuous functions from $[0,T]$ into $L^2(\Omega)$, there exists a constant $C_W > 0$ satisfying

$$\|\varphi\|_{C([0,T];L^2(\Omega))} \le C_W \|\varphi\|_{W(0,T)} \quad \text{for all } \varphi \in W(0,T); \tag{2.2}$$

see [6, p. 287].

Suppose that d_1, d_2, d_3 and k_1, k_2, k_3 are positive constants. Moreover, let $\alpha \in L^\infty(0,T;L^2(\Gamma_c))$ denote a shape function with $\alpha \ge 0$ on Σ_c almost everywhere (a.e.). We consider the following system of semilinear parabolic equations, where c_i denotes the concentration of the ith substance:

$$(c_1)_t(t,x) = d_1\Delta c_1(t,x) - k_1 c_1(t,x)c_2(t,x) \qquad \text{for all } (t,x) \in Q, \tag{2.3a}$$

$$(c_2)_t(t,x) = d_2\Delta c_2(t,x) - k_2 c_1(t,x)c_2(t,x) \qquad \text{for all } (t,x) \in Q, \tag{2.3b}$$

$$(c_3)_t(t,x) = d_3\Delta c_3(t,x) + k_3 c_1(t,x)c_2(t,x) \qquad \text{for all } (t,x) \in Q \tag{2.3c}$$

together with the Neumann boundary conditions

$$d_1\frac{\partial c_1}{\partial n}(t,x) = 0 \qquad \text{for all } (t,x) \in \Sigma, \tag{2.3d}$$

$$d_2\frac{\partial c_2}{\partial n}(t,x) = u(t)\alpha(t,x) \qquad \text{for all } (t,x) \in \Sigma_c, \tag{2.3e}$$

$$d_2\frac{\partial c_2}{\partial n}(t,x) = 0 \qquad \text{for all } (t,x) \in \Sigma_n = \Sigma \setminus \Sigma_c, \tag{2.3f}$$

$$d_3\frac{\partial c_3}{\partial n}(t,x) = 0 \qquad \text{for all } (t,x) \in \Sigma \tag{2.3g}$$

and the initial conditions

$$c_1(0,x) = c_{10}(x) \quad \text{for all } x \in \Omega, \tag{2.3h}$$

$$c_2(0,x) = c_{20}(x) \quad \text{for all } x \in \Omega, \tag{2.3i}$$

$$c_3(0,x) = c_{30}(x) \quad \text{for all } x \in \Omega, \tag{2.3j}$$

where $c_{i0} \in L^2(\Omega)$ for $i = 1,2,3$.

The control $u \in L^2(0,T)$ enters the right-hand side of (2.3e) in the inhomogeneous Neumann condition. For instance, the function α models a spray nozzle moving over the control part Γ_c, and $u(t)$ denotes the intensity of the spray.

Remark 2.1. The parabolic problem for c_3, i.e., (2.3c) together with the Neumann boundary condition (2.3g) and initial condition (2.3j) can be solved independently of the problem for (c_1,c_2). Therefore, we will focus on the computation of c_1 and c_2 and, in particular, we are interested in weak solutions for c_1 and c_2. □

DEFINITION 2.2. *The two functions c_1 and c_2 in $W(0,T)$ are called* weak solutions *to systems* (2.3a), (2.3b), (2.3d)–(2.3f), (2.3h), *and* (2.3i) *provided the initial conditions*

$$c_1(0) = c_{10} \quad \textit{and} \quad c_2(0) = c_{20} \quad \textit{in } L^2(\Omega) \tag{2.4a}$$

hold and

$$\langle (c_1)_t(t), \varphi \rangle_{H^1(\Omega)', H^1(\Omega)} + \int_\Omega d_1 \nabla c_1(t) \cdot \nabla \varphi + k_1 c_1(t) c_2(t) \varphi \, \mathrm{d}x = 0, \tag{2.4b}$$

$$\begin{aligned} &\langle (c_2)_t(t), \varphi \rangle_{H^1(\Omega)', H^1(\Omega)} + \int_\Omega d_2 \nabla c_2(t) \cdot \nabla \varphi + k_2 c_1(t) c_2(t) \varphi \, \mathrm{d}x \\ &\quad = u(t) \int_{\Gamma_c} \alpha(t) \varphi \, \mathrm{d}x \end{aligned} \tag{2.4c}$$

for all $\varphi \in H^1(\Omega)$ *and almost all* $t \in [0,T]$. *In* (2.4b) *and* (2.4c), $\langle \cdot, \cdot \rangle_{H^1(\Omega)', H^1(\Omega)}$ *denotes the duality pairing between* $H^1(\Omega)$ *and its dual* $H^1(\Omega)'$.

The following theorem ensures that (2.4) possesses a unique solution. For a proof, which is based on Leray–Schauder's fixed point theorem and variational techniques, we refer the reader to [9, Theorem 2.3].

THEOREM 2.3. *For every control* $u \in L^2(0,T)$, *there exists a unique pair* $(c_1, c_2) \in W(0,T) \times W(0,T)$ *satisfying* (2.4). *Moreover, the estimate*

$$\|c_1\|_{W(0,T)} + \|c_2\|_{W(0,T)} \le C \left(1 + \|c_{10}\|_{L^2(\Omega)} + \|c_{20}\|_{L^2(\Omega)} + \|u\|_{L^2(0,T)}\right) \tag{2.5}$$

holds for a constant $C > 0$.

Theorem 2.3 also implies the unique solvability of the partial differential equation for the reaction product (2.3c), (2.3g), and (2.3j). This is formulated in the following corollary, which is proved in [9, Corollary 2.4].

COROLLARY 2.4. *Let* $c_{10}, c_{20} \in L^2(\Omega)$ *and* $u \in L^2(0,T)$ *be given and let* $(c_1, c_2) \in W(0,T) \times W(0,T)$ *denote the solution pair to* (2.4). *Then there exists a unique* $c_3 \in W(0,T)$ *satisfying*

$$c_3(0) = c_{30} \quad \text{in } L^2(\Omega) \tag{2.6a}$$

and

$$\langle (c_3)_t(t), \varphi \rangle_{H^1(\Omega)', H^1(\Omega)} + \int_\Omega d_3 \nabla c_3(t) \cdot \nabla \varphi \, \mathrm{d}x = \int_\Omega k_3 c_1(t) c_2(t) \varphi \, \mathrm{d}x \tag{2.6b}$$

for all $\varphi \in H^1(\Omega)$ *and almost all* $t \in [0,T]$.

To write the state equations as a nonlinear operator equation, we introduce the two Hilbert product spaces

$$\begin{aligned} X &= W(0,T) \times W(0,T) \times L^2(0,T), \\ Y &= L^2(0,T; H^1(\Omega)) \times L^2(0,T; H^1(\Omega)) \times L^2(\Omega) \times L^2(\Omega) \end{aligned}$$

endowed with their product topology and identify

$$Y' \equiv L^2(0,T; H^1(\Omega)') \times L^2(0,T; H^1(\Omega)') \times L^2(\Omega) \times L^2(\Omega).$$

Then we introduce the mapping $e : X \to Y'$ by

$$e(x) = \begin{pmatrix} e_1(x) \\ e_2(x) \\ c_1(0) - c_{10} \\ c_2(0) - c_{20} \end{pmatrix} \quad \text{for } x = (c_1, c_2, u) \in X,$$

where the operators $e_1, e_2 \colon X \times L^2(0,T;H^1(\Omega)) \to L^2(0,T;H^1(\Omega)')$ involve the variational formulation of the partial differential equations for c_1 and c_2, respectively, i.e.,

$$\begin{aligned}&\langle e_1(x), \varphi\rangle_{L^2(0,T;H^1(\Omega)'),L^2(0,T;H^1(\Omega))}\\ &\quad= \int_0^T \Big(\langle (c_1)_t(t), \varphi(t)\rangle_{H^1(\Omega)',H^1(\Omega)} + \int_\Omega d_1 \nabla c_1 \cdot \nabla \varphi + k_1 c_1 c_2 \varphi \,\mathrm{d}x \Big) \mathrm{d}t\end{aligned}$$

and

$$\begin{aligned}\langle e_2(x), \varphi\rangle_{L^2(0,T;H^1(\Omega)'),L^2(0,T;H^1(\Omega))} &= \int_0^T \Big(\langle (c_2)_t(t), \varphi(t)\rangle_{H^1(\Omega)',H^1(\Omega)} \,\mathrm{d}t\\ &\quad + \int_\Omega d_2 \nabla c_2 \cdot \nabla \varphi + k_2 c_1 c_2 \varphi \,\mathrm{d}x - u \int_{\Gamma_c} \alpha \varphi \,\mathrm{d}x \Big) \mathrm{d}t\end{aligned}$$

for $\varphi \in L^2(0,T;H^1(\Omega))$. Now, (2.4) is equivalent to the operator equation $e(x) = 0$ in Y' for $x = (c_1, c_2, u) \in X$.

2.2. The optimal control problem. Our goal is to drive the reaction-diffusion system from the given initial state near a desired terminal state. Hence, we introduce the cost functional

$$J(c_1, c_2, u) = \frac{1}{2} \int_\Omega \beta_1 \,|c_1(T) - c_{1T}|^2 + \beta_2 \,|c_2(T) - c_{2T}|^2 \,\mathrm{d}x + \frac{\gamma}{2} \int_0^T |u - u_d|^2 \,\mathrm{d}t,$$

where $\beta_1, \beta_2 \geq 0$, $\beta_1 + \beta_2, \gamma > 0$, $c_{1T}, c_{2T} \in L^2(\Omega)$ are given desired terminal states and $u_d \in L^2(0,T)$ denotes some nominal (or expected) control.

The closed and bounded convex set of admissible control parameters involves an integral constraint as well as bilateral control constraints:

$$\bar{U}_{\mathsf{ad}} = \left\{ u \in L^2(0,T) : \int_0^T u(t) \,\mathrm{d}t \leq u_c \text{ and } u_a \leq u \leq u_b \text{ in } [0,T] \right\} \subset L^\infty(0,T),$$

where u_a and u_b are given functions in $L^\infty(0,T)$ satisfying $u_a \leq u_b$ in $[0,T]$ a.e., and u_c is a positive constant.

Furthermore, let us define the closed convex set

$$\bar{K}_{\mathsf{ad}} = W(0,T) \times W(0,T) \times \bar{U}_{\mathsf{ad}}.$$

The infinite-dimensional optimal control problem can be expressed as

$$\text{(P)} \qquad \min J(x) \quad \text{such that (s.t.)} \quad x \in \bar{K}_{\mathsf{ad}} \text{ and } e(x) = 0.$$

The following theorem guarantees that (P) has a solution.

THEOREM 2.5. *Problem* (P) *possesses at least one optimal control.*

Proof. The claim follows by standard arguments: let $\{x^n\}_{n=1}^\infty$, $x^n = (c_1^n, c_2^n, u^n)$, be a minimizing sequence in $\bar{K}_{\mathsf{ad}}$ for the nonnegative cost J. Since J is radially unbounded, it follows from Theorem 2.3 that this sequence is bounded in X. Therefore,

there exists an element $x^* = (c_1^*, c_2^*, u^*) \in X$ such that

$$c_1^n \rightharpoonup c_1^* \quad \text{in } W(0,T) \text{ as } n \to \infty, \tag{2.7}$$

$$c_2^n \rightharpoonup c_2^* \quad \text{in } W(0,T) \text{ as } n \to \infty, \tag{2.8}$$

$$u^n \rightharpoonup u^* \quad \text{in } L^2(0,T) \text{ as } n \to \infty. \tag{2.9}$$

By assumption, $\alpha \in L^\infty(0,T;L^2(\Gamma_c))$ holds. Recall that there exists a constant $K_1 > 0$ such that

$$\|\psi\|_{L^2(\Gamma_c)} \le K_1 \|\psi\|_{H^1(\Omega)} \quad \text{for all } \psi \in H^1(\Omega);$$

see, e.g., [6, p. 258]. Thus, for $\varphi \in L^2(0,T;H^1(\Omega))$, the mapping $t \mapsto \int_{\Gamma_c} \alpha(t)\varphi(t)\,\mathrm{d}x$ belongs to $L^2(0,T)$ and

$$\lim_{n\to\infty} \int_0^T \big(u^n(t) - u^*(t)\big)\bigg(\int_{\Gamma_c} \alpha(t)\varphi(t)\,\mathrm{d}x\bigg)\mathrm{d}t = 0 \quad \text{for all } \varphi \in L^2(0,T;H^1(\Omega)).$$

From (2.7) and (2.8), we find for $i = 1, 2$,

$$\lim_{n\to\infty} \int_0^T \big\langle (c_i^n - c_i^*)_t(t), \varphi(t)\big\rangle_{H^1(\Omega)',H^1(\Omega)}\,\mathrm{d}t = 0 \quad \text{for all } \varphi \in L^2(0,T;H^1(\Omega))$$

and

$$\lim_{n\to\infty} \int_0^T \int_\Omega d_i \nabla\big(c_i^n - c_i^*\big)(t) \cdot \nabla\varphi(t)\,\mathrm{d}x\,\mathrm{d}t = 0 \quad \text{for all } \varphi \in L^2(0,T;H^1(\Omega)).$$

Next we consider the nonlinear terms. Using Hölder's inequality we infer that for $\varphi \in L^2(0,T;H^1(\Omega))$,

$$\begin{aligned}
\int_0^T \int_\Omega \big(c_1^n c_2^n - c_1^* c_2^*\big)\varphi\,\mathrm{d}x\,\mathrm{d}t &= \int_0^T \int_\Omega \big[\big(c_1^n - c_1^*\big)c_2^n + c_1^*\big(c_2^n - c_2^*\big)\big]\,\varphi\,\mathrm{d}x\,\mathrm{d}t \\
&\le \int_0^T \|c_1^n(t) - c_1^*(t)\|_{L^3(\Omega)} \|c_2^n(t)\|_{L^2(\Omega)} \|\varphi(t)\|_{L^6(\Omega)}\,\mathrm{d}t \\
&\quad + \int_0^T \|c_1^*(t)\|_{L^2(\Omega)} \|c_2^n(t) - c_2^*(t)\|_{L^3(\Omega)} \|\varphi(t)\|_{L^6(\Omega)}\,\mathrm{d}t \\
&\le \|c_1^n - c_1^*\|_{L^2(0,T;L^3(\Omega))} \|c_2^n\|_{C([0,T];L^2(\Omega))} \|\varphi\|_{L^2(0,T;L^6(\Omega))} \\
&\quad + \|c_1^*\|_{C([0,T];L^2(\Omega))} \|c_2^n - c_2^*\|_{L^2(0,T;L^3(\Omega))} \|\varphi\|_{L^2(0,T;L^6(\Omega))}.
\end{aligned} \tag{2.10}$$

Since $W(0,T)$ is continuously embedded into $C([0,T];L^2(\Omega))$ and compactly into $L^2(0,T;L^3(\Omega))$ (see, for instance, [33, p. 271]), the sequence $\|c_i^n\|_{C([0,T];L^2(\Omega))}$ is bounded and $\lim_{n\to\infty} \|c_i^n - c_i^*\|_{L^2(0,T;L^3(\Omega))} = 0$ for $i = 1, 2$. Thus, (2.10) yields

$$\lim_{n\to\infty} \int_0^T \int_\Omega \big(c_1^n c_2^n - c_1^* c_2^*\big)\varphi\,\mathrm{d}x\,\mathrm{d}t = 0 \quad \text{for all } \varphi \in L^2(0,T;H^1(\Omega)).$$

Using

$$\int_\Omega \big(c_i^n(0) - c_i^*(0)\big)\psi\,\mathrm{d}x = 0 \quad \text{for all } \psi \in L^2(\Omega) \text{ and } i = 1, 2,$$

we have $e(x^*) = 0$ in Y'. Since $\bar{U}_{\mathsf{ad}}$ is bounded, closed, and convex, $\bar{U}_{\mathsf{ad}}$ is weakly closed. This implies that $\bar{K}_{\mathsf{ad}}$ is also weakly closed. As J is weakly lower semicontinuous, the claim follows. ∎

2.3. The penalized optimization problem. In (P), we have two different types of inequality constraints for the control variable: a scalar integral constraint and an infinite-dimensional box-constraint. To handle the integral constraint

$$\int_0^T u \,\mathrm{d}t \le u_c \tag{2.11}$$

numerically, we introduce the penalized cost functional

$$J_\varepsilon(x) = J(x) + \frac{1}{\varepsilon} I(u) \quad \text{for all } x = (c_1, c_2, u) \in X \text{ and } \varepsilon > 0,$$

where the mapping $I : L^2(0,T) \to \mathbb{R}$ defined as

$$I(u) = g\bigg(\int_0^T u \,\mathrm{d}t - u_c \bigg),$$

where we choose $g = [\,\cdot\,]_+^3$ in $\mathbb{R}$ and $[s]_+ = \max\{0, s\}$, $s \in \mathbb{R}$, denotes the positive part function.

Remark 2.6. Of course, other choices for g are possible. To analyze second-order conditions later on, we make use of the fact that $g \in C^2$. Moreover, the property $g''(s) \ge 0$ for all $s \in \mathbb{R}$ ensures coercivity of the cost functional J_ε. □

The goal of this section is to analyze the optimal control problem with the penalized cost. The bilateral control constraints are treated explicitly and enforced numerically by primal-dual active set strategies; see sections 4 and 5. Lemma 2.7 is proved in [9, Lemmas 2.7 and 2.8].

LEMMA 2.7. *The function g is twice differentiable and its second derivative is Lipschitz-continuous. Moreover, the mapping $I : L^2(0,T) \to \mathbb{R}$ is weakly continuous. Moreover, I is twice continuously Fréchet-differentiable and its second Fréchet-derivative is Lipschitz-continuous in $L^2(0,T)$. In particular, the Fréchet-derivatives of I at $u \in L^2(0,T)$ are given by*

$$\nabla I(u)\delta u = g'\bigg(\int_0^T u \,\mathrm{d}t - u_c \bigg) \int_0^T \delta u \,\mathrm{d}t$$

and

$$\nabla^2 I(u)(\delta u, \widetilde{\delta u}) = g''\bigg(\int_0^T u \,\mathrm{d}t - u_c \bigg) \int_0^T \delta u \,\mathrm{d}t \int_0^T \widetilde{\delta u} \,\mathrm{d}t. \tag{2.12}$$

As the integral constraint is already included in the cost J_ε by a penalty term, we replace $\bar{U}_{\mathsf{ad}}$ by

$$U_{\mathsf{ad}} = \{u \in L^2(0,T) : u_a \le u \le u_b \text{ in } [0,T]\} \subset L^\infty(0,T)$$

and set $K_{\mathsf{ad}} = W(0,T) \times W(0,T) \times U_{\mathsf{ad}}$. Now the penalized optimal control problem has the form

$$\min J_\varepsilon(x) \quad \text{s.t.} \quad x \in K_{\mathsf{ad}} \text{ and } e(x) = 0. \tag{P_ε}$$

Utilizing Lemma 2.7, the next result can be proved analogously to Theorem 2.5.

THEOREM 2.8. *There exists at least one optimal solution to* (P_ε).

Proof. Because of Lemma 2.7 the mapping I is weakly continuous, so that the penalized cost functional J_ε is weakly lower semicontinuous on X. Thus, the proof is analogous to that of Theorem 2.5. ∎

In the next proposition we turn to the question whether solutions of (P_ε) converges to a solution to (P) if ε tends to zero.

PROPOSITION 2.9. *Assume that $\{\varepsilon_n\}_{n=0}^\infty \subset \mathbb{R}$ is a sequence converging to zero from above. Let $\{x_n\}_{n=0}^\infty$ denote a sequence of optimal solutions to $(\mathrm{P}_{\varepsilon_n})$. Then there exists at least one weak accumulation point $x^* \in X$ for $\{x_n\}_{n=0}^\infty$. That is, $x_{n'} \rightharpoonup x^*$ in X as $n' \to \infty$ for some subsequence $\{x_{n'}\}_{n'=0}^\infty$. In addition, every weak accumulation point x^* solves* (P).

Proof. Since $x_n = (c_{1n}, c_{2n}, u_n) \in X$ solves $(\mathrm{P}_{\varepsilon_n})$, the sequence $\{u_n\}_{n=0}^\infty$ belongs to U_{ad}. Thus, $\|u_n\|_{L^2(0,T)}$ is bounded by a constant which does not depend on n. Because of the a priori bound (2.5), the family of pairs $\{(c_{1n}, c_{2n})\}_{n=0}^\infty$ is also bounded in $W(0,T) \times W(0,T)$. Since X is reflexive, there exists a subsequence in K_{ad}, denoted by $\{x_{n'}\}_{n'=0}^\infty$, and an element $x^* = (c_1^*, c_2^*, u^*) \in X$ such that

$$x_{n'} \rightharpoonup x^* \quad \text{in } X \text{ as } n' \to \infty. \tag{2.13}$$

Reasoning as in the proof of Theorem 2.5, we find that $x^* \in K_{\mathsf{ad}}$ and $e(x^*) = 0$ in Y'. Since $x_{n'}$ solves $(\mathrm{P}_{\varepsilon_{n'}})$, we have

$$J_{\varepsilon_{n'}}(x_{n'}) = J(x_{n'}) + \frac{1}{\varepsilon_{n'}} I(u_{n'}) \le J(x) + \frac{1}{\varepsilon_{n'}} I(u) = J(x) \tag{2.14}$$

for all $x = (c_1, c_2, u) \in \bar{K}_{\mathsf{ad}}$. Hence,

$$0 \le I(u_{n'}) \le \varepsilon_{n'} \big(J(x) - J(x_{n'})\big) \quad \text{for all } x \in \bar{K}_{\mathsf{ad}}. \tag{2.15}$$

Choosing $x \in \bar{K}_{\mathsf{ad}}$ arbitrarily and using the boundedness of $x_{n'}$ and thus of $J(x_{n'})$, we obtain $I(u^*) = 0$ from passing to the limit in (2.15), applying Lemma 2.7. Thus, the integral constraint (2.11) is satisfied for u^*; hence $x^* \in K_{\mathsf{ad}}$ holds. Finally, it follows from weak lower semicontinuity of J and from (2.14) that

$$J(x^*) \le \liminf_{n' \to \infty} J(x_{n'}) \le \liminf_{n' \to \infty} J(x_{n'}) + \frac{1}{\varepsilon_{n'}} I(u_{n'}) \le J(x) \quad \text{for all } x \in \bar{K}_{\mathsf{ad}}$$

so that x^* solves (P). ∎

3. Optimality conditions. In section 2 we have introduced the optimal control problem (P_ε) and proved existence of optimal controls. This section is devoted to the analysis of necessary and sufficient optimality conditions for (P_ε).

3.1. Smoothness properties for *J* and *e*. We start by investigating differentiability properties of the cost functional as well as of the mapping describing the equality constraints. The proof of the following result is based on Lemma 2.7. For more details we refer the reader to [9, Proposition 3.1].

PROPOSITION 3.1. *The penalized cost functional J_ε and the mapping e are twice continuously Fréchet-differentiable and their second Fréchet-derivatives are Lipschitz-continuous on X.*

The linear operator $\nabla_{(c_1,c_2)} e(x) : W(0,T) \times W(0,T) \to Y'$ has the following property.

PROPOSITION 3.2. *For all $x \in X$, the linearization $\nabla_{(c_1,c_2)} e(x)$ is bijective. Moreover, for all $\delta x = (\delta c_1, \delta c_2, \delta u) \in N(\nabla e(x))$, we have*

$$\|\delta c_1\|^2_{W(0,T)} + \|\delta c_2\|^2_{W(0,T)} \le C_N \|\delta u\|^2_{L^2(0,T)} \tag{3.1}$$

for all $C_N > 0$, where $N(\nabla e(x))$ denotes the null space of the operator $\nabla e(x)$.

For a proof we refer the reader to [9, Proposition 3.2].

Remark 3.3. Proposition 3.2 implies the standard constraint qualification condition for x^* (see [28], for example), which in our case has the form

$$\begin{aligned}\begin{pmatrix} 0 \\ 0 \end{pmatrix} &\in \operatorname{int}\left\{ \begin{pmatrix} X \\ \nabla e(x^*)X \end{pmatrix} - \begin{pmatrix} K_{\mathsf{ad}} - x^* \\ Y' - e(x^*) \end{pmatrix} \right\} \\ &= \operatorname{int}\{X - (K_{\mathsf{ad}} - x^*)\} \times \operatorname{int}\{\nabla e(x^*)X\},\end{aligned} \tag{3.2}$$

where int S denotes the interior of a set S and $e'(x^*)$ is the Fréchet-derivative of the operator e at x^*. It follows from (3.2) that the set of Lagrange multipliers is nonempty and bounded; see [25], for instance. □

For every $\varepsilon > 0$, the Lagrange functional $L_\varepsilon : X \times Y \to \mathbb{R}$ associated with (P_ε) is given by

$$L_\varepsilon(x,p) = J_\varepsilon(x) + \langle e(x), p\rangle_{Y',Y}$$

for $x = (c_1, c_2, u) \in X$ and $p = (\lambda_1, \lambda_2, \mu_1, \mu_2) \in Y$. From Proposition 3.1 we conclude that L_ε is twice continuously Fréchet-differentiable and its second Fréchet-derivative is Lipschitz-continuous.

3.2. First-order necessary optimality conditions. This subsection is devoted to the presentation of the first-order necessary optimality conditions for (P_ε). Problem (P_ε) is a nonconvex programming problem so that different local minima might occur. Numerical methods will produce a local minimum close to their starting point. Therefore, we do not direct our investigation to global solutions of (P_ε). We will assume that a fixed reference solution $\hat{x}^\varepsilon = (\hat{c}_1^\varepsilon, \hat{c}_2^\varepsilon, \hat{u}^\varepsilon) \in K_{\mathsf{ad}}$ is given satisfying first- and second-order optimality conditions. Let us define the active sets at $\hat{x}^\varepsilon$ by $\hat{A}^\varepsilon = \hat{A}^\varepsilon_- \cup \hat{A}^\varepsilon_+$, where

$$\hat{A}^\varepsilon_- = \{t \in [0,T] : \hat{u}^\varepsilon(t) = u_a(t) \text{ a.e.}\} \quad \text{and} \quad \hat{A}^\varepsilon_+ = \{t \in [0,T] : \hat{u}^\varepsilon(t) = u_b(t) \text{ a.e.}\}.$$

The corresponding inactive set at x^ε is given by $\hat{I}^\varepsilon = [0,T] \setminus \hat{A}^\varepsilon$. First-order necessary optimality conditions are presented in the next theorem.

THEOREM 3.4. *Let $\hat{x}^\varepsilon = (\hat{c}_1^\varepsilon, \hat{c}_2^\varepsilon, \hat{u}^\varepsilon) \in K_{\mathsf{ad}}$ be a local solution to (P_ε). Then there exists a unique Lagrange multiplier $\hat{p}^\varepsilon = (\hat{\lambda}_1^\varepsilon, \hat{\lambda}_2^\varepsilon, \hat{\mu}_1^\varepsilon, \hat{\mu}_2^\varepsilon) \in W(0,T) \times W(0,T) \times L^2(\Omega) \times L^2(\Omega) \subsetneq Y$ such that the pair $(\hat{\lambda}_1^\varepsilon, \hat{\lambda}_2^\varepsilon)$ is weak solutions to the adjoint (or dual) equations*

$$-(\hat{\lambda}_1^\varepsilon)_t - d_1\Delta\hat{\lambda}_1^\varepsilon = -k_1\hat{c}_2^\varepsilon\hat{\lambda}_1^\varepsilon - k_2\hat{c}_2^\varepsilon\hat{\lambda}_2^\varepsilon \qquad \text{in } Q, \tag{3.3a}$$

$$-(\hat{\lambda}_2^\varepsilon)_t - d_2\Delta\hat{\lambda}_2^\varepsilon = -k_1\hat{c}_1^\varepsilon\hat{\lambda}_1^\varepsilon - k_2\hat{c}_1^\varepsilon\hat{\lambda}_2^\varepsilon \qquad \text{in } Q, \tag{3.3b}$$

$$d_1\frac{\partial\hat{\lambda}_1^\varepsilon}{\partial n} = 0 \qquad \text{on } \Sigma, \tag{3.3c}$$

$$d_2\frac{\partial\hat{\lambda}_2^\varepsilon}{\partial n} = 0 \qquad \text{on } \Sigma, \tag{3.3d}$$

$$\hat{\lambda}_1^\varepsilon(T) = -\beta_1(\hat{c}_1^\varepsilon(T) - c_{1T}) \qquad \text{in } \Omega, \tag{3.3e}$$

$$\hat{\lambda}_2^\varepsilon(T) = -\beta_2(\hat{c}_2^\varepsilon(T) - c_{2T}) \qquad \text{in } \Omega, \tag{3.3f}$$

and for $i = 1, 2$, we have

$$\hat{\mu}_i^\varepsilon = \hat{\lambda}_i^\varepsilon(0) \quad \text{in } \Omega. \tag{3.4}$$

Moreover, there is a Lagrange multiplier $\hat{\xi}^\varepsilon \in L^2(0,T)$ associated with the bilateral inequality constraint satisfying

$$\hat{\xi}^\varepsilon|_{\hat{A}_-^\varepsilon} \le 0, \quad \hat{\xi}^\varepsilon|_{\hat{A}_+^\varepsilon} \ge 0, \tag{3.5}$$

and the optimality condition

$$\gamma(\hat{u}^\varepsilon(t) - u_d(t)) + \frac{1}{\varepsilon}\, g'\bigg(\int_0^T \hat{u}^\varepsilon \,\mathrm{d}t - u_c\bigg) - \int_{\Gamma_c} \alpha(t)\hat{\lambda}_2^\varepsilon(t)\,\mathrm{d}x + \hat{\xi}^\varepsilon(t) = 0 \tag{3.6}$$

holds for almost all $t \in [0,T]$.

Proof. Because of Remark 3.3, there exists a Lagrange multiplier $\hat{p}^\varepsilon = (\hat{\lambda}_1^\varepsilon, \hat{\lambda}_2^\varepsilon, \hat{\mu}_1^\varepsilon, \hat{\mu}_2^\varepsilon) \in Y = [L^2(0,T;H^1(\Omega))]^2 \times [L^2(\Omega)]^2$ such that

$$\begin{aligned} \nabla_{(c_1,c_2)} L_\varepsilon(\hat{x}^\varepsilon, \hat{p}^\varepsilon) &= \nabla_{(c_1,c_2)} J(\hat{x}^\varepsilon) + \nabla_{(c_1,c_2)} e(\hat{x}^\varepsilon)^* \hat{p}^\varepsilon \\ &= 0 \quad \text{in } W(0,T) \times W(0,T). \end{aligned} \tag{3.7}$$

In (3.7) the operator $\nabla_{(c_1,c_2)} e(\hat{x}^\varepsilon)^*$ denotes the adjoint of $\nabla_{(c_1,c_2)} e(\hat{x}^\varepsilon)$. By Proposition 3.2, $\nabla_{(c_1,c_2)} e(\hat{x}^\varepsilon)^*$ is injective, so $\hat{p}^\varepsilon$ is unique. Next we prove that $\hat{\lambda}_1^\varepsilon$ and $\hat{\lambda}_2^\varepsilon$ are more regular than suggested by Y and belong to $W(0,T)$. Condition (3.7) is equivalent to

$$\begin{aligned} 0 &= \nabla_{(c_1,c_2)} L_\varepsilon(\hat{x}^\varepsilon, \hat{p}^\varepsilon)(\delta c_1, \delta c_2) \\ &= \int_\Omega \beta_1(\hat{c}_1^\varepsilon(T) - c_{1T})\delta c_1(T) + \beta_2(\hat{c}_2^\varepsilon(T) - c_{2T})\delta c_2(T)\,\mathrm{d}x \\ &\quad + \int_0^T \langle (\delta c_1)_t(t), \hat{\lambda}_1^\varepsilon(t)\rangle_{H^1(\Omega)',H^1(\Omega)} \mathrm{d}t \\ &\quad + \int_0^T \int_\Omega d_1 \nabla \delta c_1 \cdot \nabla \hat{\lambda}_1^\varepsilon + k_1(\delta c_1 \hat{c}_2^\varepsilon + \hat{c}_1^\varepsilon \delta c_2)\hat{\lambda}_1^\varepsilon \,\mathrm{d}x\,\mathrm{d}t \\ &\quad + \int_0^T \langle (\delta c_2)_t(t), \hat{\lambda}_2^\varepsilon(t)\rangle_{H^1(\Omega)',H^1(\Omega)} \mathrm{d}t \\ &\quad + \int_0^T \int_\Omega d_2 \nabla \delta c_2 \cdot \nabla \hat{\lambda}_2^\varepsilon + k_2(\delta c_1 \hat{c}_2^\varepsilon + \hat{c}_1^\varepsilon \delta c_2)\hat{\lambda}_2^\varepsilon \,\mathrm{d}x\,\mathrm{d}t \\ &\quad + \int_\Omega \delta c_1(0)\hat{\mu}_1^\varepsilon + \delta c_2(0)\hat{\mu}_2^\varepsilon \,\mathrm{d}x \end{aligned} \tag{3.8}$$

for all $(\delta c_1, \delta c_2) \in W(0,T) \times W(0,T)$, in particular for $\delta c_i(t) = \chi(t)\varphi$, where $\chi \in C_c^\infty(0,T)$ and $\varphi \in H_0^1(\Omega)$. Here, $C_c^\infty(0,T)$ denotes the space of infinitely differentiable functions on $(0,T)$ with compact support. We find

$$\begin{aligned} \int_0^T \langle (\delta c_i)_t(t), \hat{\lambda}_i^\varepsilon(t)\rangle_{H^1(\Omega)',H^1(\Omega)} \mathrm{d}t &= \bigg\langle \int_0^T \hat{\lambda}_i^\varepsilon(t)\dot{\chi}(t)\,\mathrm{d}t, \varphi \bigg\rangle_{L^2(\Omega)} \\ &= -\bigg\langle \int_0^T (\hat{\lambda}_i^\varepsilon)_t(t)\chi(t)\,\mathrm{d}t, \varphi \bigg\rangle_{H^1(\Omega)',H^1(\Omega)} \end{aligned}$$

for $i = 1, 2$, where $(\hat\lambda_i^\varepsilon)_t$ denotes the distributional derivative of $\hat\lambda_i^\varepsilon$ with respect to t. The remaining terms in (3.8) lead to

$$\begin{aligned}
&\int_0^T \int_\Omega d_1 \nabla \delta c_1 \cdot \nabla \hat\lambda_1^\varepsilon + k_1(\delta c_1 \hat c_2^\varepsilon + \hat c_1^\varepsilon \delta c_2)\hat\lambda_1^\varepsilon \,\mathrm{d}x\,\mathrm{d}t \\
&\quad + \int_0^T \int_\Omega d_2 \nabla \delta c_2 \cdot \nabla \hat\lambda_2^\varepsilon + k_2(\delta c_1 \hat c_2^\varepsilon + \hat c_1^\varepsilon \delta c_2)\hat\lambda_2^\varepsilon \,\mathrm{d}x\,\mathrm{d}t \\
&= \left\langle \int_0^T -d_1 \Delta \hat\lambda_1^\varepsilon - d_2 \Delta \hat\lambda_2^\varepsilon + (\hat c_1^\varepsilon + \hat c_2^\varepsilon)\big(k_1 \hat\lambda_1^\varepsilon + k_2 \hat\lambda_2^\varepsilon\big)\chi \,\mathrm{d}t, \varphi \right\rangle_{H^1(\Omega)', H^1(\Omega)}.
\end{aligned}$$

Inserting these expressions into (3.8) yields

$$\begin{aligned}
&\left\langle \int_0^T (\hat\lambda_1^\varepsilon + \hat\lambda_2^\varepsilon)_t \chi \,\mathrm{d}t, \varphi \right\rangle_{H^1(\Omega)', H^1(\Omega)} \\
&\quad = \left\langle \int_0^T d_1 \Delta \hat\lambda_1^\varepsilon + d_2 \Delta \hat\lambda_2^\varepsilon - (\hat c_1^\varepsilon + \hat c_2^\varepsilon)\big(k_1 \hat\lambda_1^\varepsilon + k_2 \hat\lambda_2^\varepsilon\big)\chi \,\mathrm{d}t, \varphi \right\rangle_{H^1(\Omega)', H^1(\Omega)}
\end{aligned}$$

for all $\chi \in C_c^\infty(0,T)$ and $\varphi \in H_0^1(\Omega)$. Since

$$d_1 \Delta \hat\lambda_1^\varepsilon + d_2 \Delta \hat\lambda_2^\varepsilon - (\hat c_1^\varepsilon + \hat c_2^\varepsilon)\big(k_1 \hat\lambda_1^\varepsilon + k_2 \hat\lambda_2^\varepsilon\big) \in L^2(0,T; H^1(\Omega))$$

holds and the set

$$\{\delta c : \delta c(t) = \chi(t)\varphi \text{ for } \chi \in C_c^\infty(0,T) \text{ and } \varphi \in H_0^1(\Omega)\}$$

is dense in $L^2(0,T; H^1(\Omega))$, we conclude that $(\hat\lambda_i^\varepsilon)_t \in L^2(0,T; H^1(\Omega)')$ and consequently $\hat\lambda_i^\varepsilon \in W(0,T)$ for $i = 1, 2$. Hence, (3.3a) and (3.3b) are proved. We notice that for all $\delta c_i \in W(0,T)$, $i = 1, 2$, we have

$$\begin{aligned}
&\int_0^T \langle (\hat\lambda_i^\varepsilon)_t(t), \delta c_i(t)\rangle_{H^1(\Omega)', H^1(\Omega)} \,\mathrm{d}t + \int_0^T \langle (\delta c_i)_t(t), \hat\lambda_i^\varepsilon(t)\rangle_{H^1(\Omega)', H^1(\Omega)} \,\mathrm{d}t \\
&\quad = \int_0^T \frac{\mathrm{d}}{\mathrm{d}t} \langle \hat\lambda_i^\varepsilon(t), \delta c_i(t)\rangle_{L^2(\Omega)} \,\mathrm{d}t \\
&\quad = \langle \hat\lambda_i^\varepsilon(T), \delta c_i(T)\rangle_{L^2(\Omega)} - \langle \hat\lambda_i^\varepsilon(0), \delta c_i(0)\rangle_{L^2(\Omega)}.
\end{aligned} \tag{3.9}$$

Choosing appropriate test functions in $W(0,T)$, we infer from (3.3a), (3.3b), (3.8), and (3.9) that (3.3c)–(3.3f) as well as (3.4) are satisfied. Because of optimality the following optimality inequality holds:

$$\nabla_u L_\varepsilon(\hat x^\varepsilon, \hat p^\varepsilon)(u - \hat u^\varepsilon) \geq 0 \quad \text{for all } u \in U_{\mathsf{ad}}.$$

Setting

$$\begin{aligned}
-\langle \hat\xi^\varepsilon, u - \hat u^\varepsilon\rangle_{L^2(0,T)} &= \nabla_u L_\varepsilon(\hat x^\varepsilon, \hat p^\varepsilon)(u - \hat u^\varepsilon) \\
&= \left\langle \gamma(\hat u^\varepsilon - u_d) - \int_{\Gamma_c} \alpha \hat\lambda_2^\varepsilon \,\mathrm{d}x, u - \hat u^\varepsilon \right\rangle_{L^2(0,T)} \\
&\quad + \frac{1}{\varepsilon} g'\bigg(\int_0^T \hat u^\varepsilon \,\mathrm{d}t - u_c \bigg) \int_0^T u - \hat u^\varepsilon \,\mathrm{d}t \\
&= \left\langle \gamma(\hat u^\varepsilon - u_d) + \frac{1}{\varepsilon} g'\bigg(\int_0^T \hat u^\varepsilon \,\mathrm{d}t - u_c \bigg) - \int_{\Gamma_c} \alpha \hat\lambda_2^\varepsilon \,\mathrm{d}x, u - \hat u^\varepsilon \right\rangle_{L^2(0,T)}
\end{aligned} \tag{3.10}$$

for all $u \in U$, we obtain (3.6). For the proof of (3.5) we refer the reader to [13]. ∎

Remark 3.5. From Remark 3.13, uniqueness of the Lagrange multiplier $\hat{\xi}^\varepsilon$ will follow later on. ∎

The following corollary provides an a priori estimate for the Lagrange multipliers $\hat{\lambda}_1^\varepsilon$ and $\hat{\lambda}_2^\varepsilon$ that will be used for the second-order optimality conditions; see section 3.3 and Theorem 3.15.

COROLLARY 3.6. *There exists a constant $C_\lambda > 0$ such that*

$$\begin{aligned} &\|\hat{\lambda}_1^\varepsilon\|_{L^2(0,T;L^4(\Omega))} + \|\hat{\lambda}_2^\varepsilon\|_{L^2(0,T;L^4(\Omega))} \\ &\quad \le C_\lambda\big(\|\hat{c}_1^\varepsilon(T) - c_{1T}\|_{L^2(\Omega)} + \|\hat{c}_2^\varepsilon(T) - c_{2T}\|_{L^2(\Omega)}\big). \end{aligned} \tag{3.11}$$

For the proof we refer the reader to [9, Corollary 3.6].

3.3. Second-order optimality conditions. In section 3.2 we have investigated the first-order necessary optimality conditions for (P_ε). To ensure that a solution $(\hat{x}^\varepsilon, \hat{p}^\varepsilon)$ satisfying (2.4), $\hat{x}^\varepsilon \in K_{\mathsf{ad}}$, (3.3), and (3.6) indeed solves (P_ε), we have to guarantee second-order sufficient optimality. This is the focus of this section. In order to introduce the critical cone in Definition 3.9, we recall the notions of tangent and normal cones.

DEFINITION 3.7. *Let K be a convex subset of a Hilbert space Z and $z \in K$. The set*

$$T_K(z) = \{\tilde{z} \in Z : \text{ there exists } z(\sigma) = z + \sigma\tilde{z} + o(\sigma) \in K \text{ as } \sigma \searrow 0\}$$

is called the tangent cone at the point z. *Moreover, the* normal cone N_K at the point z *is given by*

$$N_K(z) = \{\tilde{z} \in Z : \langle \tilde{z}, \hat{z} - z\rangle_Z \le 0 \text{ for all } \hat{z} \in K\}.$$

In the case of $z \notin K$ these two cones are set equal to the empty set.

For $K = K_{\mathsf{ad}}$ we have the following characterizations.

LEMMA 3.8. *Let $x = (c_1, c_2, u) \in K_{\mathsf{ad}}$.*

(a) $T_{K_{\mathsf{ad}}}(x) = W(0,T) \times W(0,T) \times T_{U_{\mathsf{ad}}}(u)$, *where*

$$T_{U_{\mathsf{ad}}}(u) = \{\tilde{u} \in L^2(0,T) : \tilde{u}(t) \in T_{[u_a(t),u_b(t)]}(u(t)) \text{ for } t \in [0,T] \text{ a.e.}\},$$

where for $a, b, s \in \mathbb{R}$ with $a \le b$

$$T_{[a,b]}(s) = \begin{cases} \mathbb{R}^+ = \{t \in \mathbb{R} : t \ge 0\} & \text{if } s = a, \\ \mathbb{R}^- = \{t \in \mathbb{R} : t \le 0\} & \text{if } s = b, \\ \mathbb{R} & \text{otherwise.} \end{cases}$$

(b) $N_{K_{\mathsf{ad}}}(x) = \{0\} \times \{0\} \times N_{U_{\mathsf{ad}}}(u)$, *where*

$$N_{U_{\mathsf{ad}}}(u) = \{\tilde{u} \in L^2(0,T) : \tilde{u}(t) \in N_{[u_a(t),u_b(t)]}(u(t)) \text{ for } t \in [0,T] \text{ a.e.}\}.$$

That is,

$$\tilde{u}(t) \in \begin{cases} \mathbb{R}^+ = \{t \in \mathbb{R} : t \ge 0\} & \text{if } u(t) = u_a(t), \\ \mathbb{R}^- = \{t \in \mathbb{R} : t \le 0\} & \text{if } u(t) = u_b(t), \\ \mathbb{R} & \text{otherwise.} \end{cases}$$

(c) *Moreover,*

(3.12)
$$T_{U_{\mathsf{ad}}}(\hat{u}^\varepsilon) \cap \{\hat{\xi}^\varepsilon\}^\perp = \{u \in L^2(0,T) : u \geq 0 \text{ on } \hat{A}^\varepsilon_-,\ u \leq 0 \text{ on } \hat{A}^\varepsilon_+, \text{ and } u = 0 \text{ on } \hat{A}^\varepsilon_\pm\},$$

where $\hat{\xi}^\varepsilon \in N_{U_{\mathsf{ad}}}$ *is the Lagrange multiplier introduced in Theorem* 3.4. $S^\perp$ *denotes the orthogonal complement of a set* S, *and* $\hat{A}^\varepsilon_\pm = \{t \in [0,T] : \hat{\xi}^\varepsilon > 0 \text{ or } \hat{\xi}^\varepsilon < 0 \text{ a.e.}\} \subset \hat{A}^\varepsilon$.

Proof. The characterization of the tangent and normal cones is a classical result. For a proof we refer the reader to [29]. What remains to show is (3.12). Because of (3.10) we have $\hat{\xi}^\varepsilon \in N_{U_{\mathsf{ad}}}(\hat{u}^\varepsilon)$ satisfying $\hat{\xi}^\varepsilon = 0$ on the set $[0,T] \setminus \hat{A}^\varepsilon_\pm$. Suppose that $t \in \hat{A}^\varepsilon_-$. We conclude $T_{[u_a(t),u_b(t)]}(\hat{u}^\varepsilon(t)) = \mathbb{R}^+$. Thus, $u \in T_{U_{\mathsf{ad}}}(\hat{u}^\varepsilon)$ implies $u \geq 0$ on $\hat{A}^\varepsilon_-$ by part (a). Analogously, $u \leq 0$ on $\hat{A}^\varepsilon_+$ holds. Hence,

$$T_{U_{\mathsf{ad}}}(\hat{u}^\varepsilon) \cap \{\hat{\xi}^\varepsilon\}^\perp = \{\tilde{u} \in L^2(0,T) : \tilde{u}(t) \in T_{[u_a(t),u_b(t)]}(\hat{u}^\varepsilon(t)) \text{ for } t \in [0,T] \text{ a.e.}\} \cap \left\{u \in L^2(0,T) : \int_{\hat{A}^\varepsilon_\pm} \hat{\xi}^\varepsilon u \,\mathrm{d}t = 0\right\}.$$

Since $\hat{\xi}^\varepsilon > 0$ and $u \geq 0$ on $\hat{A}^\varepsilon_- \cap \hat{A}^\varepsilon_\pm$, and $\hat{\xi}^\varepsilon < 0$ and $u \leq 0$ on $\hat{A}^\varepsilon_+ \cap \hat{A}^\varepsilon_\pm$, (3.12) holds, which completes the proof. ∎

Suppose that the point $\bar{x} = (\bar{c}_1, \bar{c}_2, \bar{u}) \in X$ satisfies the first-order necessary optimality conditions. By Proposition 3.2, there exist unique Lagrange multipliers $\bar{p} = (\bar{\lambda}_1, \bar{\lambda}_2, \bar{\mu}_1, \bar{\mu}_2) \in Y$ and $\bar{\xi} \in N_{U_{\mathsf{ad}}}$ satisfying the first-order necessary optimality conditions

$$\nabla_x L_\varepsilon(\bar{x}, \bar{p}) + (0,0,\bar{\xi})^{\mathsf{T}} = 0, \quad \bar{x} \in K_{\mathsf{ad}} \text{ and } e(\bar{x}) = 0. \tag{3.13}$$

Now we introduce the critical cone at $\bar{x}$.

Definition 3.9. *The* critical cone at $\bar{x}$ *is defined by*

$$C(\bar{x}) = \{\delta x \in T_{K_{\mathsf{ad}}}(\bar{x}) \cap \{(0,0,\bar{\xi})\}^\perp : \delta x \in N(\nabla e(\bar{x}))\}.$$

The critical cone at $\bar{x}$ is the set of directions of nonincrease of the cost that are tangent to the feasible set $\{x \in K_{\mathsf{ad}} : e(x) = 0\}$. This is formulated in the next lemma.

Lemma 3.10. *It follows that* $\nabla J_\varepsilon(\bar{x})\delta x = 0$ *for all* $\delta x \in C(\bar{x})$.

Proof. Let $\delta x = (\delta c_1, \delta c_2, \delta u) \in C(\bar{x})$. From (3.13), $\delta x \in N(\nabla e(\bar{x}))$ and $\delta x \in \{(0,0,\bar{\xi})\}^\perp$ we infer that

$$0 = \big(\nabla_x L_\varepsilon(\bar{x}, \bar{p}) + (0,0,\bar{\xi})^{\mathsf{T}}\big)\delta x = \nabla J_\varepsilon(\bar{x})\delta x,$$

which completes the proof. ∎

Now we turn to the *second-order necessary optimality conditions.* Let $\delta x = (\delta c_1, \delta c_2, \delta u) \in X$. We find

$$\begin{aligned} \nabla^2_x L_\varepsilon(\bar{x}, \bar{p})(\delta x, \delta x) &= \int_\Omega \beta_1 |\delta c_1(T)|^2 + \beta_2 |\delta c_2(T)|^2 \,\mathrm{d}x \\ &\quad + \frac{1}{\varepsilon}\nabla^2 I(\bar{u})(\delta u, \delta u) + \int_0^T \gamma |\delta u|^2 \,\mathrm{d}t \\ &\quad + \int_0^T \int_\Omega (2k_1\bar{\lambda}_1 + 2k_2\bar{\lambda}_2)\delta c_1 \delta c_2 \,\mathrm{d}x \,\mathrm{d}t. \end{aligned} \tag{3.14}$$

In Theorem 2.8 we have denoted by $\hat{x}^\varepsilon$ the local solution to (P_ε). The associated unique Lagrange multipliers are $\hat{p}^\varepsilon$ and $\hat{\xi}^\varepsilon$; see Theorem 3.4.

DEFINITION 3.11. *The* second-order necessary optimality conditions *are defined as*

$$\nabla_x^2 L_\varepsilon(\hat{x}^\varepsilon, \hat{p}^\varepsilon)(\delta x, \delta x) \geq 0 \quad \textit{for all } \delta x \in C(\hat{x}^\varepsilon). \tag{3.15}$$

Now let $\bar{x} = \hat{x}^\varepsilon$ be a local solution to (P_ε).

THEOREM 3.12. *The point $(\hat{x}^\varepsilon, \hat{p}^\varepsilon)$ satisfies the second-order necessary optimality condition* (3.15).

Proof. We apply analogous arguments as in the proof of Theorem 2.7 in [4]. The equality constraints can be written as

$$e(\hat{x}^\varepsilon) \in K_Y = \{0\} \subset Y,$$

where, of course, K_Y is a closed convex set. Thus, the result follows provided the following strict semilinearized qualification condition

$$0 \in \operatorname{int} \{\nabla e(\hat{x}^\varepsilon)((K_{\mathsf{ad}} - \hat{x}^\varepsilon) \cap \{(0,0,\hat{\xi}^\varepsilon)\}^\perp)\} \subset Y \tag{CQA}$$

holds. In our case, we have

$$(K_{\mathsf{ad}} - \hat{x}^\varepsilon) \cap \{(0,0,\hat{\xi}^\varepsilon)\}^\perp = W(0,T) \times W(0,T) \times \big((U_{\mathsf{ad}} - \hat{u}^\varepsilon) \cap \{\hat{\xi}^\varepsilon\}^\perp\big).$$

Let $z \in Y$ be arbitrary, close enough to zero. Then (CQA) follows if there exists an element $\delta x = (\delta c_1, \delta c_2, \delta u) \in W(0,T) \times W(0,T) \times (U_{\mathsf{ad}} - \hat{u}^\varepsilon) \cap \{\hat{\xi}^\varepsilon\}^\perp$ satisfying

$$\nabla e(\hat{x}^\varepsilon)\delta x = z. \tag{3.16}$$

Because of Proposition 3.2 the operator $\nabla_{(c_1,c_2)} e(\hat{x}^\varepsilon)$ is bijective. Thus, there exists even a unique pair $(\delta c_1, \delta c_2) \in W(0,T) \times W(0,T)$ such that

$$\nabla_{(c_1,c_2)} e(\hat{x}^\varepsilon)(\delta c_1, \delta c_2) = z - \nabla_u e(\hat{x}^\varepsilon)\delta u$$

for arbitrary $\delta u \in L^2(0,T)$. This gives (3.16) so that the claim follows. ∎

Remark 3.13. As is proved in [4], condition (CQA) implies uniqueness of the Lagrange multipliers $\hat{p}^\varepsilon$ and $\hat{\xi}^\varepsilon$. ∎

DEFINITION 3.14. *Suppose that $\bar{x}$ satisfies the first-order necessary optimality conditions with the associated unique Lagrange multipliers $\bar{p} \in Y$ and $\bar{\xi} \in N_{U_{\mathsf{ad}}}(\bar{u})$. At $(\bar{x}, \bar{p})$, the* second-order sufficient optimality condition *holds if there exists $\kappa > 0$ such that*

$$\nabla_x^2 L_\varepsilon(\bar{x}, \bar{p})(\delta x, \delta x) \geq \kappa \|\delta x\|_X^2 \quad \textit{for all } \delta x \in C(\bar{x}).$$

THEOREM 3.15. *If $\|\hat{c}_1^\varepsilon(T) - c_{1T}\|_{L^2(\Omega)} + \|\hat{c}_2^\varepsilon(T) - c_{2T}\|_{L^2(\Omega)}$ is sufficiently small, then the second-order sufficient optimality condition is satisfied.*

Proof. Let $\delta x = (\delta c_1, \delta c_2, \delta u) \in C(\hat{x}^\varepsilon) \setminus \{0\}$. Then, we have

$$\nabla^2 I(\hat{u}^\varepsilon)(\delta u, \delta u) = 6 \left[\int_0^T \hat{u}^\varepsilon \,\mathrm{d}t - u_c \right]_+ \left(\int_0^T \delta u \,\mathrm{d}t \right)^2 \geq 0.$$

Recall *Gagliardo–Nierenberg's inequality* (see, e.g., [32, p. 81]): for $\Omega \subset \mathbb{R}^3$ there exists a constant $C_{GN} > 0$ such that

$$\|\varphi\|_{L^4(\Omega)} \le C_{GN} \|\varphi\|_{H^1(\Omega)}^{3/4} \|\varphi\|_{L^2(\Omega)}^{1/4} \quad \text{for all } \varphi \in H^1(\Omega).$$

Hence, using Hölder's and Gagliardo–Nierenberg's inequality and (2.2) we find

$$\begin{aligned}
&\int_0^T \int_\Omega \big(k_1 \hat\lambda_1^\varepsilon + k_2 \hat\lambda_2^\varepsilon\big) \delta c_1 \delta c_2 \,\mathrm{d}x \,\mathrm{d}t \\
&\le \int_0^T \big(k_1 \|\hat\lambda_1^\varepsilon(t)\|_{L^4(\Omega)} + k_2 \|\hat\lambda_2^\varepsilon(t)\|_{L^4(\Omega)}\big) \|\delta c_1(t)\|_{L^4(\Omega)} \|\delta c_2(t)\|_{L^2(\Omega)} \,\mathrm{d}t \\
&\le \int_0^T C_{GN}\big(k_1 \|\hat\lambda_1^\varepsilon(t)\|_{L^4(\Omega)} + k_2 \|\hat\lambda_2^\varepsilon(t)\|_{L^4(\Omega)}\big) \\
&\quad \times \|\delta c_1(t)\|_{L^2(\Omega)}^{1/4} \|\delta c_1(t)\|_{H^1(\Omega)}^{3/4} \|\delta c_2(t)\|_{L^2(\Omega)} \,\mathrm{d}t \\
&\le C_{GN}\big(k_1 \|\hat\lambda_1^\varepsilon\|_{L^2(0,T;L^4(\Omega))} + k_2 \|\hat\lambda_2^\varepsilon\|_{L^2(0,T;L^4(\Omega))}\big) \\
&\quad \times \|\delta c_1\|_{C([0,T];L^2(\Omega))}^{1/4} \|\delta c_1\|_{L^2(0,T;H^1(\Omega))}^{3/4} \|\delta c_2\|_{C([0,T];L^2(\Omega))} \\
&\le C_{GN} C_W^{5/4} \big(k_1 \|\hat\lambda_1^\varepsilon\|_{L^2(0,T;L^4(\Omega))} + k_2 \|\hat\lambda_2^\varepsilon\|_{L^2(0,T;L^4(\Omega))}\big) \|\delta c_1\|_{W(0,T)} \|\delta c_2\|_{W(0,T)}.
\end{aligned}$$

Thus, (3.1) and (3.14) yield

$$\begin{aligned}
&\nabla_x^2 L_\varepsilon(\hat x^\varepsilon, \hat p^\varepsilon)(\delta x, \delta x) \\
&\ge \frac{\gamma}{2} \|\delta u\|_{L^2(0,T)}^2 + \frac{\gamma}{2C_N^2}\big(\|\delta c_1\|_{W(0,T)}^2 + \|\delta c_2\|_{W(0,T)}^2\big) \\
&\quad - 2C_{GN}^{5/4} C_W^2 \big(k_1 \|\hat\lambda_1^\varepsilon\|_{L^2(0,T;L^4(\Omega))} + k_2 \|\hat\lambda_2^\varepsilon\|_{L^2(0,T;L^4(\Omega))}\big) \|\delta c_1\|_{W(0,T)} \|\delta c_2\|_{W(0,T)} \\
&\ge \frac{\gamma}{2} \|\delta u\|_{L^2(0,T)}^2 + \big(\|\delta c_1\|_{W(0,T)}^2 + \|\delta c_2\|_{W(0,T)}^2\big) \\
&\quad \times \Big(\frac{\gamma}{2C_N^2} - K_1\big(\|\hat\lambda_1^\varepsilon\|_{L^2(0,T;L^4(\Omega))} + \|\hat\lambda_2^\varepsilon\|_{L^2(0,T;L^4(\Omega))}\big)\Big),
\end{aligned}$$

where we set $K_1 = \max(k_1, k_2) C_{GN}^{5/4} C_W^2 > 0$. Because of (3.11) we find

$$\begin{aligned}
&\nabla_x^2 L_\varepsilon(\hat x^\varepsilon, \hat p^\varepsilon)(\delta x, \delta x) \\
&\ge \frac{\gamma}{2} \|\delta u\|_{L^2(0,T)}^2 + \big(\|\delta c_1\|_{W(0,T)}^2 + \|\delta c_2\|_{W(0,T)}^2\big) \\
&\quad \times \Big(\frac{\gamma}{2C_N} - K_2\big(\|\hat c_1^\varepsilon(T) - c_{1T}\|_{L^2(\Omega)} + \|\hat c_2^\varepsilon(T) - c_{2T}\|_{L^2(\Omega)}\big)\Big)
\end{aligned}$$

with the constant $K_2 = K_1 C_\lambda > 0$. Now, for instance, if

$$\|\hat c_1^\varepsilon(T) - c_{1T}\|_{L^2(\Omega)} + \|\hat c_2^\varepsilon(T) - c_{2T}\|_{L^2(\Omega)} \le \frac{\gamma}{4K_2 C_N}$$

holds, the second-order sufficient optimality condition is satisfied for the coercivity constant $\kappa = \gamma \min(1, 1/C_N)/2$. ∎

Remark 3.16.

(1) Notice that smallness of the two terms $\|\hat c_i^\varepsilon(T) - c_{iT}\|_{L^2(\Omega)}$, $i = 1, 2$, ensures that

$$\nabla_x^2 L_\varepsilon(\hat x^\varepsilon, \hat p^\varepsilon)(\delta x, \delta x) \ge \kappa \|\delta x\|_X^2 \quad \text{for all } \delta x \in N(\nabla e(\hat x^\varepsilon)).$$

Since $C(\hat{x}^\varepsilon) \subset N(\nabla e(\hat{x}^\varepsilon))$ holds, the second-order sufficient optimality condition is satisfied; see Definition 3.14.

(2) Arguing as in the proof of Theorem 4.12 in [36] it follows that the second-order sufficient optimality condition is equivalent to the *strong quadratic growth condition*, i.e., equivalent to the existence of two constants $\varrho, \zeta > 0$ such that

$$J_\varepsilon(x) \geq J_\varepsilon(\hat{x}^\varepsilon) + \varrho \, \|x - \hat{x}^\varepsilon\|_X^2 + o(\|x - \hat{x}^\varepsilon\|_X^2) \quad \text{for all } x \in \mathcal{K}_\zeta(\hat{x}^\varepsilon)$$

with $\mathcal{K}_\zeta(\hat{x}^\varepsilon) = \{x \in K_{\mathsf{ad}} : e(x) = 0 \text{ and } \|x - \hat{x}^\varepsilon\|_X \leq \zeta\}$. ∎

4. The primal-dual active set method. In this section, we describe the primal-dual active set strategy for nonlinear problems and review convergence results. For more details and for the proofs we refer the reader to [19].

Because of Theorem 2.3 we can define the solution operator

$$\mathcal{S} : L^2(0,T) \to W(0,T) \times W(0,T)$$

by $(c_1, c_2) = \mathcal{S}(u)$ for $u \in L^2(0,T)$, where the pair $(c_1, c_2) \in W(0,T) \times W(0,T)$ is the solution of (2.4). Introducing the reduced cost functional

$$\hat{J}_\varepsilon(u) = J_\varepsilon(\mathcal{S}(u), u),$$

problem (P_ε) can be expressed as

$$(\hat{\mathrm{P}}_\varepsilon) \qquad \min \hat{J}_\varepsilon(u) \quad \text{s.t.} \quad u \in U_{\mathsf{ad}}.$$

Notice that $(\hat{\mathrm{P}}_\varepsilon)$ is a minimization problem with bilateral control constraints but with no equality constraints. The gradient of $\hat{J}_\varepsilon$ at a point $\hat{u}^\varepsilon \in L^2(0,T)$ is given

$$\nabla \hat{J}_\varepsilon(\hat{u}^\varepsilon) = \gamma(\hat{u}^\varepsilon - u_d) + \frac{1}{\varepsilon} \, g'\bigg(\int_0^T \hat{u}^\varepsilon \, \mathrm{d}t - u_c \bigg) - \int_{\Gamma_c} \alpha \hat{\lambda}_2^\varepsilon \, \mathrm{d}x \in L^2(0,T), \tag{4.1}$$

where $(\hat{\lambda}_1^\varepsilon, \hat{\lambda}_2^\varepsilon) \in W(0,T) \times W(0,T)$ solves (3.3) for the state pair $(\hat{c}_1^\varepsilon, \hat{c}_2^\varepsilon)$, which in turn is the solution of (2.4) for the control input $\hat{u}^\varepsilon$.

From Theorem 3.4 we derive that the first-order necessary optimality conditions

$$\langle \nabla \hat{J}_\varepsilon(\hat{u}^\varepsilon), u - \hat{u}^\varepsilon \rangle_{L^2(0,T)} \geq 0 \quad \text{for all } u \in U_{\mathsf{ad}}$$

are equivalent to

$$\gamma(\hat{u}^\varepsilon - u_d) + \frac{1}{\varepsilon} \, g'\bigg(\int_0^T \hat{u}^\varepsilon \, \mathrm{d}t - u_c \bigg) - \int_{\Gamma_c} \alpha \hat{\lambda}_2^\varepsilon \, \mathrm{d}x + \hat{\xi}^\varepsilon = 0 \quad \text{in } L^2(0,T), \tag{4.2a}$$

where the Lagrange multiplier $\hat{\xi}^\varepsilon \in N_{U_{\mathsf{ad}}}(\hat{u}^\varepsilon)$ associated with the bilateral control constraints satisfies

$$\hat{\xi}^\varepsilon = \max\big\{0, \hat{\xi}^\varepsilon + (\hat{u}^\varepsilon - u_b)\big\} + \min\big\{0, \hat{\xi}^\varepsilon + (\hat{u}^\varepsilon - u_a)\big\} \quad \text{in } L^2(0,T). \tag{4.2b}$$

In (4.2b) the functions max and min are interpreted as pointwise a.e. operations. We next specify the primal-dual active set method.

ALGORITHM 4.1 (primal-dual active set strategy).

(1) Choose $(u^0, \xi^0) \in L^2(0,T) \times L^2(0,T)$, $\sigma > 0$ and set $k = 0$.

(2) Determine the active sets

$$A_-^k = \Big\{ t \in [0,T] : u^k(t) + \frac{\xi^k(t)}{\sigma} < u_a(t) \Big\},$$
$$A_+^k = \Big\{ t \in [0,T] : u^k(t) + \frac{\xi^k(t)}{\sigma} > u_b(t) \Big\}$$

and set $I^k = [0,T] \setminus (A_-^k \cup A_+^k)$.

(3) If $k \geq 1$ and $A_+^k = A_+^{k-1}$, $A_-^k = A_-^{k-1}$, then STOP.

(4) Solve for $(c_1, c_2, u, \lambda_1, \lambda_2) \in W(0,T) \times W(0,T) \times L^2(I^k) \times W(0,T) \times W(0,T)$ the coupled nonlinear system

$$\text{(4.3)} \quad \begin{cases} (c_1)_t = d_1 \Delta c_1 - k_1 c_1 c_2 & \text{in } Q, \\ (c_2)_t = d_2 \Delta c_2 - k_2 c_1 c_2 & \text{in } Q, \\ d_1 \frac{\partial c_1}{\partial n} = 0 & \text{on } \Sigma, \\ d_2 \frac{\partial c_2}{\partial n} = u_a \alpha & \text{on } A_-^k \times \Gamma_c, \\ d_2 \frac{\partial c_2}{\partial n} = u_b \alpha & \text{on } A_+^k \times \Gamma_c, \\ d_2 \frac{\partial c_2}{\partial n} = u \alpha & \text{on } I^k \times \Gamma_c, \\ d_2 \frac{\partial c_2}{\partial n} = 0 & \text{on } \Sigma_n, \\ c_1(0) = c_{10} & \text{in } \Omega, \\ c_2(0) = c_{20} & \text{in } \Omega, \\ -(\lambda_1)_t = d_1 \Delta \lambda_1 - k_1 c_2 \lambda_1 - k_2 c_2 \lambda_2 & \text{in } Q, \\ -(\lambda_2)_t = d_2 \Delta \lambda_2 - k_1 c_1 \lambda_1 - k_2 c_1 \lambda_2 & \text{in } Q, \\ d_1 \frac{\partial \lambda_1}{\partial n} = 0 & \text{on } \Sigma, \\ d_2 \frac{\partial \lambda_2}{\partial n} = 0 & \text{on } \Sigma, \\ \lambda_1(T) = -\beta_1 (c_1(T) - c_{1T}) & \text{in } \Omega, \\ \lambda_2(T) = -\beta_2 (c_2(T) - c_{2T}) & \text{in } \Omega, \\ \gamma(u - u_d) = \int_{\Gamma_c} \alpha \lambda_2 \, \mathrm{d}x - \frac{1}{\varepsilon} g'\big(\int_0^T \bar{u}^k \, \mathrm{d}t - u_c \big) & \text{in } I^k \end{cases}$$

with $\bar{u}^k = u_a$ in A_-^k, $\bar{u}^k = u_b$ in A_+^k, and $\bar{u}^k = u$ in I^k.

(5) Set $(c_1^{k+1}, c_2^{k+1}, \lambda_1^{k+1}, \lambda_2^{k+1}) = (c_1, c_2, \lambda_1, \lambda_2)$, $u^{k+1} = u_a$ in A_-^k, $u^{k+1} = u_b$ in A_+^k, and $u^{k+1} = u$ in I^k and

$$\text{(4.4)} \quad \xi^{k+1} = \int_{\Gamma_c} \alpha \lambda_2^{k+1} \, \mathrm{d}x - \frac{1}{\varepsilon} g'\bigg(\int_0^T u^{k+1} \, \mathrm{d}t - u_c \bigg) - \gamma(u^{k+1} - u_d) \quad \text{in } [0,T].$$

Set $k = k+1$ and go back to step (2).

Remark 4.2.

(1) Notice that (4.3) are the first-order necessary optimality conditions for

$$(\hat{\mathrm{P}}_\varepsilon^k) \qquad \min \hat{J}_\varepsilon(u) \quad \text{s.t.} \quad u = u_a \text{ in } A_-^k \text{ and } u = u_b \text{ in } A_+^k,$$

which is a minimization problem without any inequality constraints.

(2) In section 5 we solve (4.3) using an inexact Newton method for $(\hat{\mathrm{P}}_\varepsilon^k)$. The inexact Newton step is computed by applying the conjugate gradient (CG) method with negative curvature test (see, e.g., [26, section 6.2]) to the Newton system (restricted to the inactive set I^k)

$$\nabla^2 \hat{J}_\varepsilon(u_i)_{|I^k}\, \delta u_{i|I^k} = -\nabla \hat{J}_\varepsilon(u_i)_{|I^k} \quad \text{for } i \geq 0 \tag{4.5}$$

with zero initial guess $\delta u_i \equiv 0$ in $A_-^k \cup A_+^k$ for all i. More precisely, if $R : L^2(I^k) \to L^2(0,T)$ is the linear extension-by-zero operator from the inactive set I^k to $(0,T)$, then (4.5) reads

$$\nabla^2 \hat{J}_\varepsilon(u_i)(R\,\delta u_{i|I^k}, R\,\cdot) = -\nabla \hat{J}_\varepsilon(u_i)(R\,\cdot) \quad \text{in } L^2(I^k) \text{ for } i \geq 0. \tag{4.6}$$

To compute the right-hand side in (4.5) for a given current iterate u_i we have to solve the state and adjoint equations. For each CG step, the solutions of one linearized state and one adjoint problem have to be computed.

(3) The strategy for choosing the current active sets A_-^k and A_+^k is based on convex analysis techniques (see [2, 18]). Because of the simple nature of the box constraints in $(\hat{\mathrm{P}}_\varepsilon)$ this strategy is related to strategies already used in [3, 11]. □

In the case of $g \equiv 0$, sufficient conditions for global convergence of Algorithm 4.1 were given in [19]. The proof is based theoretically on descent properties of the merit function $\Phi : L^2(0,T) \times L^2(0,T) \to \mathbb{R}$ defined as

$$\Phi(u,\xi) = \gamma^2 \int_0^T ([u-u_a]_+^2 + [u_b-u]_+^2)\,\mathrm{d}t + \int_{A_-(u)} [\xi]_-^2\,\mathrm{d}t + \int_{A_+(u)} [\xi]_+^2\,\mathrm{d}t,$$

where $[x]_+ = \max\{0,x\}$ and $[x]_- = -\min\{0,x\}$ denote the positive and negative part functions, respectively, and $A_-(u) = \{t \in [0,T] : u \leq u_a\}$, $A_+(u) = \{t \in [0,T] : u \geq u_b\}$.

By expressing the primal-dual active set method as a partial semismooth Newton algorithm for (4.2), sufficient conditions for superlinear convergence were also derived in [19]. Since only the nonlinearity due to the max and min operations is linearized, whereas $u \mapsto \mathcal{S}(u)$ is not, the method is called a partial semismooth Newton algorithm. Next we specify the nonlinear equation, which is the starting point for proving superlinear convergence of Algorithm 4.1. First notice that (4.2b) is equivalent to

$$\hat{\xi}^\varepsilon = \max\left\{0, \hat{\xi}^\varepsilon + \sigma(\hat{u}^\varepsilon - u_b)\right\} + \min\left\{0, \hat{\xi}^\varepsilon + \sigma(\hat{u}^\varepsilon - u_a)\right\} \quad \text{for every } \sigma > 0. \tag{4.7}$$

Choosing $\sigma = \gamma$ (an essential prerequisite in proving superlinear convergence) in (4.7) we find

$$-\hat{\xi}^\varepsilon + \max\left\{0, \hat{\xi}^\varepsilon + \gamma(\hat{u}^\varepsilon - u_b)\right\} + \min\left\{0, \hat{\xi}^\varepsilon + \gamma(\hat{u}^\varepsilon - u_a)\right\} = 0. \tag{4.8}$$

Inserting the optimality condition (4.2a) into (4.8) we derive

$$\begin{aligned} 0 \;&= \gamma(\hat{u}^\varepsilon - u_d) + \frac{1}{\varepsilon}\, g'\Big(\int_0^T \hat{u}^\varepsilon\,\mathrm{d}t - u_c\Big) - \int_{\Gamma_c} \alpha\hat{\lambda}_2^\varepsilon\,\mathrm{d}x \\ &\quad + \max\left\{0, \int_{\Gamma_c} \alpha\hat{\lambda}_2^\varepsilon\,\mathrm{d}x - \frac{1}{\varepsilon}\, g'\Big(\int_0^T \hat{u}^\varepsilon\,\mathrm{d}t - u_c\Big) + \gamma(u_d - u_b)\right\} \\ &\quad + \min\left\{0, \int_{\Gamma_c} \alpha\hat{\lambda}_2^\varepsilon\,\mathrm{d}x - \frac{1}{\varepsilon}\, g'\Big(\int_0^T \hat{u}^\varepsilon\,\mathrm{d}t - u_c\Big) - \gamma(u_a - u_d)\right\}. \end{aligned} \tag{4.9}$$

Notice that

$$a - b + \max\{0, b - c\} = a - c + \max\{0, c - b\} \quad \text{for all } a, b, c \in \mathbb{R}.$$

Taking

$$a = \gamma(\hat{u}^\varepsilon - u_d), \quad b = \int_{\Gamma_c} \alpha \hat{\lambda}_2^\varepsilon \, \mathrm{d}x - \frac{1}{\varepsilon} g'\Big(\int_0^T \hat{u}^\varepsilon \, \mathrm{d}t - u_c \Big), \quad c = \gamma(u_b - u_d)$$

we infer from (4.9) that (4.2) is equivalent to

$$\mathcal{F}(\hat{u}^\varepsilon) = 0 \quad \text{in } L^2(0,T) \tag{4.10}$$

with the nonlinear mapping $\mathcal{F} : L^2(0,T) \to L^2(0,T)$ given by

$$\begin{aligned} \mathcal{F}(u) \quad &= \gamma(u - u_b) \\ &\quad + \max\left\{0, \gamma(u_b - u_d) - \int_{\Gamma_c} \alpha\lambda_2 \, \mathrm{d}x + \frac{1}{\varepsilon} g'\Big(\int_0^T u \, \mathrm{d}t - u_c \Big) \right\} \\ &\quad + \min\left\{0, \int_{\Gamma_c} \alpha\lambda_2 \, \mathrm{d}x - \frac{1}{\varepsilon} g'\Big(\int_0^T u \, \mathrm{d}t - u_c \Big) - \gamma(u_a - u_d) \right\} \end{aligned}$$

for $u \in L^2(0,T)$, where $\lambda_2 = \lambda_2(u)$ depends on u via the nonlinear state and the linear adjoint equations.

Remark 4.3. Note that $\mathcal{F}$ is generalized differentiable in the sense of [15]; see also [35] for a related approach. This infinite-dimensional generalized differentiability concept of the max and min functions requires a norm gap, which is guaranteed by the smoothing properties of the mapping $u \mapsto \lambda_2(u)$: for each $u \in L^2(0,T)$, the argument under the max and min functions which depend on u is an element of $L^4(0,T)$, which can be seen as follows. First of all, $g'(\int_0^T u \, \mathrm{d}t - u_c)$ is a scalar. Secondly, with $\alpha \in L^\infty(0,T;L^2(\Gamma_c))$, we have

$$\int_0^T \left| \int_{\Gamma_c} \alpha\lambda_2 \, \mathrm{d}x \right|^p \, \mathrm{d}t \le \|\alpha\|^p_{L^\infty(0,T;L^2(\Gamma_c))} \int_0^T \|\lambda_2(t)\|^p_{L^2(\Gamma_c)} \, \mathrm{d}t.$$

As the trace operator is continuous from $H^{1/2}(\Omega)$ to $L^2(\Gamma_c)$, there exists a constant $C_T > 0$ such that

$$\|\varphi\|_{L^2(\Gamma_c)} \le C_T \, \|\varphi\|_{X^{1/2}(\Omega)} \quad \text{for all } \varphi \in H^{1/2}(\Omega).$$

Furthermore, by the interpolation inequality $\|\varphi\|_{H^{1/2}(\Omega)} \le C_I \|\varphi\|^{1/2}_{H^1(\Omega)} \|\varphi\|^{1/2}_{L^2(\Omega)}$ for all $\varphi \in H^1(\Omega)$ holds; see, e.g., [23]. Hence, we have

$$\int_0^T \left| \int_{\Gamma_c} \alpha\lambda_2 \, \mathrm{d}x \right|^p \, \mathrm{d}t \le (C_T C_I)^p \, \|\alpha\|^p_{C([0,T];L^2(\Gamma_c))} \int_0^T \|\lambda_2(t)\|^{p/2}_{H^1(\Omega)} \|\lambda_2(t)\|^{p/2}_{L^2(\Omega)} \, \mathrm{d}t.$$

Since λ_2 is an element of $W(0,T)$, we can estimate the second term in the integral on the right-hand side by $\|\lambda_2\|^{p/2}_{L^\infty(0,T;L^2(\Omega))}$ and find that the right-hand side remains finite for $1 \le p \le 4$. ∎

Next we turn to the semismooth Newton method, which—in contrast to Algorithm 4.1—also linearizes the mapping $u \mapsto \mathcal{S}(u)$. Moreover, in every iteration of the semismooth Newton method, only one Newton step toward the solution of step (4) is carried out. For details, we refer to [10]. Commonly, condition (4.2a) is linear in u and λ (and y), so δu^k does not appear in its linearization; see, for example, [19].

Finally, we mention that the choice of $\sigma = \gamma$ is only of theoretical interest. In the numerical implementation, usually σ is set to a larger value in order to prevent optimization variables to jump from the upper to the lower bound (or vice versa) in consecutive iterations; see [31].

5. Numerical examples. In this section, we describe the behavior of the primal-dual and the semismooth Newton methods by means of some examples of our penalized problem (P_ε). All coding is done in MATLAB using routines from the FEMLAB 2.2 package concerning the finite element implementation. The given CPU times were obtained on a standard 1700-MHz desktop PC. They include only the run time for the core algorithm, excluding the generation of the mesh, precomputing integrals, and incomplete Cholesky decompositions. The three-dimensional geometry of the problem is given by the annular cylinder between the planes $z = 0$ and $z = 0.5$ with inner radius 0.4 and outer radius 1.0 whose rotational axis is the z-axis (Figure 5.2). The control boundary Γ_c is the upper annulus, and we use the control shape function

$$\alpha(t,x) = \exp(-5[(x - 0.7\cos(2\pi t))^2 + (y - 0.7\sin(2\pi t))^2]); \tag{5.1}$$

see Figure 5.1. Note that α corresponds to a nozzle circling for $t \in [0,1]$ once around in counterclockwise direction at a radius of 0.7. For fixed t, α is a function which decays exponentially with the square of the distance from the current location of the nozzle.

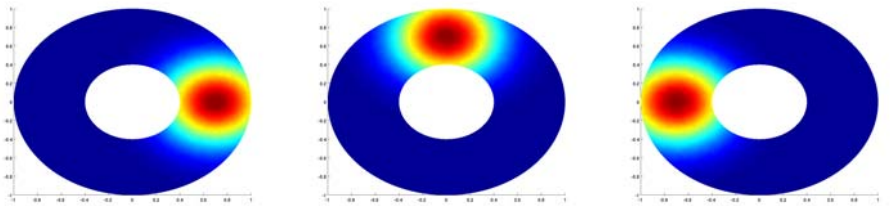


FIG. 5.1. *Control shape function* $\alpha(t,x)$ *at* $t = 0.0$, $t = 0.25$, *and* $t = 0.5$.

The "triangulation" of the domain Ω by tetrahedra is also shown in Figure 5.2. It was created using an initial mesh obtained from meshinit(fem, "Hmax," 0.4). As the geometry suggests that much of the reaction will take place near the top surface Γ_c of the annular cylinder, we refine this initial mesh near the top using the command meshinit(fem, "Hexpr," "0.4*(0.5 - z)+0.10,"...). The final mesh consists of 1797 points and 7519 tetrahedra. In the time direction, we use $T = 1$ and partition the interval into 100 subintervals of equal lengths. The values of the control u at the interval endpoints serve as optimization variables. We use the semi-implicit Euler time integration scheme for the state equations (2.3a)–(2.3b), where the nonlinearities are treated as explicit terms, i.e., they are always taken from the previous time step. In the adjoint equation, we use the same semi-implicit scheme, i.e., the right-hand side terms in (3.3a)–(3.3b) are taken from the previously computed time step. The elliptic problem which arises on each time level in the state and adjoint equations is solved

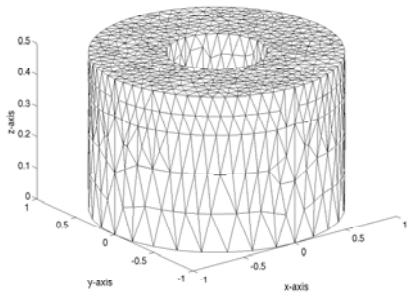


FIG. 5.2. *Domain $\Omega \subset \mathbb{R}^3$ and its triangulation with tetrahedra.*

using the conjugate gradient method with incomplete Cholesky preconditioning. Note that the preconditioner needs to be computed only once since the coefficient matrices are the same in each time step, provided the time step lengths are all identical.

The gradient of the reduced cost functional $\nabla \hat{J}_\varepsilon$ given in (4.1) is then assembled using the precomputed expressions $\int_{\Gamma_c} \alpha(t^i, x)\,\mathrm{d}x$. This strategy clearly follows the paradigm of optimize-then-discretize. Consequently, on the discrete level, the gradient and also the Hessian of the reduced cost functional are evaluated only to a certain accuracy, which depends on the level of discretization. Hence, even if the reduced Hessian system (4.5) was solved to full machine precision, the discrete solution of (3.6) would in general only be found up to a residual whose size depends on the level of discretization. For the discretization parameters given above, we used $\|\nabla \hat{J}_\varepsilon(u^n)_{|I^n}\|_{L^2(0,T)} \le 3 \times 10^{-3}$ as a termination criterion, which is evaluated by setting to zero the components of $\nabla \hat{J}_\varepsilon(u^n)$ which correspond to either of the active sets A_+^n or A_-^n. Of course, coincidence of the active sets on two consecutive iterations is also required for the algorithm to terminate.

5.1. Example 1. In the first example, we use the uniform control bounds

$$u_a \equiv 1, \qquad u_b \equiv 5. \tag{5.2}$$

Controlling the second substance, we wish to steer the concentration of the first substance to zero at the terminal time $T = 1$, i.e., we choose

$$\beta_1 = 1, \quad \beta_2 = 0, \quad c_{1T} \equiv 0. \tag{5.3}$$

The control cost parameter is $\gamma = 10^{-2}$, and the desired control is $u_d = 0$.

The chemical reaction is governed by (2.3a)–(2.3b) with parameters

$$d_1 = 0.15, \quad d_2 = 0.20, \quad k_1 = 1.0, \quad k_2 = 1.0. \tag{5.4}$$

As initial concentrations, we use

$$c_{10} \equiv 1.0, \qquad c_{20} \equiv 0.0. \tag{5.5}$$

The discrete optimal solution without integral constraint (2.11) yields

$$\int_0^T \hat{u}^\varepsilon(t)\,\mathrm{d}t = 4.2415, \qquad \hat{J}_\varepsilon(\hat{u}^\varepsilon) = 0.3186. \tag{5.6}$$

TABLE 5.1
Example 1 *for* $\varepsilon = 1$ *solved with the primal-dual active set method.*

n	$\lvert A_+^n\rvert$	$\lvert A_-^n\rvert$	$\lVert\nabla\hat{J}_\varepsilon(u^n)_{\lvert I^n}\rVert$	#CG	$\lVert r\rVert$	α
1	0	0	5.8320E–02	2	2.3835E–03	5.0000E–01
			3.2408E–02	3	7.2951E–04	1.0000E+00
			4.6224E–03	4	1.7267E–05	1.0000E+00
			2.3110E–04			
2	15	15	7.3888E–03	3	4.0316E–05	1.0000E+00
			1.0929E–03			
3	19	15	3.3424E–03	3	1.4936E–06	1.0000E+00
			1.9500E–04			
Run time: 527 s						Objective: 0.3256
						$\int_0^T u(t)\,\mathrm{d}t = 3.5803$

In order for this constraint to become relevant, we choose $u_c = 3.5$ and enforce it using the penalization parameter $\varepsilon = 1$. Below, we also study the dependence on ε of the solution and of the performance of the algorithms.

The numerical solution is obtained once using the primal-dual (Table 5.1) and once using the semismooth Newton method (Table 5.2), from a feasible initial guess $u^0 \equiv 3$. In both cases, the reduced Hessian system (4.5) was solved using the conjugate gradient method, although due to the use of an optimize-then-discretize strategy, the matrix we obtain as an approximation to the reduced Hessian $\nabla^2\hat{J}_\varepsilon(u^n)$ is only approximately symmetric. The discrete analogue of (4.6) is

$$R^\top\nabla^2\hat{J}_\varepsilon(u_i)R\,\delta u_{i|I^n} = -R^\top\nabla\hat{J}_\varepsilon(u_i) \quad \text{for } i \geq 0. \tag{5.7}$$

This linear system of equations is of size 100 minus the cardinality of the active sets, $|A_+^n|$ and $|A_-^n|$. Here, R is derived from the identity matrix by canceling the columns whose indices belong to either of the active sets. Our CG algorithm solves (5.7) by evaluating $\nabla^2\hat{J}_\varepsilon(u_i)\,\delta u_i$ and then disregarding the active components. It also respects the discrete $L^2(0,T)$-inner product. As the solution of (5.7) required only a few CG steps (Tables 5.1 and 5.2), no preconditioning was used here. Our CG method also features a coercivity check: should a direction of negative curvature be encountered, the CG iteration is terminated prematurely and we continue with the line search (see below). Additionally, we refer the reader to [12] for modification techniques of the Hessian matrix in the absence of sufficient coercivity.

The parameter σ which enters step (2) of Algorithm 4.1 (determining the active sets) was chosen as $\sigma = 10^2$. To determine the active and inactive sets A_-^k, A_+^k, and I^k, respectively, we check the corresponding inequalities pointwise at each time gridpoint. The solution is shown in Figures 5.3 and 5.4. Recall that without penalization, the optimal solution yielded an integral value of $\int_0^T \hat{u}^\varepsilon(t)\,\mathrm{d}t \approx 4.24$, while with the penalization parameter $\varepsilon = 1$, we are down to $\int_0^T \hat{u}^\varepsilon(t)\,\mathrm{d}t \approx 3.58$ which only marginally violates the bound of $u_c = 3.5$.

The primal-dual algorithm. In order to achieve local quadratic convergence, the conjugate gradient method inside the primal-dual algorithm for (4.5) was terminated when $\|r\| \leq \eta \cdot \|\nabla\hat{J}_\varepsilon(u^n)\|$, where r denotes the residual, $\|\cdot\|$ stands for the L^2-norm on $[0,T]$, and $\eta = \min\{0.5, \|\nabla\hat{J}_\varepsilon(u^n)\|\}$. Finally, an Armijo backtracking line search was employed in the direction δu^n obtained from (4.5) with trial step

TABLE 5.2
Example 1 *for* $\varepsilon = 1$ *solved with the semismooth Newton method.*

n	$\lvert A_+^n \rvert$	$\lvert A_-^n \rvert$	$\lVert \nabla \hat{J}_\varepsilon(u^n)_{\lvert I^n} \rVert$	#CG	$\lVert r \rVert$
1	0	0	5.8320E–02	4	6.8630E–05
2	15	8	5.4427E–01	3	9.1017E–05
3	22	12	2.3929E–01	2	8.3430E–05
4	19	13	6.4804E–02	2	6.0758E–05
5	19	14	1.3049E–02	1	8.9123E–05
6	19	14	1.2480E–03		
7	18	14	1.2489E–03		
Run time: 674 s					Objective: 0.3256
					$\int_0^T u(t)\,\mathrm{d}t = 3.5828$

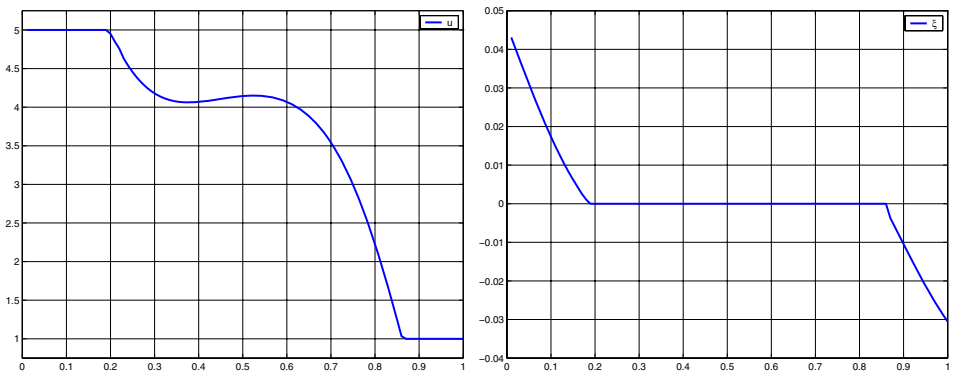


FIG. 5.3. *Example* 1*: optimal control* u *(left) and control constraint multiplier* ξ *for* $\varepsilon = 1$.

lengths $\alpha_k = 2^{-k}$, $k = 0, 1, 2$, etc. A step of length α_k was accepted whenever $\hat{J}_\varepsilon(u^n + \alpha_k \delta u^n) \le \hat{J}_\varepsilon(u^n) + 10^{-4} \cdot \frac{\mathrm{d}}{\mathrm{d}\alpha} \hat{J}_\varepsilon(u^n + \alpha \delta u^n)|_{\alpha=0}$.

The semismooth Newton method. In the case of the semismooth Newton method (Table 5.2), the reduced Hessian system (4.5) was also solved inexactly. Here, the conjugate gradient iteration was terminated when $\lVert r \rVert \le 3 \times 10^{-4}$. That is, we use a constant threshold here which does not depend on the progress in the outer iteration.

Comparing the solutions. Both methods obtained the solution depicted in Figures 5.3 and 5.4 within 15 CG iterations in the case of the primal-dual active set method and 12 CG iterations in the case of the semismooth Newton method. Note, however, that the final residual $\lVert \nabla \hat{J}_\varepsilon(u^n)_{\lvert I^n} \rVert$ achieved in the primal-dual run is smaller since the residual only scarcely falls short of the desired tolerance of 3×10^{-3} in the next to last Newton step. The primal-dual and the semismooth Newton methods appear equally well-suited for this problem. They proved to be robust with respect to the choice of the initial guess.

Dependence on the penalty parameter ε. For our chosen $\gamma = 10^{-2}$ and $\varepsilon = 200$ the terms in the cost functional involving the control variable and the penalized integral constraint have the same weights. In the case of smaller values for $\varepsilon > 0$ (see our tests below) the penalization term becomes more significant. As mentioned

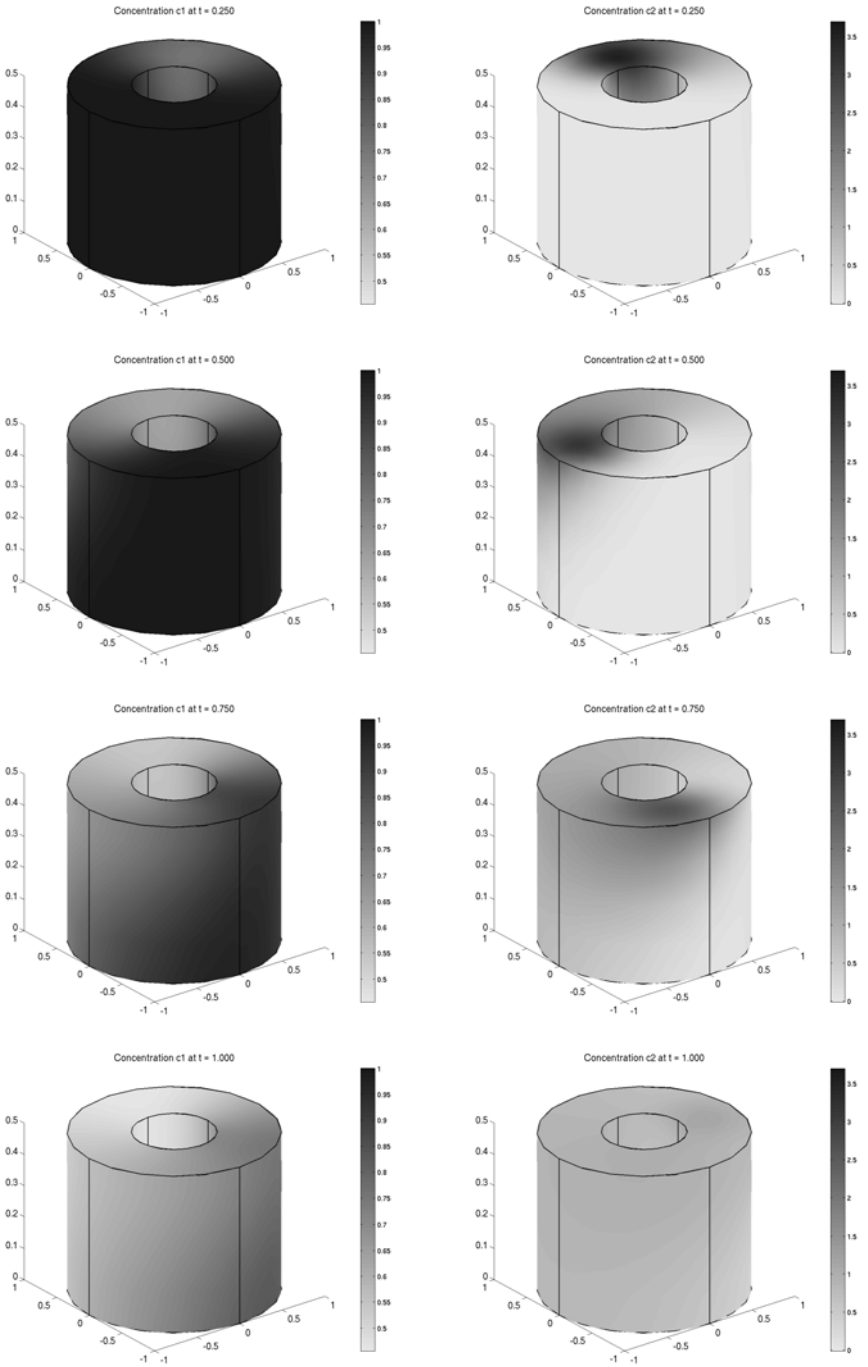


FIG. 5.4. *Example* 1 *for* $\varepsilon = 1$*: concentrations of substances* 1 *(left) and* 2 *(right) at times* $t = 0.25$*,* $t = 0.50$*,* $t = 0.75$*, and* $t = 1.00$*.*

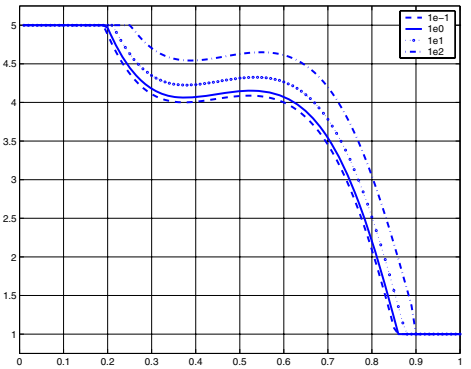


FIG. 5.5. *Example* 1*: optimal control* u *for penalty parameters* $\varepsilon = 100, 10, 1, 0.1$.

TABLE 5.3
Example 1 *for various values of* ε *solved with both methods.*

	Primal-dual active set		Semismooth Newton	
ε	$\int_0^T u(t)\,\mathrm{d}t$	#CG	$\int_0^T u(t)\,\mathrm{d}t$	#CG
100	3.9811	8	3.9865	9
10	3.7203	12	3.7214	11
1	3.5803	15	3.5828	12
0.1	3.5281	18	3.5266	15
0.01	3.5086	21	3.5091	16

in Proposition 2.9, the solutions of the penalized problem (P_ε) converge weakly to a feasible solution for the problem with strict integral constraint (P) as $\varepsilon \searrow 0$. This can be observed from Figure 5.5 and Table 5.3. One also notices that the problem becomes numerically more challenging as ε approaches zero.

5.2. Example 2. The second example was designed to be numerically more challenging. Recall that in the first example, the integral constraint was chosen such that it required a rather moderate modification of the optimal solution without integral constraint. To obtain a contrasting situation, we now use $\gamma = 10^{-3}$ as a weight for the control cost and $u_c = 1.8$ as a bound in the integral constraint. All other data are taken over from Example 1, including $\varepsilon = 1$. Note that the smaller γ leads to an increased control action, while the lowered integral constraint bound has the opposite effect. The optimal solution is depicted in Figure 5.6. It has been obtained with the primal-dual active set method as described for Example 1. One immediately observes that the active sets have an interesting structure in this example. In particular, when the control enters the lower bound for the first time, the constraint is only very weakly active. That is, the corresponding multiplier is close to zero, which makes proper identification of the active sets a challenge for both the primal-dual and semismooth Newton methods.

Table 5.4 shows the convergence history of the primal-dual active set algorithm again from an initial guess of $u_0 \equiv 3$. The angle between the search direction and the negative reduced objective's gradient is also given. The step length in the line search algorithm was always 1.

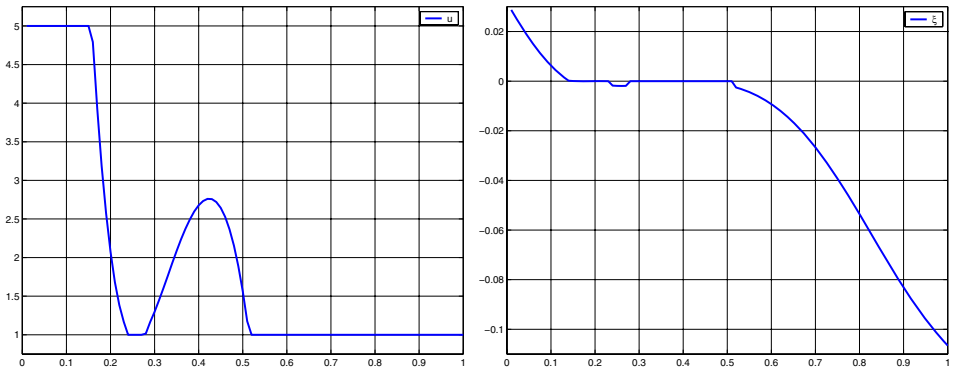


FIG. 5.6. *Example* 2*: optimal control* u *(left) and control constraint multiplier* ξ *for* $\varepsilon = 1$.

TABLE 5.4
Example 2 *with* $\varepsilon = 1$ *solved with the primal-dual active set method.*

n	$\lvert A_+^n\rvert$	$\lvert A_-^n\rvert$	$\lVert\nabla \hat{J}_\varepsilon(u^n)_{\lvert I^n}\rVert$	#CG	$\lVert r\rVert$	Angle	Objective	Penalty
1	0	0	4.2473E+00	1	4.1407E–02	0.00°	6.0826E–01	2.2882E–01
			1.0379E+00	1	4.4998E–02	0.00°	4.4013E–01	3.6003E–02
			2.3973E–01	1	4.7176E–02	0.00°	4.2334E–01	9.0226E–03
			5.8519E–02	6	2.3392E–03	52.52°	2.4420E–01	2.0718E–03
			1.2424E–02	7	7.7037E–05	36.58°	2.2946E–01	1.2367E–03
			2.2656E–03					
2	52	21	5.3326E+00	1	1.5598E–02	0.00°	1.1638E+00	8.2217E–01
			1.3115E+00	1	2.3360E–02	0.00°	5.1497E–01	1.1913E–01
			3.0751E–01	1	2.7113E–02	0.00°	4.4900E–01	2.5718E–02
			6.4022E–02	2	2.4967E–03	62.19°	3.8344E–01	7.5459E–03
			8.0378E–03	5	5.4365E–06	61.35°	3.8051E–01	5.9262E–03
			1.6748E–04					
3	17	38	4.9186E+00	1	8.4897E–03	0.00°	8.2642E–01	4.9814E–01
			1.2121E+00	1	9.4278E–03	0.00°	4.3181E–01	7.1600E–02
			2.8619E–01	1	9.9664E–03	0.00°	3.9097E–01	1.4939E–02
			5.7817E–02	2	2.7543E–03	76.73°	3.8068E–01	6.5030E–03
			6.4547E–03	7	1.5756E–05	66.44°	3.7873E–01	5.5492E–03
			5.4716E–05					
4	18	35	7.0924E–01	1	5.0034E–03	0.00°	4.0114E–01	3.6886E–02
			1.5946E–01	1	5.6519E–03	0.00°	3.8642E–01	1.0484E–02
			2.7040E–02	3	5.5633E–04	76.80°	3.8303E–01	6.3845E–03
			1.6097E–03					
5	14	50	2.1543E–01	1	1.2975E–03	0.00°	3.8592E–01	1.3592E–02
			4.0717E–02	1	1.3730E–03	0.00°	3.8428E–01	7.1675E–03
			3.9585E–03	7	3.8834E–06	72.65°	3.8402E–01	6.5111E–03
			3.7167E–05					
6	15	52	3.5305E–02	1	6.8442E–04	0.00°	3.8416E–01	7.0777E–03
			3.1016E–03	5	8.4937E–07	75.66°	3.8410E–01	6.5734E–03
			2.8178E–05					
7	15	53	6.0990E–03	2	2.2282E–05	49.47°	3.8410E–01	6.6022E–03
			1.2872E–04					
Run time: 2278 s							Objective: 0.3841	
							$\int_0^T u(t)\,\mathrm{d}t = 1.9876$	

We observe that significantly more outer iterations are necessary in order to determine the active sets. Also, in every first Newton step, the primal-dual method starts over with a fairly large contribution of the penalty part $I(u)/\varepsilon$ to the objective. The first Newton steps taken all aim to reduce mainly this part of the objective. This can be seen from the zero angle between the search direction and the negative objective gradient. In other words, the reduced Hessian matrix is a multiple of the identity matrix, being dominated by $\nabla^2 I(u)/\varepsilon$ (see (2.12)).

Acknowledgment. The authors would like to thank Georg Stadler for helpful comments.

REFERENCES

[1] M. Bergounioux, M. Haddou, M. Hintermüller, and K. Kunisch, *A comparison of interior point methods and a Moreau–Yosida based active set strategy for constrained optimal control problems,* SIAM J. Optim., 11 (2000), pp. 495–521.

[2] M. Bergounioux, K. Ito, and K. Kunisch, *Primal-dual strategy for constrained optimal control problems,* SIAM J. Control Optim., 37 (1999), pp. 1176–1194.

[3] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods,* Academic Press, New York, 1982.

[4] J. F. Bonnans and H. Zidani, *Optimal control problems with partially polyhedric constraints,* SIAM J. Control Optim., 37 (1999), pp. 1726–1741.

[5] E. Casas, J.-P. Raymond, and H. Zidani, *Pontryagin's principle for local solutions of controlproblems with mixed control-state constraints,* SIAM J. Control Optim., 39 (2000), pp. 1182–1203.

[6] L. C. Evans, *Partial Differential Equations,* Grad. Stud. Math. 19, American Mathematical Society, Providence, RI, 1998.

[7] C. Geiger and C. Kanzow, *Theorie und Numerik restringierter Optimierungsaufgaben,* Springer-Verlag, Berlin, 2002.

[8] R. Griesse, *Parametric sensitivity analysis in optimal control of a reaction diffusion system—Part* II*: Practical methods and examples,* Optim. Methods Softw., 19 (2004), pp. 217–242.

[9] R. Griesse and S. Volkwein, *Analysis for Optimal Boundary Control for a Three-Dimensional Reaction-Diffusion System,* Technical report 286, Special Research Center F 003 Optimization and Control, Project area Continuous Optimization and Control, University of Graz & Technical University of Graz, 2003 (see http://www.uni-graz.at/imawww/volkwein/publist.html).

[10] R. Griesse and S. Volkwein, *A semi-smooth Newton method for optimal boundary control of a nonlinear reaction-diffusion system*, in Proceedings of the Sixteenth International Symposium on Mathematical Theory of Networks and Systems (MTNS), Leuven, Belgium, 2004.

[11] W. W. Hager and G. Ianculescu, *Dual approximations in optimal control,* SIAM J. Control Optim., 22 (1984), pp. 423–465.

[12] M. Hintermüller, *On a globalized augmented Lagrangian-SQP algorithm for nonlinear optimal control problems with box constraints,* in Fast Solution of Discretized Optimization Problems, K.-H. Hoffmann, R. H. W. Hoppe, and V. Schulz, eds., Internat. Ser. Numer. Math. 138, Birkhäuser, Basel, 2001, pp. 139–153.

[13] M. Hintermüller, *A primal-dual active set algorithm for bilaterally control constrained optimal control problems,* Quart. Appl. Math., 61 (2003), pp. 131–160.

[14] M. Hintermüller and M Hinze, *Globalization of SQP methods in control of the instationary Navier–Stokes equations,* Math. Model. Numer. Anal., 36 (2002), pp. 725–746.

[15] M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semi-smooth Newton method,* SIAM J. Optim., 13 (2003), pp. 865–888.

[16] M. Hintermüller and G. Stadler, *A semi-smooth Newton method for constrained linear-quadratic control problems,* ZAMM Z. Angew. Math. Mech., 83 (2003), pp. 219–237.

[17] D. Hömberg and J. Sokolowski, *Optimal control of laser hardening,* Adv. Math. Sci. Appl., 8 (1998), pp. 911–928.

[18] K. Ito and K. Kunisch, *Augmented Lagrangian methods for nonsmooth convex optimization in Hilbert spaces,* Nonlinear Anal., 41 (2000), pp. 573–589.

[19] K. Ito and K. Kunisch, *The primal-dual active set method for nonlinear problems with bilateral constraints*, SIAM J. Control Optim., 43 (2004), pp. 357–376.

[20] K. Kunisch and J. C. de los Reyes, *A semismooth Newton method for control-constrained boundary optimal control of the Navier–Stokes equations*, Nonlinear Anal., to appear.
[21] K. Kunisch and A. Rösch, *Primal-dual strategy for constrained parabolic optimal control problems,* SIAM J. Optim., 13 (2002), pp. 321–334.
[22] F. Leibfritz and E. W. Sachs, *Inexact SQP interior point methods and large scale optimal control problems,* SIAM J. Control Optim., 38 (1999), pp. 272–299.
[23] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications,* Vol. 1, Springer-Verlag, New York, 1972.
[24] Z. Q. Luo, J. S. Pang, and D. Ralph, *Mathematical Programs with Equilibrium Constraints,* Cambridge University Press, Cambridge, UK, 1996.
[25] H. Maurer and J. Zowe, *First and second order necessary and sufficient optimality conditions for infinite-dimensional programming problems,* Math. Program., 16 (1979), pp. 98–110.
[26] J. Nocedal and S. J. Wright, *Numerical Optimization,* Springer Ser. Oper. Res., Springer-Verlag, New York, 1999.
[27] J. P. Raymond and H. Zidani, *Hamiltonian Pontyagin's principles for control problems governed by semilinear equations,* Appl. Math. Optim., 39 (1999), pp. 143–177.
[28] S. M. Robinson, *Stability theorems for systems of inequalities, Part* II*: Differentiable nonlinear systems,* SIAM J. Numer. Anal., 13 (1976), pp. 497–513.
[29] R. T. Rockafellar, *Conjugate Duality and Optimization,* Regional Conf. Ser. in Appl. Math., 1974.
[30] A. Rösch and F. Tröltzsch, *Sufficient second-order optimality conditions for a parabolic optimal control problem with pointwise control-state constraints,* SIAM J. Control Optim., 42 (2003), pp. 138–154.
[31] G. Stadler, *Semi-smooth Newton methods and augmented Lagrangian methods for a simplified friction problem*, SIAM J. Optim., 15 (2004), pp. 39–62.
[32] H. Tanabe, *Functional Analytic Methods for Partial Differential Equations,* Dekker, New York, 1996.
[33] R. Temam, *Navier–Stokes Equations,* North-Holland, Amsterdam, 1979.
[34] F. Tröltzsch and S. Volkwein, *The SQP method for bilaterally control constrained optimal control of the Burgers equation,* ESAIM Control Optim. Calc. Var., 6 (2001), pp. 649–674.
[35] M. Ulbrich, *Semismooth Newton methods for operator equations in function spaces,* SIAM J. Optim., 13 (2003), pp. 805–842.
[36] S. Volkwein, *Second-order conditions for boundary control problems of the Burgers equation,* Control Cybernet., 30 (2001), pp. 249–278.
[37] S. Volkwein, *Optimal Control of Laser Surface Hardening by Utilizing a Nonlinear Primal-Dual Active Set Strategy,* Technical report 277, Special Research Center F 003 Optimization and Control, Project area Continuous Optimization and Control, University of Graz & Technical University of Graz, 2003, submitted.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 2, pp. 495–514

INDIVIDUAL Q-LEARNING IN NORMAL FORM GAMES*

DAVID S. LESLIE† AND E. J. COLLINS‡

Abstract. The single-agent multi-armed bandit problem can be solved by an agent that learns the values of each action using reinforcement learning. However, the multi-agent version of the problem, the iterated normal form game, presents a more complex challenge, since the rewards available to each agent depend on the strategies of the others. We consider the behavior of value-based learning agents in this situation, and show that such agents cannot generally play at a Nash equilibrium, although if smooth best responses are used, a Nash distribution can be reached. We introduce a particular value-based learning algorithm, which we call individual Q-learning, and use stochastic approximation to study the asymptotic behavior, showing that strategies will converge to Nash distribution almost surely in 2-player zero-sum games and 2-player partnership games. Player-dependent learning rates are then considered, and it is shown that this extension converges in some games for which many algorithms, including the basic algorithm initially considered, fail to converge.

Key words. reinforcement learning, normal form games, stochastic approximation, multi-agent learning, player-dependent learning rates

AMS subject classifications. 93E35, 91A26, 68T05, 62L20

DOI. 10.1137/S0363012903437976

1. Introduction.

We will study value-based learning agents in the multi-agent multi-armed bandit problem (more commonly known as an iterated normal form game). These simple agents adapt in a very similar manner to the reinforcement learning algorithm used by Sutton and Barto (1998) in the single-agent case. However, the multi-agent setting presents a significantly more complex problem, since the rewards available to each agent depend on the strategies of all the other learners and hence are not stationary.

In this paper we will show how to study the asymptotic behavior of these agents using the ODE method of stochastic approximation. The resulting dynamical systems can be analyzed to prove that convergence to equilibrium must occur for 2-player zero-sum games and 2-player partnership games (Proposition 4.2). However, it is shown that convergence does not occur for some specific games known to cause difficulties for all learning algorithms; player-dependent learning rates are then introduced, and convergence can then be proved to occur in a larger class of games (Proposition 5.4 and Corollary 5.6).

One of the best studied models of adaptation in iterated normal form games is fictitious play (Brown (1951), Fudenberg and Levine (1998)). However, this paradigm requires each player to know her own reward function, to observe the actions of all players, and to calculate expected rewards from this information. While this is realistic in some situations, it is clear that players of a game such as the stock

*Received by the editors December 1, 2003; accepted for publication (in revised form) December 8, 2004; published electronically August 31, 2005.
http://www.siam.org/journals/sicon/44-2/43797.html
†School of Economics, The University of New South Wales, Sydney, NSW 2052, Australia (d.leslie@unsw.edu.au). This research was completed while this author was at the University of Bristol, and was supported by CASE Research Studentship 00317214 from the UK Engineering and Physical Sciences Research Council in cooperation with BAE SYSTEMS.
‡Department of Mathematics, University Walk, Bristol BS8 1TW, UK (e.j.collins@bristol.ac.uk).

market, or animals involved in evolutionary games (Maynard Smith (1982)), do not know the relevant reward structures, while in potential applications of multi-agent learning (e.g., Boyan and Littman (1994), Singh and Bertsekas (1997), Crites and Barto (1998)) these requirements are also frequently not satisfied.

Therefore we study models of adaptation under which agents simply respond to observations of the rewards they receive for playing actions, as in the simple and successful approach used by Sutton and Barto (1998) to solve the single-agent multi-armed bandit problem. Agents estimate the value of each action, and play a strategy that is a simple function of these estimates; we will call such agents *value-based learners*. These learners take no account of the presence of other players of the game—the only information used is the reward received for each action played. Recent neurophysiological data from rhesus monkeys trying to perform a repetitive task (Glimcher, Dorris, and Bayer (2005)) suggests that value-based learning is actually employed in nature, adding extra interest to our investigations.

Other current approaches to learning in games include actor-critic learning, hypothesis-based learning, consistent learning, and alternative reinforcement learning models based on Roth and Erev (1995). Our value-based learning scheme compares favorably with these in terms of simplicity, form of convergence, and structure, respectively.

The Roth and Erev (1995) model of learning is the standard model of reinforcement learning in the game theory community. However, it has the disadvantages that it requires all rewards to be positive, and it uses rewards as "reinforcement signals" instead of information about the values of actions, despite the fact that strategies and values are "dimensionally different" quantities—strategies are probability distributions whereas rewards are arbitrary real numbers (the same criticism can be made of stimulus-response learning (Börgers and Sarin (1997))).

Actor-critic learning is a more sophisticated framework that explicitly links rewards and strategies. In this model, agents maintain separate value functions and strategies and map the value function to strategy space in order to update the strategy. It has been used in several recent approaches to learning in games (Borkar (2001), Bowling and Veloso (2002), Leslie and Collins (2003)), but, in contrast with the value-based approach considered in this paper, the extra complication involved in separating the values and the strategies is unnecessary and overly sophisticated when applied in a single-agent setting.

A different approach to the problem of learning in games has been to develop consistent (Hannan (1957)) reinforcement learning procedures (Baños (1968), Megido (1980), Auer et al. (1995), Hart and Mas-Colell (2001)). Hart and Mas-Colell (2000) show that the long-run average actions of players using a consistent algorithm will converge to a correlated equilibrium of the game (Aumann (1974)) (as opposed to a classical Nash equilibrium). However, it is often difficult to characterize the correlated equilibria (Fudenberg and Tirole (1991)), and the convergence is in the sense of the average action played, instead of convergence of actual play as developed in this paper.

An even more sophisticated framework is the hypothesis-testing formulation recently proposed by Foster and Young (2003). Here the players formulate hypotheses about opponent strategies and repeatedly test these hypotheses against observed play. It is shown that play is close to a Nash equilibrium of the repeated game formulation most of the time. However, the sophistication required from the players is greater than for virtually all other algorithms in the literature, and the scheme is clearly not applicable in the minimal information setting we consider here.

It is appropriate to comment here on a recent result of Hart and Mas-Colell (2003) showing that no "uncoupled" adaptive dynamics can converge to equilibrium in all games. The dynamics we consider in this paper are certainly uncoupled, and indeed the dynamics resulting from any system where players do not observe opponent payoffs must necessarily be uncoupled. However, as we shall see later, convergence to Nash distribution may be a more attainable goal than convergence to Nash equilibrium (see section 2 for details). Furthermore, the player-dependent learning rates introduced in section 5 are a further step away from the framework studied by Hart and Mas-Colell (2003), thus allowing convergence to be proved for a larger class of games than has been previously achieved.

This paper is therefore an analysis of a simple and intuitive learning procedure applied in the setting of normal form games with minimal information available to the players. In section 2 we discuss some difficulties faced by value-based learners in games, and propose the use of smooth best responses (also known as softmax action selection). This motivates our individual Q-learners, introduced in section 3, where we show how to characterize their behavior using stochastic approximation (Benaïm (1999)). The behavior of these learners in 2-player games is analyzed in section 4, where we show that strategy evolution is closely related to the smooth best response dynamics (Hofbauer and Hopkins (2005)); this is the same dynamical system that characterizes stochastic fictitious play (Benaïm and Hirsch (1999)), despite the fact that individual Q-learning uses significantly less information than stochastic fictitious play. However, previous work (Leslie and Collins (2003)) suggests that convergence to a fixed point can be proved to occur in a larger class of games if player-dependent learning rates are used to break symmetry between the players; this is studied in section 5. A problem with player-dependent learning rates is that the analytical methods so far developed do not apply for all games; section 6 uses graphical representations of games to investigate classes of games in which player-dependent learning rates can be analyzed.

2. Value-based players in games. In this section we will introduce our notation and discuss some problems faced by value-based players of games. In particular, we will show that value-based players cannot generally play at a Nash equilibrium, but if smooth best responses are used, equilibrium play becomes possible (although this will no longer be at the classical Nash equilibrium).

We start by introducing our notation and presenting a familiar example. A normal form game consists of N players, where each player $i \in \{1, \dots, N\}$ has a finite set A^i of actions and a reward function $r^i : A^1 \times \cdots \times A^N \to \mathbb{R}$. When the game is played, each player $i \in \{1, \dots, N\}$ selects an action $a^i \in A^i$ and then receives a reward which has expected value $r^i(a^1, \dots, a^N)$; each player tries to maximize her expected reward. A traditional 2-player example is rock-scissors-paper, where the action set for each player is {Rock, Scissors, Paper}, and the reward functions are given in the payoff matrix

$$\begin{array}{c} \\ \text{Rock} \\ \text{Scissors} \\ \text{Paper} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Rock} & \text{Scissors} & \text{Paper} \end{array} \\ \begin{pmatrix} (0,0) & (1,-1) & (-1,1) \\ (-1,1) & (0,0) & (1,-1) \\ (1,-1) & (-1,1) & (0,0) \end{pmatrix} \end{array}, \tag{2.1}$$

where player 1's action determines the row, player 2's action determines the column, and an entry (x, y) means that player 1 receives reward x and player 2 receives reward y. Note that this is a 2-player zero-sum game, where for any joint action (a^1, a^2) the

rewards satisfy $r^1(a^1, a^2) + r^2(a^1, a^2) = 0$. We will also have occasion to consider partnership games, where for any joint action the reward given to each player is the same.

A mixed strategy for player i is an element $\pi^i \in \Delta^i$, where Δ^i is the set of probability distributions over the action space A^i; we will write $\pi^i(a^i)$ for the probability that player i selects action a^i when using strategy π^i. There are unique multilinear extensions of the reward functions, also denoted r^i, to the joint strategy space $\Delta = \Delta^1 \times \cdots \times \Delta^N$, and in further abuse of notation we will write $r^i(a^i, \pi^{-i})$ (resp., $r^i(\pi^i, \pi^{-i})$) for the expected reward that player i will receive if she plays action a^i (resp., strategy π^i) and the other players select actions according to the opponent mixed strategy $\pi^{-i} = (\pi^1, \ldots, \pi^{i-1}, \pi^{i+1}, \ldots, \pi^N)$.

Nash (1950) defined the classical solution concept for normal form games by observing that a rational player will not play a strategy π^i that does not maximize her expected reward given the opponent strategy π^{-i}. Thus a Nash equilibrium is a joint strategy $\tilde{\pi} \in \Delta$ satisfying, for each i,

$$r^i(\tilde{\pi}) \geq r^i(\pi^i, \tilde{\pi}^{-i}) \quad \text{for any } \pi^i \in \Delta^i.$$

Nash (1950) showed that at least one Nash equilibrium exists for any normal form game. In our rock-scissors-paper example, it is well known that there is a unique Nash equilibrium where each player plays the mixed strategy $(1/3, 1/3, 1/3)$, but in general there can be many Nash equilibria of a game. The problem of equilibrium selection (i.e., when faced with a game, how should players decide which Nash equilibrium strategy to play when many might exist) can be seen as a motivating factor for the study of learning in games (Fudenberg and Levine (1998)). An alternative perspective is to consider the Nash equilibria of a game as the only points to which a sensible learning procedure should converge—note that if only one player is present, then a Nash equilibrium is simply the action which returns the highest expected reward.

However, value-based approaches, as used by Sutton and Barto (1998) in the single-agent problem, encounter problems with Nash equilibria. At a Nash equilibrium, any action played by player i with positive probability will receive expected reward $\max_{a^i \in A^i} r^i(a^i, \tilde{\pi}^{-i})$, yet the equilibrium strategy might well require these maximizing actions to be played with specific and possibly unequal probabilities. For example, consider the game with payoff matrix

$$\begin{pmatrix} (2,0) & (0,1) \\ (0,2) & (1,0) \end{pmatrix},$$

which has a unique equilibrium where $\pi^1 = (2/3, 1/3)$ and $\pi^2 = (1/3, 2/3)$. At this equilibrium, both actions of both players get expected reward $2/3$, yet player 1 must favor action 1, and player 2 must favor action 2. Thus play at an equilibrium is not possible when the strategies played are restricted to be simple functions of the expected rewards (unless these functions are asymmetric under reordering of the actions).

A solution to this problem lies in using smooth best responses, which can also be considered as arising from a Bayesian uncertainty about the rewards (Harsanyi (1973)) and are closely related to the softmax exploration method of reinforcement learning (Sutton and Barto (1998)). If player i has estimates $Q^i(a^i)$ of the values of actions $a^i \in A^i$, then a smooth best response $\beta^i(Q^i) \in \Delta^i$ to these estimates is

given by

$$\beta^i(Q^i) = \underset{\pi^i \in \Delta^i}{\operatorname{argmax}} \left\{ \sum_{a^i \in A^i} \pi^i(a^i) Q^i(a^i) + \tau v^i(\pi^i) \right\}.$$

Here $\tau > 0$ is a temperature parameter, and $v^i : \Delta^i \to \mathbb{R}$ is a player-dependent smoothing function, which is a smooth, strictly differentiably concave function such that as π^i approaches the boundary of Δ^i, the slope of v^i becomes infinite (Fudenberg and Levine (1998)). As the temperature parameter $\tau \to 0$, $\beta^i(Q^i)$ approaches the set of best responses (i.e., strategies that select only actions a^i maximizing $Q^i(a^i)$). However, while there may be many best responses (e.g., suppose all the Q values are equal), the conditions on v^i imply that there is a unique smooth best response given τ and v^i (Fudenberg and Levine (1998)).

A familiar example of a smooth best response is Boltzmann action selection. Under this scheme, the smoothing functions are

$$v^i(\pi^i) = -\sum_{a^i \in A^i} \pi^i(a^i) \log \pi^i(a^i),$$

resulting in the smooth best response function

$$\beta^i(Q^i)(a^i) = \frac{\mathrm{e}^{Q^i(a^i)/\tau}}{\sum_{b^i \in A^i} \mathrm{e}^{Q^i(b^i)/\tau}}. \tag{2.2}$$

The use of smooth best responses means that, in general, Nash equilibria are no longer fixed points in strategy space, and an alternative equilibrium concept must be defined. (For example, consider a 1-player game with a unique optimal action; the conditions on v^1 mean that all actions are played with positive probability, which is clearly not a Nash equilibrium.) Given a set of smooth best response functions, β^i, we define a *Nash distribution* to be a joint mixed strategy $\pi \in \Delta$ such that, for each i,

$$\pi^i = \beta^i(r^i(\cdot, \pi^{-i})); \tag{2.3}$$

i.e., each player plays a smooth best response to the rewards arising from opponent play. Brouwer's fixed point theorem shows that such distributions must exist; Govindan, Reny, and Robson (2003) show that for small temperatures τ, the Nash equilibria of a game are approximated by Nash distributions (the proof of Harsanyi (1973) is insufficient in this case, because it relies on perturbations of the rewards having compact support). Note that the Nash distributions depend on the smooth best response functions β^i (through the particular choices of τ and v^i) as well as on the reward functions r^i—for the remainder of this paper we will assume that any particular player uses a fixed smooth best response function for all time.

Thus we have shown that a value-based approach cannot result in Nash equilibrium play in general games but can result in strategies that are close to a Nash equilibrium if players use smooth best responses. In the next section we will introduce a value-based learning algorithm incorporating this idea, which can therefore converge to a Nash distribution.

3. Individual *Q*-learning. Sutton and Barto (1998) show that a simple reinforcement learning scheme can be used to estimate action values in a single-agent task. The basic algorithm is given by

$$Q_{n+1}(a) = Q_n(a) + \lambda_{n+1} \mathbb{I}_{\{a_n = a\}} \{R_n - Q_n(a)\} \quad \text{for each } a \in A, \tag{3.1}$$

TABLE 1
Individual Q-learning.

Each player i selects an action a_n^i using the strategy $\beta^i(Q_n^i)$, receives reward R_n^i, and then updates Q_n^i according to

$$Q_{n+1}^i(a^i) = Q_n^i(a^i) + \lambda_{n+1}\mathbb{I}_{\{a_n^i = a^i\}}\frac{R_n^i - Q_n^i(a_n^i)}{\beta^i(Q_n^i)(a_n^i)} \quad \text{for each } a^i \in A^i, \tag{3.2}$$

where $\{\lambda_n\}_{n\geq 1}$ is a deterministic sequence of learning parameters satisfying

$$\sum_{n\geq 1}\lambda_n = \infty, \qquad \sum_{n\geq 1}(\lambda_n)^2 < \infty. \tag{3.3}$$

where a_n is the action selected at time n, and R_n is the subsequent reward. A similar scheme appears to fit recent neurophysiological data measured from the brains of rhesus monkeys learning to perform a repetitive task (Glimcher, Dorris, and Bayer (2005)). In applying (3.1), actions are selected in such a way that each action is played infinitely often, but as $n \to \infty$ the probability of playing any action which does not have a maximal Q value tends to 0. We will study an analogous system in the equivalent multi-agent task; N players are faced with a normal form game, which they play repeatedly, learning Q values by observing rewards. The algorithm we study is given in Table 1.

This scheme was originally suggested by Fudenberg and Levine (1998); it will be noticed that it is very similar to the standard reinforcement learning model (3.1). The first difference is that a player's strategy at any stage of the game is determined by the algorithm (as opposed to the single-agent case, where there is no need to be specific about action choices at any particular play). This is because the rewards observed by a particular player depend crucially on the strategies of the other players, so these strategies must be carefully specified. The second difference is that the reward prediction error $(R_n^i - Q_n^i(a_n^i))$ (Sutton and Barto (1998)) is divided by $\beta^i(Q_n^i)(a_n^i)$, the probability with which a_n^i was selected. This can be viewed as compensating for the fact that actions played with low probability do not receive frequent updates of their Q values, so when they are played any reward prediction error must have greater influence on the Q value than if frequent updates occur. Further, we will see in section 4 that this division by $\beta^i(Q_n^i)(a_n^i)$ results in a system that is closely related to the (well-studied) smooth best response dynamics (Hofbauer and Hopkins (2005)).

The conditions (3.3) on the learning parameters $\{\lambda_n\}_{n\geq 1}$ mean that standard theorems of stochastic approximation can be used (Benaïm (1999)). The interested reader may wish to consult Benaïm and Hirsch (1999) for an introduction to the ODE method of stochastic approximation in the context of game theory. Writing $Q_n = (Q_n^1, \ldots, Q_n^N)$, and writing $\beta^{-i}(Q_n)$ for the opponent mixed strategies resulting from the values Q_n, note that for each i and a^i

$$\begin{aligned}\mathbb{E}[Q_{n+1}^i(a^i) - Q_n^i(a^i) \mid Q_n] &= \lambda_{n+1} \times \beta^i(Q_n^i)(a^i) \times \frac{r^i(a^i, \beta^{-i}(Q_n)) - Q_n^i(a^i)}{\beta^i(Q_n^i)(a^i)} \\ &= \lambda_{n+1}\left\{r^i(a^i, \beta^{-i}(Q_n)) - Q_n^i(a^i)\right\}.\end{aligned}$$

The following lemma follows immediately from the results of Benaïm (1999).

LEMMA 3.1. *The values Q_n resulting from the individual Q-learning algorithm* (3.2) *converge almost surely to a connected internally chain-recurrent set*[1] *of the flow defined by the Q-learning ODE*

$$\frac{\mathrm{d}}{\mathrm{d}t}q_t^i(a^i) = r^i(a^i, \beta^{-i}(q_t)) - q_t^i(a^i) \quad \textit{for each } i \textit{ and } a^i, \tag{3.4}$$

provided that the Q_n remain bounded for all n.

This assumption of boundedness could be dropped if we used either fixed truncation to a bounded space (Kushner and Yin (1997)) or randomly varying truncations (Chen and Zhu (1986)); the former will change the differential equations to be studied a little, while the latter will have no consequence on the analysis of the asymptotic behavior. However, the purpose of this paper is not to investigate such issues, and in the numerical experiments carried out there were no problems with values growing large. Therefore in the interests of space we will be content to keep this as an assumption throughout the rest of the paper.

The first thing to notice about this algorithm is that convergence to a point can occur only at Nash distribution values. This follows from what is essentially a law of large numbers result, saying that convergence to a point can occur only if the expected change at that point is zero. Thus such a point must satisfy $Q^i(a^i) = r^i(a^i, \beta^{-i}(Q))$ for each i and a^i. However, if we write $\pi^i = \beta^i(Q^i)$, this translates to $\pi^i = \beta^i(r^i(\cdot, \pi^{-i}))$, which is precisely the definition of a Nash distribution (2.3). The learning algorithm can therefore be applied blindly in any game, and if convergence occurs, a Nash distribution must have been reached.

Our next step in analyzing this system is to consider whether the values of the players are ever consistent with the structure of the game; for example, if in a zero-sum game the values ever actually sum to zero. Writing

$$\mathcal{B} = \{(r^1(\cdot, \pi^{-1}), \ldots, r^N(\cdot, \pi^{-N})) : \pi \in \Delta\}$$

for the set of values that could arise from a joint mixed strategy, we call values Q_n *asymptotically belief-based* if the limit set of the values is contained in $\mathcal{B}$.

LEMMA 3.2. *The values Q_n resulting from the individual Q-learning algorithm* (3.2) *are almost surely asymptotically belief-based, provided that the Q_n remain bounded for all time.*

Proof. We will rewrite the Q-learning ODE (3.4) to show that $\mathcal{B}$ is a global attractor of the resulting flow, which suffices to show that the values Q_n resulting from (3.2) are asymptotically belief-based (Benaïm (1999)). Start by writing $q_t = (q_t^1(\cdot), \ldots, q_t^N(\cdot))$ and $r_t = (r^1(\cdot, \beta^{-1}(q_t)), \ldots, r^N(\cdot, \beta^{-N}(q_t)))$, so that

$$\frac{\mathrm{d}}{\mathrm{d}t}q_t = r_t - q_t.$$

This can be rewritten as

$$q_t = \mathrm{e}^{-t}q_0 + (1 - \mathrm{e}^{-t})\bar{r}_t,$$

[1] The concept of a chain-recurrent set is central to the ODE method of stochastic approximation and is developed fully in Benaïm (1999). In the interests of space, this theory is therefore not repeated here. Loosely, an internally chain-recurrent set of a flow is an invariant set of the flow such that a trajectory starting at any point of the set can return to its start point without ever leaving the set, when small "jumps" are allowed to be made. It will suffice to know that a connected internally chain-recurrent set of a flow is an invariant set of the flow containing no proper attractors.

where $\bar{r}_t = (\mathrm{e}^t - 1)^{-1} \int_0^t \mathrm{e}^s r_s \, \mathrm{d}s$ is a weighted average of r_s, $0 \le s \le t$. Since $r_s \in \mathcal{B}$ for all s, it follows immediately that $\bar{r}_t \in \mathcal{B}$ for all t. Therefore $q_t \to \mathcal{B}$ for any q_0, and $\mathcal{B}$ is a global attractor of the flow defined by (3.4). □

As well as being interesting in itself, this is crucial to the analysis of the next section, where we relate this value-based learning to the smooth best response dynamics, usually considered to arise from models of learning in which players use significantly more information on the structure of the game and observations of opponent play than is necessary for individual Q-learning.

4. 2-player games. In this section we will relate the Q-learning ODE (3.4) to the smooth best response dynamics in 2-player games, as suggested by Fudenberg and Levine (1998). This will allow us to characterize the limiting behavior of the individual Q-learning algorithm in certain classes of games.

The smooth best response (SBR) dynamics are defined by the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}\pi_t^i = \beta^i(r^i(\cdot, \pi_t^{-i})) - \pi_t^i \quad \text{for each } i \tag{4.1}$$

and have been shown to characterize stochastic fictitious play, in the same sense that (3.4) characterizes individual Q-learning (Benaïm and Hirsch (1999)). Fudenberg and Levine (1998) observe that for 2-player games, if strategies evolve according to these dynamics, the resultant rewards evolve according to the Q-learning ODE (3.4). This will be used in the proof of Lemma 4.1, where we show that a connected internally chain-recurrent set of the flow defined by the Q-learning ODE (3.4) corresponds to a connected internally chain-recurrent set of the SBR dynamics. By Lemma 3.1, this relates the limiting behavior of individual Q-learning with that of stochastic fictitious play, despite the fact that individual Q-learners have no information on the structure of the game and do not observe opponent play, both of which are necessary for stochastic fictitious play.

LEMMA 4.1. *For 2-player games, any connected internally chain-recurrent set of the Q-learning ODE* (3.4) *is of the form*

$$r(\mathcal{C}) := \{(r^1(\cdot, \pi^{-1}), \ldots, r^N(\cdot, \pi^{-N})) : \pi \in \mathcal{C}\},$$

where $\mathcal{C} \subset \Delta$ is a connected internally chain-recurrent set of the flow defined by the SBR dynamics (4.1).

Proof. Let $\mathcal{D}$ denote an arbitrary connected internally chain-recurrent set of the Q-learning ODE (3.4). Benaïm (1999, Proposition 5.3) shows that a set is connected internally chain-recurrent if and only if it is a compact invariant set admitting no proper attractor. Therefore, since the set $\mathcal{B}$ of belief-based values is a global attractor we must have $\mathcal{D} \subset \mathcal{B}$, and $q \in \mathcal{D}$ means that there exists $\pi \in \Delta$ such that $q^i(a^i) = r^i(a^i, \pi^{-i})$ for each i and a^i.

However, suppose π evolves according to the SBR dynamics (4.1), so that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} r^i(a^i, \pi_t^{-i}) &= \sum_{a^{-i} \in A^{-i}} \frac{\partial r^i(a^i, \pi^{-i})}{\partial \pi^{-i}(a^{-i})} \frac{\mathrm{d}}{\mathrm{d}t} \pi_t^{-i}(a^{-i}) \\
&= \sum_{a^{-i} \in A^{-i}} r^i(a^i, a^{-i}) \{\beta^{-i}(r^{-i}(\cdot, \pi_t^i))(a^{-i}) - \pi_t^{-i}(a^{-i})\} \\
&= r^i(a^i, \beta^{-i}(r^{-i}(\cdot, \pi_t^i))) - r^i(a^i, \pi_t^{-i}),
\end{aligned}$$

and the $r^i(a^i, \pi^{-i})$ evolve according to the same ODE as the $q^i(a^i)$. Note that this calculation is valid only for 2-player games.

Therefore trajectories of the Q-learning ODE (3.4) in $\mathcal{B}$ correspond to trajectories of the SBR dynamics (4.1) in Δ, and the invariant set $\mathcal{D}$ of (3.4) must be of the form $r(\mathcal{C})$, where $\mathcal{C} \subset \Delta$ is an invariant set of (4.1). Now since $\mathcal{D}$ admits no proper attractor, it follows that $\mathcal{C}$ admits no proper attractor. $\mathcal{C}$ may consist of several connected components, but any one of them will be a connected internally chain-recurrent set of the flow defined by the SBR dynamics (4.1) with $\mathcal{D} = r(\mathcal{C})$. □

This result allows us to characterize the limit set of the individual Q-learning algorithm in terms of the chain-recurrent sets of the SBR dynamics. These chain-recurrent sets are well studied, since they are the limit set of a stochastic fictitious play process (Benaïm and Hirsch (1999)), and in particular Hofbauer and Hopkins (2005) provide Lyapunov functions for 2-player zero-sum games and 2-player partnership games. This allows us to prove our first main proposition, which gives convergence results for individual Q-learning in these situations.

PROPOSITION 4.2. *In either 2-player zero-sum games, or 2-player partnership games with countably many Nash distributions (given the smooth best responses β^i), strategies of players using the individual Q-learning algorithm* (3.2) *will converge almost surely to a Nash distribution.*

Proof. Lemma 3.1 shows that the Q values converge to a connected internally chain-recurrent set of the Q-learning ODE (3.4). From Lemma 4.1 we know that this is of the form $r(\mathcal{C})$, where $\mathcal{C} \subset \Delta$ is a connected internally chain-recurrent set of flow defined by the SBR dynamics (4.1). In the games we consider, Hofbauer and Hopkins (2005) provide Lyapunov functions for the set of Nash distributions under the SBR dynamics, and this set is isolated (by assumption in the case of partnership games and by a result of Hofbauer and Hopkins (2005) for zero-sum games). Therefore Benaïm (1999, Corollary 6.6) shows that we can assume $\mathcal{C} = \{\tilde{\pi}\}$ where $\tilde{\pi}$ is a Nash distribution. Therefore $Q_n^i \to r^i(\cdot, \tilde{\pi}^{-i})$ for each i, and so $\beta^i(Q_n^i) \to \beta^i(r^i(\cdot, \tilde{\pi}^{-i}))$ by continuity. But $\tilde{\pi}$ is a Nash distribution, so $\beta^i(r^i(\cdot, \tilde{\pi}^{-i})) = \tilde{\pi}^i$, and we have shown that the strategies converge to a Nash distribution. □

We illustrate this convergence with the rock-scissors-paper game (2.1). This is a 2-player zero-sum game, and therefore strategies converge to the unique Nash distribution where all actions are played with probability 1/3 (this is the same as the Nash equilibrium, although for general games the Nash equilibria and Nash distributions do not coincide). In Figure 1 we see that despite erratic initial strategy shifts, the strategy of player 1 appears to be converging to the Nash distribution.

However, this convergent behavior does not occur for all games. Two classic examples of games that cause problems for learning algorithms are Shapley's variant of rock-scissors-paper (Shapley (1964)) and Jordan's 3-player matching pennies game (Jordan (1993)). In both of these games, the SBR dynamics (4.1) admit a unique linearly unstable fixed point, and an asymptotically stable limit cycle, for certain smooth best response functions β^i (Cowan (1992), Benaïm and Hirsch (1999)). We shall not reproduce all of these results for the Q-learning ODE but will show that in Shapley's game, using Boltzmann action selection (2.2), a Hopf bifurcation occurs at the unique Nash distribution as the temperature parameter tends to 0. This shows that for sufficiently small τ the Nash distribution is linearly unstable and a periodic orbit is an attractor. We use the symmetric formulation of Shapley's game, with payoff matrix

$$\begin{pmatrix} (0,0) & (1,0) & (0,1) \\ (0,1) & (0,0) & (1,0) \\ (1,0) & (0,1) & (0,0) \end{pmatrix}, \tag{4.2}$$

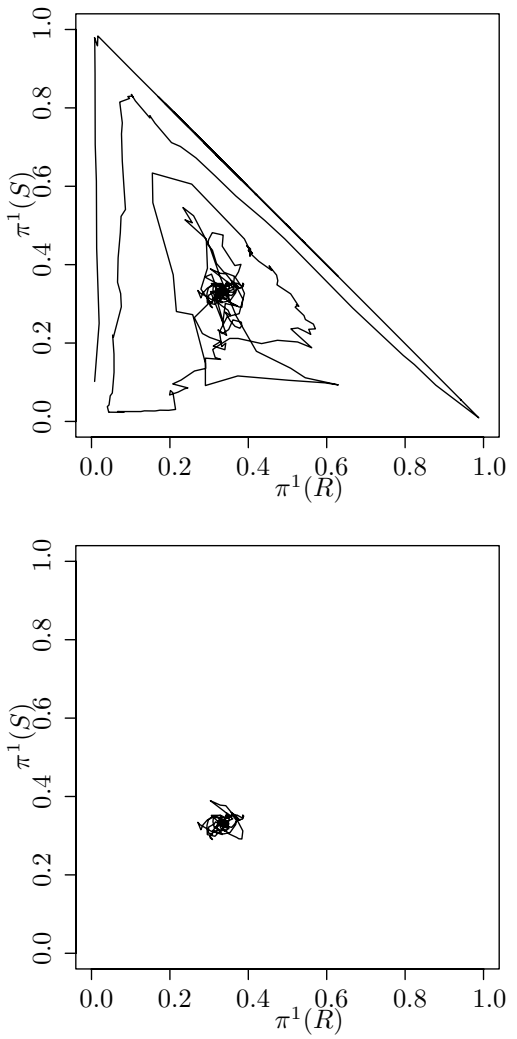


FIG. 1. *Strategies of player* 1 *in the rock-scissors-paper game* (2.1) *with Boltzmann action selection* ($\tau = 0.1$) *over* 5×10^5 *iterations of basic individual Q-learning* (3.2) *with* $\lambda_n = (n+100)^{-0.9}$. *The top diagram shows the entire learning run, whereas the bottom diagram omits the first* 10^4 *iterations, showing that strategies are converging to the unique Nash distribution.*

and therefore, for $i = 1, 2$,

$$\frac{\mathrm{d}}{\mathrm{d}t}Q^i(R) = \pi^{-i}(S) - Q^i(R),$$
$$\frac{\mathrm{d}}{\mathrm{d}t}Q^i(S) = \pi^{-i}(P) - Q^i(S),$$
$$\frac{\mathrm{d}}{\mathrm{d}t}Q^i(P) = \pi^{-i}(R) - Q^i(P).$$

The Jacobian for this system, evaluated at the unique Nash distribution where all Q

values have the value $1/3$, is given by

$$\begin{pmatrix} -1 & 0 & 0 & \frac{-1}{9\tau} & \frac{2}{9\tau} & \frac{-1}{9\tau} \\ 0 & -1 & 0 & \frac{-1}{9\tau} & \frac{-1}{9\tau} & \frac{2}{9\tau} \\ 0 & 0 & -1 & \frac{2}{9\tau} & \frac{-1}{9\tau} & \frac{-1}{9\tau} \\ \frac{-1}{9\tau} & \frac{2}{9\tau} & \frac{-1}{9\tau} & -1 & 0 & 0 \\ \frac{-1}{9\tau} & \frac{-1}{9\tau} & \frac{2}{9\tau} & 0 & -1 & 0 \\ \frac{2}{9\tau} & \frac{-1}{9\tau} & \frac{-1}{9\tau} & 0 & 0 & -1 \end{pmatrix},$$

which has eigenvalues

$$\frac{1}{6\tau}(1 - 6\tau \pm \sqrt{3}i), \quad \frac{1}{6\tau}(-1 - 6\tau \pm \sqrt{3}i), \quad -1, \quad -1.$$

As $\tau \to 0$, the real part of the first conjugate pair crosses the imaginary axis from the negative half-plane to the positive half-plane, resulting in a Hopf bifurcation, so for sufficiently small τ the fixed point is linearly unstable. Therefore by the results of Pemantle (1990), convergence to this fixed point is a probability zero event; in Figure 2 we see that in fact play cycles, as it would under the SBR dynamics, and as implied by the Hopf bifurcation. Previous work (Leslie and Collins (2003)) suggests that using player-dependent learning rates helps to break the symmetry that allows this cycling to occur. We will apply this idea to individual Q-learning in the next section.

5. Player-dependent learning rates. Returning to general N-player games, Leslie and Collins (2003) introduce player-dependent learning rates (PDLR) to break the symmetry that allows strategies to cycle under the SBR dynamics (4.1). Under this paradigm, each player's learning parameters decay to zero at different rates, resulting in a process which is a stochastic approximation of a singularly perturbed dynamical system. The algorithm we study is shown in Table 2; it is a simple extension of individual Q-learning (3.2) which incorporates PDLR. In fact the only difference between this and the individual Q-learning algorithm (3.2) is that each player uses her own sequence of learning parameters $\{\lambda_n^i\}_{n\geq 1}$.

Note that condition (5.2) is used for ease of exposition but is equivalent to the condition that either $\lambda_n^i/\lambda_n^j \to 0$ or $\lambda_n^j/\lambda_n^i \to 0$ whenever $i \neq j$, since if this latter condition is true we can assume that the players are indexed in such a way that (5.2) is true. A suitable choice of learning parameters would be to choose $\lambda_n^i = (n+C)^{-\rho^i}$, where the rate $\rho^i \in (0.5, 1]$ is chosen differently for each player; indeed if players are thought of as selecting their own learning rate ρ^i independently using any continuous distribution on $(0.5, 1]$, then the necessary conditions will be met with probability 1.

The more slowly a player's learning parameters decrease to 0, the more "responsive" that player will be, since greater emphasis is placed on recent observations when estimating an action's value. In contrast, players with more rapidly decreasing learning parameters are more "cautious," since their value estimates take greater account of the entire history of observed rewards. Condition (5.2) means that players with higher indices i are more responsive (and hence less cautious) than those with lower indices.

As observed in Leslie and Collins (2003), in order to analyze an algorithm incorporating PDLR theoretically, we need to make an assumption about what would happen to the more responsive players if the strategies of the $i-1$ most cautious players were fixed.

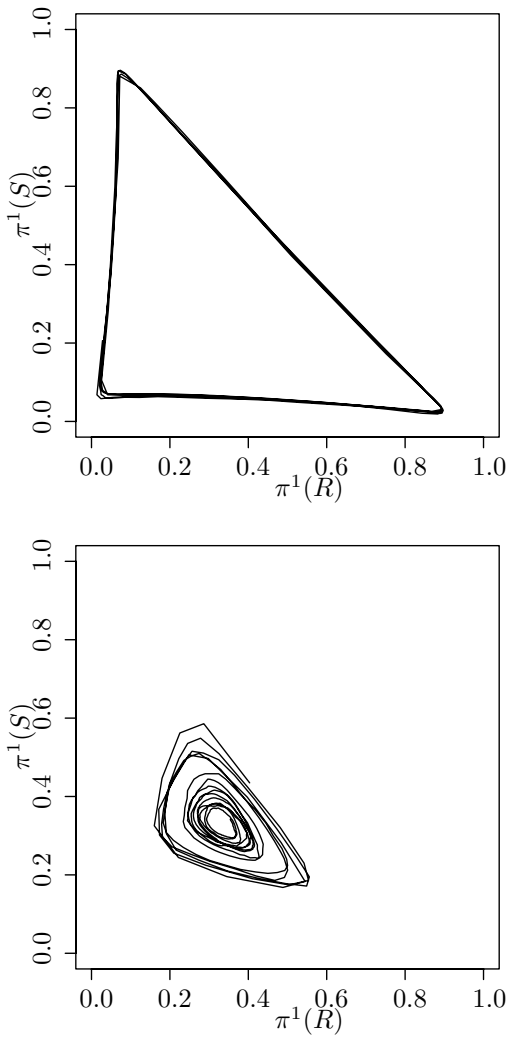


FIG. 2. *Strategies of player* 1 *in Shapley's game* (4.2) *with Boltzmann action selection* ($\tau = 0.1$) *over* 5×10^5 *iterations of basic individual Q-learning* (3.2) *with* $\lambda_n = (n + 100)^{-0.9}$ *(top) and of individual Q-learning with PDLR* (5.1)*, with* $\lambda_n^1 = (n + 100)^{-0.9}$ *and* $\lambda_n^2 = (n + 100)^{-0.7}$ *(bottom). The first* 1×10^4 *iterations are omitted in each case. For basic individual Q-learning* (3.2) *the strategies follow a limit cycle, while for individual Q-learning with PDLR* (5.1) *the strategies are spiraling anticlockwise toward the unique Nash distribution.*

ASSUMPTION 5.1. *For each* $i \in \{2, \ldots, N\}$ *there exists a function* $\tilde{q}^i : \Delta^1 \times \cdots \times \Delta^{i-1} \to \mathbb{R}^{|A^i|}$ *such that, for arbitrary fixed* $(Q^1, \ldots, Q^{i-1})$*, the ODE*

$$\frac{\mathrm{d}}{\mathrm{d}t} q_t^i(a^i) = r^i(a^i, [\pi^{(<i)}, B^{(>i)}[\pi^{(<i)}, \beta^i(q_t^i)]]) - q_t^i(a^i) \quad \textit{for each } a^i \in A^i$$

has the globally attracting fixed point $\tilde{q}^i(\pi^{(<i)})$*, where*

$$\pi^{(<i)} = (\beta^1(Q^1), \ldots, \beta^{i-1}(Q^{i-1}))$$

TABLE 2
Individual Q-learning with PDLR.

Each player i selects an action a_n^i using the strategy $\beta^i(Q_n^i)$, receives reward R_n^i, and then updates Q_n^i according to

$$Q_{n+1}^i(a^i) = Q_n^i(a^i) + \lambda_{n+1}^i \mathbb{I}_{\{a_n^i = a^i\}} \frac{R_n^i - Q_n^i(a_n^i)}{\beta^i(Q_n^i)(a_n^i)} \quad \text{for each } a^i \in A^i, \tag{5.1}$$

where for each i, $\{\lambda_n^i\}_{n\geq 1}$ is a deterministic sequence of learning parameters satisfying the conditions (3.3), and additionally

$$\lambda_n^i / \lambda_n^{i+1} \longrightarrow 0 \quad \text{as } n \longrightarrow \infty. \tag{5.2}$$

and $B^{(>i)} : \Delta^1 \times \cdots \times \Delta^i \to \Delta^{i+1} \times \cdots \times \Delta^N$ *is defined recursively by*

$$B^{(>(N-1))}(\pi^{(<N)}) = \beta^N(\tilde{q}^N(\pi^{(<N)})),$$
$$B^{(>(i-1))}(\pi^{(<i)}) = (\beta^i(\tilde{q}^i(\pi^{(<i)})), B^{(>i)}[\pi^{(<i)}, \beta^i(\tilde{q}^i(\pi^{(<i)}))]).$$

Although this looks technical, it can be expressed relatively simply: for any i, if the values (and hence strategies) of players $(1, \ldots, i-1)$ were fixed, the strategies of the more responsive players $(i, \ldots, N)$ would converge to a unique fixed point determined by the functions $\tilde{q}^i$ and $B^{(>i)}$. This assumption is satisfied for any 2-player game: for Q^1 fixed, player 2 simply faces a multi-armed bandit problem. However, it is clearly not always satisfied for general N-player games: in a 3-player partnership game, for Q^1 fixed the other two players still face a partnership game, which might well have more than one Nash distribution, and therefore more than one potential limit point for the more responsive players. We will investigate when Assumption 5.1 is satisfied using a graphical analysis in section 6, but in this section we will retain it as an assumption.

As with the basic individual Q-learning algorithm of section 3, it is immediate that convergence of algorithm (5.1) to a fixed point can occur only at Nash distribution values. Also analogously, we can prove the following lemma.

LEMMA 5.2. *Under Assumption* 5.1, *the values* Q_n^1 *resulting from individual Q-learning with PDLR* (5.1) *converge almost surely to a connected internally chain-recurrent set of the flow defined by the singularly perturbed Q-learning ODE*

$$\frac{\mathrm{d}}{\mathrm{d}t} q_t^1(a^1) = r^1(a^1, B^{(>1)}[\beta^1(q_t^1)]) - q_t^1(a^1) \quad \textit{for each } a^1 \in A^1, \tag{5.3}$$

provided that the Q_n *remain bounded for all time, where* $B^{(>1)}$ *is the function defined in Assumption* 5.1. *Additionally,*

$$\|Q_n^i - r^i(\cdot, [\beta^1(Q_n^1), B^{>1}[\beta^1(Q_n^1)]])\|_\infty \longrightarrow 0$$
$$\textit{almost surely for each } i > 1 \textit{ as } n \longrightarrow \infty.$$

Proof. This is immediate from the results of Leslie and Collins (2003) and Assumption 5.1. □

Note that Lemma 5.2 tells us we can analyze the asymptotic behavior of the algorithm as if the values of the remaining players have all converged to the fixed

point determined by the current strategy of the most cautious player, despite the fact that in reality all players adjust their values after every play of the game.

We will proceed to analyze the dynamical system (5.3) in two different ways: in section 5.1 we proceed as in section 4 and relate (5.3) to the singularly perturbed SBR dynamics (Leslie and Collins (2003)), while in section 5.2 we perform a direct analysis in a more restricted class of games.

5.1. Relating the singularly perturbed Q-learning ODE to the singularly perturbed SBR dynamics. Leslie and Collins (2003) study the singularly perturbed SBR dynamics, defined by

$$\frac{\mathrm{d}}{\mathrm{d}t}\pi_t^1 = \beta^1[r^1(\cdot, B^{(>1)}[\pi_t^1])] - \pi_t^1. \tag{5.4}$$

As in section 4, we will relate the singularly perturbed Q-learning ODE (5.3) to the singularly perturbed SBR dynamics (5.4).

LEMMA 5.3. *Any connected internally chain-recurrent set of the singularly perturbed Q-learning ODE* (5.3) *is of the form*

$$r^1(\mathcal{C}^1) := \{r^1(\cdot, B^{(>1)}(\pi^1)) : \pi^1 \in \mathcal{C}^1\},$$

where $\mathcal{C}^1 \subset \Delta^1$ is a connected internally chain-recurrent set of flow defined by the singularly perturbed SBR dynamics (5.4)*, and $B^{(>1)}$ is the function defined in Assumption* 5.1*.*

Proof. This lemma's proof is identical to that of Lemma 4.1, and will not be repeated here. □

As in section 4 this enables us to use previous results (Leslie and Collins (2003)) on more sophisticated forms of learning to prove convergence of the individual Q-learning algorithm with PDLR (5.1) to Nash distribution in certain classes of games.

PROPOSITION 5.4. *The strategies of players using the individual Q-learning algorithm with PDLR* (5.1) *will converge almost surely to a Nash distribution, provided that the Q_n remain bounded for all time, in the following games:*

(i) *2-player zero-sum games,*

(ii) *2-player partnership games,*

(iii) *Shapley's game* (4.2) *(if Boltzmann action selection is used), and*

(iv) *the N-player matching pennies game (Leslie and Collins* (2003)*) (if the smooth best responses are symmetric under a reordering of the actions).*

Proof. Leslie and Collins (2003) show that Assumption 5.1 holds for these games. The result therefore follows immediately from Lemmas 5.2 and 5.3 and the results of Leslie and Collins (2003) on the connected internally chain-recurrent sets of the singularly perturbed SBR dynamics. □

Thus the individual Q-learning algorithm with PDLR (5.1) is proved to converge in the same games as basic individual Q-learning (3.2). In addition, we have proved that individual Q-learning with PDLR will converge in two games (Shapley's game and the N-player matching pennies game) for which most learning algorithms fail to converge. Indeed the authors know of no adaptive algorithm that converges to either Nash equilibrium or Nash distribution in these games without using PDLR.

We illustrate these results with some numerical experiments using Shapley's game (4.2). As observed in section 4, the basic individual Q-learning algorithm (3.2) will cycle in this game. On the other hand, we have shown that the individual Q-learning algorithm with PDLR (5.1) will converge to the unique Nash distribution values. This is confirmed in Figure 2.

5.2. Direct analysis of the singularly perturbed Q-learning ODE. We now change our emphasis away from the SBR dynamics and instead analyze the singularly perturbed Q-learning ODE (5.3) directly in games where player 1 has 2 actions. We show here that convergence to Nash distribution occurs if Boltzmann action selection is used.

PROPOSITION 5.5. *Suppose player* 1 *has* 2 *actions and uses Boltzmann action selection* (2.2)*. Suppose further that the game and smooth best responses are such that there are countably many Nash distributions. Then, under Assumption* 5.1*, the strategies of players using the individual Q-learning algorithm with PDLR* (5.1) *converge almost surely to a Nash distribution, provided that the Q_n remain bounded for all time.*

Proof. By Lemma 5.2, the Q^1 values converge to a connected internally chain-recurrent set of the singularly perturbed Q-learning ODE (5.3). To ease notation, for this proof we will write

$$\pi_t(1) = \beta^1(q_t^1)(1) = \frac{\mathrm{e}^{q_t^1(1)/\tau}}{\mathrm{e}^{q_t^1(1)/\tau} + \mathrm{e}^{q_t^1(2)/\tau}} = (1 + \mathrm{e}^{\{q_t^1(2) - q_t^1(1)\}/\tau})^{-1} = 1 - \pi_t(2),$$
$$\hat{r}_t(a) = r^1(a, B^{>1}[\beta^1(q_t^1)]), \quad a = 1, 2.$$

Since $\beta^1(q_t^1) = (\pi_t(1), 1 - \pi_t(1))$, $\hat{r}_t(a)$ is a function of the scalar variable $\pi_t(1)$, which is in turn a function of $q_t^1(1)$ and $q_t^1(2)$. Hence

$$\begin{aligned}\frac{\mathrm{d}\hat{r}_t(a)}{\mathrm{d}t} &= \frac{\mathrm{d}\hat{r}_t(a)}{\mathrm{d}\pi_t(1)} \left\{ \frac{\partial \pi_t(1)}{\partial q_t^1(1)} \frac{\mathrm{d}q_t^1(1)}{\mathrm{d}t} + \frac{\partial \pi_t(1)}{\partial q_t^1(2)} \frac{\mathrm{d}q_t^1(2)}{\mathrm{d}t} \right\} \\ &= \frac{\mathrm{d}\hat{r}_t(a)}{\mathrm{d}\pi_t(1)} \pi_t(1)\pi_t(2) \frac{\mathrm{d}}{\mathrm{d}t} \{q_t^1(1) - q_t^1(2)\}.\end{aligned}$$

From this, it follows that

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}t^2} \{q_t^1(1) - q_t^1(2)\} &= \frac{\mathrm{d}}{\mathrm{d}t} \{\hat{r}_t(1) - \hat{r}_t(2) - q_t^1(1) + q_t^1(2)\} \\ &= \left[\pi_t(1)\pi_t(2) \left\{ \frac{\mathrm{d}\hat{r}_t(1)}{\mathrm{d}\pi_t(1)} - \frac{\mathrm{d}\hat{r}_t(2)}{\mathrm{d}\pi_t(1)} \right\} - 1 \right] \frac{\mathrm{d}}{\mathrm{d}t} \{q_t^1(1) - q_t^1(2)\}.\end{aligned}$$

Therefore $\frac{\mathrm{d}}{\mathrm{d}t}\{q_t^1(1) - q_t^1(2)\}$ does not change sign, and $\{q_t^1(1) - q_t^1(2)\}$ acts as a Lyapunov function. The result follows from Benaïm (1999, Corollary 6.6). □

Note that Proposition 5.5 provides an independent proof of the convergence of the individual Q-learning algorithm with PDLR (5.1) for the N-player matching pennies game that does not rely on the smooth best responses being symmetric under a reordering of the actions. Furthermore the following immediate corollary shows that the result can be applied directly in a wide class of games.

COROLLARY 5.6. *In a* 2*-player game where player* 1 *has* 2 *actions and uses Boltzmann action selection* (2.2)*, the strategies of players using the individual Q-learning algorithm with PDLR* (5.1) *converge almost surely to a Nash distribution, provided that the Q_n remain bounded for all time.*

Proof. As already noted, Assumption 5.1 holds automatically in 2-player games. Hence this is immediate from Proposition 5.5. □

This is comparable with a recent result (Berger (2005)) showing that fictitious play approaches equilibrium in nondegenerate $2 \times n$ games, although clearly more information is required for the players in a fictitious play process.

5.3. PDLR and Stackelberg equilibria. Here we discuss a point raised by an anonymous referee, concerning the relationship between PDLR and Stackelberg equilibria (equilibria in which one player selects her strategy first, in the knowledge that the other player will play an optimal action in response). If players used individual Q-learning with PDLR, it might be anticipated that a cautious player could gain an advantage by initially selecting a strategy and then relying on the fact that the responsive player will adapt quickly to this strategy. For example, in the battle of the sexes game with payoff matrix

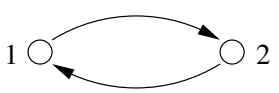
$$\begin{pmatrix} (10,1) & (0,0) \\ (0,0) & (1,10) \end{pmatrix},$$

player 1 could initially select row 1 with high probability, forcing a responsive player 2 to learn to play column 1; player 1 is acting as a Stackelberg leader.

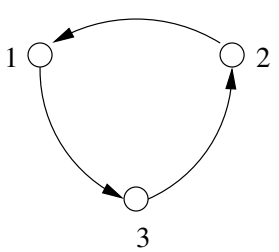
However, in the minimal information situation we consider here, player 1 cannot know in advance which strategy to select to achieve this effect, since the payoff matrix is not known. In the battle of the sexes game, ignoring stochastic effects, if player 1's initial strategy has $\pi^1(2) > 1/11$, then a responsive player 2 will play a strategy in which $\pi^2(2)$ is very nearly 1; the cautious player 1 will then slowly increase $\pi^1(2)$ until joint play reaches the Nash distribution where both players select action 2 with high probability. (Note that this is actually the equilibrium that would have been selected if player 2 was a Stackelberg leader.) Since $\pi^1(2) > 1/11$ with high probability, under reasonable assumptions about the initial Q values of the players, there is a high probability that play will converge to a Nash distribution different from that suggested by the Stackelberg theory. Thus the relationship between PDLR and Stackelberg equilibria is much less clear than might be expected.

6. Graphical analysis. The results of section 5 on individual Q-learning with PDLR rely on Assumption 5.1, which states that there is a unique limit point of the strategies of the more responsive players for any fixed values of the more cautious players $(1, \ldots, i-1)$. Although we have observed that this is always true for 2-player games, in this section we develop a graphical approach to analyzing when this is satisfied in general N-player games, building on previous graphical representations of games (Littman, Kearns, and Singh (2001), Koller and Milch (2003)).

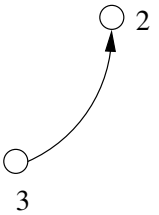
Given a game, we construct a graph by taking a node for each player and drawing a directed arc $\vec{ij}$ if the actions of player i directly affect the rewards of player j. Thus the graph of a (generic) 2-player game is given by



whereas the graph of the 3-player matching pennies game (Jordan (1993)), in which the rewards of player i are affected only by the actions of player $i+1$ (modulo 3), is given by



These graphs can be used to investigate Assumption 5.1 by removing node 1 (corresponding to the most cautious player) from the graph, then considering the behavior of the other players. The intuition behind this is that the fixed strategy of the most cautious player in Assumption 5.1 results in the other players facing a reduced game. For example, in the 3-player matching pennies game, removal of node 1 (and connected arcs) results in the graph



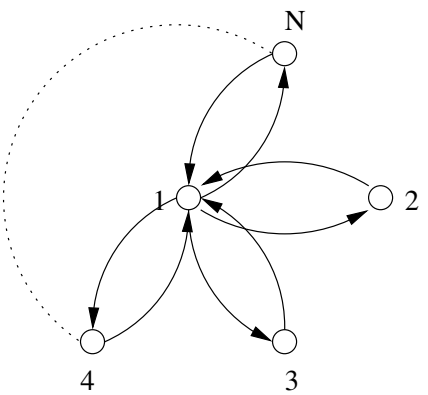
Thus for fixed Q^1, player 3's rewards are not affected by any other player, so Q_n^3 converges to unique values. But because of this, player 2's values must also converge to unique values, and so for fixed Q^1, the values (Q_n^2, Q_n^3) converge to a unique point, as required by Assumption 5.1.

In fact, this example generalizes very simply. If, after removal of a node, a graph has no directed cycle, then the players corresponding to nodes left in the graph will converge to a unique fixed point.

PROPOSITION 6.1. *Suppose that a game is such that removal of node* 1 *from the game graph results in a subgraph containing no directed cycles. Then Assumption* 5.1 *holds for this game.*

Proof. If the subgraph remaining after removal of node 1 has no directed cycle, then there exists at least one node with in-degree 0 (i.e., no arcs terminate at this node). The values of a player associated with such a node do not depend on the strategies of any other players (except perhaps the fixed strategy of player 1), and so the Q values of that player will converge almost surely to a unique point. Since the rewards, and hence strategies, of any player corresponding to a node with in-degree 0 are uniquely determined, given the fixed strategy of player 1, these nodes can be removed from the graph. This again results in a graph with no directed cycles, and we can proceed recursively to show that the rewards of all the players are uniquely determined by the values of player 1, as required in Assumption 5.1. □

From this proposition, it is immediate that Assumption 5.1 holds in games which have a graph with no directed cycle even before removal of a node, and games with a single directed cycle will clearly become acyclic if the node to be removed is part of that cycle. Consider also games with a star graph:



Here there is a distinguished player (the hub) connected to all others, but all non-distinguished players are disconnected from each other. This graph will become completely disconnected (and so obviously will have no directed cycle) if the node corresponding to the distinguished player is removed. A game such as this could have applications in computing, for example, where the distinguished player is a central resource, such as a server or router, and the other nodes correspond to users of the resource.

Although useful, this graphical approach is not sufficient in all situations. For example, consider a 3-player game in which the rewards of players 2 and 3 always sum to 0 for any fixed Q^1. Thus, the graph after removal of node 1 is the same as that of a 2-player game, and so (generically) contains a directed cycle. However, the resulting game is a 2-player zero-sum game, in which the players converge to a unique Nash distribution (Proposition 5.4 and Hofbauer and Hopkins (2005)). Therefore Assumption 5.1 is satisfied, even though the conditions of Proposition 6.1 are not.

7. Conclusion. We have shown that value-based learning agents cannot generally converge to a Nash equilibrium of a game, but if smooth best responses are used, a Nash distribution can be reached. Although Nash distributions are not generally the same as Nash equilibria, they are close if the temperature parameter of the smooth best responses is sufficiently small, and therefore we proposed that value-based learning agents should use smooth best responses to allow equilibrium play, even if this is not a classical Nash equilibrium.

Our value-based learning algorithm, individual Q-learning (3.2), is very similar to the simple and successful algorithm used by Sutton and Barto (1998) in the single-agent multi-armed bandit problem. We showed that convergence to a point that is not a Nash distribution is not possible and that the value estimates are asymptotically belief-based. Further, by relating the limiting behavior of the individual Q-learning algorithm to the SBR dynamics (a system previously used to characterize more sophisticated models of learning), it was shown that strategies of players converge almost surely to a Nash distribution for 2-player zero-sum games and 2-player partnership games.

The nonconvergence of strategies for certain games motivates the introduction of individual Q-learning with PDLR (5.1), resulting in cautious and responsive players. This modified algorithm converges to Nash distribution for the same games as basic individual Q-learning (3.2), and also for Shapley's game and the N-player matching pennies game. Moreover, convergence was shown to occur for any game in which player 1, the most cautious, has only 2 actions.

Finally, since the results on PDLR rely on an assumption about the behavior of the more responsive players if the values of player 1 were fixed, a simple graphical method was introduced to help determine when this assumption holds.

Acknowledgments. The authors thank three anonymous referees for helpful comments and Prof. Josef Hofbauer and Dr. Andy Wright for helpful discussions in the course of the research leading to this paper.

REFERENCES

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire (1995), *Gambling in a rigged casino: The adversarial multi-armed bandit problem*, in 36th Annual Symposium on Foundations of Computer Science, IEEE Computer Society Press, Los Alamitos, CA, pp. 322–331.

R. J. Aumann (1974), *Subjectivity and correlation in randomized strategies*, J. Math. Econom., 1, pp. 67–96.

A. Baños (1968), *On pseudo-games*, Ann. Math. Statist., 39, pp. 1932–1945.

M. Benaïm (1999), *Dynamics of stochastic approximation algorithms*, in Le Séminaire de Probabilités, Lecture Notes in Math. 1709, Springer-Verlag, Berlin, pp. 1–68.

M. Benaïm and M. W. Hirsch (1999), *Mixed equilibria and dynamical systems arising from fictitious play in perturbed games*, Games Econom. Behav., 29, pp. 36–72.

U. Berger (2005), *Fictitious play in $2 \times n$ games*, J. Econom. Theory, 120, pp. 139–154.

T. Börgers and R. Sarin (1997), *Learning through reinforcement and replicator dynamics*, J. Econom. Theory, 77, pp. 1–14.

V. S. Borkar (2001), *Reinforcement learning in Markovian evolutionary games*, available at http://www.tcs.tifr.res.in/∼borkar/game.ps.

M. Bowling and M. Veloso (2002), *Multiagent learning using a variable learning rate*, Artificial Intelligence, 136, pp. 215–250.

J. A. Boyan and M. L. Littman (1994), *Packet routing in dynamically changing networks: A reinforcement learning approach*, in Advances in Neural Information Processing Systems, Vol. 6, J. D. Cowan, G. Tesauro, and J. Alspector, eds., Morgan Kaufmann, San Francisco, pp. 671–678.

G. W. Brown (1951), *Iterative solution of games by fictitious play*, in Activity Analysis of Production and Allocation, T. C. Koopmans, ed., John Wiley & Sons, New York, pp. 374–376.

H.-F. Chen and Y.-M. Zhu (1986), *Stochastic approximation procedures with randomly varying truncations*, Sci. China Ser. A, 29, pp. 914–926.

S. Cowan (1992), *Dynamical Systems Arising from Game Theory*, Ph.D. thesis, University of California, Berkeley.

R. H. Crites and A. G. Barto (1998), *Elevator group control using multiple reinforcement learning agents*, Machine Learning, 33, pp. 235–262.

D. P. Foster and H. P. Young (2003), *Learning, hypothesis testing, and Nash equilibrium*, Games Econom. Behav., 45, pp. 73–96.

D. Fudenberg and D. K. Levine (1998), *The Theory of Learning in Games*, MIT Press, Cambridge, MA.

D. Fudenberg and J. Tirole (1991), *Game Theory*, MIT Press, Cambridge, MA.

P. W. Glimcher, M. C. Dorris, and H. M. Bayer (2005), *Physiological utility theory and the neuroeconomics of choice*, Games Econom. Behav., to appear.

S. Govindan, P. J. Reny, and A. J. Robson (2003), *A short proof of Harsanyi's purification theorem*, Games Econom. Behav., 45, pp. 369–374.

J. Hannan (1957), *Approximation to Bayes risk in repeated play*, in Contributions to the Theory of Games, Vol. 3, Ann. of Math. Stud. 39, M. Drescher, A. W. Tucker, and P. Wolfe, eds., Princeton University Press, Princeton, NJ, pp. 97–139.

J. C. Harsanyi (1973), *Games with randomly disturbed payoffs: A new rationale for mixed-strategy equilibrium points*, Internat. J. Game Theory, 2, pp. 1–23.

S. Hart and A. Mas-Colell (2000), *A simple adaptive procedure leading to correlated equilibrium*, Econometrica, 68, pp. 1127–1150.

S. Hart and A. Mas-Colell (2001), *A reinforcement procedure leading to correlated equilibrium*, in Economic Essays: A Festschrift for Werner Hildenbrand, W. N. G. Debreu and W. Trockel, eds., Springer-Verlag, Berlin, pp. 181–200.

S. Hart and A. Mas-Colell (2003), *Uncoupled dynamics cannot lead to Nash equilibrium*, Amer. Econom. Rev., 93, pp. 1830–1836.

J. Hofbauer and E. Hopkins (2005), *Learning in perturbed asymmetric games*, Games Econom. Behav., 52, pp. 133–152.

J. S. Jordan (1993), *Three problems in learning mixed strategy equilibria*, Games Econom. Behav., 5, pp. 368–386.

D. Koller and B. Milch (2003), *Multi-agent influence diagrams for representing and solving games*, Games Econom. Behav., 45, pp. 181–221.

H. J. Kushner and G. G. Yin (1997), *Stochastic Approximation Algorithms and Applications*, Appl. Math. 35, Springer-Verlag, New York.

D. S. Leslie and E. J. Collins (2003), *Convergent multiple-timescales reinforcement learning algorithms in normal form games*, Ann. Appl. Probab., 13, pp. 1231–1251.

M. L. Littman, M. Kearns, and S. Singh (2001), *An efficient, exact algorithm for solving tree-structured graphical games*, in Advances in Neural Information Processing Systems, Vol. 14, T. G. Dietterich, S. Becker, and Z. Ghahramani, eds., MIT Press, Cambridge, MA.

J. Maynard Smith (1982), *Evolution and the Theory of Games*, Cambridge University Press, Cambridge, UK.

N. Megiddo (1980), *On repeated games with incomplete information played by non-Bayesian players*, Internat. J. Game Theory, 9, pp. 157–167.

J. Nash (1950), *Equilibrium points in n-person games*, Proc. Natl. Acad. Sci. USA, 36, pp. 48–49.

R. Pemantle (1990), *Nonconvergence to unstable points in urn models and stochastic approximations*, Ann. Probab., 18, pp. 698–712.

A. E. Roth and I. Erev (1995), *Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term*, Games Econom. Behav., 8, pp. 164–212.

L. S. Shapley (1964), *Some topics in two person games*, in Advances in Game Theory, M. Drescher, L. S. Shapley, and A. W. Tucker, eds., Princeton University Press, Princeton, NJ, pp. 1–28.

S. Singh and D. Bertsekas (1997), *Reinforcement learning for dynamic channel allocation in cellular telephone systems*, in Advances in Neural Information Processing Systems, Vol. 9, M. C. Mozer, M. I. Jordan, and T. Petsche, eds., MIT Press, Cambridge, MA, p. 974.

R. S. Sutton and A. G. Barto (1998), *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 2, pp. 515–530

H_2 AND H_∞ FILTERING DESIGN SUBJECT TO IMPLEMENTATION UNCERTAINTY*

MAURÍCIO C. DE OLIVEIRA† AND JOSÉ C. GEROMEL‡

Abstract. This paper presents new filtering design procedures for discrete-time linear systems. It provides a solution to the problem of linear filtering design, assuming that the filter is subject to parametric uncertainty. The problem is relevant, since the proposed filter design incorporates real world implementation constraints that are always present in practice. The transfer function and the state space realization of the filter are simultaneously computed. The design procedure can also handle plant parametric uncertainty. In this case, the plant parameters are assumed not to be exactly known but belonging to a given convex and closed polyhedron. Robust performance is measured by the H_2 and H_∞ norms of the transfer function from the noisy input to the filtering error. The results are based on the determination of an upper bound on the performance objectives. All optimization problems are linear with constraint sets given in the form of LMI (linear matrix inequalities). Global optimal solutions to these problems can be readily computed. Numerical examples illustrate the theory.

Key words. robust filtering, implementation uncertainty, fragility, linear matrix inequalities

AMS subject classifications. 93E11, 93E25, 60G35

DOI. 10.1137/S0363012903424721

1. Introduction.

In the 1980s, a great deal of effort was dedicated to the study of implementation issues of filters and controllers [1, 2, 3, 4]. The motivation was to devise design techniques that would lead to filters and controllers that could perform well when implemented on a digital computer. The main objectives were (a) to minimize the degradation of performance caused by computation of signals in a finite precision computational architecture, and (b) to minimize the impact of truncation and rounding on the coefficients of the filter or controller. These objectives were addressed using many different techniques (see [1] for details). Among these techniques, a popular approach to dealing with degradation of the signals was to model rounding and truncation as noise [5], whereas rounding and truncation of the filter or controller coefficients was addressed by studying the *sensitivity* of these parameters to variations [1]. The great development of the computer industry in the 1990s brought to the signal processing and control practitioner processors with more and more bits of precision at very low cost, which somewhat dimmed the importance of the topic. The fact that every few years the computer industry provides processors with longer wordlength is used by some to justify the design of filters and controllers with little or no regard to finite precision perturbation effects. In fact, for many simple systems, this increase in wordlength means that the quantization effects can be practically ignored. However, faster and more precise computers also provide the opportunity to increase the complexity of the systems, in terms of both more sophisticated algorithms

*Received by the editors March 18, 2003; accepted for publication (in revised form) December 15, 2004; published electronically August 31, 2005. This work was supported in part by grants from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) Brazil.

http://www.siam.org/journals/sicon/44-2/42472.html

†Department of Mechanical and Aerospace Engineering, University of California, San Diego, La Jolla, CA 92093-0411 (mauricio@ucsd.edu).

‡DSCE, School of Electrical and Computer Engineering, UNICAMP, CP 6101, 13083-970, Campinas, SP, Brazil (geromel@dsce.fee.unicamp.br).

and number of devices. As observed in [6], this increased complexity will eventually face limitations in bandwidth, that is, the speed at which the devices communicate, reducing the sampling rates (relative to the available processor wordlength). In this scenario, a careful analysis of perturbation effects on filters and controllers certainly will be required. Also, in many consumer electronics products, inexpensive processors (say, fixed point digital signal processors (DSPs)) are usually preferred. These processors often impose nontrivial wordlength limitations, and thus better design algorithms are needed to deal with them. Some recent efforts along this line are reported in [7].

In fact, the importance of robustness to filter and control parametric perturbations seems to have been *rediscovered* by the end of the 1990s with the paper [8]. This work, despite the controversy it raised [9], showed that many robust control design methods, which were targeted to deal with plant uncertainty, could be particularly sensitive to parameter uncertainty on the controller. The authors use a series of numerical examples to illustrate that a very small perturbation on the coefficients of controllers could lead to a loss of stability of the closed-loop system [8]. Since then, many authors have addressed the problem of robustness to parametric control or filter perturbation under the label of *fragility* [10, 11, 12, 13, 14].

While many works on filter *sensitivity* are more concerned with the problem of choosing an appropriate realization for a given filter transfer function [2, 3], many works on *fragility* seem to focus more on the robustness of the filter transfer function rather than its realization [12]. The approach developed in this paper blends these two issues by simultaneously designing the optimal filter transfer function *and* its realization. The strategy is to modify the filtering procedure introduced in [15] to take into account robustness with respect to filter parametric variations. Variations of the filter parameters are allowed inside a region specified by a quadratic matrix inequality. The maximum allowed norm of the filter uncertainty is specified as a percentage of the norm of the nominal filter parameters. The ability to specify the uncertainty in the filter parameters relative to the size of the nominal filter is especially important when the transfer function and the state space realization of the filter are to be designed simultaneously. This model is also very appropriate to model perturbations on the parameters coming from truncation on a floating-point computational architecture, where rounding and truncation introduce errors relative to the size of the original numbers.

In this paper, guaranteed cost functions are developed to provide upper bounds on the maximum value of the H_2 or H_∞ norm of the uncertain transfer function from an exogenous noise input to the filtering error on the filter uncertainty region. This paper introduces and completely solves these H_2 and H_∞ guaranteed cost filtering design problems. The design conditions are expressed as linear matrix inequalities (LMIs), and hence numerical solutions can be readily computed [16]. In contrast to [15, 17], the results specify not only the transfer function of the filter but also its realization. Illustrative examples show the effectiveness of the proposed approach. An interesting feature observed in the examples is that the filters designed by the proposed technique have less round-off gain than the standard Kalman filter [2, 5], although such a performance measure is not directly addressed in the optimization process. The design procedures introduced in this paper admit straightforward extensions to simultaneously handle plant parameter uncertainty specified in terms of convex bounded polyhedrons. These extensions can be derived to contemplate both the quadratic stability [17] and the extended stability [18] approaches. In the former, a single quadratic Lyapunov function is used to evaluate the performance on the

uncertainty region, while in the latter a parameter dependent Lyapunov function [19] is built.

The notation is standard. Lowercase letters denote vectors while capital letters represent matrices. The symbol $(^T)$ is used to indicate the transpose of vectors and matrices. If a symmetric matrix X is positive definite, this is indicated by $X > 0$.

2. Preliminary results on filtering. Consider the linear discrete-time time-invariant system

$$x(k+1) = Ax(k) + Bw(k), \tag{1}$$

$$z(k) = C_z x(k) + D_z w(k), \tag{2}$$

$$y(k) = C_y x(k) + D_y w(k), \tag{3}$$

where all matrices and vectors are assumed to have appropriate dimensions. The *optimal filtering* problem consists of designing a linear filter

$$x_f(k+1) = A_f x_f(k) + B_f y(k), \tag{4}$$

$$z_f(k) = C_f x_f(k) + D_f y(k), \tag{5}$$

which makes use of the plant output $y(k)$ to produce the filtered output $z_f(k)$, with the objective of minimizing a norm of the transfer function from the noise input $w(k)$ to the filtering error $e(k) := z(k) - z_f(k)$. Collecting the filter parameters in the matrix

$$\mathcal{F} := \begin{bmatrix} D_f & C_f \\ B_f & A_f \end{bmatrix}, \tag{6}$$

we can state the optimal filtering problem as the optimization problem

$$\min_{\mathcal{F}} \|H_{we}(z;\mathcal{F})\|_p. \tag{7}$$

The values of $p = \{2, \infty\}$ are the choices usually found in the literature. The next lemmas revisit the solutions of the optimal filtering problems given in [17]. The solution is given as LMI conditions formulated in terms of the transformed set of filter parameters

$$\mathcal{K} := \begin{bmatrix} R & L \\ F & Q \end{bmatrix}, \tag{8}$$

defined with respect to the above partitioning.

LEMMA 1 (H_2 filtering). *There exist a matrix $\mathcal{K}$, partitioned as in* (8)*, and symmetric matrices Y, Z, W such that the LMI*

$$\begin{bmatrix} Z & \bullet & \bullet & \bullet & \bullet \\ Z & Y & \bullet & \bullet & \bullet \\ A^TZ & A^TY + C_y^TF^T + Q^T & Z & \bullet & \bullet \\ A^TZ & A^TY + C_y^TF^T & Z & Y & \bullet \\ B^TZ & B^TY + D_y^TF^T & \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix} > 0, \tag{9}$$

$$\begin{bmatrix} W & \bullet & \bullet & \bullet \\ C_z^T - C_y^TR^T - L^T & Z & \bullet & \bullet \\ C_z^T - C_y^TR^T & Z & Y & \bullet \\ D_z^T - D_y^TR^T & \mathbf{0} & \mathbf{0} & I \end{bmatrix} > 0, \tag{10}$$

$$\operatorname{trace}(W) < \mu \tag{11}$$

have a feasible solution if and only if the filter

$$\mathcal{F} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & V^{-1} \end{bmatrix} \mathcal{K} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & Z^{-1}U^{-1} \end{bmatrix}, \tag{12}$$

where U and V are nonsingular otherwise arbitrary matrices chosen to satisfy $Y + VUZ = Z$, is such that

$$\|H_{we}(z;\mathcal{F})\|_2^2 < \mu. \tag{13}$$

LEMMA 2 (H_∞ filtering). *There exist a matrix $\mathcal{K}$, partitioned as in* (8)*, and symmetric matrices Y, Z such that the LMI*

(14)
$$\begin{bmatrix} Z & \bullet & \bullet & \bullet & \bullet & \bullet \\ Z & Y & \bullet & \bullet & \bullet & \bullet \\ A^TZ & A^TY + C_y^TF^T + Q^T & Z & \bullet & \bullet & \bullet \\ A^TZ & A^TY + C_y^TF^T & Z & Y & \bullet & \bullet \\ B^TZ & B^TY + D_y^TF^T & \mathbf{0} & \mathbf{0} & \mu\mathbf{I} & \bullet \\ \mathbf{0} & \mathbf{0} & C_z - RC_y - L & C_z - RC_y & D_z - RD_y & \mu\mathbf{I} \end{bmatrix} > 0$$

have a feasible solution if and and only if the filter $\mathcal{F}$ given in (12) *is such that*

$$\|H_{we}(z;\mathcal{F})\|_\infty < \mu. \tag{15}$$

The above lemmas are generalizations of the results obtained in [17]. Here, the assumptions that the filter(4)–(5) is strictly proper and that the matrix D_z is null have been removed. There is virtually no change from the proofs presented in [15, 17] to the ones required to prove Lemmas 1 and 2. These proofs are omitted for brevity and the interested reader is referred to [15, 17] for more details. The constraints stated in Lemmas 1 and 2 are all LMI, and hence solutions to the optimization problem (7) can be obtained by minimizing the scalar μ subject to the given inequalities. The resulting problems are convex and their global optimal solutions can be obtained via convex programming techniques [16].

Once a solution to the inequalities stated in Lemmas 1 or 2 has been found, the user is asked to pick an arbitrary nonsingular matrix U and then solve for V to satisfy $Y + VUZ = Z$ (or choose V and solve for U). This will produce the optimal filter parameters $\mathcal{F}$ through (12). Notice that this is done a posteriori, and that this arbitrary choice does not affect the optimality of the solution. In fact, it is possible to show that the transfer function of the filter associated with the parameters (12) is not affected by the choice of U and V (see [17] for details). The main role of these matrices is to parameterize a particular state space realization of the filter, a fact that will be explored in the next sections.

3. Problem statement. The main purpose of this paper is to derive conditions for the design of filters subject to parametric perturbations. More specifically, it is assumed that the parameters of the filter(4)–(5) are subject to an additive perturbation of the form

$$\mathcal{F} = \mathcal{F}_0 + \Delta_{\mathcal{F}}. \tag{16}$$

The symbol $\mathcal{F}_0$ denotes *nominal* filter parameters, and the unknown perturbation $\Delta_{\mathcal{F}}$ is assumed to be in the set

$$\mathbb{F}_{\mathcal{R}}(\mathcal{F}_0) := \left\{\Delta_{\mathcal{F}} \,:\, \Delta_{\mathcal{F}}^T\mathcal{R}^{-1}\Delta_{\mathcal{F}} \le \gamma^2\mathcal{F}_0^T\mathcal{R}^{-1}\mathcal{F}_0\right\}. \tag{17}$$

In contrast to the conventional norm bounded uncertainty model, where the right-hand side of the inequality given in (17) is usually constant, the uncertainty set $\mathbb{F}_{\mathcal{R}}(\mathcal{F}_0)$ relates the size of the parametric perturbation $\Delta_{\mathcal{F}}$ to the size of the nominal filter parameters $\mathcal{F}_0$. These factors are weighted by the inverse of an arbitrary positive definite matrix $\mathcal{R}$. In this way, by setting the scalar $0 \leq \gamma \leq 1$, the size of the perturbation $\Delta_{\mathcal{F}}$ can be specified *relative* to the size of the nominal filter parameters $\mathcal{F}_0$, which are yet to be determined. The inequality (17) can also be translated as a more standard norm bound relation of the kind

$$\Delta_{\mathcal{F}} \in \mathbb{F}_{\mathcal{R}}(\mathcal{F}_0) \quad \Rightarrow \quad \|\Delta_{\mathcal{F}}\|_{\mathcal{R}^{-1}} \leq \gamma \|\mathcal{F}_0\|_{\mathcal{R}^{-1}}, \tag{18}$$

where $\|\cdot\|_{\mathcal{R}}$ denotes a weighted Frobenius or two norm. This inequality is evidence that the norm of $\Delta_{\mathcal{F}} \in \mathbb{F}_{\mathcal{R}}(\mathcal{F}_0)$ is limited to being a fraction of the norm of the nominal filter $\mathcal{F}_0$. Another interpretation is obtained in terms of a norm bound on the amplitude of the noise signal

$$w_{\Delta_{\mathcal{F}}}(k) = \begin{pmatrix} w_y(k) \\ w_{x_f}(k) \end{pmatrix} := \Delta_{\mathcal{F}} \begin{pmatrix} y(k) \\ x_f(k) \end{pmatrix}, \tag{19}$$

for which

$$\Delta_{\mathcal{F}} \in \mathbb{F}_{\mathcal{R}}(\mathcal{F}_0) \quad \Rightarrow \quad \|w_{\Delta_{\mathcal{F}}}(k)\|_{\mathcal{R}^{-1}} \leq \gamma \left\| \mathcal{F}_0 \begin{pmatrix} y(k) \\ x_f(k) \end{pmatrix} \right\|_{\mathcal{R}^{-1}}. \tag{20}$$

The above interpretation relates the uncertainty set $\mathbb{F}_{\mathcal{R}}(\mathcal{F}_0)$ to the uncertainty models considered in the recent work [20].

The weighting factor $\mathcal{R}$ plays an interesting role in the definition of $\mathbb{F}_{\mathcal{R}}(\mathcal{F}_0)$ and can have a major impact on the reduction of the conservatism of the design conditions to be derived. Roughly speaking, the matrix $\mathcal{R}$ can play the same role as a *scaling matrix*[1] in robust H_∞ analysis [21]. Using the LMI conditions to be derived in the next section, one can simultaneously perform the design of both the filter parameters $\mathcal{F}$ and the scaling matrix $\mathcal{R}$. If desired, one can also set $\mathcal{R}$ to a constant value without destroying the linearity of the design conditions. However, notice that, if the objective of fixing $\mathcal{R}$ is to establish a certain fixed weight on (17–18) and (20), say, $\mathcal{R} = \bar{\mathcal{R}}$, then one can still use a scaling matrix $\mathcal{R} = \lambda\bar{\mathcal{R}}$, where λ is a positive scalar to be determined. Leaving the scalar λ as a variable can be of much help in reducing conservatism (see the numerical example in section 6).

Throughout the rest of this paper, the norm minimization problem defined in (7) is replaced with

$$\min_{\mathcal{F}_0} \rho_p\left(\mathcal{F}_0\right), \tag{21}$$

where the function ρ_p is a *guaranteed cost* function, that is, it satisfies the inequality

$$\|H_{we}(z; \mathcal{F}_0 + \Delta_{\mathcal{F}})\|_p \leq \rho_p(\mathcal{F}_0) \quad \forall \Delta_{\mathcal{F}} \in \mathbb{F}_{\mathcal{R}}(\mathcal{F}_0). \tag{22}$$

In other words, the function ρ_p is an upper bound to the H_p norm of the uncertain transfer function $H_{we}(z; \mathcal{F}_0 + \Delta_{\mathcal{F}})$ that holds for all $\Delta_{\mathcal{F}} \in \mathbb{F}_{\mathcal{R}}(\mathcal{F}_0)$.

[1] Notice that when $\mathcal{R}$ is a scalar it can be canceled on both sides of (17) and (18).

4. Main result. We are not aware of any available design technique that can effectively solve the filter design problems stated in the previous section, where the filter parameter is subject to uncertainties $\Delta_{\mathcal{F}} \in \mathbb{F}_{\mathcal{R}}(\mathcal{F}_0)$. In the following paragraphs we will show that a much simpler design problem, that is, one that can be stated as a set of LMI, can be obtained if uncertainties are introduced in the transformed set of parameters $\mathcal{K}$ defined in (8). That is, we will consider the filter design problem, where the transformed set of parameters $\mathcal{K}$ is perturbed as

$$\mathcal{K} = \mathcal{K}_0 + \Delta_{\mathcal{K}}, \qquad \Delta_{\mathcal{K}} \in \mathbb{F}_{\mathcal{W}}(\mathcal{K}_0), \tag{23}$$

where the scaling $\mathcal{W}$ will be chosen to maintain equivalence between $\mathbb{F}_{\mathcal{R}}(\mathcal{F}_0)$ and $\mathbb{F}_{\mathcal{W}}(\mathcal{K}_0)$. More specifically, $\mathcal{W}$ will be chosen to ensure that we can find the optimal solution to problem (21)–(22) by solving an equivalent but simpler problem, where the perturbations act on the transformed set of filter variables. This is made possible due to the result in the following lemma.

LEMMA 3. *Let $\mathcal{S}$ and $\mathcal{T}$ be any square and nonsingular matrices of appropriate dimensions. Then $\Delta_{\mathcal{F}} \in \mathbb{F}_{\mathcal{R}}(\mathcal{F}_0)$ if and only if $\mathcal{S}\Delta_{\mathcal{F}}\mathcal{T} \in \mathbb{F}_{\mathcal{S}\mathcal{R}\mathcal{S}^T}(\mathcal{S}\mathcal{F}_0\mathcal{T})$.*

Proof. The proof follows immediately from using the assumption that matrices $\mathcal{S}$ and $\mathcal{T}$ are nonsingular and properly factorizing the variables and matrices appearing in the definition of $\mathbb{F}_{\mathcal{R}}(\mathcal{F}_0)$. □

Lemma 3 deserves two remarks. The first is that it makes explicit how the scaling matrix $\mathcal{W}$ must be chosen to cope with the one to one change of variables in the form $\mathcal{K} = \mathcal{S}\mathcal{F}\mathcal{T}$ that will be used to parameterize the transformed set of filter parameters. Notice that the corresponding "transformed" scaling $\mathcal{W} = \mathcal{S}\mathcal{R}\mathcal{S}^T$ depends exclusively on $\mathcal{S}$. Second, equivalence between $\mathbb{F}_{\mathcal{R}}(\mathcal{F}_0)$ and $\mathbb{F}_{\mathcal{W}}(\mathcal{K}_0)$ is achieved when the change of variables is performed simultaneously on the nominal filter $\mathcal{F}_0$ and on the parametric uncertainty $\Delta_{\mathcal{F}}$. These properties enables us to determine a solution to problem (21) by equivalently rewriting the inequality that defines the guaranteed cost function (22) in the form

$$\|H_{we}(z; \mathcal{K}_0 + \Delta_{\mathcal{K}})\|_p \le \rho_p(\mathcal{K}_0) \quad \forall \Delta_{\mathcal{K}} \in \mathbb{F}_{\mathcal{S}\mathcal{R}\mathcal{S}^T}(\mathcal{K}_0), \tag{24}$$

which is expressed entirely in terms of the transformed variables $(\mathcal{K}_0, \Delta_{\mathcal{K}}) = (\mathcal{S}\mathcal{F}_0\mathcal{T}, \mathcal{S}\Delta_{\mathcal{F}}\mathcal{T})$. Notice that the assumption that $\mathcal{S}$ and $\mathcal{T}$ are nonsingular and square matrices is naturally satisfied whenever the order of the filter is the same as the order of the plant.

In the following lemma we develop an inequality associated with a perturbation on the transformed set of parameters $\mathcal{K}$. This inequality will be used to derive the main result of this paper.

LEMMA 4. *If there exists a symmetric and positive definite matrix $\mathcal{W}$ such that*

$$\begin{bmatrix} \mathcal{Q} + \mathcal{B}\mathcal{K}_0\mathcal{C} + \mathcal{C}^T\mathcal{K}_0^T\mathcal{B}^T - \mathcal{B}\mathcal{W}\mathcal{B}^T & \gamma\mathcal{C}^T\mathcal{K}_0^T \\ \gamma\mathcal{K}_0\mathcal{C} & \mathcal{W} \end{bmatrix} > 0, \tag{25}$$

then

$$\mathcal{Q} + \mathcal{B}\left(\mathcal{K}_0 + \Delta_{\mathcal{K}}\right)\mathcal{C} + \mathcal{C}^T\left(\mathcal{K}_0 + \Delta_{\mathcal{K}}\right)^T\mathcal{B}^T > 0 \quad \forall\, \Delta_{\mathcal{K}} \in \mathbb{F}_{\mathcal{W}}(\mathcal{K}_0). \tag{26}$$

Proof. Applying the Schur complement on (25), one obtains that for all $\Delta_{\mathcal{K}} \in \mathbb{F}_{\mathcal{W}}(\mathcal{K}_0)$,

$$\begin{aligned} \mathcal{Q} + \mathcal{B}\mathcal{K}_0\mathcal{C} + \mathcal{C}^T\mathcal{K}_0^T\mathcal{B}^T &> \mathcal{B}\mathcal{W}\mathcal{B}^T + \gamma^2\mathcal{C}^T\mathcal{K}_0^T\mathcal{W}^{-1}\mathcal{K}_0\mathcal{C} \\ &> \mathcal{B}\mathcal{W}\mathcal{B}^T + \mathcal{C}^T\Delta_{\mathcal{K}}^T\mathcal{W}^{-1}\Delta_{\mathcal{K}}\mathcal{C} \\ &> -\mathcal{B}\Delta_{\mathcal{K}}\mathcal{C} - \mathcal{C}^T\Delta_{\mathcal{K}}^T\mathcal{B}^T, \end{aligned}$$

which recovers (26). □

The condition stated in the above lemma is only sufficient. Yet it has been extensively used in the filtering and control to characterize computable robustness conditions as, for instance, in [12, 21]. However, notice that the scaling matrix $\mathcal{W}$ enters the above condition linearly so that is can be freely optimized. This will help reduce the conservatism of this condition.

The above two lemmas will be combined to show that the optimal solution to the problem (21) subject to the transformed guaranteed cost function (24), for $p = \{2, \infty\}$, can be formulated and solved in terms of LMI conditions. We first consider the case when the multiplier $\mathcal{R}$ is a free optimization variable.

THEOREM 1 (H_2 filtering). *If there exist matrices G and $\mathcal{K}_0$, partitioned as in* (8)*, and symmetric matrices Y, Z, W, E, H such that the LMI*

$$\begin{bmatrix} Z & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ Z & Y-H & \bullet & \bullet & \bullet & \bullet & \bullet \\ A^TZ & A^TY+C_y^TF^T+Q^T & Z & \bullet & \bullet & \bullet & \bullet \\ A^TZ & A^TY+C_y^TF^T & Z & Y & \bullet & \bullet & \bullet \\ B^TZ & B^TY+D_y^TF^T & \mathbf{0} & \mathbf{0} & \mathbf{I} & \bullet & \bullet \\ \mathbf{0} & \mathbf{0} & \gamma RC_y+\gamma L & \gamma RC_y & \gamma RD_y & E & \bullet \\ \mathbf{0} & \mathbf{0} & \gamma FC_y+\gamma Q & \gamma FC_y & \gamma FD_y & G^T & H \end{bmatrix} > 0, \tag{27}$$

$$\begin{bmatrix} W-E & \bullet & \bullet & \bullet & \bullet & \bullet \\ C_z^T-C_y^TR^T-L^T & Z & \bullet & \bullet & \bullet & \bullet \\ C_z^T-C_y^TR^T & Z & Y & \bullet & \bullet & \bullet \\ D_z^T-D_y^TR^T & \mathbf{0} & \mathbf{0} & \mathbf{I} & \bullet & \bullet \\ \mathbf{0} & \gamma RC_y+\gamma L & \gamma RC_y & \gamma RD_y & E & \bullet \\ \mathbf{0} & \gamma FC_y+\gamma Q & \gamma FC_y & \gamma FD_y & G^T & H \end{bmatrix} > 0, \tag{28}$$

$$\operatorname{trace}(W) < \mu \tag{29}$$

have a feasible solution, then the nominal filter

$$\mathcal{F}_0 = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & V^{-1} \end{bmatrix} \mathcal{K}_0 \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & Z^{-1}U^{-1} \end{bmatrix}, \tag{30}$$

where U and V are nonsingular otherwise arbitrary matrices chosen to satisfy $Y + VUZ = Z$, is such that

$$\|H_{we}(z; \mathcal{F}_0 + \Delta_{\mathcal{F}})\|_2^2 \le \rho_2(\mathcal{F}_0) := \mu \quad \forall \Delta_{\mathcal{F}} \in \mathbb{F}_{\mathcal{R}}(\mathcal{F}_0), \tag{31}$$

where $\mathbb{F}_{\mathcal{R}}(\mathcal{F}_0)$ is as defined in (17) *with the scaling matrix*

$$\mathcal{R} := \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & V^{-1} \end{bmatrix} \begin{bmatrix} E & G \\ G^T & H \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & V^{-T} \end{bmatrix}. \tag{32}$$

Proof. Defining

$$\Delta_{\mathcal{K}} := \begin{bmatrix} \Delta_R & \Delta_L \\ \Delta_F & \Delta_Q \end{bmatrix}, \qquad \mathcal{S} := \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & V \end{bmatrix}, \qquad \mathcal{T} := \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & UZ \end{bmatrix},$$

the nominal filter parameters (30) and their parametric perturbations can be recovered from $(\mathcal{K}_0, \Delta_{\mathcal{K}})$, for U, V, and Z nonsingular, by the formulas

$$\mathcal{F}_0 = \mathcal{S}^{-1}\mathcal{K}_0\mathcal{T}^{-1}, \quad \Delta_{\mathcal{F}} = \mathcal{S}^{-1}\Delta_{\mathcal{K}}\mathcal{T}^{-1}.$$

Hence, it is possible to conclude from the result of Lemma 3 that the constraint $\Delta_{\mathcal{F}} \in \mathbb{F}_{\mathcal{R}}(\mathcal{F}_0)$ can be replaced without loss of generality by $\Delta_{\mathcal{K}} \in \mathbb{F}_{\mathcal{S}\mathcal{R}\mathcal{S}^T}(\mathcal{K}_0)$, as indicated in (24).

With that in mind, it suffices to show that (27)–(29) guarantee robustness with respect to all $\Delta_{\mathcal{K}} \in \mathbb{F}_{\mathcal{S}\mathcal{R}\mathcal{S}^T}(\mathcal{K}_0)$. This can be done with the help of Lemma 4. Notice that a perturbed version of (9), where $\mathcal{K}$ is replaced with $\mathcal{K}_0 + \Delta_{\mathcal{K}}$, can be written as (26) with

$$\mathcal{Q} := \begin{bmatrix} Z & Z & ZA & ZA & ZB \\ Z & Y & YA & YA & YB \\ A^TZ^T & A^TY^T & Z & Z & \mathbf{0} \\ A^TZ^T & A^TY^T & Z & Y & \mathbf{0} \\ B^TZ & B^TY & \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix}, \quad \mathcal{B} := \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \\ \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \mathcal{C}^T := \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \\ C_y^T & \mathbf{I} \\ C_y^T & \mathbf{0} \\ D_y^T & \mathbf{0} \end{bmatrix},$$

while a perturbed inequality (10) is in the form (26) with

$$\mathcal{Q} := \begin{bmatrix} W & C_z & C_z & D_z \\ C_z^T & Z & Z & \mathbf{0} \\ C_z^T & Z & Y & \mathbf{0} \\ D_z^T & \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix}, \qquad \mathcal{B} := \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \qquad \mathcal{C}^T := -\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ C_y^T & \mathbf{I} \\ C_y^T & \mathbf{0} \\ D_y^T & \mathbf{0} \end{bmatrix}.$$

Therefore we can define the variables

$$\begin{bmatrix} E & G \\ G^T & H \end{bmatrix} := \mathcal{S}\mathcal{R}\mathcal{S}^T = \mathcal{W}$$

to obtain both inequalities (27) and (28) directly from Lemma 4. □

THEOREM 2 (H_∞ filtering). *If there exist matrices G and $\mathcal{K}_0$, partitioned as in* (8)*, and symmetric matrices Y, Z, E, H such that the LMI*

(33)
$$\begin{bmatrix} Z & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ Z & Y-H & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ A^TZ & A^TY + C_y^TF^T + Q^T & Z & \bullet & \bullet & \bullet & \bullet & \bullet \\ A^TZ & A^TY + C_y^TF^T & Z & Y & \bullet & \bullet & \bullet & \bullet \\ B^TZ & B^TY + D_y^TF^T & \mathbf{0} & \mathbf{0} & \mu\mathbf{I} & \bullet & \bullet & \bullet \\ \mathbf{0} & G & C_z - RC_y - L & C_z - RC_y & D_z - RD_y & \mu\mathbf{I} - E & \bullet & \bullet \\ \mathbf{0} & \mathbf{0} & \gamma RC_y + \gamma L & \gamma RC_y & \gamma RD_y & \mathbf{0} & E & \bullet \\ \mathbf{0} & \mathbf{0} & \gamma FC_y + \gamma Q & \gamma FC_y & \gamma FD_y & \mathbf{0} & G^T & H \end{bmatrix} > 0,$$

has a feasible solution, then the nominal filter $\mathcal{F}_0$ given in (30) *is such that*

$$\|H_{we}(z; \mathcal{F}_0 + \Delta_{\mathcal{F}})\|_\infty \le \rho_\infty(\mathcal{F}_0) := \mu \quad \forall \Delta_{\mathcal{F}} \in \mathbb{F}_{\mathcal{R}}(\mathcal{F}_0), \tag{34}$$

where $\mathbb{F}_{\mathcal{R}}(\mathcal{F}_0)$ is defined with the scaling matrix $\mathcal{R}$ given by (32).

Proof. This proof follows the same pattern as the proof of Theorem 1 and is thus omitted. □

The constraints stated in Theorems 1 and 2 are all LMI. The scalar μ can be used to define the guaranteed cost function (22). The global optimal solution to the guaranteed cost problem (21) can be obtained by minimizing the scalar μ subject to the given LMI.

It is interesting to observe that under the assumption that the scaling matrix $\mathcal{R}$ is a free variable, the filter provided by Theorem 1 shares with the one proposed

in [17] the property that its state space realization is irrelevant as far as the upper bound of the estimation error is concerned. As in Lemmas 1 and 2, the state space parameterization of the optimal filter obtained in Theorems 1 and 2 can be arbitrarily chosen by changing the matrices U and V. However, notice that, from (32), the choice of V does affect the multiplier $\mathcal{R}$. For instance, for the particular choice $V = \mathbf{I}$, the sets $\mathbb{F}_{\mathcal{R}}(\mathcal{F}_0) \equiv \mathbb{F}_{\mathcal{S}\mathcal{R}\mathcal{S}^T}(\mathcal{K}_0)$. Due to the coupling condition (32), this property does not remain valid when the scaling matrix is fixed. This special case is treated in detail in the following paragraphs.

When $\mathcal{R}$ is a given constant matrix, the variable V, which is associated with the choice of filter realization, becomes part of the optimization variables by the relation (32). In general, the introduction of (32) in the form of a constraint in the optimization design problem destroys the desired convexity properties. However, in the important case when $\mathcal{R}$ is a given matrix with the block diagonal structure

$$\bar{\mathcal{R}} = \begin{bmatrix} \bar{\mathcal{R}}_1 & \mathbf{0} \\ \mathbf{0} & \bar{\mathcal{R}}_2 \end{bmatrix}, \tag{35}$$

one can show that convexity is preserved, still leading to an LMI design problem. In this case, which is possibly the most meaningful for modeling implementation uncertainty, the following corollaries to Theorems 1 and 2 apply.

COROLLARY 3. *Let $\bar{\mathcal{R}}$ be partitioned as in* (35)*. If there exist a positive scalar λ, matrix $\mathcal{K}_0$, partitioned as in* (8)*, and symmetric matrices Y, Z, W, E, H such that the LMI* (27)–(29) *with the additional linear constraints*

$$E = \lambda \bar{\mathcal{R}}_1, \quad G = \mathbf{0}, \tag{36}$$

have a feasible solution, then the nominal filter $\mathcal{F}_0$ given in (30) *with*

$$V = \lambda^{-1/2} H^{1/2} \bar{\mathcal{R}}_2^{-1/2}$$

is such that $\|H_{we}(z; \mathcal{F}_0 + \Delta_{\mathcal{F}})\|_2^2 \leq \rho_2(\mathcal{F}_0) := \mu$ for all $\Delta_{\mathcal{F}} \in \mathbb{F}_{\bar{\mathcal{R}}}(\mathcal{F}_0)$.

COROLLARY 4. *Let $\bar{\mathcal{R}}$ be partitioned as in* (35)*. If there exist a positive scalar λ, matrix $\mathcal{K}_0$, partitioned as in* (8)*, and symmetric matrices Y, Z, E, H such that the LMI* (33) *with the additional linear constraints* (36) *has a feasible solution, then the nominal filter $\mathcal{F}_0$ given in* (30) *with $V = \lambda^{-1/2} H^{1/2} \bar{\mathcal{R}}_2^{-1/2}$ is such that $\|H_{we}(z; \mathcal{F}_0 + \Delta_{\mathcal{F}})\|_\infty \leq \rho_\infty(\mathcal{F}_0) := \mu$ for all $\Delta_{\mathcal{F}} \in \mathbb{F}_{\bar{\mathcal{R}}}(\mathcal{F}_0)$.*

Proof. Corollaries 3 and 4 can be proved in the same way. If $\mathcal{R} = \lambda \bar{\mathcal{R}}$, given in (35), then from (32) and (36) we have that

$$V^{-1} H V^{-T} = \lambda \bar{\mathcal{R}}_2,$$

which is satisfied by the choice of $V = \lambda^{-1/2} H^{1/2} \bar{\mathcal{R}}_2^{-1/2}$. Also notice that $\mathbb{F}_{\bar{\mathcal{R}}}(\mathcal{F}_0) = \mathbb{F}_{\lambda \bar{\mathcal{R}}}(\mathcal{F}_0)$. ∎

In Corollaries 3 and 4, the matrix V (and, consequently, matrix U) is automatically chosen by the optimization problem and cannot be picked by the designer, as in Theorems 1 and 2. This implies that the state space realization of the optimal filter is obtained as a result of the optimization procedure. In this sense, Theorems 1 and 2 simultaneously design the optimal filter transfer function *and* its realization. This result is in accordance with the well-known fact that some realizations of the same filter transfer function can be better than others for implementation [1, 2].

Also notice that, as in [17, 15], all of the above results can be shown to reduce to the standard Kalman filter and to the central H_∞ filter when $\gamma = 0$. In fact, with

$\gamma = 0$ the scaling matrices E, G, and H can be set arbitrarily close to zero, reducing these inequalities to the ones given in [17].

An interesting comment on the technical device used to prove Theorems 1 and 2 is that, to the authors' knowledge, it is the first time that a filtering or control robustness property has been derived directly from the transformed inequalities given in Lemmas 1 and 2. The robustness analysis was performed with respect to the transformed set of filter parameters $\mathcal{K}$ instead of the actual filter parameters $\mathcal{F}$. Working with the transformed parameters $\mathcal{K}$ instead of $\mathcal{F}$ was the key that permitted us to both incorporate and keep the scaling matrix $\mathcal{R}$ as an extra variable in the obtained design inequalities.

5. Extension to plant parameter uncertainty. In this section the assumption that the plant parameters are exactly known is relaxed. Following [17], the plant parameters, collected in the matrix

$$\mathcal{M} := \begin{bmatrix} A & B \\ C_z & D_z \\ C_y & D_y \end{bmatrix}, \tag{37}$$

are allowed to be unknown but to belong to the convex hull of N given extreme matrices (see [22]). That is,

$$\mathcal{M} \in \mathbb{M} := \operatorname{co}\left\{ \mathcal{M}_i := \begin{bmatrix} A_i & (B)_i \\ (C_z)_i & (D_z)_i \\ (C_y)_i & (D_y)_i \end{bmatrix}, \quad i = 1, \dots, N \right\}. \tag{38}$$

The goal is to derive design procedures that enable one to take into account the filter parameter uncertainty as well as the plant parameter uncertainty. This can be done by defining guaranteed cost functions that satisfy the general inequality

$$\|H_{we}(z; \mathcal{F}_0 + \Delta_{\mathcal{F}}, \mathcal{M})\|_p \le \rho_p(\mathcal{F}_0) \qquad \forall\, \Delta_{\mathcal{F}} \in \mathbb{F}_{\mathcal{R}}(\mathcal{F}_0) \quad \forall\, \mathcal{M} \in \mathbb{M}. \tag{39}$$

In the case of plant parametric uncertainty, the uncertain transfer function $H_{we}(z; \mathcal{F}_0 + \Delta_{\mathcal{F}}, \mathcal{M})$ depends on both the filter perturbation $\Delta_{\mathcal{F}}$ and the uncertain plant parameters $\mathcal{M}$. The guaranteed cost ρ_p provides an upper bound to the H_p norm of $H_{we}(z; \mathcal{F}_0 + \Delta_{\mathcal{F}}, \mathcal{M})$, which holds for all $\Delta_{\mathcal{F}} \in \mathbb{F}_{\mathcal{R}}(\mathcal{F}_0)$ *and* all $\mathcal{M} \in \mathbb{M}$. Following [17, 22], a guaranteed cost function ρ_2 can be built by generating N copies of the LMI (27)–(29) whose plant parameters correspond to those of $\mathcal{M}_i$, $i = 1, \dots, N$. The same procedure can be applied to generate ρ_∞ from appropriate versions of the inequalities given in Theorem 2.

The rationale behind this procedure is that the LMI (27)–(29) and (33) are all affine on the parameters of the uncertain matrix $\mathcal{M}$. Therefore, a convex combination of feasible inequalities (27)–(29) and (33) can be used to generate appropriate feasible inequalities for each $M \in \mathbb{M}$ (see [17, 22]). It is also straightforward to generate robust filtering conditions, which use a parameter dependent Lyapunov function to test stability following the methods of [18, 19]. The derivation of these extensions and the corresponding LMI conditions are left to the interested reader.

6. Numerical example. Consider the system in the form (1)–(3) with matrices

$$\left[\begin{array}{c|c} A & B \\ \hline C_z & D_z \\ \hline C_y & D_y \end{array}\right] = \left[\begin{array}{cc|ccc} 0.8 & 0.9 & 1 & 0 & 0 \\ 0.3 & -0.5 & 0 & 1 & 0 \\ \hline 1 & 1 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 & 1 \end{array}\right].$$

TABLE 1
H_2 filter transfer functions.

γ	0.01	0.05	0.1
Design I	$\frac{0.63z(z+0.91)}{(z-0.33)(z+0.53)}$	$\frac{0.86z(z+0.74)}{(z-0.09)(z+0.60)}$	$\frac{0.62z(z+0.70)}{(z-0.02)(z+0.64)}$
Design II	$\frac{0.73z(z+0.85)}{(z-0.25)(z+0.55)}$	$\frac{1.01z(z+0.68)}{z(z+0.68)}$	$\frac{0.65z(z+0.68)}{z(z+0.68)}$

TABLE 2
H_2 filtering performance.

	Nominal cost			Guaranteed cost			Round-off gain		
γ	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
Kalman	1.36	1.36	1.36		—		13.74	13.74	13.74
Design I	1.36	1.75	3.26	1.56	3.61	5.63	10.88	1.14	0.07
Design II	1.39	1.94	3.29	1.68	3.67	5.65	7.00	0.00	0.00

In the next sections we will design filters $\mathcal{F}_0$ to minimize an upper bound to the H_p norm, $p = \{2, \infty\}$, of $H_{we}(z; \mathcal{F}_0 + \Delta_{\mathcal{F}})$, where $\Delta_{\mathcal{F}} \in \mathbb{F}_{\mathcal{R}}(\mathcal{F}_0)$ for the values of $\gamma = \{0.01, 0.05, 0.1\}$.

6.1. H_2 filtering. A standard stationary Kalman filter has been designed to serve as a template for the H_2 filtering design. The transfer function $\mathcal{F}_{\mathrm{K}}(z)$ of the Kalman filter is given by

$$\mathcal{F}_{\mathrm{K}}(z) = \frac{0.58z(z+0.96)}{(z-0.38)(z+0.52)}.$$

The following two H_2 filter design methods have been tried:

Design I: Minimize μ subject to the LMI (27)–(29) with a full variable scaling $\mathcal{R}$ (Theorem 1).

Design II: Minimize μ subject to the LMI (27)–(29) and the linear constraint (36) with a fixed scaling $\bar{\mathcal{R}} = \mathbf{I}$ (Corollary 3).

The transfer functions of Designs I and II are given in Table 1. These filters are associated with the performance measures given in Table 2. In this table the "Nominal cost" is the H_2 norm of $H_{we}(z; \mathcal{F}_0)$, and the "Guaranteed cost" is the square root of the value of μ obtained by solving the problems in Theorem 1 and Corollary 3. The "Round-off gain" shown in the third column of Table 2 is a measure that has not been directly optimized by solving the design problems of this paper. It was computed after determining the minimal round-off gain realizations for the designed filters according to [2, 5].

It is important to notice that the solution of the problems in Theorem 1 and Corollary 3 implies the simultaneous design of a filter realization. Moreover, if one is to use these results to compare performance with a given filter realization $\mathcal{F}$, it is necessary to impose an additional constraint relating $\mathcal{F}$ and $\mathcal{K}$. As noted before, such a relationship is nonlinear and destroys the convexity of the problem. For these reasons, and to be able to compare our results with other techniques, we have arbitrarily chosen $V = \mathbf{I}$, in which case the relationship between $\mathcal{F}$ and $\mathcal{K}$ becomes linear. This is, in a certain sense, equivalent to fixing the admissible filter realizations. Additionally, it also implies $\mathbb{F}_{\mathcal{R}}(\mathcal{F}_0) \equiv \mathbb{F}_{\mathcal{S}\mathcal{R}\mathcal{S}^T}(\mathcal{K}_0)$, which seems to be an appropriate choice for comparing a given realization to one obtained by the methods proposed in this paper.

TABLE 3
H_∞ filter transfer functions.

γ	0.01	0.05	0.1
Design III	$\frac{0.72(z+1.09)}{z+0.24}$	$\frac{1.16(z+0.57)}{z+0.51}$	$\frac{1.20(z+0.13)}{z+0.13}$
Design IV	$\frac{0.85(z+0.93)}{z+0.36}$	$\frac{1.11(z+0.73)}{z+0.60}$	$\frac{1.17(z+0.70)}{z+0.65}$

Using this idea, we have computed guaranteed cost for the standard Kalman filter. However, no results are shown in the table since the LMI in Theorem 1 and Corollary 3 become infeasible for $\gamma = 3.3\times10^{-4}$ and $\gamma = 1.2\times10^{-4}$, respectively. This is evidence of the importance of allowing the optimization to freely tune the filter realization.

Also notice that the optimal round-off gain realizations of the filters produced by Designs I and II have lower round-off gains than the optimal coordinates of the Kalman filter, although this measure of performance has not been directly optimized. This asserts the effectiveness of the uncertainty domain $\mathbb{F}_{\mathcal{R}}(\mathcal{F}_0)$ in producing *non-fragile* filters.

It is interesting to try to interpret the effect of the parameter perturbation on the performance measures and in the filter transfer functions. From Table 1, one can notice that, as the parameter uncertainty increases, the filter transfer function tends to a constant, with no dynamics. This trend can help explain why the round-off gains have decreased accordingly in Table 2. This effect is even accentuated in Design II, where the filter optimization has fewer parameters with which to play. It seems interesting that, to maximize the performance in the presence of an increasing implementation uncertainty, the designed filter has been made simpler by the design procedure.

6.2. H_∞ filtering. This time we have designed a standard H_∞ filter to serve as a template for the H_∞ filtering design. The transfer function $\mathcal{F}_{\rm H}(z)$ of the standard H_∞ filter is given by

$$\mathcal{F}_{\rm H}(z) = \frac{0.76(z+1.02)}{(z+0.29)}.$$

It is interesting to notice that the standard optimal H_∞ filter design already presents a pole-zero cancellation. In fact, this feature will be present in all designed filters. As before, two H_∞ filter design methods have been tried as follows:

Design III: Minimize μ subject to the LMI (33) with a full variable scaling $\mathcal{R}$ (Theorem 2).

Design IV: Minimize μ subject to the LMI (33) and the linear constraint (36) with a fixed scaling $\bar{\mathcal{R}} = \mathbf{I}$ (Corollary 4).

The transfer functions of Designs III and IV are given in Table 3 and their performance measures in Table 4. The guaranteed costs for the standard H_∞ filter have been computed by setting $V = \mathbf{I}$ and solving the design LMI for the given realization. The costs in the first line corresponds to the case when the scaling $\mathcal{R}$ has been optimized, whereas the costs in the second line have been obtained with $\bar{\mathcal{R}} = \mathbf{I}$.

The same trends observed in the H_2 filter design appear in the H_∞ design. Notice especially the tendency to simplify the filter by reducing it to a constant scaling. Interestingly enough, this tendency now appears more accentuated in Design III,

TABLE 4
H_∞ filtering performance.

	Nominal cost			Guaranteed cost			Round-off gain		
γ	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
H_∞	1.44	1.44	1.44	1.96 13.67	5.19 178.8	11.64 593.5	6.98	6.98	6.98
Design III	1.56	3.61	5.63	1.66	3.93	7.28	10.88	1.14	0.07
Design IV	1.68	3.67	5.65	1.75	4.89	7.82	7.00	0.00	0.00

TABLE 5
Actual H_2 filtering performance for the example in section 6.1.

	$\bar{\sigma}$		
γ	0.01	0.05	0.1
Kalman	1.36	1.55	2.13
Design II	1.42	2.05	3.59

where the scaling $\mathcal{R}$ has been allowed to be optimized. Notice again a significant reduction in the round-off gain.

6.3. Estimating conservativeness. In the previous section, the performances of the designed filters have been evaluated with respect to guaranteed cost functions, which are upper bounds to the norm of the filtering error system. In this section we attempt to access the filter performance by directly evaluating an estimate of the actual error system norms. The idea is to estimate the conservativeness of the method and to evaluate its practical usefulness. We restrict our attention to the case of H_2 filtering design with $\mathcal{R} = \mathbf{I}$.

For each filter design $\mathcal{F}_0$, we randomly generate a number of perturbation matrices $\Delta_{\mathcal{F}_j}$, $j = 1, \ldots, M$, in $\mathbb{F}_\mathcal{R}(\mathcal{F}_0)$. The following procedure was used in this generation:

1. Generate a square matrix Δ_j, with the same number of rows as in $\mathcal{F}_0$, where all entries are normally distributed random real numbers with zero mean and unitary variance.
2. Compute $\Delta_{\mathcal{F}j} = \frac{\gamma}{\|\Delta_j\|}\Delta_j\mathcal{F}_0$ and $\mathcal{F}_j = \mathcal{F}_0 + \Delta_{\mathcal{F}j}$.
3. If $\mathcal{F}_j$ is asymptotically stable, set $\sigma_j = \|H_{we}(z; \mathcal{F}_j)\|_2$; otherwise set $\sigma_j = \infty$.

All filter perturbations generated by the above procedure are guaranteed to be in the boundary of the set $\mathbb{F}_\mathcal{R}(\mathcal{F}_0)$, and an estimate of the error system norm can be computed as

$$\bar{\sigma} := \max_{j=1,\ldots,M} \sigma_j \approx \sup_{\Delta_\mathcal{F} \in \mathbb{F}_\mathcal{R}(\mathcal{F}_0)} \|H_{we}(z; \mathcal{F}_0 + \Delta_\mathcal{F})\|_2.$$

Strictly speaking, $\bar{\sigma}$ provides a lower bound to the error system norm, which serves as a good approximation for the worst case norm as M becomes large. In our experiments we have set $M = 1000$.

We start by evaluating the problem described in section 6.1. The results of the above numerical experiment applied to the filters previously labelled Kalman and Design II are shown in Table 5 for several values of γ. Note that the performance of Design II is, as expected, always below the designed guaranteed cost but, surprisingly, above the performance of the nominal Kalman filter design. We credit this apparently surprising behavior to a relative insensitivity of this particular example to variations on the filter parameters, rather than to an overconservativeness of our approach. We try to support this claim in the following paragraphs.

TABLE 6
Actual H_2 filtering performance for the second example in section 6.3.

	$\bar{\sigma}$		
γ	0.01	0.05	0.1
Kalman	1.30	∞	∞
Corollary 3	1.28	1.35	1.43

The previous example might leave the impression that the proposed design procedure produces robust filters at the expense of sacrificing performance; this impression is possibly due to the implicit conservativeness in the design inequalities. In order to show that the proposed procedure can indeed lead to efficient robust designs, we consider another simple example with

$$\left[\begin{array}{c|c} A & B \\ \hline C_z & D_z \\ \hline C_y & D_y \end{array}\right] = \left[\begin{array}{ccc|cc} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ -0.5 & 0.5 & 0.1 & 1 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 \\ \hline -0.5 & 0.25 & 0.5 & 0 & 1 \end{array}\right].$$

For the above example, the Kalman filter has a nominal performance of $\|H_{we}(z;\mathcal{F}_0)\|_2 = 1.28$. The results of the above numerical experiment applied to the Kalman filter for several values of γ are shown in the first row of Table 6. These results show that the Kalman filter is extremely sensitive to parameter variations: a relative perturbation of size $\gamma = 0.01$ already implies some loss of performance but, more important, for higher values of γ, the Kalman filter becomes unstable (indicated as an infinite cost). This highly sensitive system seems to provide a better benchmark for our design methodology. After computing the robust filters using Corollary 3, we run the numerical experiment and obtain the performance estimate shown in the second row of Table 6. In this example, the design procedure not only produced a filter, which performs as efficiently as the nominal Kalman filter for $\gamma = 0.01$, but also produced robust filters for $\gamma = 0.05$ and $\gamma = 0.1$, which were able to withstand large parameter perturbations without becoming unstable and without sacrificing too much performance.

In the above example, an aspect that might have contributed to the sensitivity of the Kalman filter to parameter variations is the increased order of the filter. Generally speaking, it seems natural to expect that state space realizations of filters become more sensitive to parameter variations as the order of the filter (and the associated matrix dimensions) increases. In order to verify this trend we modify the system used in section 6.1 to augment its order. More specifically, we introduce a delay on the measurement signal $y(k)$. This produces the third order system

$$\left[\begin{array}{c|c} A & B \\ \hline C_z & D_z \\ \hline C_y & D_y \end{array}\right] = \left[\begin{array}{ccc|ccc} 0.8 & 0.9 & 0 & 1 & 0 & 0 \\ 0.3 & -0.5 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ \hline 1 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 & 1 \end{array}\right].$$

Due to the introduction of the delay on the measurement, the Kalman filter now has a nominal performance of $\|H_{we}(z;\mathcal{F}_0)\|_2 = 1.72$.

Table 7 shows the results of the numerical experiment performed on both the nominal Kalman filter design and the filters produced by Corollary 3. Note that now

TABLE 7
Actual H_2 filtering performance for the example in section 6.1 with a measurement delay.

	$\bar{\sigma}$		
γ	0.01	0.05	0.1
Kalman	1.74	2.16	3.70
Corollary 3	1.75	2.15	3.62

the performances of the Kalman filter and the robust filter designed for $\gamma = 0.01$ are practically the same, whereas Corollary 3 produces filters that perform better than the Kalman filter for values of γ greater than 0.05. These results agree with the statement that we should expect the Kalman filter to become more sensitive to parameter variations as the order of the filter increases, in which case the procedure we have proposed provides an effective way to design robust filters.

Finally, note that the main source of conservatism in this design comes from the fact that the guaranteed cost functions we have used evaluate performance with respect to parameter perturbations $\Delta_{\mathcal{F}} \in \mathbb{F}_{\mathcal{R}}(\mathcal{F}_0)$, which are allowed to vary with time. Indeed, this explains the gap between the (time-varying) guaranteed cost values in Table 2 and the (time-invariant) values of $\bar{\sigma}$ in Table 5.

7. Conclusions. A new procedure has been proposed for designing filters which are robust in the presence of perturbations on the filter parameters. The filters are obtained by minimizing guaranteed H_2 and H_∞ cost functions developed by confining the filter parametric uncertainty in a region defined by a quadratic inequality. The size of this uncertainty region depends on the size of the filter parameters, and the maximum allowed parametric perturbation is specified as a percentage of the size of the filter gains. Both the transfer function and the realization of the robust filter are simultaneously designed. The optimization problems to be solved have constraints specified in terms of LMI, whose global optimal solutions can be determined using convex programming. The numerical examples suggest that the proposed technique may produce filters with reduced round-off noise gain, although this performance measure is not directly optimized in the design process.

REFERENCES

[1] M. GEVERS AND G. LI, *Parametrizations in Control, Estimation and Filtering Problems*, Springer-Verlag, London, 1993.
[2] D. WILLIAMSON, *Finite wordlength design of digital Kalman filters for state estimation*, IEEE Trans. Automat. Control, 30 (1985), pp. 930–939.
[3] D. WILLIAMSON AND K. KADIMAN, *Optimal finite wordlength linear quadratic regulators*, IEEE Trans. Automat. Control, 34 (1989), pp. 1218–1228.
[4] K. LIU, R. E. SKELTON, AND K. GRIGORIADIS, *Optimal controllers for finite wordlength implementation*, IEEE Trans. Automat. Control, 37 (1992), pp. 1294–1304.
[5] S. Y. HWANG, *Minimum uncorrelated unit noise in state-space digital filtering*, IEEE Trans. Acoustics Speech Signal Process., 25 (1977), pp. 273–281.
[6] G. AMIT AND U. SHAKED, *Minimization of roundoff errors in digital realizations of Kalman filters*, IEEE Trans. Acoustics Speech Signal Process., 37 (1989), pp. 1980–1982.
[7] M. C. DE OLIVEIRA AND R. E. SKELTON, *Synthesis of controllers with finite precision considerations*, in Digital Controller Implementation and Fragility: A Modern Perspective, R. S. H. Istepanian and J. F. Whidborne eds., Springer-Verlag, New York, 2001, pp. 229–251.
[8] L. H. KEEL AND S. P. BHATTACHARYYA, *Robust, fragile or optimal*, IEEE Trans. Automat. Control, 42 (1997), pp. 1098–1105.
[9] L. H. KEEL AND S. P. BHATTACHARYYA, *Authors' reply to: "Comments on 'Robust, fragile or optimal'" by P. M. Mäkilä*, IEEE Trans. Automat. Control, 43 (1998), p. 1268.

[10] P. Dorato, *Non-fragile controller design: An overview*, in Proceedings of the 1998 American Control Conference, (Philadelphia), vol. 5, IEEE, Piscataway, NJ, 1998, pp. 2829–2831.

[11] D. Famularo, P. Dorato, C. T. Abdallah, W. H. Haddad, and A. Jadbabaie, *Robust non-fragile LQ controllers: The static state feedback case*, Internat. J. Control, 73 (2000), pp. 159–165.

[12] G. H. Yang and J. L. Wang, *Robust nonfragile Kalman filtering for uncertain linear systems with estimator gain uncertainty*, IEEE Trans. Automat. Control, 46 (2001), pp. 343–348.

[13] W. M. Haddad and J. R. Corrado, *Robust resilient dynamic controllers for systems with parametric uncertainty and controller gain variations*, Internat. J. Control, 73 (2000), pp. 1405–1423.

[14] L. H. Keel and S. P. Bhattacharyya, *Stability margins and digital implementation of controllers*, in Proceedings of the 1998 American Control Conference, (Philadelphia), vol. 5, IEEE, Piscataway, NJ, 1998, pp. 2852–2856.

[15] J. C. Geromel, *Optimal linear filtering under parameter uncertainty*, IEEE Trans. Signal Process., 47 (1999), pp. 168–175.

[16] Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia, 1994.

[17] J. C. Geromel, J. Bernussou, G. Garcia, and M. C. de Oliveira, *H_2 and H_∞ robust filtering for discrete-time linear systems*, SIAM J. Control Optim., 38 (2000), pp. 1353–1368.

[18] J. C. Geromel, M. C. de Oliveira, and J. Bernussou, *Robust filtering of discrete-time linear systems with parameter dependent Lyapunov functions*, SIAM J. Control Optim., 41 (2002), pp. 700–711.

[19] M. C. de Oliveira, J. Bernussou, and J. C. Geromel, *A new discrete-time robust stability condition*, Systems Control Lett., 37 (1999), pp. 261–265.

[20] A. H. Sayed, *A framework for state-space estimation with uncertain models*, IEEE Trans. Automat. Control, 46 (2001), pp. 998–1013.

[21] V. Balakrishnan, Y. Huang, A. Packard, and J. C. Doyle, *Linear matrix inequalities in analysis with multipliers*, in Proceedings of the 1994 American Control Conference, vol. 2, (Baltimore, MD), IEEE, Piscataway, NJ, 1994, pp. 1228–1232.

[22] J. C. Geromel, P. L. D. Peres, and J. Bernussou, *On a convex parameter space method for linear control design of uncertain systems*, SIAM J. Control Optim., 29 (1991), pp. 381–402.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 2, pp. 531–563

STATE-SPACE FORMULAS FOR THE NEHARI–TAKAGI PROBLEM FOR NONEXPONENTIALLY STABLE INFINITE-DIMENSIONAL SYSTEMS*

JOSEPH A. BALL†, KALLE M. MIKKOLA‡, AND AMOL J. SASANE§

Abstract. We obtain state-space formulas for the solution of the Nehari–Takagi/suboptimal Hankel norm approximation problem for infinite-dimensional systems with a nonexponentially stable generator, via the method of J-spectral factorization. We make key use of a purely frequency-domain solution of the problem.

Key words. Nehari–Takagi problem, infinite-dimensional systems, state-space formulas

AMS subject classifications. 41A30, 47B35, 47B50, 47N70, 93B28

DOI. 10.1137/S0363012903433024

1. Introduction.

The Hankel norm approximation problem has received a lot of attention, both in the mathematical and engineering literature (see Adamjan, Arov, and Kreĭn [1], Ball and Helton [4], Ball and Ran [7], Glover [19], and Doyle, Glover, and Zhou [17]). Its importance in control theory is due to its connections with the model reduction problem (see [19]).

In order to state the suboptimal Hankel norm approximation problem, we will need a few preliminaries. First we recall the definition of the (frequency-domain) Hankel operator corresponding to a symbol $G \in L_\infty(i\mathbb{R}, \mathbb{C}^{p\times m})$ and the definition of its singular values. Let $\mathbb{C}_+ := \{s \in \mathbb{C} \mid \mathrm{Re}(s) > 0\}$ and $\mathbb{C}_- := \{s \in \mathbb{C} \mid \mathrm{Re}(s) < 0\}$.

Let $H_2(\mathbb{C}_+, \mathbb{C}^k)$ denote the set of all analytic functions $f : \mathbb{C}_+ \to \mathbb{C}^k$ such that

$$\|f\|_2 := \sup_{\zeta>0} \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} \|f(\zeta + i\omega)\|^2 d\omega \right)^{\frac{1}{2}} < \infty.$$

Analogously one defines $H_2(\mathbb{C}_-, \mathbb{C}^k)$. For $G \in L_\infty(i\mathbb{R}, \mathbb{C}^{p\times m})$ we define the *Hankel operator with symbol* G, denoted by H_G, acting from $H_2(\mathbb{C}_-, \mathbb{C}^m)$ to $H_2(\mathbb{C}_+, \mathbb{C}^p)$, as follows:

$$H_G f = P_{H_2(\mathbb{C}_+, \mathbb{C}^p)}(M_G f) \quad \text{for } f \in H_2(\mathbb{C}_-, \mathbb{C}^m),$$

where M_G is the multiplication map on $L_2(i\mathbb{R}, \mathbb{C}^m)$ induced by G, and $P_{H_2(\mathbb{C}_+, \mathbb{C}^p)}$ is the orthogonal projection operator from $L_2(i\mathbb{R}, \mathbb{C}^p)$ onto $H_2(\mathbb{C}_+, \mathbb{C}^p)$. The Hankel operator is bounded, that is, $H_G \in \mathcal{L}(H_2(\mathbb{C}_-, \mathbb{C}^m), H_2(\mathbb{C}_+, \mathbb{C}^p))$.

Now we recall the notion of singular values of a bounded linear operator from a Hilbert space $\mathcal{H}_1$ to a Hilbert space $\mathcal{H}_2$. For $k \in \{1, 2, \dots\}$ the kth *singular value*

*Received by the editors August 12, 2003; accepted for publication (in revised form) August 27, 2004; published electronically August 31, 2005.
http://www.siam.org/journals/sicon/44-2/43302.html

†Department of Mathematics, 460 McBryde, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0123 (ball@ math.vt.edu).

‡Institute of Mathematics, Box 1100, FIN-02015 Helsinki University of Technology, Helsinki, Finland (kalle.mikkola@iki.fi). The work of this author was supported by the Academy of Finland under grant 203946.

§Department of Mathematics, London School of Economics, Houghton St., London WC2A 2AE, UK (a.j.sasane@lse.ac.uk).

(denoted by $\sigma_k(\mathtt{H})$) of an operator $\mathtt{H} \in \mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$ is defined to be the distance with respect to the norm in $\mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$ of $\mathtt{H}$ from the set of operators in $\mathcal{L}(\mathcal{H}_1, \mathcal{H}_2)$ of rank at most $k-1$. Thus $\sigma_1(\mathtt{H}) = \|\mathtt{H}\|$, and $\sigma_1(\mathtt{H}) \geq \sigma_2(\mathtt{H}) \geq \sigma_3(\mathtt{H}) \geq \cdots \geq 0$. For $G \in L_\infty(i\mathbb{R}, \mathbb{C}^{p\times m})$, we refer to the singular values of H_G simply as the *Hankel singular values of* G.

Let $H_{\infty,k}(\mathbb{C}_-, \mathbb{C}^{p\times m})$ denote the set of all $p \times m$ matrix-valued functions K of a complex variable defined in the open left half-plane such that $K = G_{\mathtt{f}} + F$, where F is an element in $H_\infty(\mathbb{C}_-, \mathbb{C}^{p\times \mathtt{m}})$ and $G_{\mathtt{f}}$ is the transfer function of a finite-dimensional system with order at most k, with all its poles in the open left half-plane. The set $H_{\infty,k}(\mathbb{C}_-, \mathbb{C}^{p\times m})$ is a subset of $L_\infty(i\mathbb{R}, \mathbb{C}^{p\times m})$.

We recall the following well-known result of Adamjan, Arov, and Kreĭn [1], adapted here to the right half-plane setting: If $G \in L_\infty(i\mathbb{R}, \mathbb{C}^{p\times m})$, then

$$\inf_{K\in H_{\infty,k}(\mathbb{C}_-,\mathbb{C}^{p\times m})} \|G(i\cdot) + K(i\cdot)\|_\infty = \sigma_{k+1}(G).$$

We are now ready to give the statement of the suboptimal Hankel norm approximation problem, which is also known as the *Nehari–Takagi problem*. The *suboptimal Hankel norm approximation problem* is the following: Let $G(i\cdot) \in L_\infty(\mathbb{R}, \mathbb{C}^{p\times m})$. If $\sigma_{k+1} < \sigma < \sigma_k$, then find $K \in H_{\infty,k}(\mathbb{C}_-, \mathbb{C}^{p\times m})$ such that $\|G(i\cdot) + K(i\cdot)\|_\infty \leq \sigma$. In fact, the authors of [1], working with Schmidt pairs of the Hankel operator, also gave a linear-fractional description for the set of all solutions of the suboptimal Hankel norm approximation problem; later work of Ball and Helton [4] obtained such a linear-fractional description, but via an indefinite-metric Beurling–Lax theorem combined with some Kreĭn-space projective geometry.

Now suppose that G is in fact the transfer function of some well-posed linear system; that is, G is not simply an L_∞ function, but it has the special form $G(s) = C(sI - A)^{-1}B$, where (A, B, C) are the generators of the system. Then by a *state-space solution* to the suboptimal Hankel norm approximation problem we mean a K given explicitly in terms of the A, B, C operators. For the case of rational $G(s)$ with system-generators (A, B, C) equal to finite matrices, a state-space solution of the Hankel norm approximation problem has been obtained by Kung and Lin [29], Glover [19], Ball and Ran [7], and Ball, Gohberg, and Rodman [3, Chapter 20].

In Curtain and Sasane [14, 13], state-space solutions to the suboptimal Hankel norm approximation problem were given for two classes of infinite-dimensional state-linear systems, but under the assumption that A generates an *exponentially* stable, strongly continuous semigroup. Recall that a semigroup $\{T(t)\}_{t\geq 0}$ on a Hilbert space X is said to be exponentially stable if there exist positive constants M and ϵ such that

$$\|T(t)\| \leq Me^{-\epsilon t} \quad \text{for all } \ t \geq 0.$$

However, there exists an important class of systems with a transfer function $G \in H_\infty(\mathbb{C}_-, \mathbb{C}^{p\times m})$ for which A does not generate an exponentially stable semigroup (see, for example, Oostveen [33]), for example, if A is the generator of a *strongly stable semigroup*, that is, a semigroup satisfying

$$T(t)x \to 0 \ \text{ as } \ t \to \infty \quad \text{for all } \ x \in X.$$

Roughly speaking, the rate of convergence to zero is not uniform but depends on the choice of the element in the Hilbert space. An elementary example of a semigroup

which is strongly stable but not exponentially stable is given by e^{tA} on ℓ_2, where

$$A = \begin{bmatrix} -1 & & & & \\ & -\frac{1}{2} & & & \\ & & -\frac{1}{3} & & \\ & & & \ddots & \end{bmatrix} \in \mathcal{L}(\ell_2).$$

In this article, we consider an even weaker notion of stability, the so-called *nonexponentially stable semigroup*, namely, a semigroup whose generator has a nonnegative growth bound. Clearly this class encompasses both strongly stable semigroups and (hence surely) exponentially stable semigroups; thus we emphasize that the prefix "non" is really short for "not necessarily."

Earlier work on the problem for infinite-dimensional systems includes the work of Curtain and Ran [12], which handled the case of Pritchard–Salamon systems, and of Glover, Curtain, and Partington [20], where approximating solutions to the optimal Hankel norm approximation problem were obtained without assuming exponential stability, but only for the case that the Hankel operator is nuclear, a rather strong assumption. In this paper, we give solutions to the suboptimal Hankel norm approximation problem for infinite-dimensional systems having a nonexponentially stable semigroup. Our solution depends on a preliminary result which obtains the linear-fractional parameterization of the set of all solutions in purely frequency-domain terms via the solution Θ of a certain J-spectral factorization problem. The fact that Θ may be unbounded in our general setting makes the analysis much more delicate. We give three proofs of this key frequency-domain result in order to point out the close connections with results already existing in the literature. The first proof shows how the result can be reduced to the result of Adamjan, Arov, and Kreĭn in [1]. The second proof revisits the proof of Ball and Helton [4] with special care given to the details required to handle the general case where Θ may be unbounded. The third proof revisits the homotopy argument appearing in [3, 40]. The standard homotopy argument works well in case the coefficients of the linear-fractional parameterization and the free parameter are continuous up to the boundary. We show how an approximation argument can be used to reduce the general case here to the classical situation, at least for the proof that every admissible free parameter leads to a solution of the Nehari–Takagi problem. The proof that any solution of the Nehari–Takagi problem necessarily is of the linear-fractional form follows the ideas appearing in the second proof.

The outline of the paper is as follows. In section 2, we give the key frequency-domain result (the reduction of the parameterization of the set of all solutions of the suboptimal Hankel norm approximation problem to solving a certain J-spectral factorization problem), along with our three proofs of this result. In section 3 we use this frequency-domain result to parameterize all solutions to the suboptimal Hankel norm approximation problem for infinite-dimensional state-space systems for which the generator is not necessarily exponentially stable. Finally in the last section, we give state-space solutions for well-posed linear systems by applying the result in section 3 to the associated reciprocal system.

2. The key frequency-domain result.

THEOREM 2.1. *Let $G \in H_\infty(\mathbb{C}_+, \mathbb{C}^{p\times m})$ and let $H_G\colon H_2(\mathbb{C}_-, \mathbb{C}^m) \mapsto H_2(\mathbb{C}_+, \mathbb{C}^p)$ denote the corresponding Hankel operator, with the singular values $\sigma_1 \geq \sigma_2 \geq \cdots (\geq 0)$. Suppose that $\sigma_k > \sigma > \sigma_{k+1}$. Then there exists a matrix function $\Lambda\colon \mathbb{C}_- \mapsto$*

$\mathbb{C}^{(p+m)\times(p+m)}$, *uniquely determined up to a* $(p+m)\times(p+m)$*-matrix right constant factor* U *satisfying* $U^*\left[\begin{smallmatrix} I_p & 0 \\ 0 & -I_m \end{smallmatrix}\right]U = \left[\begin{smallmatrix} I_p & 0 \\ 0 & -I_m \end{smallmatrix}\right]$, *such that*

S1. $\Lambda(i\omega)^*\left[\begin{smallmatrix} I_p & 0 \\ 0 & -I_m \end{smallmatrix}\right]\Lambda(i\omega) = \left[\begin{smallmatrix} I_p & G(i\omega) \\ 0 & I_m \end{smallmatrix}\right]^*\left[\begin{smallmatrix} I_p & 0 \\ 0 & -\sigma^2 I_m \end{smallmatrix}\right]\left[\begin{smallmatrix} I_p & G(i\omega) \\ 0 & I_m \end{smallmatrix}\right]$ *for* $\omega \in \mathbb{R}$;

S2. $\frac{1}{\cdot - 1}\Lambda \in H_2(\mathbb{C}_-, \mathbb{C}^{(p+m)\times(p+m)})$;

S3. Λ *is invertible* (*i.e., there exists a* $V\colon \mathbb{C}_- \mapsto \mathbb{C}^{(p+m)\times(p+m)}$ *such that* $\Lambda(s)V(s) = I_{p+m}$ *for* $s \in \mathbb{C}_-$) *and* $\frac{1}{\cdot - 1}V \in H_2(\mathbb{C}_-, \mathbb{C}^{(p+m)\times(p+m)})$.

Define

$$\Theta(i\omega)\left(= \begin{bmatrix} \Theta_{11}(i\omega) & \Theta_{12}(i\omega) \\ \Theta_{21}(i\omega) & \Theta_{22}(i\omega) \end{bmatrix}\right) = \begin{bmatrix} I_p & G(i\omega) \\ 0 & I_m \end{bmatrix} V(i\omega) \quad \text{for } \omega \in \mathbb{R}.$$

Then we have the following: $K\colon \mathbb{C}_- \mapsto \mathbb{C}^{p\times m}$ *such that* $K \in H_{\infty,k}(\mathbb{C}_-, \mathbb{C}^{p\times m})$ *and* $\|G(i\cdot) + K(i\cdot)\|_\infty \le \sigma$ *if and only if*

(2.1)
$$G(i\omega) + K(i\omega) = (\Theta_{11}(i\omega)Q(i\omega) + \Theta_{12}(i\omega))(\Theta_{21}(i\omega)Q(i\omega) + \Theta_{22}(i\omega))^{-1} \quad \text{for } \omega \in \mathbb{R}$$

for some $Q\colon \mathbb{C}_- \mapsto \mathbb{C}^{p\times m}$ *such that* $Q \in H_\infty(\mathbb{C}_-, \mathbb{C}^{p\times m})$ *and* $\|Q(i\cdot)\|_\infty \le 1$.

For the application of Theorem 2.1 in section 3, we note that a sufficient condition for the validity of S2 is the existence of a constant $\Lambda(\infty) \in \mathbb{C}^{(p+m)\times(p+m)}$ such that $\Lambda - \Lambda(\infty) \in H_2(\mathbb{C}_-, \mathbb{C}^{(p+m)\times(p+m)})$.

By using the transformation

$$f(s) \mapsto \widetilde{f}(z) := f\left(\frac{1-z}{1+z}\right)$$

and observing (via the Jacobi change-of-variable formula) that

$$\int_{\mathbb{T}} |\widetilde{f}(z)|^2 \, |dz| = \int_{i\mathbb{R}} |f(s)|^2 \, \frac{|ds|}{1+|s|^2},$$

we see that Theorem 2.1 is exactly equivalent to the following discrete-time version. Here $\mathbb{D}$ denotes the unit disk, $\mathbb{D}_e$ denotes the exterior of the unit disk (including the point at infinity), and $\mathbb{T}$ denotes the unit torus (equal to the boundary of $\mathbb{D}$).

THEOREM 2.2. *Let* $G \in H_\infty(\mathbb{D}_e)$ *and let* $H_G\colon H_2(\mathbb{D}, \mathbb{C}^m) \mapsto H_2(\mathbb{D}, \mathbb{C}^p)^\perp$ *be the associated Hankel operator, with singular values* $\sigma_1 \ge \sigma_2 \ge \cdots (\ge 0)$. *Suppose that* $\sigma_k > \sigma > \sigma_{k+1}$. *Then there exists a unique matrix function* $\Lambda\colon \mathbb{D} \mapsto \mathbb{C}^{(p+m)\times(p+m)}$, *uniquely determined up to a* $(p+m)\times(p+m)$*-matrix right constant factor* U *satisfying* $U^*\left[\begin{smallmatrix} I_p & 0 \\ 0 & -I_m \end{smallmatrix}\right]U = \left[\begin{smallmatrix} I_p & 0 \\ 0 & -I_m \end{smallmatrix}\right]$, *such that*

S′1. $\Lambda(\zeta)^*\left[\begin{smallmatrix} I_p & 0 \\ 0 & -I_m \end{smallmatrix}\right]\Lambda(\zeta) = \left[\begin{smallmatrix} I_p & G(\zeta) \\ 0 & I_m \end{smallmatrix}\right]^*\left[\begin{smallmatrix} I_p & 0 \\ 0 & -\sigma^2 I_m \end{smallmatrix}\right]\left[\begin{smallmatrix} I_p & G(\zeta) \\ 0 & I_m \end{smallmatrix}\right]$ *for* $\zeta \in \mathbb{T}$;

S′2. $\Lambda \in H_2(\mathbb{D}, \mathbb{C}^{(p+m)\times(p+m)})$;

S′3. Λ *is invertible* (*i.e., there exists a* $V\colon \mathbb{D} \mapsto \mathbb{C}^{(p+m)\times(p+m)}$ *such that* $\Lambda(z)V(z) = I_{p+m}$ *for* $z \in \mathbb{D}$) *and* $V \in H_2(\mathbb{D}, \mathbb{C}^{(p+m)\times(p+m)})$.

Define

$$\Theta(\zeta)\left(= \begin{bmatrix} \Theta_{11}(\zeta) & \Theta_{12}(\zeta) \\ \Theta_{21}(\zeta) & \Theta_{22}(\zeta) \end{bmatrix}\right) = \begin{bmatrix} I_p & G(\zeta) \\ 0 & I_m \end{bmatrix} V(\zeta) \quad \text{for } \zeta \in \mathbb{T}. \tag{2.2}$$

Then we have the following: $K\colon \mathbb{D} \mapsto \mathbb{C}^{p\times m}$ *is such that* $K \in H_{\infty,k}(\mathbb{D}, \mathbb{C}^{p\times m})$ *and* $\|(G+K)|_{\mathbb{T}}\|_\infty \le \sigma$ *if and only if*

$$G(\zeta) + K(\zeta) = (\Theta_{11}(\zeta)Q(\zeta) + \Theta_{12}(\zeta))(\Theta_{21}(\zeta)Q(\zeta) + \Theta_{22}(\zeta))^{-1} \quad \text{for } \zeta \in \mathbb{T} \tag{2.3}$$

for some $Q\colon \mathbb{D} \mapsto \mathbb{C}^{p\times m}$ *such that* $Q \in H_\infty(\mathbb{D}, \mathbb{C}^{p\times m})$ *and* $\|Q\|_\infty \le 1$.

We next indicate several proofs of Theorem 2.2 based on various different points of view. We first need to lay out a few preliminaries.

2.1. Preliminaries. For p and m positive integers we let $\mathbb{C}^{p\times m}$ be the space of complex $p \times m$ matrices M with norm $\|M\|$ equal to the induced operator norm:

$$\|M\| = \sup_{x\in\mathbb{C}^m:\ \|x\|_2\le 1} \|Mx\|_2,$$

where $\|x\|_2$ is the standard Euclidean 2-norm on $\mathbb{C}^m$. The trace norm $\mathrm{Tr}(M)$ of a $p \times m$ matrix M is defined by

$$\mathrm{Tr}(M) = \mathrm{tr}(M^*M)^{1/2},$$

where the *trace* $\mathrm{tr}(A)$ of an $m \times m$ matrix A is defined by

$$\mathrm{tr}(A) = \sum_{k=1}^{p} \langle Ae_k, e_k\rangle,$$

where $\{e_1, \dots, e_n\}$ is any orthonormal basis for $\mathbb{C}^p$—a good reference for the natural infinite-dimensional setting for this material is [21, Chapter VII]. We let $L_\infty(\mathbb{T}, \mathbb{C}^{p\times m})$ denote the space of measurable $p \times m$ matrix-valued functions on the unit circle $\mathbb{T}$ with finite essential supremum (supremum up to sets of measure zero) norm uniformly bounded:

$$\|F\|_\infty = \operatorname*{ess\text{-}sup}_{\zeta\in\mathbb{T}} \|F(\zeta)\| < \infty.$$

We let $L_1(\mathbb{T}, \mathbb{C}^{m\times p})$ be the space of measurable $m \times p$ matrix-valued functions f on $\mathbb{T}$ with integrable trace norm:

$$\|f\|_1 = \frac{1}{2\pi}\int_{\mathbb{T}} \mathrm{Tr}(f(\zeta))|d\zeta|.$$

It is well known (see, e.g., [38, page 197]) that the Banach space $L_\infty(\mathbb{T}, \mathbb{C}^{p\times m})$ can be identified as the dual of the Banach space $L_1(\mathbb{T}, \mathbb{C}^{m\times p})$ under the duality pairing

$$[F, f] = \frac{1}{2\pi}\int_{\mathbb{T}} \mathrm{tr}(F(\zeta)f(\zeta))\ d|\zeta| \quad \text{for } F \in L_\infty(\mathbb{T}, \mathbb{C}^{p\times m}) \text{ and } f \in L_1(\mathbb{T}, \mathbb{C}^{m\times p}).$$

Therefore, in addition to its norm topology, $L_\infty(\mathbb{T}, \mathbb{C}^{p\times m})$ carries a *weak-$*$ topology* induced by its duality with respect to $L_1(\mathbb{T}, \mathbb{C}^{p\times m})$. We shall have use of the following facts concerning this weak-$*$ topology.

Proposition 2.3.

(1) *A subspace $\mathcal{S}$ of $L_\infty(\mathbb{T}, \mathbb{C}^{p\times m})$ is closed in the weak-$*$ topology of $L_\infty(\mathbb{T}, \mathbb{C}^{p\times m})$ if and only if whenever $\{F_n\}_{n=1,2,\dots}$ is a sequence of elements of $\mathcal{S}$ converging weak-$*$ to $F \in L_\infty(\mathbb{T}, \mathbb{C}^{p\times m})$, then in fact $F \in \mathcal{S}$.*

(2) *Suppose that $\{F_n\}_{n=1,2,\dots}$ is a sequence of elements of $L_\infty(\mathbb{T}, \mathbb{C}^{p\times m})$ converging* pointwise boundedly *to the element $F \in L_\infty(\mathbb{T}, \mathbb{C}^{p\times m})$, i.e.,*

$$\lim_{n\to\infty} F_n(\zeta) = F(\zeta) \quad \text{for almost all } \zeta \in \mathbb{T}, \text{ and}$$
$$\|F_n(\zeta)\| \le M \quad \text{for some } M < \infty \text{ for all } n = 1, 2, \dots.$$

Then F_n converges to F in the weak-$*$ topology of $L_\infty(\mathbb{T}, \mathbb{C}^{p\times m})$.

Proof. By the Kreĭn–Šmulian theorem (see [42, Theorem 10.1, page 173]), a subspace $\mathcal{S}$ of $L_\infty(\mathbb{T}, \mathbb{C}^{p\times m})$ (or more generally, a convex subset) is weak-$*$ closed if and only if $\mathcal{S}\cap\{F\colon \|F\|_\infty \le r\}$ is weak-$*$ closed for each $r > 0$. Since $\mathcal{S}$ is a subspace, by homogeneity it suffices to consider only the case $r = 1$. As $L_\infty(\mathbb{T}, \mathbb{C}^{p\times m})$ is the dual of the *separable* space $L_1(\mathbb{T}, \mathbb{C}^{m\times p})$, it follows that the weak-$*$ topology on the unit ball of $L_\infty(\mathbb{T}, \mathbb{C}^{p\times m})$ is metrizable (see [18, Theorem 102, page 174]). Hence, to show that $\mathcal{S}$ is closed in the weak-$*$ topology, it suffices to show that $\mathcal{S}$ is closed under sequential weak-$*$ limits as asserted. This proves part (1) of Proposition 2.3.

Suppose now that $\{F_n\}_{n=1,2,\dots}$ is a sequence of elements of $L_\infty(\mathbb{T}, \mathbb{C}^{p\times m})$ converging pointwise boundedly to F. To show that F_n converges to F in the weak-$*$ topology, we must show that

$$\lim_{n\to\infty}[F_n, f] = \lim_{n\to\infty}\frac{1}{2\pi}\int_{\mathbb{T}}\operatorname{tr}(F_n(\zeta)f(\zeta))|d\zeta| = \frac{1}{2\pi}\int_{\mathbb{T}}\operatorname{tr}(F(\zeta)f(\zeta))|d\zeta| \tag{2.4}$$

for each choice of $f \in L_1(\mathbb{T}, \mathbb{C}^{m\times p})$. Note that the assumptions imply that

$$\lim_{n\to\infty}\operatorname{tr}(F_n(\zeta)f(\zeta)) = \operatorname{tr}(F_n(\zeta)f(\zeta)) \quad \text{for almost all } \zeta \in \mathbb{T}.$$

By the standard trace estimate

$$|\operatorname{tr}(AB)| \le \operatorname{Tr}(AB) \le \|A\|\operatorname{Tr}(B),$$

we have

$$|\operatorname{tr}(F_n(\zeta)f(\zeta))| \le \|F_n(\zeta)\|\operatorname{Tr}(f(\zeta)) \le M\operatorname{Tr}(f(\zeta)),$$

where $M\operatorname{Tr}(f(\cdot))$ is integrable by the definition of $f \in L_1(\mathbb{T}, \mathbb{C}^{p\times m})$. It now follows from the Lebesgue dominated convergence theorem (see, e.g., [37, Theorem 16, page 91]) that (2.4) follows as required. This completes the proof of part (2) of Proposition 2.3. ∎

The subspace $H_\infty(\mathbb{D}, \mathbb{C}^{p\times m})$ can be viewed as the subspace of $L_\infty(\mathbb{D}, \mathbb{C}^{p\times m})$ consisting of functions $F \in L_\infty(\mathbb{T}, \mathbb{C}^{p\times m})$ such that the Fourier coefficients of negative index vanish:

$$\frac{1}{2\pi}\int_{\mathbb{T}}F(\zeta)\zeta^n|d\zeta| = 0 \quad \text{for } n = -1, -2, \dots.$$

The subspace $H_\infty(\mathbb{D}, \mathbb{C}^{p\times m})$ can also be viewed (via identification through nontangential-limit boundary values) as the space of analytic $p \times m$ matrix-valued functions on the unit disk which are uniformly bounded there:

$$\|F\|_\infty = \sup_{z\in\mathbb{D}}\|F(z)\| < \infty.$$

We define $H_{\infty,k}(\mathbb{D}, \mathbb{C}^{p\times m})$ as consisting of all elements G of $L_\infty(\mathbb{T}, \mathbb{C}^{p\times m})$ for which the associated Hankel operator $H_G\colon H_2(\mathbb{D}, \mathbb{C}^m) \mapsto H_2(\mathbb{D}, \mathbb{C}^p)^\perp$ given by

$$H_G\colon f \mapsto P_{H_2(\mathbb{D},\mathbb{C}^p)^\perp}M_G|_{H_2(\mathbb{D},\mathbb{C}^m)}$$

has rank equal to k. Here M_G denotes the multiplication operator associated with G. Equivalently, the *Hankel matrix*

$$[H_G] = \begin{bmatrix} g_{-1} & g_{-2} & g_{-3} & \cdots \\ g_{-2} & g_{-3} & g_{-4} & \cdots \\ g_{-3} & g_{-4} & g_{-5} & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix} : \ell^2(\mathbb{Z}_+, \mathbb{C}^m) \mapsto \ell^2(\mathbb{Z}_+, \mathbb{C}^p)$$

based on the Fourier coefficients for G,

$$G(z) \sim \sum_{n=-\infty}^{\infty} g_n z^n \quad \text{for } z \in \mathbb{T},$$

has rank equal to k. In what follows we shall use the following result.

PROPOSITION 2.4. *For a given $k \in \{0, 1, 2, \dots\}$, the set*

$$\bigcup_{k' \,:\, k' \le k} H_{\infty,k'}(\mathbb{D}, \mathbb{C}^{p\times m})$$

is closed in the weak-$*$ topology of $L_\infty(\mathbb{T}, \mathbb{C}^{p\times m})$.

Proof. Let us suppose that $G_n \in H_{\infty,k'}(\mathbb{D}, \mathbb{C}^{p\times m})$ for some $k' \le k$ for all $n = 1, 2, \dots$, and that G_n converges to $G \in L_\infty(\mathbb{T}, \mathbb{C}^{p\times m})$ in the weak-$*$ topology. By part (1) of Proposition 2.3, Proposition 2.4 follows if we are able to show that necessarily the limit G is again in $\bigcup_{k' \,:\, k' \le k} H_{\infty,k'}(\mathbb{D}, \mathbb{C}^{p\times m})$. For $f \in H_2(\mathbb{D}, \mathbb{C}^m)$ and $g \in H_2(\mathbb{D}, \mathbb{C}^p)^\perp$ we then have

$$\begin{aligned} \langle H_{G_n} f, g\rangle_{H_2^\perp} &= \frac{1}{2\pi}\int_{\mathbb{T}} g(\zeta)^* G_n(\zeta) f(\zeta) |d\zeta| \\ &= \frac{1}{2\pi}\int_{\mathbb{T}} \operatorname{tr}\left(G_n(\zeta) f(\zeta) g(\zeta)^*\right) |d\zeta|. \end{aligned} \tag{2.5}$$

As $f(\zeta)g(\zeta)^* \in L_1(\mathbb{T}, \mathbb{C}^{m\times p})$ and G_n converges weak-$*$ to G by assumption, we conclude from (2.5) that

$$\begin{aligned} \lim_{n\to\infty} \langle H_{G_n} f, g\rangle_{H_2^\perp} &= \frac{1}{2\pi}\int_{\mathbb{T}} \operatorname{tr}\left(G(\zeta) f(\zeta) g(\zeta)^*\right) |d\zeta| \\ &= \langle H_G f, g\rangle_{H_2^\perp}, \end{aligned} \tag{2.6}$$

i.e., H_{G_n} *converges to* H_G *in the weak operator topology of* $\mathcal{L}(H_2(\mathbb{D}, \mathbb{C}^m), H_2(\mathbb{D}, \mathbb{C}^p)^\perp)$. The fact that $G_n \in H_{\infty,k'}(\mathbb{D}, \mathbb{C}^{p\times m})$ with $k' \le k$ means that

$$\det[\langle H_{G_n} e_j, f_\ell\rangle_{H_2^\perp}]_{j,l=1,\dots,k+1} = 0 \tag{2.7}$$

for all $n = 1, 2, 3, \dots$ for any choice of $k+1$ linearly independent vectors $\{e_1, \dots, e_{k+1}\}$ in $H_2(\mathbb{D}, \mathbb{C}^m)$ and $k+1$ linearly independent vectors $\{f_1, \dots, f_{k+1}\}$ in $H_2(\mathbb{D}, \mathbb{C}^p)^\perp$. Using (2.5) and taking limits in (2.7) then implies that

$$\det[\langle H_G e_j, f_l\rangle_{H_2^\perp}]_{j,l=1,\dots,k+1} = 0 \tag{2.8}$$

for all such $\{e_1, \dots, e_{k+1}\}$ and $\{f_1, \dots, f_{+1}\}$. This then implies that H_G has rank at most k, or, by definition, $G \in H_{\infty,k'}(\mathbb{D}, \mathbb{C}^{p\times m})$ for some $k' \le k$. ∎

Sometimes it is of interest to focus on the "unit ball" of $H_{\infty,k}(\mathbb{D}, \mathbb{C}^{p\times m})$, namely, the set of functions $G \in H_{\infty,k}(\mathbb{D}, \mathbb{C}^{p\times m})$ with $\|G\|_\infty \le 1$. This class is often given a special name, namely, the *generalized Schur class of index* k, denoted as $\mathcal{S}_k(\mathbb{D}, \mathbb{C}^{p\times m})$. The following result concerning the class $\mathcal{S}_k(\mathbb{D}, \mathbb{C}^{p\times m})$ originates in the work of Kreĭn and Langer (see [27, 28]).

PROPOSITION 2.5. *Let $G \in L^\infty(\mathbb{T}, \mathbb{C}^{p\times m})$. Then the following are equivalent:*

(1) $G \in \mathcal{S}_k(\mathbb{D}, \mathbb{C}^{p\times m})$.

(2) *G has a factorization $G = F \cdot B^{-1}$, where F is in $H_\infty(\mathbb{D}, \mathbb{C}^{p\times m})$ with $\|F\|_\infty \leq 1$ and B is an $m \times m$ Blaschke–Potapov product of degree k, and no such representation $G = f' \cdot B'^{-1}$ is possible with B' an $m \times m$ matrix Blaschke–Potapov product of degree $k' < k$.*

(3) *G has meromorphic continuation to $\mathbb{D}$ and, for any choice of vectors $x_1, \ldots, x_N \in \mathbb{C}^p$, points $z_1, \ldots, z_N \in \Omega_G$ (where $\Omega_G \subset \mathbb{D}$ is the domain of analyticity for G), and $N = 1, 2, 3, \ldots$, the Hermitian matrix*

$$\left[\frac{x_i^* x_j - x_i^* G(z_i) G(z_j)^* x_j}{1 - z_i \overline{z_j}}\right] \tag{2.9}$$

has at most k negative eigenvalues, and there is at least one choice of $x_1, \ldots, x_N$, $z_1, \ldots, z_N$, and N for which (2.9) has exactly k negative eigenvalues.

We shall also need an asymptotic version of the maximum modulus theorem for the generalized Schur class $\mathcal{S}_k(\mathbb{D}, \mathbb{C}^{p\times m})$ (with $k < \infty$).

PROPOSITION 2.6. *Suppose G is in the generalized Schur class $\mathcal{S}_k(\mathbb{D}, \mathbb{C}^{p\times m})$, where $k < \infty$, and let $s > 1$. Then there exists an $r < 1$ so that*

$$z \in \mathbb{D}, \quad r < |z| < 1 \Rightarrow z \in \Omega_G, \quad \text{and} \quad \|G(z)\| \leq s.$$

Proof. Let $G = F \cdot B^{-1}$ be the Kreĭn–Langer factorization G and suppose that we are given a number $s > 1$. As $F \in H_\infty(\mathbb{D}, \mathbb{C}^{p\times m})$ with $\|F\|_\infty \leq 1$, we have $\|F(z)\| \leq 1$ for all $z \in \mathbb{D}$ by the maximum modulus theorem for H_∞. As B is a finite matrix Blaschke–Potapov product, B is uniformly continuous on the closed disk $\overline{\mathbb{D}}$, and B^{-1} is uniformly continuous on any annulus $\mathbb{A}_r = \{z \colon r \leq |z| \leq 1\}$ which misses the zeros of B. As B^{-1} has norm 1 on the unit circle, we can therefore guarantee that $\|B^{-1}(z)\| \leq s$ (for any preassigned $s > 1$) as long as we restrict to an annulus $\mathbb{A}_r$ with r sufficiently close to 1. The result now follows. ∎

We also need the following elementary result.

PROPOSITION 2.7. *Suppose that $G \in H_\infty(\mathbb{D}_e, \mathbb{C}^{p\times m})$ with Hankel singular values $\sigma_1 \geq \sigma_2 \geq \cdots (\geq 0)$. If $Q \in H_{\infty,k}(\mathbb{D}, \mathbb{C}^{p\times m})$, then $\|G + Q\|_\infty \geq \sigma_{k+1}$.*

Proof. The Hankel singular values are characterized by

$$\sigma_{k+1}(H_G) = \inf_{X \colon \operatorname{rank} X \leq k} \|H_G - X\|,$$

where X here is an operator from $H_2(\mathbb{D}, \mathbb{C}^m)$ to $H_2(\mathbb{D}, \mathbb{C}^p)$ (see, e.g., [21, Chapter VI, Theorem 1.5, page 98]). In particular, if $K \in H_{\infty,k}(\mathbb{D}, \mathbb{C}^{p\times m})$, then $X = H_K$ has rank equal to k. Hence

$$\|G + K\|_\infty \geq \|H_{G+K}\|_{op} = \|H_G + H_K\|_{op} \geq \inf_{X \colon \operatorname{rank} X \leq k} \|H_G + X\|_{op} = \sigma_{k+1},$$

and the assertion follows. ∎

2.2. Existence of Λ and Θ in Theorem 2.2. In this section we point out some general considerations which guarantee the existence of a function $\Lambda \colon \mathbb{D} \mapsto \mathbb{C}^{(p+m)\times(p+m)}$ satisfying conditions S′1, S′2, and S′3. It then remains to prove that such a Λ leads to a parameterization of the set of all solutions of the Nehari–Takagi problem as in Theorem 2.2. In practice, it then remains to compute Λ (and Θ) in some explicit form in terms of known parameters in the application; this is what we do in section 3 (for the setting of the left half-plane rather than of the unit disk), where $G(s) = C(sI - A)^{-1}B$ is assumed to be the transfer function of a continuous-time linear system having certain (nonexponential) stability properties.

First we need to make a few general observations. The invertibility of $H_G^*H_G - \sigma^2 I$ on $L_2(\mathbb{T}, \mathbb{C}^m)$ is equivalent to σ being in the resolvent set of $[H_G^*H_G]^{1/2}$, i.e., of σ being in a gap of the spectrum of $[H_G^*H_G]^{1/2}$. It is well known that the singular values $\sigma_1 > \sigma_2 > \cdots$ of H_G consist of the points of the spectrum of $[H_G^*H_G]^{1/2}$ which are isolated eigenvalues of finite multiplicity positioned to the right of the continuous spectrum. The condition that σ is in a gap between Hankel singular values implies in particular that σ is in a gap of the spectrum of $[H_G^*H_G]^{1/2}$, and hence implies the invertibility of $H_G^*H_G - \sigma^2 I$ on $L_2(\mathbb{T}, \mathbb{C}^m)$. Further details on singular values in general are given in Lemma 6.2 in Appendix B. For a given matrix function $G \in L_\infty(\mathbb{T}, \mathbb{C}^{p\times m})$, in addition to the notation $H_G\colon H_2(\mathbb{D}, \mathbb{C}^m) \mapsto H_2(\mathbb{D}, \mathbb{C}^p)^\perp$ for the Hankel operator $H_G\colon f \mapsto P_{H_2(\mathbb{D},\mathbb{C}^p)^\perp}(G \cdot f)$ associated with G, we let $T_G\colon H_2(\mathbb{D}, \mathbb{C}^m) \mapsto H_2(\mathbb{D}, \mathbb{C}^p)$ denote the *Toeplitz operator* associated with G,

$$T_G\colon f \mapsto P_{H_2(\mathbb{D},\mathbb{C}^p)}(G \cdot f) \quad \text{for } f \in H_2(\mathbb{D}, \mathbb{C}^m),$$

and we let $M_G\colon L_2(\mathbb{T}, \mathbb{C}^m) \mapsto L_2(\mathbb{T}, \mathbb{C}^p)$ denote the *multiplication* (sometimes also called the *Laurent*) *operator* associated with G,

$$M_G\colon f \mapsto G \cdot f.$$

The next proposition gives a number of conditions equivalent to the invertibility of $H_G^*H_G - \sigma^2 I$ on $H_2(\mathbb{D}, \mathbb{C}^m)$.

PROPOSITION 2.8. *Let $G \in L_\infty(\mathbb{T}, \mathbb{C}^{p\times m})$ and set $A = \left[\begin{smallmatrix} I & G \\ 0 & I \end{smallmatrix}\right] \in L_\infty(\mathbb{T}, \mathbb{C}^{(p+m)\times(p+m)})$. Then the following conditions are equivalent:*

(1) *$H_G^*H_G - \sigma^2 I$ is invertible.*
(2) *The Toeplitz operator $T_{A^*J_\sigma A}$ is invertible on $H_2(\mathbb{D}, \mathbb{C}^{p+m})$.*
(3) *The singular integral operator $S := M_{A^*J_\sigma A}P_{H_2(\mathbb{D},\mathbb{C}^{p+m})} + P_{H_2(\mathbb{D},\mathbb{C}^{p+m})^\perp}$ is invertible on $L_2(\mathbb{T}, \mathbb{C}^m)$.*

Proof. To see that (1) $\Rightarrow$ (2), note that

$$A^*J_\sigma A = \begin{bmatrix} I_p & G \\ 0 & I_m \end{bmatrix}^* \begin{bmatrix} I_p & 0 \\ 0 & -\sigma^2 I_m \end{bmatrix} \begin{bmatrix} I_p & G \\ 0 & I_m \end{bmatrix} = \begin{bmatrix} I_p & G \\ G^* & G^*G - \sigma^2 I_m \end{bmatrix}.$$

Taking Schur complements, we see that invertibility of $T_{A^*J_\sigma A}$ is equivalent to invertibility of

$$\begin{aligned} T_{G^*G-\sigma^2 I_m} - T_{G^*}T_G &= P_{H_2}\left(M_{G^*}M_G - M_{G^*}P_{H_2}M_G\right)|_{H_2} - \sigma^2 I_{H_2} \\ &= P_{H_2}M_{G^*}P_{H_2^\perp}M_G|_{H_2} - \sigma^2 I_{H_2} \\ &= H_G^*H_G - \sigma^2 I_{H_2}, \end{aligned}$$

and (1) $\iff$ (2) follows.

If we decompose $L_2(\mathbb{C}^{p+m})$ in the form $L_2(\mathbb{C}^{p+m}) = \left[\begin{smallmatrix} H_2(\mathbb{D},\mathbb{C}^{p+m}) \\ H_2(\mathbb{D},\mathbb{C}^{p+m})^\perp \end{smallmatrix}\right]$, then the singular integral operator $S := M_{A^*J_\sigma A}P_{H_2} + P_{H_2^\perp}$ has the operator-block representation

$$S = \begin{bmatrix} T_{A^*J_\sigma A} & 0 \\ H_{A^*J_\sigma A} & I_{H_2^\perp} \end{bmatrix}.$$

From the triangular form of this block operator matrix, we see (2) $\iff$ (3). ∎

Theorem VII.2.1 combined with Theorem VIII.4.1 from [10], adapted to our setting, gives the following.

THEOREM 2.9 (see [10, Theorem VII.2.1 and Theorem VIII.4.1] or [31]). *Let $G \in L_\infty(\mathbb{D}, \mathbb{C}^{p\times m})$. Then the following are equivalent:*

(1) *Any of the equivalent conditions* (1), (2), *or* (3) *in Proposition* 2.8 *holds.*
(2) *There exists a function* $\Lambda \in H_2(\mathbb{D}, \mathbb{C}^{(p+m)\times(p+m)})$ *meeting the conditions* S′1, S′2, *and* S′3 *of Theorem* 2.2 *and satisfying the additional condition:* The operator $M_V P_{H_2} M_\Lambda$ $(= M_{\Lambda^{-1}} P_{H_2} M_\Lambda)$ defines a bounded projection operator on $L_2(\mathbb{T}, \mathbb{C}^{p+m})$. *Moreover,* Λ *is uniquely determined up to a* $(p+m)\times(p+m)$*-matrix right constant factor* U *satisfying* $U^* \left[\begin{smallmatrix} I_p & 0 \\ 0 & -I_m \end{smallmatrix}\right] U = \left[\begin{smallmatrix} I_p & 0 \\ 0 & -I_m \end{smallmatrix}\right]$.

We point out that in fact conditions S′1, S′2, and S′3 already determine Λ uniquely up to a constant (without the additional condition on the boundedness of $M_V P_{H_2} M_\Lambda$). Indeed if Λ and Λ' satisfy S′1, S′2, and S′3, then $\Lambda\Lambda'^{-1}(z)$ is analytic on $\mathbb{D}$ and satisfies

$$(\Lambda\Lambda'^{-1})^*(\zeta) \begin{bmatrix} I_p & 0 \\ 0 & -I_m \end{bmatrix} (\Lambda\Lambda'^{-1})(\zeta) = \begin{bmatrix} I_p & 0 \\ 0 & -I_m \end{bmatrix} \quad \text{for } \zeta \in \mathbb{T}. \tag{2.10}$$

We then use the formula $\left[\begin{smallmatrix} I_p & 0 \\ 0 & -I_m \end{smallmatrix}\right] (\Lambda\Lambda'^{-1})^{*-1}(1/\overline{z}) \left[\begin{smallmatrix} I_p & 0 \\ 0 & -I_m \end{smallmatrix}\right]$ to analytically continue $\Lambda\Lambda'^{-1}$ to the exterior of the unit disk. From (2.10) we see that the nontangential boundary values from outside the disk agree with the nontangential boundary values from inside the disk. By using Lemma 6.6 from [32, page 223], we see that the analytic continuation passes through the unit circle as well. Then by Liouville's theorem we see that $\Lambda\Lambda'^{-1}$ must be an invertible constant matrix U. Since Λ and Λ' both satisfy S′1, we see next that the constant matrix U must also satisfy $U^* \left[\begin{smallmatrix} I_p & 0 \\ 0 & -I_m \end{smallmatrix}\right] U = \left[\begin{smallmatrix} I_p & 0 \\ 0 & -I_m \end{smallmatrix}\right]$.

We remark that the version of Theorem 2.9 as formulated in [10] uses the invertibility of the singular integral operator (condition (3) in Proposition 2.8) as the operator theory condition equivalent to the existence of the so-called canonical generalized factorization with respect to L_2.

2.3. Proof of Theorem 2.2 via the Adamjan–Arov–Kreĭn (AAK) theorem. The following result of Adamjan, Arov, and Kreĭn (see [1]) also gives a parameterization of the set of all solutions of the (discrete-time) Nehari–Takagi problem under the assumption that $\sigma_k > \sigma > \sigma_{k+1}$. A thorough recent treatment of the AAK approach can be found in Peller [34].

THEOREM 2.10. *Let* $G \in H_\infty(\mathbb{D}_e)$ *and let* $H_G \colon H_2(\mathbb{D}, \mathbb{C}^m) \mapsto H_2(\mathbb{D}, \mathbb{C}^p)^\perp$ *be the associated Hankel operator, with singular values* $\sigma_1 \geq \sigma_2 \geq \cdots (\geq 0)$. *Suppose that* $\sigma_k > \sigma > \sigma_{k+1}$. *Define*

$$\Theta(\zeta) = \begin{bmatrix} \Theta_{11}(\zeta) & \Theta_{12}(\zeta) \\ \Theta_{21}(\zeta) & \Theta_{22}(\zeta) \end{bmatrix} \in L_2(\mathbb{T}, \mathbb{C}^{(p+m)\times(p+m)})$$

by (*viewed as an operator from* $\mathbb{C}^{p+m}$ *into* $L_2(\mathbb{T}, \mathbb{C}^{p+m})$)

$$\Theta = \begin{bmatrix} \zeta \cdot Z_* e_*^* \gamma_* & H_G Z e^* \gamma \\ \zeta \cdot H_G^* Z_* e_*^* \gamma_* & Z e^* \gamma \end{bmatrix}, \tag{2.11}$$

where $e_* \colon L^2(\mathbb{T}, \mathbb{C}^p) \mapsto \mathbb{C}^p$, $e \colon L^2(\mathbb{T}, \mathbb{C}^m) \mapsto \mathbb{C}^m$, $Z \colon H_2(\mathbb{D}, \mathbb{C}^m) \mapsto H_2(\mathbb{D}, \mathbb{C}^m)$, $Z_* \colon H_2(\mathbb{D}, \mathbb{C}^p) \mapsto H_2(\mathbb{D}, \mathbb{C}^p)$, $\gamma \colon \mathbb{C}^m \mapsto \mathbb{C}^m$, *and* $\gamma_* \colon \mathbb{C}^p \mapsto \mathbb{C}^p$ *are given by*

$$\begin{aligned} &e_* \colon \textstyle\sum_{j=-\infty}^{\infty} \zeta^j f_j \mapsto f_{-1}, && e \colon \textstyle\sum_{j=-\infty}^{\infty} \zeta^j g_j \mapsto g_0, \\ &Z = (I - \sigma^{-2} H_G^* H_G)^{-1}, && Z_* = (I - \sigma^{-2} H_G H_G^*)^{-1}, \\ &\gamma = (e Z e^*)^{-1/2}, && \gamma_* = (e_* Z_* e_*^*)^{-1/2}. \end{aligned}$$

Then the conclusion of Theorem 2.2 *holds with* Θ *given by* (2.11) *rather than by* (2.2).

In [4] it is argued that one way to compute a function $\Theta \in L_2(\mathbb{D}, \mathbb{C}^{(p+m)\times(p+m)})$ meeting the requirement in Theorem 2.2 is as follows: Θ should satisfy the conditions

C′1. $\Theta(\zeta)^*\left[\begin{smallmatrix} I_p & 0 \\ 0 & -\sigma^2 I_m \end{smallmatrix}\right]\Theta(\zeta) = \left[\begin{smallmatrix} I_p & 0 \\ 0 & I_q \end{smallmatrix}\right]$ and

C′2. the columns of Θ should form a basis for the "wandering subspace" $\mathcal{L}$ associated with the problem

$$\Theta \cdot \mathbb{C}^{p+m} = \mathcal{L} := \mathcal{M} \ominus_{J_\sigma} \zeta \cdot \mathcal{M},$$

where we have set

$$\mathcal{M} := \begin{bmatrix} I_p & G \\ 0 & I_m \end{bmatrix} \cdot H_2(\mathbb{D}, \mathbb{C}^{p+m})$$

and where the notation $\ominus_{J_\sigma}$ refers to the orthogonal difference in the indefinite inner product $\langle \cdot, \cdot \rangle_{J_\sigma}$ on $L_2(\mathbb{T}, \mathbb{C}^{p+m})$ induced by J_σ:

$$\langle f, g \rangle_{J_\sigma} = \frac{1}{2\pi} \int_{\mathbb{T}} \langle J_\sigma f(\zeta), g(\zeta) \rangle_{\mathbb{C}^{p+m}} \, |d\zeta|.$$

In fact, this construction is very close to that in [10] for the construction of Wiener–Hopf factors under the assumption that the associated singular integral operator is invertible as discussed in section 2.2. Furthermore, in [2] it is verified that Θ as defined in (2.11) meets the criteria C′1 and C′2. In this way we have a proof of Theorem 2.2 which ultimately rests on the main result from [1].

2.4. Proof of Theorem 2.2 via Kreĭn-space projective geometry: The Ball–Helton approach. This approach, originating in [4] (see also [39] and [2]), relies on a projective geometry of Kreĭn spaces. The method is reasonably straightforward in case the spectral factor Λ and its inverse V are bounded (and hence also Θ is bounded), but there are some extra complications for the general case. Since these extra complications remained a little obscure in the original exposition [4], we now revisit the ideas there in an attempt to make them more accessible for the system-theory community. For basic background concerning Kreĭn spaces, we refer to [9].

We first observe that the space $L_2(\mathbb{T}, \mathbb{C}^{p+m})$ is a Kreĭn space in the J_σ inner product given by

$$\langle f, g \rangle_{J_\sigma} = \frac{1}{2\pi} \int_{\mathbb{T}} \langle J_\sigma f(\zeta), g(\zeta) \rangle_{\mathbb{C}^{p+m}} \, |d\zeta|.$$

A key role is played by the subspace $\mathcal{M}$ given by

$$\mathcal{M} = \begin{bmatrix} I & G \\ 0 & I \end{bmatrix} \cdot H_2(\mathbb{D}, \mathbb{C}^{p+m}) \subset L_2(\mathbb{T}, \mathbb{C}^{p+m}). \tag{2.12}$$

In general a subspace $\mathcal{M}$ of a Kreĭn space $\mathcal{K}$ is said to be *regular* if it has a good orthogonal complement in the Kreĭn space inner product (i.e., if $\mathcal{K} = \mathcal{M} \dot{+} \mathcal{M}^{[\perp]}$, where $\dot{+}$ indicates direct-sum decomposition), where $\mathcal{M}^{[\perp]}$ indicates the orthogonal complement in the indefinite Kreĭn space inner product. We have the following characterization of when the subspace $\mathcal{M}$ given by (2.12) is a regular subspace of the Kreĭn spaces $(L_2(\mathbb{T}, \mathbb{C}^{p+m}), \langle \cdot, \cdot \rangle_{J_\sigma})$.

PROPOSITION 2.11. *The subspace $\mathcal{M}$ given by* (2.12) *is a regular subspace of the Kreĭn space $(L_2(\mathbb{T}, \mathbb{C}^{p+m}), \langle \cdot, \cdot \rangle_{J_\sigma})$ if and only if any one of the equivalent conditions in Proposition* 2.8 *holds.*

Proof. Note that, for $f, g \in H_2(\mathbb{D}, \mathbb{C}^{p+m})$, we have

$$\left\langle \begin{bmatrix} I_p & G \\ 0 & I_m \end{bmatrix} f, g \right\rangle_{J_\sigma} = \left\langle J_\sigma \begin{bmatrix} I_p & G \\ 0 & I_m \end{bmatrix} f, g \right\rangle_{L_2(\mathbb{T}, \mathbb{C}^{p+m})} = \langle T_{A^* J_\sigma A} f, g \rangle_{L_2(\mathbb{T}, \mathbb{C}^{p+m})},$$

where $A = \left[\begin{smallmatrix} I & G \\ 0 & I \end{smallmatrix}\right]$ is as in Proposition 2.8. Thus the map $U\colon f \mapsto \left[\begin{smallmatrix} I_p & G \\ 0 & I_m \end{smallmatrix}\right] \cdot f$ is unitary from $H_2(\mathbb{D}, \mathbb{C}^{p+m})$ with the inner product induced by the Toeplitz operator $T_{A^* J_\sigma A}$ to $\mathcal{M}$ with the inner product induced by J_σ. A standard fact concerning Kreĭn spaces (see, e.g., [9]) is that a subspace of a Kreĭn space $\mathcal{K}$ is regular if and only if it is itself a Kreĭn space in the inner product inherited from $\mathcal{K}$. In the case at hand, by the indefinite-metric unitary property of U, this happens if and only if $H_2(\mathbb{T}, \mathbb{C}^{p+m})$ is a Kreĭn space in the inner product induced by $T_{A^* J_\sigma A}$; this in turn is equivalent to the invertibility of the Toeplitz operator $T_{A^* J_\sigma A}$, i.e., condition (2) in Proposition 2.8. Proposition 2.11 now follows. ∎

When $\mathcal{M}$ is a regular subspace of $(L_2(\mathbb{T}, \mathbb{C}^{p+m}), \langle \cdot, \cdot \rangle_{J_\sigma})$, we denote by $P_{\mathcal{M}}$ the projection of $L_2(\mathbb{T}, \mathbb{C}^{p+m})$ onto $\mathcal{M}$ along $\mathcal{M}^{[\perp]}$. Then $P_{\mathcal{M}}$ is bounded as an operator on $L_2(\mathbb{T}, \mathbb{C}^{p+m})$ and is self-adjoint in the J_σ-inner product:

$$\langle P_{\mathcal{M}} f, g \rangle_{J_\sigma} = \langle f, P_{\mathcal{M}} g \rangle_{J_\sigma}.$$

A key result from [4] is that when $\mathcal{M}$ is regular, then $\mathcal{M}$ has the following Beurling–Lax-type representation.

THEOREM 2.12 (see [4, 5]). *Assume that the subspace $\mathcal{M}$ as in* (2.12) *is a regular subspace of $(L_2(\mathbb{T}, \mathbb{C}^{p+m}), \langle \cdot, \cdot \rangle_{J_\sigma})$. Then there is a matrix function $\Theta \in L_2(\mathbb{T}, \mathbb{C}^{p+m})$ such that*

(1) *$\mathcal{M} = L_2(\mathbb{T}, \mathbb{C}^{p+m})$-closure of $\Theta \cdot H_\infty(\mathbb{D}, \mathbb{C}^{p+m})$;*
(2) *$\Theta(\zeta)^* J_\sigma \Theta(\zeta) = J_1 := \left[\begin{smallmatrix} I_p & 0 \\ 0 & -I_m \end{smallmatrix}\right]$ for almost all $\zeta \in \mathbb{T}$;*
(3) *the operator $M_\Theta P_{H_2(\mathbb{D}, \mathbb{C}^{p+m})} M_\Theta^{-1}$ defines a bounded operator, namely, the J_σ-orthogonal projection $P_{\mathcal{M}}$ of $L_2(\mathbb{T}, \mathbb{C}^{p+m})$ onto $\mathcal{M}$ along $\mathcal{M}^{[\perp]}$.*

Moreover, Θ is uniquely determined up to a constant J-unitary factor on the right, and in principle can be computed from the (J_1, J_σ)-unitary property (2) *above, along with the condition that*

$$\Theta \cdot \mathbb{C}^{p+m} = \mathcal{M} \ominus_{J_\sigma} \zeta \cdot \mathcal{M}.$$

Alternatively, Θ arises as $\Theta = \left[\begin{smallmatrix} I_p & G \\ 0 & I_m \end{smallmatrix}\right] V$, where $V = \Lambda^{-1}$ and Λ is the spectral factor for $A^* J_\sigma A$ as in Theorem 2.9.

Remark 2.13. Note that if we set $A = \left[\begin{smallmatrix} I_p & G \\ 0 & I_m \end{smallmatrix}\right]$ (with then $A^{-1} = \left[\begin{smallmatrix} I_p & -G \\ 0 & I_m \end{smallmatrix}\right]$), we have

$$M_\Theta P_{H_2(\mathbb{D}, \mathbb{C}^{p+m})} M_{\Theta^{-1}} = M_A M_V P_{H_2(\mathbb{D}, \mathbb{C}^{p+m})} M_\Lambda M_A^{-1},$$

where M_A and its inverse M_A^{-1} are bounded on $L_2(\mathbb{T}, \mathbb{C}^{p+m})$. In this way we see that the last part of condition (2) in Theorem 2.9 fits with condition (3) in Theorem 2.12.

The next step is to reformulate the Nehari–Takagi problem itself in terms of a certain graph subspace of $L_2(\mathbb{T}, \mathbb{C}^{p+m})$ instead of in terms of the matrix function $K \in H_{\infty,k}(\mathbb{T}, \mathbb{C}^{p \times m})$. We shall work with the Kreĭn–Langer representation for an element K of $H_{\infty,k}(\mathbb{D}, \mathbb{C}^{p \times m})$. Specifically, a matrix function $K \in L_\infty(\mathbb{T}, \mathbb{C}^{p \times m})$ is in the class $H_{\infty,k}(\mathbb{D}, \mathbb{C}^{p \times m})$ if and only if K has a representation as $K = F \cdot B^{-1}$, where $F \in H_\infty(\mathbb{D}, \mathbb{C}^{p \times m})$ and $B \in H_\infty(\mathbb{D}, \mathbb{C}^{m \times m})$ is a Blaschke–Potapov product of degree k, and k is the smallest nonnegative integer for which such a representation is possible. Then we have the following reformulation of the Nehari–Takagi problem.

PROPOSITION 2.14 (see [4, 2]). *The angle-operator–graph correspondence induces a one-to-one correspondence between solutions $K \in H_{\infty,k'}(\mathbb{D}, \mathbb{C}^{p \times m})$ of the Nehari–Takagi problem with datum $G \in L_\infty(\mathbb{T}, \mathbb{C}^{p \times m})$ and with index $k' \le k$, on the one*

hand, and subspaces $\mathcal{G}$ of the Kreĭn space

$$\mathcal{K} = \left(\begin{bmatrix} L_2(\mathbb{T}, \mathbb{C}^p) \\ H_2(\mathbb{D}, \mathbb{C}^m) \end{bmatrix}, \langle \cdot, \cdot \rangle_{J_\sigma} \right) \tag{2.13}$$

such that

(1) $\mathcal{G} \subset \mathcal{M} := \left[\begin{smallmatrix} I_p & 0 \\ 0 & I_m \end{smallmatrix} \right] \cdot H_2(\mathbb{D}, \mathbb{C}^{p+m})$,

(2) $\mathcal{G}$ *has codimension* k *in a maximal negative subspace of* $\mathcal{K}$, *and*

(3) $\mathcal{G}$ *is shift invariant, i.e.,* $\zeta \cdot \mathcal{G} \subset \mathcal{G}$,

on the other hand, as follows. If $K \in H_{\infty,k'}(\mathbb{D}, \mathbb{C}^{p\times m})$ *has a representation as* $K = FB^{-1}$, *with* $F \in H_\infty(\mathbb{D}, \mathbb{C}^{p\times m})$ *and with* $B \in H_\infty(\mathbb{D}, \mathbb{C}^{m\times m})$ *a Blaschke–Potapov product of degree* k, *and is such that* $\|G + K\|_\infty \le \sigma$, *and if we set*

$$\mathcal{G}_K = \begin{bmatrix} G + K \\ I \end{bmatrix} B \cdot H_2(\mathbb{T}, \mathbb{C}^{p\times m}) = \begin{bmatrix} GB + F \\ B \end{bmatrix} \cdot H_2(\mathbb{D}, \mathbb{C}^{p\times m}), \tag{2.14}$$

then $\mathcal{G}_K$ *satisfies conditions* (1), (2), *and* (3) *listed above. Conversely, if* $\mathcal{G}$ *satisfies conditions* (1), (2), *and* (3) *listed above, then necessarily there is a* $K \in H_{\infty,k'}(\mathbb{D}, \mathbb{C}^{p\times m})$ *with* $k' \le k$ *and with* $K = FB^{-1}$ *for a Blaschke–Potapov product of degree* k' *such that* $\|G + K\|_\infty \le \sigma$ *and* $\mathcal{G}$ *has the form* $\mathcal{G}_K$ *as in* (2.14).

Proof. We defer the proof to Appendix A (see section 5.1). ∎

The next step is to note the geometric significance of the fact that $\sigma_k > \sigma > \sigma_{k+1}$.

PROPOSITION 2.15. *Assume that* $G \in L_\infty(\mathbb{T}, \mathbb{C}^{p\times m})$ *has Hankel singular values* $\sigma_1 \ge \sigma_2 \ge \cdots$ *with* $\sigma_k > \sigma > \sigma_{k+1}$, *and define subspaces* $\mathcal{M}$ *and* $\mathcal{K}$ *of* $L_2(\mathbb{T}, \mathbb{C}^{p+m})$ *as in* (2.12) *and* (2.13). *Then* $\mathcal{M}$ *is a regular subspace of* $\mathcal{K}$, *and the* J_σ*-orthogonal complement* $\mathcal{K} \ominus_{J_\sigma} \mathcal{M}$ *of* $\mathcal{M}$ *inside* $\mathcal{K}$ *has* k *negative squares.*

Proof. One can compute that the relative J_σ-orthogonal complement $\mathcal{K} \ominus_{J_\sigma} \mathcal{M}$ is given by

$$\mathcal{K} \ominus_{J_\sigma} \mathcal{M} = \begin{bmatrix} I \\ \sigma^{-2} H_G^* \end{bmatrix} H_2(\mathbb{D}, \mathbb{C}^p)^\perp.$$

Hence, the negative signature of $\mathcal{K} \ominus_{J_\sigma} \mathcal{M}$ is equal to the number of negative eigenvalues of the self-adjoint operator

$$\begin{bmatrix} I & \sigma^{-2} H_G \end{bmatrix} \begin{bmatrix} I_p & 0 \\ 0 & -\sigma^2 I_m \end{bmatrix} \begin{bmatrix} I \\ \sigma^{-2} H_G^* \end{bmatrix} = I - \sigma^{-2} H_G H_G^*.$$

From the definition of singular values, we see that this quantity in turn is equal to k if $\sigma_k > \sigma > \sigma_{k+1}$, and the assertion follows. ∎

Proposition 2.15 enables us to adjust Proposition 2.14 to a more useful form as follows.

PROPOSITION 2.16. *Assume that* $G \in L_\infty(\mathbb{T}, \mathbb{C}^{p\times m})$ *has Hankel singular values* $\sigma_1 \ge \sigma_2 \ge \cdots$ *satisfying* $\sigma_k > \sigma > \sigma_{k+1}$ *as in Proposition* 2.15. *Then the angle-operator–graph correspondence as sketched in Proposition* 2.14 *induces a one-to-one correspondence between solutions* $K \in H_{\infty,k}(\mathbb{D}, \mathbb{C}^{p\times m})$ *of the Nehari–Takagi problem with datum* $G \in L_\infty(\mathbb{T}, \mathbb{C}^{p\times m})$ *and with index* k, *on the one hand, and subspaces* $\mathcal{G}$ *of the Kreĭn space* $\mathcal{K}$ *as in* (2.13) *such that*

(1) $\mathcal{G} \subset \mathcal{M} := \left[\begin{smallmatrix} I_p & G \\ 0 & I_m \end{smallmatrix} \right] \cdot H_2(\mathbb{D}, \mathbb{C}^{p+m})$,

(2) $\mathcal{G}$ *is a maximal negative subspace as a subspace of* $\mathcal{M}$*, and*
(3) $\mathcal{G}$ *is shift invariant, i.e.,* $\zeta \cdot \mathcal{G} \subset \mathcal{G}$*,*
on the other hand.

Proof. We defer the proof to Appendix A (see section 5.2). ∎

Proposition 2.16 reduces the description of all solutions K of the Nehari–Takagi problem to a description of all shift-invariant subspaces $\mathcal{G}$ of $\mathcal{M}$ which are maximal negative as a subspace of $\mathcal{M}$. The next proposition gives a characterization of these subspaces; it is at this point that we use the Beurling–Lax representation of $\mathcal{M}$ given in Theorem 2.12.

PROPOSITION 2.17. *Assume that* $K \in L_\infty(\mathbb{T}, \mathbb{C}^{p\times m})$ *with Hankel singular values* $\sigma_1 > \sigma_2 > \cdots$ *satisfying* $\sigma_k > \sigma > \sigma_{k+1}$ *as in Proposition* 2.16*. As in* (2.12)*, let* $\mathcal{M}$ *be considered as a Kreĭn space in the* J_σ*-inner product, and let* $\Theta \in L_2(\mathbb{T}, \mathbb{C}^{p\times m})$ *be the* J_σ*-Beurling–Lax representer for* $\mathcal{M}$ *as in Theorem* 2.12*. Then a subspace* $\mathcal{G}$ *of* $\mathcal{M}$ *satisfies conditions* (1)*,* (2)*, and* (3) *in Proposition* 2.16 *if and only if there is a matrix function* $Q \in H_\infty(\mathbb{T}, \mathbb{C}^{p\times m})$ *such that*

$$\mathcal{G} = L_2(\mathbb{T}, \mathbb{C}^{p+m})\text{-closure of } \Theta \begin{bmatrix} Q \\ I \end{bmatrix} \cdot H_\infty(\mathbb{D}, \mathbb{C}^m) \tag{2.15}$$

for a uniquely determined matrix function $Q \in H_\infty(\mathbb{D}, \mathbb{C}^{p\times m})$ *with* $\|Q\|_\infty \le 1$*.*

Proof. The proof is deferred to Appendix A (see section 5.3). ∎

We are now ready to put all the pieces together to complete the proof of Theorem 2.2

Proof of Theorem 2.2. By combining Propositions 2.14 and 2.16 with Proposition 2.17, we see that K solves the Nehari–Takagi problem if and only if K has a Kreĭn–Langer factorization $Q = FB^{-1}$ (where $F \in H_\infty(\mathbb{D}, \mathbb{C}^{p\times m})$ and $B \in H_\infty(\mathbb{D}, \mathbb{C}^{m\times m})$ is a Blaschke–Potapov product of degree k) such that

$$\begin{bmatrix} G+K \\ I_m \end{bmatrix} B \cdot H_2(\mathbb{D}, \mathbb{C}^m) \cap \Theta \cdot H_\infty(\mathbb{D}, \mathbb{C}^{p+m}) = \Theta \begin{bmatrix} Q \\ I_m \end{bmatrix} \cdot H_\infty(\mathbb{D}, \mathbb{C}^m)$$

for a uniquely determined $Q \in H_\infty(\mathbb{D}, \mathbb{C}^{p\times m})$ with $\|Q\|_\infty \le 1$. In particular, we see that for each of the standard basis vectors $e_1, \ldots, e_m$ in $\mathbb{C}^m$ there must be corresponding vector functions $f_1, \ldots, f_m \in BH_2(\mathbb{D}, \mathbb{C}^m)$ so that

$$\begin{bmatrix} G+K \\ I_m \end{bmatrix} f_j = \Theta \begin{bmatrix} Q \\ I_m \end{bmatrix} e_j,$$

or, in operator form,

$$\begin{bmatrix} G+K \\ I_m \end{bmatrix} F = \begin{bmatrix} \Theta_{11}Q + \Theta_{12} \\ \Theta_{21}Q + \Theta_{22} \end{bmatrix}.$$

From the bottom component we read off that $F = \Theta_{21}Q + \Theta_{22}$; then the top component gives

$$(G+K)(\Theta_{21}Q + \Theta_{22}) = \Theta_{11}Q + \Theta_{12}.$$

Once we confirm that $F(\zeta)^{-1} = (\Theta_{21}(\zeta)Q(\zeta) + \Theta_{22}(\zeta))^{-1}$ makes sense for almost all $\zeta \in \mathbb{T}$, we can solve for $G+K$ and arrive at the formula (2.3) for $G+K$. As all the analysis is necessary and sufficient, this will then complete the proof of Theorem 2.2.

We can see that $\Theta_{21}(\zeta)Q(\zeta) + \Theta_{22}(\zeta)$ is invertible for almost all $\zeta \in \mathbb{T}$ by the following geometric argument; for those readers who would prefer an analytic argument, we also give an analytic proof of the same point in the next section. By our construction we have that the linear manifold

$$\begin{bmatrix} G+K \\ I_m \end{bmatrix} F \cdot H_\infty(\mathbb{D}, \mathbb{C}^m)$$

is dense in a shift-invariant subspace $\mathcal{G}_{G+K}$ of $\mathcal{K} = \left[\begin{smallmatrix} L_2(\mathbb{T},\mathbb{C}^p) \\ H_2(\mathbb{D},\mathbb{C}^m) \end{smallmatrix}\right]$ which has codimension k in a maximal J_σ-negative subspace of $\mathcal{K}$. By the angle-operator–graph correspondence for J_σ-negative subspaces, equivalently $\|G+K\|_\infty \le \sigma$ and the L_2-closure of $FH_\infty(\mathbb{D}, \mathbb{C}^m)$ has the form $BH_2(\mathbb{D}, \mathbb{C}^m)$ for a Blaschke–Potapov product in $H_\infty(\mathbb{D}, \mathbb{C}^{m\times m})$ of degree k. For this to occur, it is necessarily the case that $\det F(\zeta) \neq 0$ for almost all $\zeta \in \mathbb{T}$. The proof of Theorem 2.2 (via Kreĭn-space projective geometry) is now complete. □

2.5. Proof of Theorem 2.2 via a winding number argument. It is also possible to bypass the Kreĭn-space geometry ideas and give a more analytic, less geometric proof for most of the content of Theorem 2.2, as we now show. The main idea for this approach comes from [6]; it can also be considered as a purely frequency-domain version of the state-space solution given in [3] for the rational case. One key point of Theorem 2.2 is that every solution of the Nehar–Takagi problem arises from a contractive H_∞-free parameter via the linear-fractional map; for the proof of this part we translate the ideas from the Grassmannian approach to the more analytic setting here.

The starting point for this alternative derivation is still the Beurling–Lax representation for the subspace $\mathcal{M}$ given in Theorem 2.12. Under the assumption that $\mathcal{M}$ is a *regular* subspace of $(L_2(\mathbb{T}, \mathbb{C}^{p+m}), \langle\cdot,\cdot\rangle_{J_\sigma})$, we know that $L_2(\mathbb{T}, \mathbb{C}^{p+m})$ has a direct sum decomposition

$$L_2(\mathbb{T}, \mathbb{C}^{p+m}) = \mathcal{M}^{[\perp]} \dot{+} \mathcal{M},$$

and hence there is a bounded projection operator $P_{\mathcal{M}}$ from $L_2(\mathbb{T}, \mathbb{C}^{p+m})$ onto $\mathcal{M}$ along the J_σ-orthogonal complement $\mathcal{M}^{[\perp]}$ of $\mathcal{M}$. In addition, in this setup the projection operator $P_{\mathcal{M}}$ is J_σ-self-adjoint in the sense that

$$\langle P_{\mathcal{M}} f, g\rangle_{J_\sigma} = \langle f, P_{\mathcal{M}} g\rangle_{J_\sigma} \quad \text{for all } f, g \in L_2(\mathbb{T}, \mathbb{C}^{p+m}).$$

In addition, we shall have use for the operator $P_-^* P_{\mathcal{M}} P_-$ on $H_2(\mathbb{D}, \mathbb{C}^m)$, where we have set

$$P_- = \begin{bmatrix} 0_{p\times m} \\ I_m \end{bmatrix} : H_2(\mathbb{D}, \mathbb{C}^m) \mapsto L_2(\mathbb{T}, \mathbb{C}^{p+m}).$$

Note that then

$$P_-^* = \begin{bmatrix} 0_{m\times p} & P_{H_2(\mathbb{D},\mathbb{C}^m)} \end{bmatrix} : L_2(\mathbb{T}, \mathbb{C}^{p+m}) \mapsto H_2(\mathbb{D}, \mathbb{C}^m).$$

PROPOSITION 2.18. *Assume that $\mathcal{M}$ as in* (2.12) *is regular and that $\mathcal{M}$ has the J_σ-Beurling–Lax representation $\mathcal{M} = L_2$-closure of $\Theta \cdot H_\infty(\mathbb{D}, \mathbb{C}^{p+m})$ as in Theorem* 2.12. *Then the following hold:*

(1) *J_σ-orthogonal projection of $L_2(\mathbb{T},\mathbb{C}^{p+m})$ onto $\mathcal{M}$ can be computed either in terms of G as*

$$P_{\mathcal{M}} = \begin{bmatrix} P_{H_2(\mathbb{D},\mathbb{C}^p)} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} H_G \\ I \end{bmatrix} \left(H_G^* H_G - \sigma^2 I_m\right)^{-1} \begin{bmatrix} H_G^* & -\sigma^2 I_m \end{bmatrix} \tag{2.16}$$

or in terms of Θ as

$$P_{\mathcal{M}} = M_\Theta P_{H_2(\mathbb{D},\mathbb{C}^{p+m})} M_\Theta^{-1}. \tag{2.17}$$

(2) *The operator $P_-^* P_{\mathcal{M}} P_-$ can be expressed in two ways:*

$$P_-^* P_{\mathcal{M}} P_- = -\sigma^2 (H_G^* H_G - \sigma^2 I_m)^{-1} \tag{2.18}$$

$$= -\sigma^2 M_{[\Theta_{21}\ \Theta_{22}]} P_{H_2(\mathbb{D},\mathbb{C}^{p+m})} M_{[\Theta_{21}\ -\Theta_{22}]^*}. \tag{2.19}$$

(3) *The number k of negative eigenvalues of the self-adjoint operator $P_-^* P_{\mathcal{M}} P_-$ on $H_2(\mathbb{D},\mathbb{C}^m)$ can be expressed either as*

$$k = \text{ the number of Hankel singular values } > \sigma \tag{2.20}$$

or as the number of negative squares of the kernel

$$\frac{\Theta_{22}(z)\Theta_{22}(w)^* - \Theta_{21}(z)\Theta_{21}(w)^*}{1 - z\overline{w}}. \tag{2.21}$$

Consequently, the matrix function $\Theta_{22}^{-1}\Theta_{21} \in H_{\infty,k}(\mathbb{D},\mathbb{C}^{m\times p})$ with $\|\Theta_{22}^{-1} \cdot \Theta_{21}\|_\infty \le 1$, and Θ_{22} has outer-inner factorization $\Theta_{22} = F \cdot B$, where $F \in H_2(\mathbb{D},\mathbb{C}^{m\times m})$ is outer and $B \in H_\infty(\mathbb{D},\mathbb{C}^{m\times m})$ is a Blaschke–Potapov product of degree k.

Proof. To prove (2.16) note that $\mathcal{M}$ has J_σ-orthogonal decomposition

$$\mathcal{M} = \begin{bmatrix} H_2(\mathbb{D},\mathbb{C}^p) \\ 0 \end{bmatrix} \oplus_{J_\sigma} \begin{bmatrix} H_G \\ I_m \end{bmatrix} H_2(\mathbb{D},\mathbb{C}^m). \tag{2.22}$$

The J_σ-orthogonal projection onto $\operatorname{im}\left[\begin{smallmatrix} H_G \\ I_m \end{smallmatrix}\right]$ can be computed as

$$P_{\operatorname{im}\left[\begin{smallmatrix} H_G \\ I_m \end{smallmatrix}\right]} = \begin{bmatrix} H_G \\ I_m \end{bmatrix} \left(\begin{bmatrix} H_G \\ I \end{bmatrix}^{[*]} \begin{bmatrix} H_G \\ I_m \end{bmatrix} \right)^{-1} \begin{bmatrix} H_G \\ I \end{bmatrix}^{[*]}. \tag{2.23}$$

Here we view $\left[\begin{smallmatrix} H_G \\ I_m \end{smallmatrix}\right]$ as an operator acting from $H_2(\mathbb{D},\mathbb{C}^m)$ with the standard Hilbert space inner product into $L_2(\mathbb{T},\mathbb{C}^{p+m})$ with the J_σ-inner product. Hence

$$\begin{bmatrix} H_G \\ I_m \end{bmatrix}^{[*]} = \begin{bmatrix} H_G \\ I_m \end{bmatrix}^* J_\sigma = \begin{bmatrix} H_G^* & -\sigma^2 I \end{bmatrix}. \tag{2.24}$$

Substituting (2.24) into (2.23) and using (2.22) then gives the formula (2.16) for $P_{\mathcal{M}}$.

Formula (2.17) for $P_{\mathcal{M}}$ was already noted as condition (3) of Theorem 2.12.

Formula (2.18) now follows upon multiplying (2.16) on the left by $[0\ \ P_{H_2(\mathbb{D},\mathbb{C}^m)}]$ and on the right by $\left[\begin{smallmatrix} 0 \\ I_m \end{smallmatrix}\right]$ (considered as acting from $H_2(\mathbb{D},\mathbb{C}^m)$ into $L_2(\mathbb{T},\mathbb{C}^{p+m})$).

Formula (2.20) for the number of negative eigenvalues of $P_-^* P_{\mathcal{M}} P_-$ can now be read off immediately from formula (2.18) for $P_- P_{\mathcal{M}} P_-$. To get formula (2.21) for the number of negative eigenvalues of $P_-^* P_{\mathcal{M}} P_-$, we use (2.19) to compute, where we set

$k_w(\zeta) = \frac{1}{1-\zeta\overline{w}}$ equal to the kernel function for $H_2(\mathbb{D},\mathbb{C})$, for any $w_1,\dots,w_N \in \mathbb{D}$ and $x_1,\dots,x_N \in \mathbb{C}^m$,

$$\begin{aligned}
&\left\langle P_- P_{\mathcal{M}} P_- \left(\sum_{j=1}^N k_{w_j} x_j \right), \sum_{i=1}^N k_{w_i} x_i \right\rangle_{H_2(\mathbb{D},\mathbb{C}^m)} \\
&\quad = -\sigma^2 \sum_{i,j=1}^N \left\langle \left(M_{\Theta_{21}} P_{H_2} M_{\Theta_{21}^*} - M_{\Theta_{22}} P_{H_2} M_{\Theta_{22}^*} \right) k_{w_j} x_j, k_{w_i} x_i \right\rangle_{H_2(\mathbb{D},\mathbb{C}^m)} \\
&\quad = -\sigma^2 \sum_{i,j=1}^N x_i^* \frac{\Theta_{21}(w_i)\Theta_{21}(w_j)^* - \Theta_{22}(w_i)\Theta_{22}(w_j)^*}{1 - w_i\overline{w_j}} x_j.
\end{aligned}$$

By the density of the span of the kernel functions $\{k_w x \colon w \in \mathbb{D},\ x \in \mathbb{C}^m\}$, the formula (2.21) for the number of negative eigenvalues for $P_-^* P_{\mathcal{M}} P_-$ now follows.

Finally, from the (J, J_σ)-unitary property of Θ we know that $\Theta(\zeta)^{-1} = J_1 \Theta(\zeta)^* J_\sigma$ for almost all $\zeta \in \mathbb{T}$, and hence

$$\Theta(\zeta) J_1 \Theta(\zeta)^* = J_{\sigma^{-1}}.$$

In particular,

$$\Theta_{21}(\zeta)\Theta_{21}(\zeta)^* - \Theta_{22}(\zeta)\Theta_{22}(\zeta)^* = -\sigma^{-2} I_m$$

or

$$\Theta_{22}(\zeta)\Theta_{22}(\zeta)^* = \Theta_{21}(\zeta)\Theta_{21}(\zeta)^* + \sigma^{-2} I_m \geq \sigma^{-2} I_m \tag{2.25}$$

for almost all $\zeta \in \mathbb{T}$. Hence, for all such ζ, $\Theta_{22}(\zeta)$ is invertible and

$$0 \leq \Theta_{22}(\zeta)^{-1}\Theta_{21}(\zeta)\Theta_{21}(\zeta)^*\Theta_{22}(\zeta)^{*-1} = I_m - \sigma^{-2}\Theta_{22}(\zeta)^{-1}\Theta_{22}(\zeta)^{*-1} \leq I_m. \tag{2.26}$$

We conclude that $\Theta_{22}^{-1}\Theta_{21} \in L_\infty(\mathbb{T}, \mathbb{C}^{m\times p})$ with $\|\Theta_{22}^{-1}\Theta_{21}\| \leq 1$. Moreover, by conjugating the kernel in (2.21) by Θ_{22}^{-1} (multiplying by $\Theta_{22}(z)^{-1}$ on the left and by $\Theta_{22}(w)^{*-1}$ on the right for the generic sets of z and w for which these are defined), we see that the kernel

$$\frac{I_m - (\Theta_{22}^{-1}\Theta_{21})(z)(\Theta_{22}^{-1}\Theta_{21})(w)^*}{1 - z\overline{w}}$$

also has k negative squares on $\mathbb{D}\times\mathbb{D}$, i.e., $\Theta_{22}^{-1}\Theta_{21} \in H_{\infty,k}(\mathbb{D}, \mathbb{C}^{m\times p})$ with $\|\Theta_{22}^{-1}\Theta_{21}\|_\infty \leq 1$. Thus $\Theta_{22}^{-1}\Theta_{21}$ has a left Kreĭn–Langer factorization $\Theta_{22}^{-1}\Theta_{21} = B^{-1}F$ with B an $m \times m$ Blaschke–Potapov product of degree k and $F \in H_\infty(\mathbb{D}, \mathbb{C}^{m\times p})$. From the fact that $\mathcal{M} = L_2$-closure of $\Theta \cdot H_\infty(\mathbb{D}, \mathbb{C}^{m+p})$ and the fact that $\begin{bmatrix} 0 & I_m \end{bmatrix} \mathcal{M} \subset H_2(\mathbb{D}, \mathbb{C}^m)$, we see that the matrix entries of both Θ_{21} and Θ_{22} are in H_2. From $\Theta_{22}^{-1}\Theta_{21} = B^{-1}F$ we conclude that Θ_{22} must have outer-inner factorization of the form $\Theta_{22} = \Theta_{22,o} \cdot B$ with $\Theta_{22,o} \in H_2(\mathbb{D}, \mathbb{C}^{m\times m})$ outer and $B \in H_\infty(\mathbb{D}, \mathbb{C}^{m\times m})$ a Blaschke–Potapov product of degree k. This completes the proof of Proposition 2.18. □

Winding number proof of Theorem 2.2. Suppose that $Q \in H_\infty(\mathbb{D}, \mathbb{C}^{p\times m})$ with $\|Q\|_\infty \leq 1$. From (2.25) and (2.26) we see that $\Theta_{22}(\zeta)$ is invertible and that $\|(\Theta_{22}^{-1}\Theta_{21})(\zeta)\| < 1$ for almost all $\zeta \in \mathbb{T}$. Hence the quantity

$$\Theta_{21}(\zeta)Q(\zeta) + \Theta_{22}(\zeta) = \Theta_{22}(\zeta)(I_m + \Theta_{22}(\zeta)^{-1}\Theta_{21}(\zeta)Q(\zeta))$$

is invertible for almost all $\zeta \in \mathbb{T}$. We may then define a $p \times m$ matrix-valued function K on $\mathbb{T}$ by

$$K = (V_{11}Q + V_{12})(\Theta_{21}Q + \Theta_{22})^{-1}. \tag{2.27}$$

We verify that $K \in L_\infty(\mathbb{T}, \mathbb{C}^{p\times m})$ and in fact that $\|G + K\|_\infty \le \sigma$ as follows. Note that

$$\begin{aligned}\begin{bmatrix} G+K \\ I_m \end{bmatrix}(\Theta_{21}Q + \Theta_{22}) &= \begin{bmatrix} I_p & G \\ 0 & I_m \end{bmatrix}\begin{bmatrix} K \\ I_m \end{bmatrix}(\Theta_{21}Q + \Theta_{22}) = \begin{bmatrix} I_p & G \\ 0 & I_m \end{bmatrix}\begin{bmatrix} V_{11}Q + V_{12} \\ \Theta_{21}Q + \Theta_{22} \end{bmatrix} \\ &= \begin{bmatrix} I_p & G \\ 0 & I_m \end{bmatrix} V \begin{bmatrix} Q \\ I_m \end{bmatrix} = \Theta \begin{bmatrix} Q \\ I_m \end{bmatrix}. \end{aligned} \tag{2.28}$$

(Here we use that $\left[\begin{smallmatrix} I_p & G \\ 0 & I_m \end{smallmatrix}\right] \cdot V = \Theta$ and thus also $\begin{bmatrix} V_{21} & V_{22} \end{bmatrix} = \begin{bmatrix} \Theta_{21} & \Theta_{22} \end{bmatrix}$.) Consequently, considering the various expressions below as functions on $\mathbb{T}$, we have

$$\begin{aligned}(G+K)^*(G+K) - \sigma^2 I_m &= \begin{bmatrix} (G+K)^* & I_m \end{bmatrix} J_\sigma \begin{bmatrix} G+K \\ I_m \end{bmatrix} \\ &= (\Theta_{21}Q + \Theta_{22})^{*-1} \begin{bmatrix} Q^* & I_m \end{bmatrix} \Theta^* J_\sigma \Theta \begin{bmatrix} Q \\ I_M \end{bmatrix} (\Theta_{21}Q + \Theta_{22})^{-1} \\ &= (\Theta_{21}Q + \Theta_{22})^{*-1}(Q^*Q - I_m)(\Theta_{21}Q + \Theta_{22})^{-1} \le 0, \end{aligned}$$

where the last step follows from the assumption that $\|Q\|_\infty \le 1$. In particular, it follows that

$$\|K\|_\infty \le \sigma + \|G\|_\infty \text{ whenever } K = (V_{11}Q + V_{12})(\Theta_{21}Q + \Theta_{22})^{-1} \text{ with } \|Q\|_\infty \le 1. \tag{2.29}$$

Moreover, from (2.28) we see that $G + K$ is given in terms of Q, as in (2.3).

For the discussion in this paragraph we consider the special case $\|Q\|_\infty < 1$; in the end we shall use this special case to arrive at the general case by an approximation argument. We observed at the end of the proof of Proposition 2.18 that the matrix entries of Θ_{21} and Θ_{22} are all in H_2, and by Proposition 2.18 we know that Θ_{22} has outer-inner factorization $\Theta_{22} = F \cdot B$ with F outer and the inner factor B equal to a Blaschke–Potapov product of degree k. For any function f analytic on the disk $\mathbb{D}$ (with possibly finitely many exceptional points), set $f_r(z) = f(rz)$ for each $r < 1$. Then $\Theta_{22,r}$ still has the form $F_r' \cdot B_r'$ with F_r' outer and B_r' a Blaschke–Potapov product of degree k, as long as we take $r < 1$ sufficiently close to 1. Moreover, by Proposition 2.6, we know that there is an $r_0 < 1$ such that, for all r subject to $r_0 \le r < 1$, we have $\|\Theta_{22,r}^{-1}\Theta_{21,r}\|_\infty \le \frac{1}{2}(1 + \|Q\|_\infty^{-1})$, with the consequence that

$$\begin{aligned}\|\Theta_{22,r}^{-1}\Theta_{21,r}Q_r\|_\infty &\le \frac{1}{2}\|Q_r\|_\infty(1 + \|Q\|_\infty^{-1}) \\ &\le \frac{1}{2}(\|Q\|_\infty + 1) < 1 \quad \text{for all } r_0 \le r < 1. \end{aligned} \tag{2.30}$$

By the Neumann series estimate, it follows that $(I + \Theta_{22,r}^{-1}\Theta_{21,r}Q_r)$ is invertible in $L_\infty(\mathbb{T}, \mathbb{C}^{m\times m})$ with

$$\|(I + \Theta_{22,r}^{-1}\Theta_{21,r}Q_r)^{-1}\|_\infty \le \frac{1}{1 - \frac{1}{2}(1 + \|Q\|_\infty)} = \frac{2}{1 - \|Q\|_\infty}. \tag{2.31}$$

Another consequence of the estimate (2.30) is that the determinant of $(I+\Theta_{22,r}^{-1}\Theta_{21,r}Q_r)$ has winding number around the unit circle equal to zero. As $\det\Theta_{22,r}=\det(F_r')\cdot\det B_r'$ has winding number equal to k (since F_r' is outer and B_r' is a matrix Blaschke–Potapov product of degree k), it follows that the determinant of

$$\Theta_{21,r}Q_r+\Theta_{22,r}=\Theta_{22,r}\left(\Theta_{22,r}^{-1}\Theta_{21,r}Q_r+I_m\right)$$

has winding number equal to k around the unit circle. As $\Theta_{21,r}Q_r+\Theta_{22,r}$ is in the disk algebra (analytic on the open disk and continuous on the closed disk), we conclude that $(\Theta_{21,r}Q_r+\Theta_{22,r})^{-1}$ is in $H_{\infty,k}(\mathbb{D},\mathbb{C}^{m\times m})$.

We now return to the case of a general $Q\in H_\infty(\mathbb{D},\mathbb{C}^{p\times m})$ with $\|Q\|\le 1$. Let s be a number with $0<s<1$. Then the above analysis applies to the situation where we have $s\cdot Q$ in place of Q. Thus

$$\|(\Theta_{22,r}^{-1}\Theta_{21,r}(sQ_r)+I)^{-1}\|_\infty\le\frac{2}{1-s\|Q\|_\infty}\le\frac{2}{1-s} \tag{2.32}$$

for all $r<1$ sufficiently close to 1. Also, from (2.26) we read off that $\|\Theta_{22}^{-1}\|_\infty\le\sigma$; by Proposition 2.6 we then have $\|\Theta_{22,r}^{-1}\|_\infty\le\sigma+\epsilon$ for any given $\epsilon>0$ as long as we take $r<1$ sufficiently close to 1. Hence, for all $r<1$ but sufficiently close to 1, we have

$$\begin{aligned}\|(\Theta_{21,r}(sQ_r)+\Theta_{22,r})^{-1}\|_\infty&=\|(\Theta_{22,r}^{-1}\Theta_{21,r}(sQ_r)+I)^{-1}\Theta_{22,r}^{-1}\|_\infty\\&\le\left(\frac{2}{1-s}\right)\cdot(\sigma+\epsilon)<\infty.\end{aligned}$$

We conclude that $(\Theta_{21,r}(sQ_r)+\Theta_{22,r})^{-1}$ converges pointwise boundedly, and hence, by part (2) of Proposition 2.3, also in the $L_\infty(\mathbb{T},\mathbb{C}^{m\times m})$-weak-$*$ topology, to $(\Theta_{21}(sQ)+\Theta_{22})^{-1}$ as $r\to 1$. As we have seen above, each $(\Theta_{21,r}(sQ_r)+\Theta_{22,r})^{-1}$ is in $H_{\infty,k}(\mathbb{D},\mathbb{C}^{m\times m})$. By Proposition 2.4, we conclude that $(\Theta_{21}(sQ)+\Theta_{22})^{-1}\in H_{\infty,k'}(\mathbb{D},\mathbb{C}^{m\times m})$ with $k'\le k$. But then $K_s:=(V_{11}(sQ)+V_{12})(\Theta_{21}(sQ)+\Theta_{22})^{-1}$ is in $(H_2(\mathbb{D},\mathbb{C}^{p\times m})\cdot H_{\infty,k'})\cap L_\infty$, and hence is in fact in $H_{\infty,k'}(\mathbb{D},\mathbb{C}^{p\times m})$ for some $k'\le k$. By (2.29), we know that $\|K_s\|_\infty\le\sigma+\|G\|_\infty$ for all $s<1$. Hence, by another application of part (2) of Proposition 2.3 combined with Proposition 2.4, we may let $s\to 1$ and conclude that $K=(V_{11}Q+V_{12})(\Theta_{21}Q+\Theta_{22})^{-1}$ is in $H_{\infty,k'}(\mathbb{D},\mathbb{C}^{p\times m})$ for some $k'\le k$. Since we have already verified that $\|G+K\|_\infty\le\sigma$ and we know by Proposition 2.7 that k is the smallest possible index for a solution to the Nehari–Takagi problem to exist for level σ if $\sigma_k>\sigma>\sigma_{k+1}$, we conclude that necessarily $k'=k$. We have now verified that the formula (2.3) provides a solution K to the Nehari–Takagi problem as asserted in Theorem 2.2.

Conversely, suppose that $K\in H_{\infty,k}(\mathbb{D},\mathbb{C}^{p\times m})$ provides a solution of the Nehari–Takagi problem. Then K has a Kreĭn–Langer factorization $K=F'B'^{-1}$, where $F'\in H_\infty(\mathbb{D},\mathbb{C}^{p+m})$ and $B'\in H_\infty(\mathbb{D},\mathbb{C}^{m\times m})$ is a Blaschke–Potapov product of degree k. Then

$$\begin{aligned}\begin{bmatrix}G+K\\I_m\end{bmatrix}B'&=\begin{bmatrix}I_p&G\\0&I_m\end{bmatrix}\begin{bmatrix}K\\I_m\end{bmatrix}B'\\&=\Theta\cdot\Lambda\cdot\begin{bmatrix}F'\\B'\end{bmatrix}\\&=\Theta\cdot\begin{bmatrix}\Lambda_{11}F'+\Lambda_{12}B'\\\Lambda_{21}F'+\Lambda_{22}B'\end{bmatrix}=:\Theta\cdot\begin{bmatrix}Q_1\\Q_2\end{bmatrix},\end{aligned} \tag{2.33}$$

where $Q_1 := \Lambda_{11}F' + \Lambda_{12}B' \in H_2(\mathbb{D}, \mathbb{C}^{p\times m})$ and $Q_2 := \Lambda_{21}F' + \Lambda_{22}B' \in H_2(\mathbb{D}, \mathbb{C}^{m\times m})$. Since $\|G + K\|_\infty \le \sigma$ by assumption,

$$\begin{aligned} 0 &\ge B'^* \left((G+K)^*(G+K) - \sigma^2 I_m\right) B' \\ &= B'^* \begin{bmatrix} (G+K)^* & I_m \end{bmatrix} J_\sigma \begin{bmatrix} G+K \\ I_m \end{bmatrix} B' \\ &= \begin{bmatrix} Q_1^* & Q_2^* \end{bmatrix} \Theta^* J_\sigma \Theta \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} \\ &= Q_1^* Q_1 - Q_2^* Q_2 \end{aligned} \tag{2.34}$$

a.e. on $\mathbb{T}$. We conclude that

$$Q_2(\zeta)x(\zeta) = 0 \Rightarrow Q_1(\zeta)x(\zeta) = 0. \tag{2.35}$$

From the definition of Q_1 and Q_2 in (2.33) we see that

$$B' = \Theta_{21}Q_1 + \Theta_{22}Q_2. \tag{2.36}$$

Hence (2.35) forces $B'(\zeta)x(\zeta) = 0$ as well, and hence $x(\zeta) = 0$ for almost all $\zeta \in \mathbb{T}$. We conclude that $Q_2(\zeta)$ is invertible a.e. on $\mathbb{T}$ and $Q(\zeta) := Q_1(\zeta)Q_2(\zeta)^{-1}$ makes sense. The calculation (2.34) then implies that $\|Q\|_\infty \le 1$, while (2.33) shows that we recover $G + K$ from Q as in the representation (2.3).

It remains to show that $Q \in H_\infty(\mathbb{D}, \mathbb{C}^{p\times m})$. For this piece of the argument we borrow some ideas from the Grassmannian approach. If $Q_2H_\infty(\mathbb{C}, \mathbb{C}^m)$ is not dense in $H_2(\mathbb{D}, \mathbb{C}^m)$, we may choose a nonzero $h_0 \in H_\infty(\mathbb{D}, \mathbb{C}^m)$ lying in $H_2(\mathbb{D}, \mathbb{C}^m) \ominus \overline{Q_2H_\infty(\mathbb{D}, \mathbb{C}^m)}$. Then

$$\begin{bmatrix} 0 \\ h_0 \end{bmatrix} \perp_{J_1} \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} H_\infty(\mathbb{D}, \mathbb{C}^m).$$

Since $\Theta^* J_\sigma \Theta = J_1$ on $\mathbb{T}$, it then follows that

$$\Theta \begin{bmatrix} 0 \\ h_0 \end{bmatrix} = \begin{bmatrix} \Theta_{12}h_0 \\ \Theta_{22}h_0 \end{bmatrix} \perp_{J_\sigma} \Theta \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} H_\infty(\mathbb{D}, \mathbb{C}^m) = \begin{bmatrix} G+K \\ I_m \end{bmatrix} B'H_\infty(\mathbb{D}, \mathbb{C}^m) \text{ (by (2.33))}.$$

Hence

$$\|\Theta_{12}h_0 + (G+K)B'h\|_2^2 \le \|\Theta_{22}h_0 + B'h\|_2^2$$

for all $h \in H_2(\mathbb{D}, \mathbb{C}^m)$. Therefore there is a contraction operator X from $\mathcal{D}_0 := \operatorname{span}\{\Theta_{22}h_0\} + B'H_2(\mathbb{D}, \mathbb{C}^m)$ into $L_2(\mathbb{T}, \mathbb{C}^p)$ such that

$$\begin{bmatrix} X \\ I_m \end{bmatrix} (h_0 + B'h) = \begin{bmatrix} \Theta_{12}h_0 \\ \Theta_{22}h_0 \end{bmatrix} + \begin{bmatrix} G+K \\ I_m \end{bmatrix} B'h \in \begin{bmatrix} I_p & G \\ 0 & I_m \end{bmatrix} H_2(\mathbb{D}, \mathbb{C}^{p+m}) \tag{2.37}$$

for all $h \in H_2(\mathbb{D}, \mathbb{C}^m)$. Note that $\left[\begin{smallmatrix} I_p & G \\ 0 & I_m \end{smallmatrix}\right] H_2(\mathbb{D}, \mathbb{C}^{p+m})$ has a J_σ-orthogonal splitting

$$\begin{bmatrix} I_p & G \\ 0 & I_m \end{bmatrix} H_2(\mathbb{D}, \mathbb{C}^{p+m}) = \begin{bmatrix} H_2(\mathbb{D}, \mathbb{C}^p) \\ \{0\} \end{bmatrix} \oplus_{J_\sigma} \begin{bmatrix} H_G \\ I_m \end{bmatrix} \mathcal{E}_+ \oplus_{J_\sigma} \begin{bmatrix} H_G \\ I_m \end{bmatrix} \mathcal{E}_-, \tag{2.38}$$

where $\mathcal{E}_+ = \operatorname{im} E((\sigma, +\infty))$ and $\mathcal{E}_- = \operatorname{im} E([0, \sigma))$ and where we have set $E(\cdot)$ equal to the spectral projection for the self-adjoint operator $(H_G^* H_G)^{1/2}$. Note that the first

two direct summands in (2.38) are uniformly J_σ-positive, while the last is uniformly J_σ-negative. As $\|X\| \le 1$, the equality (2.37) forces the existence of a subspace $\mathcal{E}_-^0$ of $\mathcal{E}_-$ so that

$$\begin{bmatrix} X \\ I_m \end{bmatrix} \mathcal{D}_0 = \left\{ \begin{bmatrix} Y_2 e \\ 0 \end{bmatrix} + \begin{bmatrix} H_G \\ I_m \end{bmatrix} Y_1 e + \begin{bmatrix} H_G \\ I_m \end{bmatrix} e \colon e \in \mathcal{E}_-^0 \right\} \tag{2.39}$$

for operators $Y_1 \colon \mathcal{E}_-^0 \mapsto \mathcal{E}_+$ and $Y_2 \colon \mathcal{E}_-^0 \mapsto H_2(\mathbb{D}, \mathbb{C}^p)$. In particular,

$$\mathcal{D}_0 = \{Y_1 e + e \colon e \in \mathcal{E}_-^0\}. \tag{2.40}$$

But the subspace $\mathcal{D}_0 = \operatorname{span}\{\Theta_{22} h_0\} + B' H_2(\mathbb{D}, \mathbb{C}^m)$ has codimension $k-1$ in $H_2(\mathbb{D}, \mathbb{C}^m)$, while the subspace on the right in (2.40) has the same codimension in $H_2(\mathbb{D}, \mathbb{C}^m)$ as does $\mathcal{E}_-^0$. As $\mathcal{E}_-^0$ is a subspace of $\mathcal{E}_-$ which has codimension k in $H_2(\mathbb{D}, \mathbb{C}^m)$, we conclude that the right-hand side of (2.40) has codimension at most k in $H_2(\mathbb{D}, \mathbb{C}^m)$. In this way we arrive at a contradiction and conclude that necessarily Q_2 is outer. It now follows that $Q = Q_1 Q_2^{-1}$ is of bounded type with no inner factor in the denominator. This together with $Q \in L_\infty(\mathbb{T}, \mathbb{C}^{p\times m})$ gives us finally that $Q \in H_\infty(\mathbb{D}, \mathbb{C}^{p\times m})$, as wanted. ∎

Remark 2.19. *The band method.* A very flexible method for solving a variety of interpolation and extension problems which has evolved into increasing levels of sophistication over the past two decades is the so-called *band method* (see [22] for an excellent overview and [35] for one of the latest variations). Recent work (see [26]) enhances this abstract scheme to handle the Nehari–Takagi problem ($\sigma_{k+1} < \sigma < \sigma_k$ with $k \ge 1$) rather than merely the suboptimal Nehari problem ($\sigma_1 < \sigma$). However, the core of the method involves solving equations in a Wiener-like algebra; this limitation forces the spectral factor Λ and its inverse $\Lambda^{-1} = V$ (in the discrete-time setting) to be in $H_\infty(\mathbb{D}, \mathbb{C}^{(p+m)\times(p+m)})$ rather than merely in $H_2(\mathbb{D}, \mathbb{C}^{(p+m)\times(p+m)})$. A remaining open issue appears to be the extension of this abstract framework to include the situation studied in this paper.

3. State-space solutions. Let X be an arbitrary Hilbert space, and let A be the infinitesimal generator of a strongly continuous semigroup $\{T(t)\}_{t\ge 0}$. Let $B \in \mathcal{L}(\mathbb{C}^m, X)$, $C \in \mathcal{L}(X, \mathbb{C}^p)$. Assume that the triple (A, B, C) satisfies

A1. $B^*(\cdot I - A^*)^{-1} x \in H_2(\mathbb{C}_+, \mathbb{C}^m)$ (input stable),

A2. $C(\cdot I - A)^{-1} x \in H_2(\mathbb{C}_+, \mathbb{C}^p)$ (output stable),

A3. $C(\cdot I - A)^{-1} B u \in H_\infty(\mathbb{C}_+, \mathbb{C}^p)$ (input-output stable)

for all $x \in X$, $u \in \mathbb{C}^m$. Condition A3 holds if and only if $\mathcal{D} \in \mathcal{L}(L^2(\mathbb{R}_+; \mathbb{C}^m), L^2(\mathbb{R}_+; \mathbb{C}^p))$, where

$$(\mathcal{D}u)(t) = C \int_0^t T(t-s) B u(s)\, ds \quad (u \in L^2(\mathbb{R}_+; \mathbb{C}^m)). \tag{3.1}$$

Equation (3.1) is equivalent to $\widehat{\mathcal{D}u} = G\widehat{u}$, where $G(s) := C(sI - A)^{-1} B$ and $\widehat{u}$ denotes the Laplace transform of u ($\widehat{u}(s) := \int_0^\infty \mathrm{e}^{-st} u(t)\, du$). It is well known that $\|\mathcal{D}\| = \|G\|_{H_\infty}$. By Plancherel's theorem (and the closed graph theorem), A2 means that $\mathcal{C} : X \to L^2(\mathbb{R}_+; \mathbb{C}^p)$ is bounded, where $(\mathcal{C}x)(t) := CT(t)x$, $t \ge 0$. Hence it also follows that $L_C := \mathcal{C}^*\mathcal{C} \in \mathcal{L}(X)$. Similarly, A1 implies that $\mathcal{B}^{\mathrm{d}} : X \to L^2(\mathbb{R}_+; \mathbb{C}^m)$ is bounded, where $(\mathcal{B}^{\mathrm{d}} x)(t) := B^* T(t)^* x$, $t \ge 0$. Thus, $L_B := \mathcal{B}\mathcal{B}^* \in \mathcal{L}(X)$, where $\mathcal{B}^* := \mathcal{R}\mathcal{B}^{\mathrm{d}}$, $(\mathcal{R}f)(t) := f(-t)$ (the reflection). (See, for instance, Curtain and Zwart [16] or Mikkola [30] for details.)

It is easy to see that if the system is *exponentially stable* (that is, there are $\epsilon > 0$, $M < \infty$ such that $\|T(t)\|_{\mathcal{L}(X)} \leq M\mathrm{e}^{-\epsilon t}$ for all $t > 0$), then A1–A3 are satisfied (and $\overline{\mathbb{C}_+} \subset \rho(A)$). However, there are several important systems that are not exponentially stable but for which A1–A3 hold. In this section we shall derive the state-space formulas for the factors Λ and V for such systems; we use the additional assumption that the open right half-plane $\mathbb{C}_+$ is contained in the resolvent set $\rho(A)$, but this assumption can be relaxed (for example, a zero-measurable spectrum on each vertical line on $\mathbb{C}_+$ is not a problem; see Remark 3.4 below).

In Lemma 3.1 below we show that if A1–A3 hold, then the systems $(A, -, B^*L_C)$ and $(A^*, -, CL_B)$ are output stable, that is, $B^*L_CT \in \mathcal{L}(X, L^2(\mathbb{R}_+; \mathbb{C}^m))$, $CL_BT^* \in \mathcal{L}(X, L^2(\mathbb{R}_+; \mathbb{C}^p))$. We use the following notation:

$$(\pi_+ f)(t) := \begin{cases} f(t) & \text{if } t \geq 0, \\ 0 & \text{if } t < 0, \end{cases} \quad \text{and} \quad \mathcal{D}^{\mathrm{d}} := \mathcal{R}\mathcal{D}^*\mathcal{R}$$

is the input-output map of (A^*, C^*, B^*) (see [30, Lemma 6.2.9(b)]).

LEMMA 3.1. *If* A1*,* A2*, and* A3 *hold, then* $\pi_+\mathcal{D}^*\mathcal{C}x = B^*L_CT(\cdot)x$ *and* $\mathcal{R}\pi_-\mathcal{D}\mathcal{B}^*x = \pi_+(\mathcal{D}^{\mathrm{d}})^*\mathcal{B}^{\mathrm{d}}x = CL_BT(\cdot)^*x$ *for each* $x \in X$*. In particular, there is* $M < \infty$ *such that* $\|B^*L_C(\cdot I - A)^{-1}x\|_{H_2(\mathbb{C}_+,\mathbb{C}^m)} \leq M\|x\|_X$ *and* $\|CL_B(\cdot I - A^*)^{-1}x\|_{H_2(\mathbb{C}_+,\mathbb{C}^p)} \leq M\|x\|_X$ *for all* $x \in X$*.*

Proof. By Lemma 4.2.6 of [33], we have $\pi_+\mathcal{D}^*\mathcal{C}x = B^*L_CTx$ (everywhere, by continuity). The first inequality is obtained from Plancherel's theorem with $M := \|\mathcal{D}\| \max\{\|\mathcal{C}\|, \|\mathcal{B}\|\}$. Applying the above to (A^*, C^*, B^*), we obtain the second equality and inequality (because $\|(\mathcal{D}^{\mathrm{d}})^*\| = \|\mathcal{D}^{\mathrm{d}}\| = \|B^*(\cdot I - A^*)^{-1}C^*\|_\infty = \|G^*\|_\infty = \|G\|_\infty = \|\mathcal{D}\|$). ∎

Now we are ready to give the state-space formulas for the factors Λ and V. The case where $\overline{\mathbb{C}_+} \subset \rho(A)$ is simple, and the general case will be reduced to that by using the results given in section 6.

LEMMA 3.2. *Assume that the triple* (A, B, C) *satisfies* A1*,* A2*,* A3 *and that* $\mathbb{C}_+ \subset \rho(A)$*. Let* $G(s) = C(sI - A)^{-1}B$ *be the associated transfer function, with associated Hankel singular values* $\sigma_1 \geq \sigma_2 \geq \cdots$*, and let* σ *be such that* $\sigma_{k+1} < \sigma < \sigma_k$*. Let* Λ *be defined as follows:*

(3.2)
$$\Lambda(s) = \begin{bmatrix} I_p & 0 \\ 0 & \sigma I_m \end{bmatrix} + \frac{1}{\sigma^2}\begin{bmatrix} -CL_B \\ \sigma B^* \end{bmatrix}\left(I - \frac{1}{\sigma^2}L_CL_B\right)^{-1}(sI + A^*)^{-1}\begin{bmatrix} C^* & L_CB \end{bmatrix},$$

$s \in \mathbb{C}_-$*. Then* Λ *has the following properties:*

(1) $\Lambda(i\omega)^*\left[\begin{smallmatrix} I_p & 0 \\ 0 & -I_m \end{smallmatrix}\right]\Lambda(i\omega) = \left[\begin{smallmatrix} I_p & G(i\omega) \\ 0 & I_m \end{smallmatrix}\right]^*\left[\begin{smallmatrix} I_p & 0 \\ 0 & -\sigma^2 I_m \end{smallmatrix}\right]\left[\begin{smallmatrix} I_p & G(i\omega) \\ 0 & I_m \end{smallmatrix}\right]$ *for almost all* $\omega \in \mathbb{R}$*.*

(2) $\Lambda(s)$ *is invertible for each* $s \in \mathbb{C}_-$*, and its inverse is given by*

(3.3)
$$V(s) = \begin{bmatrix} I_p & 0 \\ 0 & \frac{1}{\sigma}I_m \end{bmatrix} - \frac{1}{\sigma^2}\begin{bmatrix} -CL_B \\ \sigma B^* \end{bmatrix}(sI + A^*)^{-1}\left(I - \frac{1}{\sigma^2}L_CL_B\right)^{-1}\begin{bmatrix} C^* & \frac{1}{\sigma}L_CB \end{bmatrix},$$

$s \in \mathbb{C}_-$*.*

(3) $\Lambda(\cdot) - \left[\begin{smallmatrix} I_p & 0 \\ 0 & \sigma I_m \end{smallmatrix}\right] \in H_2(\mathbb{C}_-, \mathbb{C}^{(p+m)\times(p+m)})$*.*

(4) $V(\cdot) - \left[\begin{smallmatrix} I_p & 0 \\ 0 & \frac{1}{\sigma}I_m \end{smallmatrix}\right] \in H_2(\mathbb{C}_-, \mathbb{C}^{(p+m)\times(p+m)})$*.*

Proof. 1° *Case* $\overline{\mathbb{C}_+} \subset \rho(A)$*:* The proofs of parts (1) and (2) go in a way similar to the suboptimal Nehari problem addressed in Curtain and Zwart [16, section 8.3].

The new part is to show that parts (3) and (4) hold. Set

$$\widehat{g}(s) := \begin{bmatrix} C \\ B^*L_C \end{bmatrix} (sI - A)^{-1} \quad (s \in \rho(A)). \tag{3.4}$$

By Lemma 3.1, we have $\widehat{g}x \in H_2(\mathbb{C}_+, \mathbb{C}^{p+m})$ for all $x \in X$, and so

$$\widehat{f} := \widehat{g}\left(I - \frac{1}{\sigma^2} L_B L_C\right)^{-1} \begin{bmatrix} -L_B C^* \\ \sigma B \end{bmatrix} \tag{3.5}$$

satisfies $\widehat{f}z \in H_2(\mathbb{C}_+, \mathbb{C}^{p+m})$ for all $z \in \mathbb{C}^{p+m}$. Thus $\widehat{f} \in H_2(\mathbb{C}_+, \mathbb{C}^{(p+m)\times(p+m)})$. Since $-\widehat{f}(-\bar{s})^* = \Lambda(s) - \left[\begin{smallmatrix} I_p & 0 \\ 0 & \sigma I_m \end{smallmatrix}\right]$, we obtain that (3) holds.

Part (4) can be proved in an analogous way.

2° *The general case* $\mathbb{C}_+ \subset \rho(A)$*:* The proof in 1° establishes (2)–(4). However, (1) is more complicated: Now (3.2) defines Λ on $\mathbb{C}_-$ only, and on $i\mathbb{R}$ it is defined a.e. as the radial (or nontangential) limit or, equivalently, as the Fourier (Laplace) transform of the inverse Laplace transform of Λ. This follows from (3) (see below).

Nevertheless, the triple $(A-\epsilon, B, C)$ satisfies the assumptions of 1° (cf. Lemma 6.1). Therefore, the corresponding functions Λ_ϵ and G_ϵ satisfy (1) in place of Λ and G. (Note that $G_\epsilon(i\omega) := C(i\omega I - (A-\epsilon))^{-1}B = G(i\omega + \epsilon)$.) Repeat (3.4) and (3.5) with $\widehat{g}_\epsilon, \widehat{f}_\epsilon, A-\epsilon, L_{C,\epsilon}, L_{B,\epsilon}$ in place of $\widehat{g}, \widehat{f}, A, L_C, L_B$, respectively.

By (2) and (3) of Lemma 6.1 (and Lemma A.3.1(j3) of Mikkola [30]), we have $g_\epsilon x := \pi_+ \left[\begin{smallmatrix} I \\ \mathcal{D}_\epsilon^* \end{smallmatrix}\right] \mathcal{C}_\epsilon x \to \pi_+ \left[\begin{smallmatrix} I \\ \mathcal{D}^* \end{smallmatrix}\right] \mathcal{C}x$ in $L^2(\mathbb{R}_+; \mathbb{C}^{p+m})$, and so $\widehat{g}_\epsilon x \to \widehat{g}x$ in $L^2(i\mathbb{R}; \mathbb{C}^{p+m})$, as $\epsilon \to 0+$, for all $x \in X$. Therefore, $\widehat{f}_\epsilon z \to \widehat{f}z$ in $L^2(i\mathbb{R}; \mathbb{C}^{p+m})$ for all $z \in \mathbb{C}^{p+m}$ (here we also need (7) and (4) of Lemma 6.1); hence a subsequence converges a.e. on $i\mathbb{R}$. But, similarly, $G_\epsilon(i\omega)z = G(i\omega + \epsilon)z \to G(i\omega)z$, as $\epsilon \to 0+$, for almost every $\omega \in \mathbb{R}$, for each $z \in \mathbb{C}^{p+m}$ (use the standard H^∞ boundary function result, such as Theorem 3.3.1(c1) of Mikkola [30]).

We already noted above that $\langle \Lambda_\epsilon(i\omega)\tilde{z}, \left[\begin{smallmatrix} I_p & 0 \\ 0 & -I_m \end{smallmatrix}\right] \Lambda_\epsilon(i\omega)z \rangle = \left\langle \left[\begin{smallmatrix} I_p & G_\epsilon(i\omega) \\ 0 & I_m \end{smallmatrix}\right] \tilde{z}, \left[\begin{smallmatrix} I_p & 0 \\ 0 & -\sigma^2 I_m \end{smallmatrix}\right] \left[\begin{smallmatrix} I_p & G_\epsilon(i\omega) \\ 0 & I_m \end{smallmatrix}\right] z \right\rangle$ for almost every $\omega \in \mathbb{R}$, for any $z, \tilde{z} \in \mathbb{C}^{p+m}$. By the above, we can remove the ϵ's (just let $\epsilon \to 0$). Since $\mathbb{C}^{p+m}$ has a finite basis, (1) holds. ∎

In light of Lemma 3.2, we now obtain our main theorem by invoking the key frequency-domain result, namely, Theorem 2.1. The following theorem gives explicit formulas (in terms of the state-space parameters) for all solutions to the suboptimal Hankel norm approximation problem in the case of infinite-dimensional systems which do not necessarily have an exponentially stable semigroup.

THEOREM 3.3. *Assume that the triple* (A, B, C) *satisfies* A1, A2, A3 *and that* $\mathbb{C}_+ \subset \rho(A)$*. Let* $G(s) = C(sI - A)^{-1}B$ *be the associated transfer function, and let the Hankel singular values be denoted by* $\sigma_1 \geq \sigma_2 \geq \cdots$*. Suppose that* σ *is such that* $\sigma_{k+1} < \sigma < \sigma_k$*, and let* V *be given by* (3.3)*.*

Then K *is such that* $K(\cdot) \in H_{\infty,k}(\mathbb{C}_-, \mathbb{C}^{p\times m})$ *and* $\|G(i\cdot) + K(i\cdot)\|_\infty \leq \sigma$ *if and only if* K *is given by* $K(i\omega) = (V_{11}(i\omega)Q(i\omega)+V_{12}(i\omega))(V_{21}(i\omega)Q(i\omega)+V_{22}(i\omega))^{-1}, \omega \in \mathbb{R}$*, for some* $Q \in H_\infty(\mathbb{C}_-, \mathbb{C}^{p\times m})$ *such that* $\|Q\|_\infty \leq 1$*.*

This follows from Theorem 2.1 and Lemma 3.2.

Remark 3.4.

(a) The assumption $\mathbb{C}_+ \subset \rho(A)$ can be weakened in all our results, including the above. Indeed, it suffices that, for instance, the Lebesgue measure of $\{r+\omega i \in \sigma(A) : \omega \in \mathbb{R}\}$ is zero for all small $r > 0$, as one can verify from the proofs.

(b) Finally, we remark that in Chapter 6 of Sasane [40], using another approach, state-space formulas were given in the nonexponentially stable case. However, these were in terms of the parameters of the shifted system Σ_ϵ and only guaranteed that, for a small enough shift, they are also solutions to the original system. Also, while only the existence of some solutions was demonstrated in [40], here we give a complete parameterization of *all* solutions.

4. An application to the case of well-posed linear systems. Finally, in this last section we give an application of Theorem 3.3 to obtain state-space formulas for the suboptimal Hankel norm approximation problem for *well-posed* linear systems. This was done using the idea of reciprocal systems in Curtain and Sasane [15], but there, instead of Theorem 3.3, a weaker result from Chapter 6 of Sasane [40] (which was mentioned in Remark 3.4) was used. Here, using Theorem 3.3, we obtain a different solution to the problem, where, as opposed to Curtain and Sasane [15], we now obtain a parameterization of the set of *all* solutions to the suboptimal Hankel norm approximation problem for well-posed linear systems.

We consider the suboptimal Hankel norm approximation problem for a well-posed linear system Σ on a Hilbert space X, with input space $\mathbb{C}^m$, output space $\mathbb{C}^p$, generating operators A, B, C, semigroup $\{T(t)\}_{t\geq 0}$, and transfer function G, under the following assumptions:

B1. $0 \in \rho(A)$ and $\mathbb{C}_+ \subset \rho(A)$.

B2. Σ is input-stable.

B3. Σ is output-stable.

B4. $G \in H_\infty(\mathbb{C}_+, \mathbb{C}^{p\times m})$.

(Condition B1 can be relaxed; for example, it suffices to assume that $0 \in \rho(A)$ and $\sigma(A) \cap \mathbb{C}_+$ is at most countable (see Remark 3.4(a)). Moreover, instead of $0 \in \rho(A)$ it suffices to assume that $ir \in \rho(A)$ for some $r \in \mathbb{R}$, but then one must replace A by $A - ir$ in the formulas, so that the new G equals the old $G(ir + \cdot)$.)

The *reciprocal system* of such a well-posed linear system is defined as the well-posed linear system Σ_r with the bounded generating operators $A^{-1}, A^{-1}B, -CA^{-1}$. In Curtain and Sasane [15], it was established that if Σ satisfies B1–B4 above, then its reciprocal system is such that

1. A1, A2, A3 from the previous section are satisfied;
2. $\mathbb{C}_+ \subset \rho(A^{-1})$;
3. the controllability and observability Gramians of Σ_r are the same as the controllability and observability Gramians of Σ;
4. $K_r \in H_{\infty,k}(\mathbb{C}_-; \mathbb{C}^{p\times m})$ is a solution to the suboptimal Hankel norm approximation problem of the reciprocal system Σ_r if and only if

$$K(s) := K_r\left(\frac{1}{s}\right) - G(0) \quad \text{for } s \in \mathbb{C}_- \tag{4.1}$$

is a solution[1] to the suboptimal Hankel norm approximation problem of the original system Σ.

[1]Note that from equation (4.1), it follows that $K \in H_{\infty,k}(C_-, \mathbb{C}^{p\times m})$ if and only if $K_r \in H_{\infty,k}(C_-, \mathbb{C}^{p\times m})$.

In light of these remarks, we have thus proved the following theorem.

THEOREM 4.1. *Suppose that the well-posed linear system Σ with transfer function G satisfies assumptions* B1–B4. *Let σ be such that $\sigma_{k+1} < \sigma < \sigma_k$, where $\sigma_1 \geq \sigma_2 \geq \cdots$ are the Hankel singular values of G. Let V be given by*

$$V(s) = \begin{bmatrix} I_p & 0 \\ 0 & \frac{1}{\sigma} I_m \end{bmatrix} - \frac{1}{\sigma^2} \begin{bmatrix} CA^{-1}L_B \\ \sigma(A^{-1}B)^* \end{bmatrix} \big(s + (A^{-1})^*\big)^{-1} \left(I - \frac{1}{\sigma^2} L_C L_B\right)^{-1} \cdot \begin{bmatrix} -(CA^{-1})^* & \frac{1}{\sigma} L_C B A^{-1} \end{bmatrix},$$

$s \in \mathbb{C}_-$, where L_B and L_C denote the controllability Gramian and the observability Gramian, respectively, of the system Σ, and $N_\sigma := (I - \frac{1}{\sigma^2} L_B L_C)^{-1}$. Then $K \in H_{\infty,k}(\mathbb{C}_-, \mathbb{C}^{p\times m})$ satisfies $\|G(i\cdot) + K(i\cdot)\|_\infty \leq \sigma$ if and only if

$$K(s) = K_r\left(\frac{1}{s}\right) - G(0),$$

where $K_r(i\omega) = (V_{11}(i\omega)Q(i\omega) + V_{12}(i\omega))(V_{21}(i\omega)Q(i\omega) + V_{22}(i\omega))^{-1}$, $\omega \in \mathbb{R}$, for some $Q \in H_\infty(\mathbb{C}_-, \mathbb{C}^{p\times m})$ such that $\|Q\|_\infty \leq 1$.

This solves the suboptimal Hankel norm approximation problem for well-posed linear systems.

5. Appendix A. In this appendix we present the proofs that were deferred in section 2.

5.1. Proof of Proposition 2.14.

Proof. Suppose that the matrix-valued function K has a Kreĭn–Langer factorization $K = F \cdot B^{-1}$ with $F \in H_\infty(\mathbb{D}, \mathbb{C}^{p\times m})$ and with $B \in H_\infty(\mathbb{D}, \mathbb{C}^{m\times m})$ a Blaschke–Potapov function of degree k. Then the graph of the multiplication operator M_{G+K} restricted to the subspace $B \cdot H_2(\mathbb{D}, \mathbb{C}^m)$ satisfies

$$\begin{aligned} \mathcal{G}_{M_{G+K}} &:= \begin{bmatrix} G+K \\ I \end{bmatrix} BH_2(\mathbb{D}, \mathbb{C}^m) \\ &= \begin{bmatrix} I & G \\ 0 & I \end{bmatrix} \begin{bmatrix} K \\ I \end{bmatrix} BH_2(\mathbb{D}, \mathbb{C}^m) \\ &\subset \begin{bmatrix} I & G \\ 0 & I \end{bmatrix} H_2(\mathbb{D}, \mathbb{C}^{m+p}) =: \mathcal{M}. \end{aligned} \tag{5.1}$$

If also $\|G+K\|_\infty \leq \sigma$, then we see that $\mathcal{G}_{G+K}$ is a *negative subspace* in the Kreĭn-space inner product

$$\langle f, g\rangle_{J_\sigma} = \frac{1}{2\pi} \int_{\mathbb{T}} \langle J_\sigma f(\zeta), g(\zeta)\rangle_{\mathbb{C}^{m+p}} \, |d\zeta|$$

on $L_2(\mathbb{T}, \mathbb{C}^{m+p})$; i.e., each function $f \in \mathcal{G}_{M_{G+K}}$ has negative J_σ-self-inner product

$$\langle f, f\rangle_{J_\sigma} \leq 0 \quad \text{for } f \in \mathcal{G}_{M_{G+K}}. \tag{5.2}$$

Since $B \cdot H_2(\mathbb{D}, \mathbb{C}^m)$ has codimension k in $H_2(\mathbb{D}, \mathbb{C}^m)$, we see in addition that $\mathcal{G}_{M_{G+K}}$ *has codimension k in a maximal negative subspace of the Kreĭn space* $\mathcal{K} := \big(\left[\begin{smallmatrix} L_2(\mathbb{T}, \mathbb{C}^p) \\ H_2(\mathbb{D}, \mathbb{C}^m) \end{smallmatrix}\right], \langle\cdot,\cdot\rangle_{J_\sigma}\big)$. In addition, since $\mathcal{G} := \mathcal{G}_{G+K}$ is the graph of a multiplication operator M_{G+K}, we see that $\mathcal{G}$ *is invariant for the shift operator* $M_\zeta \colon f(\zeta) \mapsto \zeta f(\zeta)$. We have

thus verified the following: *If $K = FB^{-1}$ is a solution of the Nehari–Takagi problem, then the subspace $\mathcal{G} = \mathcal{G}_{G+K} := \left[\begin{smallmatrix} G+K \\ I \end{smallmatrix}\right] \cdot B \cdot H_2(\mathbb{D}, \mathbb{C}^m)$ (where B is the Blaschke–Potapov product of degree k chosen so that $K \cdot B \in H_\infty(\mathbb{D}, \mathbb{C}^{p\times m})$) satisfies conditions* (1)–(3) *in the statement of Proposition* 2.14.

Conversely, if $\mathcal{G}$ is a subspace of $\mathcal{K}$ which satisfies conditions (1)–(3) in Proposition 2.14, one can reverse the steps and come up with a $K \in H_{\infty,k'}(\mathbb{D}, \mathbb{C}^{p+m})$ ($k' \le k$) which solves the Nehari–Takagi problem as follows. Since $\mathcal{G}$ is a negative subspace in the J_σ-inner product, $\mathcal{G}$ necessarily has the form of a graph space

$$\mathcal{G} = \begin{bmatrix} X \\ I \end{bmatrix} \mathcal{D}(X),$$

where the angle operator $X\colon \mathcal{D}(X) \mapsto L_2(\mathbb{T}, \mathbb{C}^p)$ has domain $\mathcal{D}(X) \subset H_2(\mathbb{D}, \mathbb{C}^m)$ and norm $\|X\| \le \sigma$. Since $\mathcal{G}$ has codimension k in a maximal negative subspace, necessarily $\dim H_2(\mathbb{D}, \mathbb{C}^m) \ominus \mathcal{D}(X) = k$. Since $\mathcal{G}$ is shift invariant, we have

$$\begin{bmatrix} M_\zeta X \\ M_\zeta \end{bmatrix} \mathcal{D}(X) \subset \begin{bmatrix} X \\ I \end{bmatrix} \mathcal{D}(X).$$

Hence $\mathcal{D}(X)$ is shift invariant, and

$$M_\zeta X x = X M_\zeta x \quad \text{for } x \in \mathcal{D}(X).$$

But then, by the Beurling–Lax theorem, $\mathcal{D}(X)$ has the form $\mathcal{D}(X) = B \cdot H_2(\mathbb{D}, \mathbb{C}^m)$ for a Blaschke–Potapov factor of degree k, and the rule

$$X\colon \zeta^{-n} Bh \mapsto \zeta^{-n} X(Bh)$$

(for $h \in H_2(\mathbb{D}, \mathbb{C}^m)$ and $n = 0, 1, 2, \dots$) extends X to an operator, still called X, defined on the dense subset $\cup_{n=0}^\infty \zeta^{-n} B H_2(\mathbb{D}, \mathbb{C}^m)$ of $L_2(\mathbb{T}, \mathbb{C}^m)$, still with norm $\|X\| \le \sigma$, such that $X M_\zeta = M_\zeta X$. This forces X to be a multiplication operator $X = M_{G+K}$ for some matrix function $K \in L_\infty(\mathbb{T}, \mathbb{C}^{p+m})$ with $\|G + K\|_\infty \le \sigma$. From the fact that $\mathcal{G} \subset \mathcal{M}$, we have

$$\begin{bmatrix} G+K \\ I \end{bmatrix} B H_2(\mathbb{D}, \mathbb{C}^m) \subset \begin{bmatrix} I & G \\ 0 & I \end{bmatrix} H_2(\mathbb{D}, \mathbb{C}^m),$$

i.e.,

$$\begin{bmatrix} I & G \\ 0 & I \end{bmatrix}^{-1} \begin{bmatrix} G+K \\ I \end{bmatrix} B H_2(\mathbb{D}, \mathbb{C}^m) = \begin{bmatrix} K \\ I \end{bmatrix} B H_2(\mathbb{D}, \mathbb{C}^m) \subset H_2(\mathbb{D}, \mathbb{C}^{p+m}).$$

In particular, $K \cdot B$ maps $H_2(\mathbb{D}, \mathbb{C}^m)$ into $H_2(\mathbb{D}, \mathbb{C}^p)$, and we see that $F\colon = K \cdot B \in H_\infty(\mathbb{D}, \mathbb{C}^{p\times m})$. But then $K = F \cdot B^{-1}$ has the Kreĭn–Langer factorization form required to be in the class $H_{\infty,k'}(\mathbb{D}, \mathbb{C}^{p\times m})$ for k' at most k. Proposition 2.14 now follows. □

5.2. Proof of Proposition 2.16.

Proof. By Proposition 2.7, since $\sigma_k > \sigma > \sigma_{k+1}$ we know that the existence of a solution $K \in H_{\infty,k'}(\mathbb{D}, \mathbb{C}^{p\times m})$ with $k' \le k$ forces $k' = k$. This combined with the result in Proposition 2.14 implies that the only content to be added by Proposition 2.16 is that, under the hypothesis that $\sigma_k > \sigma > \sigma_{k+1}$, *a subspace $\mathcal{G}$ of $\mathcal{M}$ is $\mathcal{M}$-maximal negative* (i.e., maximal as a negative subspace contained in $\mathcal{M}$) *if and only*

if $\mathcal{G} \subset \mathcal{M}$ has codimension k in a $\mathcal{K}$-maximal negative subspace $\widetilde{\mathcal{G}}$ of $\mathcal{K}$. One can see this general principle as follows. As $\mathcal{M}$ is regular, $\mathcal{M}$ has a fundamental decomposition $\mathcal{M} = \mathcal{M}_+ \dot{+} \mathcal{M}_-$, where $\mathcal{M}_+$ is a uniformly positive subspace and $\mathcal{M}_-$ is a uniformly negative subspace in the Kreĭn-space inner product $\langle \cdot, \cdot \rangle_{J_\sigma}$. As $\mathcal{M}^{[\perp]}$ is also regular, $\mathcal{M}^{[\perp]}$ also has a fundamental decomposition as $\mathcal{M}^{[\perp]} = \mathcal{P} \dot{+} \mathcal{N}$, where $\mathcal{P}$ is uniformly positive and $\mathcal{N}$ is uniformly negative. We note also that, as a consequence of Proposition 2.15, $\dim \mathcal{N} = k$. Then $\mathcal{K} = \mathcal{K}_+ \dot{+} \mathcal{K}_-$ is a fundamental decomposition for $\mathcal{K}$, where

$$\mathcal{K}_+ = \mathcal{M}_+ \dot{+} \mathcal{P}, \qquad \mathcal{K}_- = \mathcal{M}_- \dot{+} \mathcal{N}.$$

By the angle-operator–graph correspondence, $\mathcal{M}$-maximal negative subspaces of $\mathcal{M}$ are of the form

$$\mathcal{G} = \{Xx + x \colon x \in \mathcal{M}_-\},$$

where X is a Hilbert space contraction operator from $(\mathcal{M}_-, -\langle \cdot, \cdot \rangle_{J_\sigma})$ into $(\mathcal{M}_+, \langle \cdot, \cdot \rangle_{J_\sigma})$. Similarly, $\mathcal{K}$-maximal negative subspaces of $\mathcal{K}$ are of the form

$$\widetilde{\mathcal{G}} = \{\widetilde{X}x + x \colon x \in \mathcal{K}_- = \mathcal{M}_- \dot{+} \mathcal{N}\},$$

where $\widetilde{X}$ is a contraction operator from $(\mathcal{K}_-, -\langle \cdot, \cdot \rangle_{J_\sigma})$ into $(\mathcal{K}_+ = \mathcal{M}_+ \dot{+} \mathcal{P}, \langle \cdot, \cdot \rangle_{J_\sigma})$. From this model, it is clear that $\mathcal{M}$-maximal negative subspaces of $\mathcal{M}$ match up exactly with those subspaces of $\mathcal{M}$ which have codimension k in a $\mathcal{K}$-maximal negative subspace of $\mathcal{K}$. This completes the proof of Proposition 2.16. ∎

5.3. Proof of Proposition 2.17. The proof of Proposition 2.17 requires a preliminary lemma.

LEMMA 5.1. *Suppose that $R \in H_2(\mathbb{D}, \mathbb{C}^{p+m})$ is outer and that $\mathcal{G}$ is a closed, shift-invariant subspace of $H_2(\mathbb{D}, \mathbb{C}^{p+m})$. Then $\mathcal{G} \cap R \cdot H_\infty(\mathbb{D}, \mathbb{C}^{p+m})$ is dense in $\mathcal{G}$.*

Proof. Let $g \in \mathcal{G}$. For $n = 1, 2, \ldots$ choose scalar outer functions r_n so that

$$|r_n(\zeta)| = \min\left\{ \frac{n}{\|R(\zeta)^{-1} g(\zeta)\|}, 1 \right\} \quad \text{for almost all } \zeta \in \mathbb{T}.$$

Then $g_n := r_n \cdot g \in \mathcal{G}$ since $\mathcal{G}$ is shift invariant. Since $\|R^{-1}(\zeta) g_n(\zeta)\| \le n$ for almost all $\zeta \in \mathbb{T}$, by construction, we see that $g_n \in R \cdot H_\infty(\mathbb{D}, \mathbb{C}^{p+n})$. Finally, since $|r_n(\zeta)| \le 1$ for almost all $\zeta \in \mathbb{T}$ and $g \in H_2(\mathbb{D}, \mathbb{C}^{p+m})$, we see that $\{g_n\}$ converges to g as $n \to \infty$ in $H_2(\mathbb{D}, \mathbb{C}^{p+m})$, and the lemma follows. ∎

Proof of Proposition 2.17. Suppose first that $\mathcal{G} \subset \mathcal{M}$ is maximal negative as a subspace of $\mathcal{M}$ in the J_σ-inner product. Then $\mathcal{G}$ has the form $\mathcal{G} = \left[\begin{smallmatrix} I_p & G \\ 0 & I_m \end{smallmatrix}\right] \cdot \mathcal{G}'$, where $\mathcal{G}'$ is a closed shift-invariant subspace of $H_2(\mathbb{D}, \mathbb{C}^{p+m})$. By Lemma 5.1 we know that $\mathcal{G}' \cap V \cdot H_\infty(\mathbb{D}, \mathbb{C}^{p+m})$ is dense in $\mathcal{G}'$. Multiplication by $\left[\begin{smallmatrix} I_p & G \\ 0 & I_m \end{smallmatrix}\right]$ then gives that $\mathcal{G} \cap \Theta \cdot H_\infty(\mathbb{D}, \mathbb{C}^{p+m})$ is dense in $\mathcal{G}$. We may write $\mathcal{G} \cap \Theta \cdot H_\infty(\mathbb{D}, \mathbb{C}^{p+m})$ in the form

$$\mathcal{G} \cap \Theta \cdot H_\infty(\mathbb{D}, \mathbb{C}^{p+m}) = \Theta \cdot \mathcal{G}_1,$$

where $\mathcal{G}_1 \subset H_\infty(\mathbb{D}, \mathbb{C}^{p+m})$.

We assert that $\mathcal{G}_1$ *is weak-$*$ closed in* $H_\infty(\mathbb{D}, \mathbb{C}^{p+m})$. By part (1) of Proposition 2.3, it suffices to consider a sequence $\{h_n\}_{n=1,2,\ldots}$ of elements of $\mathcal{G}_1$ convergent in the weak-$*$ topology to an element h of $L_\infty(\mathbb{D}, \mathbb{C}^{p+m})$ and prove that in fact $h \in \mathcal{G}_1$, i.e., that $\Theta h \in \mathcal{G} \cap \Theta \cdot H_\infty(\mathbb{D}, \mathbb{C}^{p+m})$. From the characterization of $H_\infty(\mathbb{D}, \mathbb{C}^{p+m})$

as that subspace of $L_\infty(\mathbb{T},\mathbb{C}^{m+p})$ consisting of functions F for which all the Fourier coefficients of negative index vanish,

$$\frac{1}{2\pi}\int_{\mathbb{T}} F(\zeta)\overline{\zeta}^n|d\zeta| = 0 \quad \text{for } n=-1,-2,\dots,$$

it is easily seen that $H_\infty(\mathbb{D},\mathbb{C}^{p+m})$ is weak-$*$ closed in $L_\infty(\mathbb{T},\mathbb{C}^{p+m})$ and hence $h \in H_\infty(\mathbb{D},\mathbb{C}^{p+m})$. It therefore remains only to show that $\Theta h \in \mathcal{G}$. For this purpose note that, for any $k \in L_2(\mathbb{T},\mathbb{C}^{p+m})$,

$$\langle \Theta h_n, k\rangle_{L_2} = \frac{1}{2\pi}\int_{\mathbb{T}} k(\zeta)^*\Theta h_n(\zeta)k(\zeta)|d\zeta|. \tag{5.3}$$

As $k^*\Theta \in L_2(\mathbb{T},\mathbb{C}^{1\times(m+n)}) \subset L_1(\mathbb{T},\mathbb{C}^{1\times(m+n)})$ and h_n is assumed to converge to h in the weak-$*$ topology of $L_\infty(\mathbb{T},\mathbb{C}^{(m+n)\times 1})$, we may take limits in (5.3) to get

$$\lim_{n\to\infty}\langle \Theta h_n, k\rangle_{L_2} = \langle \Theta h, k\rangle_{L_2} \quad \text{for each } k \in L_2(\mathbb{T},\mathbb{C}^{p+m}); \tag{5.4}$$

i.e., Θh_n converges to Θh in the weak topology on $L_2(\mathbb{T},\mathbb{C}^{p+m})$. As $\Theta h_n \in \mathcal{G}$ for each n and as norm-closed subspaces of a Hilbert space are also closed in the weak topology (see [42, Theorem 6.3, page 158]), it follows that $\Theta h \in \mathcal{G}$ as wanted. We conclude that $\mathcal{G}_1$ *is weak-$*$ closed as asserted.*

By the Beurling–Lax theorem for weak-$*$ closed subspaces of $H_\infty(\mathbb{D},\mathbb{C}^{m+p})$ (see, e.g., [41] or [25, page 25] for the scalar case), there is a matrix inner function $\psi = \left[\begin{smallmatrix}\psi_1\\ \psi_2\end{smallmatrix}\right]$ in $H_\infty(\mathbb{D},\mathbb{C}^{(p+m)\times m_1})$ (for some $m_1 \le m+p$) so that

$$\mathcal{G}_1 = \begin{bmatrix}\psi_1\\ \psi_2\end{bmatrix} H_\infty(\mathbb{D},\mathbb{C}^{m_1}). \tag{5.5}$$

The inner property of ψ means that

$$\psi_1(\zeta)^*\psi_1(\zeta) + \psi_2(\zeta)^*\psi_2(\zeta) = I_{m'} \quad \text{for almost all } \zeta \in \mathbb{T}. \tag{5.6}$$

From the fact that $\mathcal{G}_1$ is J_1-negative we then also get

$$\psi_1(\zeta)^*\psi_1(\zeta) - \psi_2(\zeta)^*\psi_2(\zeta) \le 0 \quad \text{for almost all } \zeta \in \mathbb{T}. \tag{5.7}$$

Hence, if we set $Q(\zeta) = \psi_2(\zeta)\psi_1(\zeta)^\ddagger$, where $\psi_2(\zeta)^\ddagger$ is the left Moore–Penrose generalized inverse of $\psi_2(\zeta)$,

$$\psi_2(\zeta)^\ddagger \colon c \mapsto \begin{cases} 0 & \text{if } c \perp \operatorname{im}\psi_2(\zeta),\\ c' & \text{if } c = \psi_2(\zeta)c', \end{cases} \tag{5.8}$$

then $Q(\zeta)$ defines a contraction operator from $\mathbb{C}^m$ into $\mathbb{C}^p$ for almost all $\zeta \in \mathbb{T}$, and we can rewrite (5.5) as

$$\mathcal{G}_1 = \begin{bmatrix} Q\\ I_m\end{bmatrix}\psi_2 H_\infty(\mathbb{D},\mathbb{C}^{m_1}). \tag{5.9}$$

We next argue that $\psi_2 H_\infty(\mathbb{D},\mathbb{C}^{m_1})$ *is weak-$*$ closed in* $H_\infty(\mathbb{D},\mathbb{C}^m)$. Indeed, suppose that $\psi_2 h_n$ converges in the L_∞-weak-$*$ topology to an element $k \in L_\infty$. Then the computation (for each $g \in L_1(\mathbb{T},\mathbb{C}^p)$)

$$[\psi_1 h_n, g] = [Q\psi_2 h_n, g] = [\psi_2 h_n, Q^* g] \to [k, Q^* g] = [Qk, g]$$

shows that $\psi_1 h_n$ tends weak-$*$ to Qk. (Here we let $[\cdot,\cdot]$ denote the duality pairing

$$[F,f]=\frac{1}{2\pi}\int_{\mathbb{T}} f(\zeta)^* F(\zeta)|d\zeta| \quad \text{for } F\in L_\infty(\mathbb{T},\mathbb{C}^{m'}) \text{ and } f\in L_1(\mathbb{T},\mathbb{C}^{m'})$$

between $L_\infty(\mathbb{T},\mathbb{C}^{m'})$ and $L_1(\mathbb{T},\mathbb{C}^{m'})$ for any fixed choice of m', and we use that $Q^*\cdot g\in L_1(\mathbb{T},\mathbb{C}^m)$ for any $g\in L_1(\mathbb{T},\mathbb{C}^p)$ since $Q(\zeta)$ is contractive a.e.) We conclude that

$$\begin{bmatrix}\psi_1\\ \psi_2\end{bmatrix} h_n \to \begin{bmatrix}Q\\ I_m\end{bmatrix} k$$

in the weak-$*$ topology. As $\mathcal{G}_1$ is closed in the weak-$*$ topology, it follows that $\left[\begin{smallmatrix}Q\\ I_m\end{smallmatrix}\right]k\in\mathcal{G}_1$, and hence, in particular, $k\in\psi_2 H_\infty(\mathbb{D},\mathbb{C}^{m_1})$. We conclude that $\psi_2 H_\infty(\mathbb{D},\mathbb{C}^{m_1})$ *is closed in the weak-$*$ topology as wanted.*

We next argue that in fact

$$\psi_2 H_\infty(\mathbb{D},\mathbb{C}^{m'})=H_\infty(\mathbb{D},\mathbb{C}^m). \tag{5.10}$$

Indeed, via a second application of the Beurling–Lax theorem for weak-$*$ closed subspaces of vector-valued H_∞, by the fact established in the previous paragraph it follows that there is an $m\times m'$ matrix inner function ϕ so that $\psi_2 H_\infty(\mathbb{D},\mathbb{C}^{m'})=\phi H_\infty(\mathbb{D},\mathbb{C}^{m'})$. If $m'<m$ (or more generally, if $\phi H_\infty(\mathbb{D},\mathbb{C}^{m'})$ does not fill up all of $H_\infty(\mathbb{D},\mathbb{C}^m)$), then we may choose a nonzero vector $f\in(H_2(\mathbb{D},\mathbb{C}^m)\ominus\psi H_2(\mathbb{D},\mathbb{C}^{m'}))\cap H_\infty(\mathbb{D},\mathbb{C}^m)$ so that the L_2-closure of $M_\Theta\cdot\left(\operatorname{span}\left[\begin{smallmatrix}0\\ f\end{smallmatrix}\right]+\mathcal{G}_1\right)$ is a larger negative subspace of $\mathcal{M}$ which includes $\mathcal{G}$ as a subspace. As $\mathcal{G}$ is assumed to be maximal negative in $\mathcal{M}$, this leads to a contradiction, and we conclude that $m'=m$ and ϕ is a unitary constant. We have now arrived at

$$\mathcal{G}\cap\Theta\cdot H_\infty(\mathbb{D},\mathbb{C}^{p+m})=\Theta\begin{bmatrix}Q\\ I\end{bmatrix}\cdot H_\infty(\mathbb{D},\mathbb{C}^m).$$

Taking closures in this identity gives the representation (2.15) for the shift-invariant subspace $\mathcal{G}$ assumed to be maximal negative in $\mathcal{M}$.

Conversely, suppose that $Q\in H_\infty(\mathbb{D},\mathbb{C}^{p\times m})$ with $\|Q\|_\infty\le 1$ and we define $\mathcal{G}\subset L_2(\mathbb{T},\mathbb{C}^{p\times m})$ by (2.15). From the factorization $\left[\begin{smallmatrix}I_p & G\\ 0 & I_m\end{smallmatrix}\right]=\Theta\cdot\Lambda$, where Λ and $V=\Lambda^{-1}$ are in $H_2(\mathbb{D},\mathbb{C}^{(p+m)\times(p+m)})$, it is clear that $\mathcal{G}\subset\mathcal{M}:=\left[\begin{smallmatrix}I_p & G\\ 0 & I_m\end{smallmatrix}\right]\cdot H_2(\mathbb{D},\mathbb{C}^{p+m})$. Since $\|Q\|_\infty\le 1$ and Θ is (J_1,J_σ)-unitary on $\mathbb{T}$, we see that necessarily $\mathcal{G}$ is negative in the J_σ-inner product. If $\mathcal{G}'$ is a J_σ-negative subspace with $\mathcal{G}\subset\mathcal{G}'\subset\mathcal{M}$, then by the same argument as in the first part of the proof we know that

$$\mathcal{G}'\cap\Theta\cdot H_\infty(\mathbb{D},\mathbb{C}^{p+m})=\mathcal{G}'_1$$

for some weak-$*$ closed subspace $\mathcal{G}'_1$ of $H_\infty(\mathbb{D},\mathbb{C}^{p+m})$. The (J_1,J_σ)-unitary property of Θ and the fact that $\mathcal{G}'$ is J_σ-negative then force $\mathcal{G}'_1$ to be J_1-negative. Hence $\mathcal{G}'_1$ can contain no elements of the form $\left[\begin{smallmatrix}h\\ 0\end{smallmatrix}\right]$ with $h\in H_\infty(\mathbb{D},\mathbb{C}^p)$ nonzero. We conclude that $\mathcal{G}'_1$ is a graph space; i.e., there is an operator X mapping some domain $\mathcal{D}(X)\subset H_\infty(\mathbb{D},\mathbb{C}^m)$ into $H_\infty(\mathbb{D},\mathbb{C}^p)$ so that $\mathcal{G}'_1=\left[\begin{smallmatrix}X\\ I_m\end{smallmatrix}\right]\mathcal{D}(X)$. Since $\mathcal{G}'\supset\mathcal{G}$, we see next that $\mathcal{G}'_1\supset\mathcal{G}_1$, i.e.,

$$\begin{bmatrix}X\\ I_m\end{bmatrix}\mathcal{D}(X)\supset\begin{bmatrix}Q\\ I_m\end{bmatrix}H_\infty(\mathbb{D},\mathbb{C}^m).$$

As $\mathcal{D}(X)\subset H_\infty(\mathbb{D},\mathbb{C}^m)$, we must have $\mathcal{D}(X)=H_\infty(\mathbb{D},\mathbb{C}^m)$, X is the operator of multiplication by Q, and $\mathcal{G}=\mathcal{G}'$ is $\mathcal{M}$-maximal negative. This concludes the proof of Proposition 2.17. ∎

6. Appendix B. In this appendix we present the proofs that were deferred in section 3.

To prove Lemma 3.2(1) we want to study how the operators determined by "shifted" triple $(A-\epsilon, B, C)$ converge to those determined by (A, B, C), as $\epsilon \to 0+$. For the shifted system, the semigroup is denoted by $\{T_\epsilon(t)\}_{t\geq 0}$, the controllability map is denoted by $\mathcal{C}_\epsilon$, and the controllability Gramian $\mathcal{C}_\epsilon^*\mathcal{C}_\epsilon$ is abbreviated by $L_{C,\epsilon}$; similarly, one uses the notation $\mathcal{B}_\epsilon^*, L_{B,\epsilon}, \mathcal{D}_\epsilon$, etc. Since the functions in A1, A2, and A3 are shifted to the left $(C(sI-(A-\epsilon I))^{-1}x = C((s+\epsilon)I-A)^{-1}x$, etc.), the assumptions A1–A3 hold a fortiori. Some further claims are straightforward, while others are more complicated.

LEMMA 6.1. *Assume that the triple (A, B, C) satisfies* A1*,* A2*, and* A3*. Then, with the above notation, as $\epsilon \to 0+$, we have the following for all $k \in \{1, 2, 3, \dots\}$, $t \geq 0$, $x \in X$, $u \in L^2(\mathbb{R}_+; \mathbb{C}^m)$, $y \in L^2(\mathbb{R}_+; \mathbb{C}^p)$:*

(1) $T_\epsilon(t) = \mathrm{e}^{-\epsilon t}T(t)$, $\mathcal{C}_\epsilon x = \mathrm{e}^{-\epsilon\cdot}\mathcal{C}x$, *and* $\mathcal{D}_\epsilon u = \mathrm{e}^{-\epsilon\cdot}\mathcal{D}\mathrm{e}^{\epsilon\cdot}u$.

(2) $\|\mathcal{C}_\epsilon\| \leq \|\mathcal{C}\|$ *and* $\|\mathcal{C}_\epsilon x_\epsilon - \mathcal{C}x\|_2 \to 0$ *whenever* $\|x_\epsilon - x\|_X \to 0$.

(3) $\|\mathcal{D}_\epsilon\| \leq \|\mathcal{D}\|$ *and* $\|\mathcal{D}_\epsilon^* y_\epsilon - \mathcal{D}^* y\|_2 \to 0$ *whenever* $\|y_\epsilon - y\|_2 \to 0$.

(4) $\|L_{C,\epsilon}\| \leq \|L_C\|$, $\|L_{B,\epsilon}\| \leq \|L_B\|$, $L_{C,\epsilon}x \to L_C x$, *and* $L_{B,\epsilon}x \to L_B x$.

(5) σ_k *is the* k*th singular value of* $\Gamma := \mathcal{C}\mathcal{B}$.

(6) $\sigma_{k,\epsilon} \to \sigma_k-$, *where* $\sigma_{k,\epsilon}$ *is the* k*th singular value of* $\Gamma_\epsilon := \mathcal{C}_\epsilon\mathcal{B}_\epsilon$.

(7) *Let* $\sigma_k > \sigma > \sigma_{k+1}$. *Then there are* $\epsilon_0 > 0$ *and* $M_0 < \infty$ *such that* $N_{\sigma,\epsilon} := (I - \sigma^{-2}L_{B,\epsilon}L_{C,\epsilon})^{-1}$ *exists and* $\|N_{\sigma,\epsilon}\| \leq M_0$ *for all* $\epsilon \in (0, \epsilon_0)$. *Moreover,* $N_{\sigma,\epsilon}x \to N_\sigma x$.

Proof. (1) This is straightforward.

(2) By (1), we have $\|\mathcal{C}_\epsilon\| \leq \|\mathcal{C}\|$. Obviously, $\|\mathcal{C}_\epsilon x - \mathcal{C}x\|_2 \to 0$. Since $\mathcal{C}_\epsilon x_\epsilon - \mathcal{C}x = \mathcal{C}_\epsilon(x_\epsilon - x) + (\mathcal{C}_\epsilon - \mathcal{C})x$, we obtain that (2) holds.

(3) We have $\widehat{\mathcal{D}_\epsilon u}(s) = (G(\cdot)u(\cdot-\epsilon))(s+\epsilon) = G(\epsilon+s)u(s)$, i.e., $G_\epsilon = G(\epsilon+\cdot)$. Therefore, $\|G_\epsilon\|_\infty \leq \|G\|_\infty$ and $G_\epsilon^* y_0 \to G^* y_0$ a.e. on $i\mathbb{R}$, for any $y_0 \in \mathbb{C}^p$ (see, e.g., Theorem 3.3.1(c1) of [30]). Consequently, $G_\epsilon^*\widehat{y} \to G^*\widehat{y}$ in $L^2(i\mathbb{R}; \mathbb{C}^m)$ for any $\widehat{y} \in L^2(i\mathbb{R}; \mathbb{C}^p)$, by the dominated convergence theorem. By Plancherel's theorem, this means that $\mathcal{D}_\epsilon^* y \to \mathcal{D}^* y$ for any $y \in L^2(\mathbb{R}; \mathbb{C}^m)$. Because the functions $\mathcal{D}_\epsilon^*$ are uniformly bounded, (3) also holds.

(4) We have

$$\|L_{C,\epsilon}\| = \sup_{\|x\|\leq 1}\langle x, L_{C,\epsilon}x\rangle = \sup_{\|x\|\leq 1}\|\mathcal{C}_\epsilon x\|_{L_2}^2 \leq \sup_{\|x\|\leq 1}\|\mathcal{C}x\|_{L_2}^2 = \sup_{\|x\|\leq 1}\langle x, L_C x\rangle = \|L_C\|.$$

Moreover, $\langle x, (L_C - L_{C,\epsilon})x\rangle = \int_0^\infty (1-\mathrm{e}^{-2\epsilon t})|(\mathcal{C}x)(t)|^2\,dt \to 0$. By duality (i.e., (A^*, C^*, B^*) in place of (A, B, C)), we obtain the claims for L_B.

(5) By Plancherel's theorem, H_G and Γ are isomorphic (see [16, Lemma 8.2.3(c), page 397]).

(6) Define $(S_\epsilon f)(t) = \mathrm{e}^{-\epsilon t}f$ $(\epsilon \in \mathbb{R})$. For any $f \in L^2(\mathbb{R}; \mathbb{C}^n)$, $n \geq 1$, we have $S_\epsilon f \to f$ in L^2 as $\epsilon \to 0$. Moreover, $\|S_\epsilon f\| \leq \|f\|$ (respectively, $\|S_{-\epsilon}f\| \leq \|f\|$) if $f = 0$ on $\mathbb{R}_-$ (respectively, $\mathbb{R}_+$). Therefore, $\Gamma_\epsilon^*\Gamma_\epsilon u \to \Gamma^*\Gamma u$ for each $u \in L^2(\mathbb{R}_-; \mathbb{C}^m)$ (see Mikkola [30, Lemma A.3.1(j3)] and note that $\mathcal{C}_\epsilon = S_\epsilon\mathcal{C}$, $\mathcal{B}_\epsilon = \mathcal{B}S_\epsilon$). Thus, we get $\liminf_{\epsilon\to 0+}\sigma_{k,\epsilon} \geq \sigma_k$ from Lemma 6.3.

Conversely, if $\operatorname{rank} K \leq k-1$, then $\operatorname{rank} K_\epsilon \leq k-1$, where $K_\epsilon := S_\epsilon K S_{-\epsilon}$, and $\|\Gamma_\epsilon - K_\epsilon\| = \|S_\epsilon(\Gamma - K)S_{-\epsilon}\| \leq \|\Gamma - K\|$. Hence $\sigma_{k+1,\epsilon} \leq \sigma_{k+1}$. Similarly, we observe that $\sigma_{k+1,\epsilon} \leq \sigma_{k+1,\epsilon'}$ when $\epsilon > \epsilon' > 0$.

(7) Let δ be as in Lemma 6.2(2), and choose ϵ_0 so that $\sigma_{l,\epsilon}^2 - \sigma_l^2 < \delta/2$ for $l = k$, $k+1$, and $\epsilon \in (0, \epsilon_0)$ (use (6)). Then Lemma 6.2(2) implies that $\|(\sigma^2 I - \Gamma_\epsilon^*\Gamma_\epsilon)^{-1}\| \leq 2/\delta$,

that is, that $\|(I-\sigma^{-2}\Gamma_\epsilon^*\Gamma_\epsilon)^{-1}\| \le 2\sigma^2/\delta$, for $\epsilon \in (0,\epsilon_0)$. Apply $(I-ST)^{-1} = I + S(I-TS)^{-1}T$ to $T = \sigma^{-2}\mathcal{B}_\epsilon^*$, $S := \mathcal{C}_\epsilon^*\mathcal{C}_\epsilon\mathcal{B}_\epsilon$ to obtain the inequality in (7) for $M_0 := 1+2\sigma^2\|\mathcal{C}\|\|\mathcal{B}\|/\delta$ (use (2) and its dual). The last claim follows from the others and (4) (see (j3)–(j5) of Lemma A.3.1 of Mikkola [30]). ∎

In the above we have used the following two lemmas. In the first one we present some sort of a singular value decomposition with k largest singular values on the diagonal and a small operator on the bottom-right corner.

LEMMA 6.2 (partial singular value decomposition). *Assume that $\{\sigma_k\}$ are the singular values of $S \in \mathcal{L}(X,Y)$ and X,Y are Hilbert spaces.*

(1) *For any $k \in \{0,1,2,\dots\}$, there is a k-dimensional subspace $X_k \subset X$ such that $S^*S = \operatorname{diag}(\sigma_1^2,\dots,\sigma_k^2;T)$ on $X_k \times X_k^\perp = X$, $\|T\| = \sigma_{k+1}^2$.*

(2) *We have $\sigma^2 \in \rho(S^*S)$ and $\|(\sigma^2 - S^*S)^{-1}\| \le \delta^{-1}$, where $\delta := \min\{\sigma_k^2 - \sigma^2, \sigma^2 - \sigma_{k+1}^2\}$.*

Claim (1) follows from pp. 212–213 of [21], alternatively, by using a resolution of the identity of S^*S. Claim (2) follows, because $(\sigma^2 - S^*S)^{-1} = \operatorname{diag}((\sigma_1^2-\sigma^2)^{-1},\dots,(\sigma_k^2-\sigma^2)^{-1};(\sigma^2-T)^{-1})$. Recall that $\sigma_k := \inf\{\|S-K\| : K \in \mathcal{L}(X,Y), \operatorname{rank} K \le k-1\} = \inf_{\dim M \le k-1}\|SP_{M^\perp}\|$, where $P_{M^\perp}$ is the orthogonal projection $X \to M^\perp$.

LEMMA 6.3 ($\liminf_n \sigma_{k,n} \ge \sigma_k$). *Let $S_n, S \in \mathcal{L}(X,Y)$ for all n, and let $S_n^*S_nx \to S^*Sx$ for all $x \in X$, where X,Y are Hilbert spaces. Then $\liminf_{n\to\infty}\sigma_{k,n} \ge \sigma_k$, where $\sigma_{k,n}$ is the kth singular value of S_n $(n \in \mathbb{N})$.*

Proof. Given $\epsilon > 0$, choose N such that $\|(S_n^*S_n - S^*S)P\| < \epsilon$ for all $n \ge N$, where $P : X \to X_{k-1}$ is the orthogonal projection and X_{k-1} is as in Lemma 6.2.

We obviously have $\sigma_k = \inf_{\dim M \le k-1}\|SP_{M^\perp}\|$, where $P_{M^\perp}$ is the orthogonal projection $X \to M^\perp$. But if $\dim M \le k-2$, then there is $x \in M^\perp \cap X_{k-1}$ such that $\|x\| = 1$ (otherwise we would have $X_{k-1} \subset \operatorname{Ker}(P_{M^\perp}) = M$). Then $\|S_nx\| = \langle Px, S_n^*S_nPx\rangle \ge \langle x, S^*Sx\rangle - \epsilon \ge \sigma_k - \epsilon$. ∎

Acknowledgments. The authors would like to thank Professor Ruth Curtain of the Department of Mathematics, University of Groningen, The Netherlands, and the anonymous referees for useful comments on the previous draft, which led to improvements in the exposition.

REFERENCES

[1] V. M. ADAMJAN, D. Z. AROV, AND M. G. KREĬN, *Infinite Hankel block matrices and related problems of extension*, Amer. Math. Soc. Transl. Ser. 2, 111 (1978), pp. 133–156 (in English); Izv. Akad. Nauk. Armjan. SSR Ser. Mat., 6 (1971), pp. 87–112 (in Russian).

[2] J. A. BALL, *Nevanlinna-Pick interpolation: Generalizations and applications*, in Surveys of Some Recent Results in Operator Theory I, J. B. Conway and B. B. Morrell, eds., Pitman, Boston, 1988, pp. 51–94.

[3] J. A. BALL, I. GOHBERG, AND L. RODMAN, *Interpolation of Rational Matrix Functions*, Oper. Theory Adv. Appl. 45, Birkhäuser-Verlag, Basel, Switzerland, 1990.

[4] J. A. BALL AND J. W. HELTON, *A Beurling-Lax theorem for the Lie group $U(m,n)$ which contains most classical interpolation theory*, J. Operator Theory, 9 (1983), pp. 107–142.

[5] J. A. BALL AND J. W. HELTON, *Factorization results related to shifts in an indefinite metric*, Integral Equations Operator Theory, 5 (1982), pp. 632–658.

[6] J. A. BALL AND J. W. HELTON, *Shift invariant subspaces, passivity, reproducing kernels and H_∞-optimization*, in Contributions to Operator Theory and Its Applications, I. Gohberg, J. W. Helton, and L. Rodman, eds., Oper. Theory Adv. Appl. 35, Birkhäuser-Verlag, Basel, Switzerland, 1988, pp. 265–310.

[7] J. A. BALL AND A. C. M. RAN, *Optimal Hankel norm model reductions and Wiener–Hopf factorization* I: *The canonical case*, SIAM J. Control Optim., 25 (1987), pp. 362–382.

[8] J. S. Baras and R. W. Brockett, *H^2-functions and infinite-dimensional realization theory*, SIAM J. Control Optim., 13 (1975), pp. 221–241.
[9] J. Bognár, *Indefinite Inner Product Spaces*, Springer-Verlag, New York, Heidelberg, Berlin, 1974.
[10] K. F. Clancey and I. Gohberg, *Factorization of Matrix Functions and Singular Integral Operators*, Oper. Theory Adv. Appl. 3, Birkhäuser-Verlag, Basel, Boston, 1981.
[11] R. F. Curtain and J. C. Oostveen, *The Nehari problem for nonexponentially stable systems*, Integral Equations Operator Theory, 31 (1998), pp. 307–320.
[12] R. F. Curtain and A. C. M. Ran, *Explicit formulas for Hankel norm approximations of infinite-dimensional systems*, Integral Equations Operator Theory, 12 (1989), pp. 455–469.
[13] R. F. Curtain and A. J. Sasane, *Sub-optimal Hankel norm approximation for the analytic class of infinite-dimensional systems*, Integral Equations Operator Theory, 43 (2002), pp. 356–377.
[14] R. F. Curtain and A. J. Sasane, *Sub-optimal Hankel norm approximation for the Pritchard–Salamon class of infinite-dimensional systems*, Integral Equations Operator Theory, 39 (2001), pp. 98–126.
[15] R. F. Curtain and A. J. Sasane, *Hankel norm approximation for well-posed linear systems*, Systems Control Lett., 48 (2003), pp. 407–414.
[16] R. F. Curtain and H. J. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Texts Appl. Math. 21, Springer-Verlag, New York, 1995.
[17] J. C. Doyle, K. Glover, and K. Zhou, *Robust and Optimal Control*, Prentice–Hall, Englewood Cliffs, NJ, 1996.
[18] N. Dunford and L. T. Schwartz, *Linear Operators Part* I: *General Theory*, John Wiley & Sons, New York, 1958.
[19] K. Glover, *All optimal Hankel-norm approximations of linear multivariable systems and their L_∞ error bounds*, Internat. J. Control, 39 (1984), pp. 1115–1193.
[20] K. Glover, R. F. Curtain, and J. R. Partington, *Realization and approximation of linear infinite-dimensional systems with error bounds*, SIAM J. Control Optim., 26 (1988), pp. 863–898.
[21] I. Gohberg, S. Goldberg, and M. A. Kaashoek, *Classes of Linear Operators*, Vol. I, Oper. Theory Adv. Appl. 49, Birkhäuser-Verlag, Basel, Switzerland, 1990.
[22] I. Gohberg, S. Goldberg, and M. A. Kaashoek, *Classes of Linear Operators*, Vol. II, Oper. Theory Adv. Appl. 63, Birkhäuser-Verlag, Basel, Switzerland, 1993.
[23] P. R. Halmos, *A Hilbert Space Problem Book*, 2nd ed., Grad. Texts in Math. 19, Springer-Verlag, New York, Berlin, 1982.
[24] S. Hansen and G. Weiss, *New results on the operator Carleson measure criterion*, IMA J. Math. Control Inform., 14 (1997), pp. 3–32.
[25] H. Helson, *Lectures on Invariant Subspaces*, Academic Press, New York, London, 1964.
[26] O. Iftime, M. A. Kaashoek, H. Sandberg, and A. Sasane, *Grassmannian Approach to the Hankel Norm Approximation Problem*, preprint, Mittag–Leffler Institute, Stockholm, Sweden, 2003.
[27] M. G. Kreĭn and H. Langer, *Über die verallgemeinerten Resolventen und die charakteristische Funktion eines isometrischen Operators im Raume Π_κ*, Colloq. Math. Soc. János Bolyai, 5 (1972), pp. 353–399.
[28] M. G. Kreĭn and H. Langer, *Über einige Fortsetzungsprobleme, die eng mit der Theorie hermitescher Operatoren im Raume Π_κ zusammenhängen.* I. *Einige Funktionenklassen und ihre Darstellungen*, Math. Nachr., 77 (1977), pp. 187–236.
[29] S. Y. Kung and D. W. Lin, *A state-space formulation for optimal Hankel-norm approximations*, IEEE Trans. Automat. Control, 26 (1981), pp. 942–946.
[30] K. M. Mikkola, *Infinite-Dimensional Linear Systems, Optimal Control and Algebraic Riccati Equations*, Ph.D. thesis, Helsinki University of Technology, Helsinki, Finland, 2002; available online at http://www.math.hut.fi/∼kmikkola/research/thesis/.
[31] A. M. Nikolaičuk and I. M. Spitkovsky, *Factorization of Hermitian matrix-functions and their applications to boundary value problems*, Ukrainian Math. J., 27 (1975), pp. 767–779.
[32] B. Sz.-Nagy and C. Foiaş, *Harmonic Analysis of Operators on Hilbert Space*, American Elsevier, New York, 1970.
[33] J. Oostveen, *Strongly Stabilizable Distributed Parameter Systems*, Frontiers Appl. Math. 20, SIAM, Philadelphia, 2000.
[34] V. V. Peller, *Hankel Operators and Their Applications*, Springer Monogr. Math., Springer-Verlag, New York, 2003.
[35] L. Rodman, I. M. Spitkovsky, and H. J. Woerdeman, *Abstract Band Method via Factorization, Positive and Band Extensions of Multivariable Almost Periodic Matrix Functions, and Spectral Estimation*, Mem. Amer. Math. Soc. 762, AMS, Providence, RI, 2002.

[36] M. Rosenblum and J. Rovnyak, *Hardy Classes and Operator Theory*, Oxford Math. Monogr., Oxford University Press, New York, 1985.
[37] H. L. Royden, *Real Analysis*, 3rd ed., Prentice–Hall, Englewood Cliffs, NJ, 1988.
[38] D. Sarason, *Generalized interpolation in H^∞*, Trans. Amer. Math. Soc., 127 (1967), pp. 179–201.
[39] D. Sarason, *Operator-theoretic aspects of the Nevanlinna-Pick interpolation problem*, in Operators and Function Theory (Lancaster, 1984), NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. 153, S. C. Power, ed., D. Reidel, Dordrecht, The Netherlands, 1985, pp. 279–314.
[40] A. J. Sasane, *Hankel Norm Approximation for Infinite-Dimensional Systems*, Lecture Notes in Control and Inform. Sci. 277, Springer-Verlag, Berlin, 2002.
[41] T. P. Srinivasan, *Simply invariant subspaces*, Bull. Amer. Math. Soc., 69 (1963), pp. 706–709.
[42] A. E. Taylor and D. C. Lay, *Introduction to Functional Analysis*, 2nd ed., John Wiley & Sons, New York, 1980.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 2, pp. 564–583

INSTANTANEOUS CLOSED LOOP CONTROL OF THE NAVIER–STOKES SYSTEM*

M. HINZE†

Abstract. Instantaneous control is applied to the control of the instationary Navier–Stokes system. This control technique is closely related to receding horizon control and allows for an interpretation as suboptimal closed loop controller, whose parameters may be adjusted so as to stabilize the nonlinear equation under consideration. Besides stability analysis for the distributed control case, numerical examples for the continuous and discrete-in-time control laws are presented.

Key words. optimal control, instantaneous control, closed loop control, Navier–Stokes equations

AMS subject classifications. 49N90, 76D05, 93B52, 93C10

DOI. 10.1137/S036301290241246X

1. Introduction.

This research is devoted to the study of instantaneous control applied to the instationary Navier–Stokes system. It thereby builds upon and extends results of [14] for the Burgers equation, where a comprehensive discussion of the method also can be found.

In primal variables the Navier–Stokes system can be written in the form

$$\text{(D)} \qquad \begin{cases} y_t - \nu\Delta y + (y\nabla)y + \nabla p = \mathcal{B}u & \text{in } Q, \\ -\operatorname{div} y = 0 & \text{in } Q, \\ y = 0 & \text{on } (0,T)\times\partial\Omega, \\ y(0) = \phi & \text{in } \Omega. \end{cases}$$

Here $\nu := 1/\mathrm{Re}$ denotes the viscosity parameter, with Re denoting the Reynolds number. Furthermore, $Q := (0,T)\times\Omega$, with $\Omega \subset \mathbb{R}^2$ denoting an open bounded domain and $T > 0$ the time horizon. The control target is to match the given desired state z in the $L^2(Q)$-sense by adjusting the body force $\mathcal{B}u$ in an appropriate manner. Above, $\mathcal{B}$ denotes an abstract control extension operator which maps controls of an abstract Hilbert space $\mathcal{U}$ to admissible right-hand sides of the Navier–Stokes system (D). In this context the method of instantaneous control serves a dual purpose—to construct a closed loop feedback control law which steers the system state y to z for t tending to ∞, and to compute open loop control policies which (hopefully) approximate optimal open loop control strategies, i.e. solutions of

$$\text{(P)} \qquad \begin{cases} \min J(y,u) = \dfrac{1}{2}\displaystyle\int_0^T \int_\Omega |y-z|^2 dx\, dt + \frac{\alpha}{2}\int_0^T |u|_{\mathcal{U}}^2\, dt \\ \text{s.t. (D).} \end{cases}$$

Optimal control problems for the Navier–Stokes system of the form of (P) are mathematically and numerically well understood [1, 2, 9, 13]. For (P) instantaneous control

*Received by the editors July 31, 2002; accepted for publication (in revised form) December 6, 2004; published electronically August 31, 2005.
http://www.siam.org/journals/sicon/44-2/41246.html
†Technische Universität Dresden, Institut für Numerische Mathematik, Zellescher Weg 12-14, D-01062 Dresden, Germany (hinze@math.tu-dresden.de).

may be regarded as a suboptimal control approach which provides control policies on the time horizon $(0, T)$. However, as is shown in [14], instantaneous control may be regarded as nonlinear feedback control policy, and it is the main aim of this work to extend the stability analysis presented there to the Navier–Stokes system.

Instantaneous control works as follows. The uncontrolled Navier–Stokes equations are discretized with respect to time. Then, at selected time slices an instantaneous version of the cost functional is approximately minimized w.r.t. a stationary quasi-Navier–Stokes system, whose structure depends on the chosen time-discretization method. The control obtained is used to steer the system to the next time slice, where the procedure is repeated.

Instantaneous control therefore is closely related to receding horizon control (rhc) or model predictive control (mpc) with finite time horizon [7, 21, 22]. It was applied to control the Burgers equation with stochastic forcing in [5] and was also successfully applied to compute suboptimal controls for a great variety of fluid mechanical control problems [3, 4, 10, 12, 19, 20], and in the recent past it was also applied as a real-time control approach to the cooling of steel [25]. Linear body force feedback control for the Navier–Stokes system is investigated in [8, 17, 18].

As far as the author knows there are only a few stability investigations for the application of the method to the control of infinite-dimensional systems [14, 16].

In the present work the results obtained in [14] are extended to the instationary Navier–Stokes system in two spatial dimensions. While for the continuous control policies developed in [14] similar stability properties can be proved as in the case of the Burgers equation, for the discrete controllers developed there only conditional stability can be shown for the Navier–Stokes system (Theorems 4.4 and 4.5). This motivates the construction of a slightly modified control policy, for which unconditional stability is proved in Theorem 4.7. The main results of the present work can be summarized as follows.

Main results.

1. Given a sufficiently smooth desired state z, instantaneous control of the instationary Navier–Stokes system can be regarded as time discretization of a closed loop feedback policy K that steers the system exponentially fast to z, i.e., with S denoting the Stokes operator and $b(y)$ the nonlinearity of the Navier–Stokes equations

$$y_t + \nu S y + b(y) = K(y),$$

the solution of this system satisfying $|y(t) - z(t)|_{H^1} \leq c \exp(-\gamma t)$ with some positive constants c and γ (Theorems 4.1 and 4.2).

2. Instantaneous control may be regarded as a discrete-in-time feedback policy that steers the dynamical system exponentially fast to the desired state z, provided that either $z(0)$ is sufficiently close to the initial value ϕ or the viscosity parameter ν is sufficiently large (Theorems 4.4 and 4.5).

3. Instantaneous control gives rise to a discrete-in-time feedback policy that steers the dynamical system exponentially fast to the desired state z (Theorem 4.7).

The paper is organized as follows. In section 2 an appropriate functional analytic framework is introduced and preliminary results are collected. In section 3 the derivation of the instantaneous control approach, the formulation of the algorithm, and its interpretation as nonlinear continuous and discrete-in-time feedback control policy are sketched. The results in this part of the work are similar to those in [14] and are stated for the convenience of the reader. In section 4 both exponential stability of the continuous controllers and conditional exponential stability of the discrete controllers

are shown. To obtain these results in consequence of the nonlinearity in the Navier–Stokes system a more subtle analysis is necessary than for the Burgers equation in [14]. Moreover, a slightly modified discrete controller is proposed for which unconditional exponential stability is proved. Finally, in section 5 numerical examples are presented, which illustrate the theoretical results and also compare the feedback policy on a time horizon $[0,T]$ to the corresponding optimal open loop control strategy for (P).

Throughout this work c and C denote global generic constants whose dependencies are mentioned when necessary.

2. Preliminaries. Set $V = \{v \in H_0^1(\Omega)^2,\ \operatorname{div} v = 0\}$, $H = \operatorname{clos}_{L^2(\Omega)^2}\{v \in C_0^\infty(\Omega)^2,\ \operatorname{div} v = 0\}$ and identify the Hilbert space H with its dual H'. On H the common inner product is used, and V is endowed with the inner product

$$(\varphi,\psi)_V = (\nabla\varphi,\nabla\psi)_H \quad \text{for } \varphi,\psi \in V.$$

Moreover, with Z denoting a Hilbert space, $L^p(Z)$ $(1 \le p \le \infty)$ denotes the space of measurable abstract functions $\varphi : (0,T) \to Z$, which are p-integrable $(1 \le p < \infty)$ or essentially bounded on $(0,T)$ $(p = \infty)$, respectively.

As control space, $L^2(\mathcal{U})$ is taken, where $\mathcal{U}$ denotes the Hilbert space of abstract controls. The space $\mathcal{U}$ also is identified with its dual. Furthermore,

$$\mathcal{B} : \mathcal{U} \to V' \tag{2.1}$$

denotes the control extension operator, which is assumed to be bounded. In order to formulate the weak form of the instationary Navier–Stokes equations, let $W := W(V) = \{\varphi \in L^2(V) : \varphi_t \in L^2(V')\}$ supplied with the common inner product. Further, introduce

$$b(u,v,w) := \int_\Omega (u\cdot\nabla)vw\,dx.$$

Then from [24],

$$b(u,v,w) \le C \begin{cases} |u|_H^{\frac12}|u|_V^{\frac12}|v|_H^{\frac12}|v|_V^{\frac12}|w|_V & \forall u,v,w \in V, \\ |u|_H^{\frac12}|u|_V^{\frac12}|v|_V^{\frac12}|\Delta v|_H^{\frac12}|w|_H & \forall u \in V,\ v \in V \cap H^2(\Omega)^2,\ w \in H, \\ |u|_H|v|_V|w|_H^{\frac12}|\Delta w|_H^{\frac12} & \forall u \in H,\ v \in V,\ w \in V \cap H^2(\Omega)^2, \\ |u|_H^{\frac12}|\Delta u|_H^{\frac12}|v|_V|w|_H & \forall u \in V \cap H^2(\Omega)^2,\ v \in V,\ w \in H, \end{cases} \tag{2.2}$$

with a positive constant C, which for the uppermost estimate can be chosen as $C = \sqrt{2}$ [23, Chap. III, eq. (3.55)]. Moreover, for $y \in L^2(V)$ the function $b(y)$ defined by

$$\langle b(y), v\rangle_{V',V} := -b(y,y,v) \quad \forall\, v \in V \tag{2.3}$$

is an element of V' for almost all $t \in (0,T)$ and $b(y) \in L^1(V)$. Now let $P : L^2(\Omega)^2 \to H$ denote the Leray projector [6, Remark 1.10]. Then, the Stokes operator S is given by

$$S : \mathcal{D}(S) \subset H \to H, \qquad S := -P\Delta, \qquad \mathcal{D}(S) = H^2(\Omega)^2 \cap V.$$

Further define

$$A := \nu S$$

and denote by B the solution operator of

$$v + hAv = f \text{ in } V', \tag{2.4}$$

where $f \in V'$ is given. The operator B is linear, bounded, and self-adjoint, and there holds $B^{-1} = I + hA$.

In this setting the Navier–Stokes system (D) may be rewritten as the Burgers equation in the space V,

$$\begin{aligned} y_t + Ay &= b(y) + \mathcal{B}u, \\ y(0) &= \phi, \end{aligned}$$

where the nonlinearity $b(y)$ is defined in (2.3). The derivation and also the interpretation of instantaneous control for the Navier–Stokes system in the following therefore are abutted to the exposition in [14].

Let $X = W \times L^2(\mathcal{U})$ and $Y = L^2(V') \times H$. Introducing the operator $e : X \to Y$ by

$$e(y,u) = (e_1(y,u), e_2(y,u)) = (y_t - \nu\Delta y - b(y) - \mathcal{B}u, y(0) - \phi),$$

the Navier–Stokes system (D) can be expressed in the form $e(y,u) = 0$ in Y, and the optimal control problem (P) can be regarded as a minimization problem with equality constraints:

$$\text{minimize } J(y,u) \quad \text{s.t.} \quad e(y,u) = 0.$$

Among other things, it is shown in [13] that this problem admits a solution $(y^*, u^*) \in X$ and that both J and e are infinitely continuously Fréchet-differentiable.

Young's inequality

$$ab \le \delta a^2 + \frac{1}{4\delta} b^2 \quad \forall\ a, b \ge 0,\ \delta > 0, \tag{2.5}$$

and the following lemmas are frequently used in the proofs of the main theorems.

LEMMA 2.1. *For $y \in V$ let $w := By$. Then $w \in V \cap H^3(\Omega)^2$ and $Sw \in V$. Moreover,*

$$|w|_H^2 \ge |y|_H^2 - 2\nu h |y|_V^2. \tag{2.6}$$

Further, let $z := B((y\nabla)y) = Bb(y)$. Then $z \in V$ and

$$|z|_H^2 \le \frac{1}{2\nu h} |y|_H^2 |y|_V^2. \tag{2.7}$$

Proof. By the definition of the operator B the regularity claim for w follows from [23, Chap. I, Prop. 2.3]. Thus, $Sw \in H^1(\Omega)^2$ [6, Remark 1.10] and, since $y \in V$, $Sw \in V$. Therefore, Sw can be utilized as test function in the equation

$$w + hAw = y.$$

This gives

$$|w|_V^2 + \nu h |Sw|_H^2 = \int_\Omega \nabla w \nabla y \, dx \Rightarrow |w|_V^2 + 2\nu h |Sw|_H^2 \le |y|_V^2.$$

Furthermore, using y as test function, the latter estimate leads to

$$|y|_H^2 = \int_\Omega yw + \nu h \nabla y \nabla w \, dx \le \frac{1}{2}|y|_H^2 + \frac{1}{2}|w|_H^2 + h\nu |y|_V |w|_V \le \frac{1}{2}|y|_H^2 + \frac{1}{2}|w|_H^2 + h\nu |y|_V^2,$$

which gives the first claim.

To prove the second claim take z as test function in the equation

$$z + hAz = b(y).$$

This gives, using the first estimate of (2.2) and Young's inequality (2.5) with $\delta = \nu h$,

$$|z|_H^2 + \nu h |z|_V^2 \le |b(y)|_{V'} |z|_V \le \sqrt{2} |y|_H |y|_V |z|_V \le \frac{1}{2\nu h} |y|_H^2 |y|_V^2 + \nu h |z|_V^2,$$

which completes the proof of the lemma. □

LEMMA 2.2. *Let $y \in V$, $\kappa := By$, and $\tau := B\kappa$. Then*

$$\int_\Omega ByBSy \, dx = |y|_V^2 - \nu h \int_\Omega (S\kappa + S\tau) Sy \, dx$$

and

$$|S\tau|_H^2, |S\kappa|_H^2 \le \frac{1}{4\nu h} |y|_V^2. \tag{2.8}$$

Proof. The definition of κ and τ implies $\kappa \in H^3(\Omega)^2 \cap V$, $\tau \in H^5(\Omega)^2 \cap V$. Moreover, $S\kappa$ and $S\tau$ are elements of V. Integration by parts gives the first part of the claim. To obtain the second claim, test the equation for κ with $S\kappa$. This gives

$$|\kappa|_V^2 + \nu h |S\kappa|_H^2 \le |y|_V |\kappa|_V \le \begin{cases} \frac{1}{4}|y|_V^2 + |\kappa|_V^2 \\ \frac{1}{2}|y|_V^2 + \frac{1}{2}|\kappa|_V^2, \end{cases}$$

where Young's inequality (2.5) was used for the upper estimate with $\delta = 1$, and for the lower estimate with $\delta = \frac{1}{2}$. Since the same estimate holds with κ replaced by τ and y replaced by κ, and $|\kappa|_V \le |y|_V$ by the lower estimate, the lemma is proved. □

3. Instantaneous control strategy. For $m \in \mathbb{N}$ an equidistant discretization of the time interval $(0, T)$ is defined by $h = \frac{T}{m}$ and $t_k = kh$, $k = 0, 1, \dots, m$. Instantaneous versions of the cost functional J in (P) are given by

$$J^k : V \times \mathcal{U} \to \mathbb{R}, \qquad (y, u) \mapsto \frac{1}{2} \, |y - z^k|_H^2 + \frac{\alpha}{2} \, |u|_{\mathcal{U}}^2,$$

where

$$z^k = \frac{1}{h} \int_{t_k - \frac{h}{2}}^{t_k + \frac{h}{2}} z(s, \cdot) ds \tag{3.1}$$

and $z(t,\cdot) = 0$ for $t > T$. Finally, for $k = 1, \dots, m$ and $i = 1, 2$, introduce the operators $e_i^k : V \times \mathcal{U} \to V'$ by

$$e_1^k(y,u) = (I + hA)y - hb(y^{k-1}) - y^{k-1} - \mathcal{B}u,$$

and for later purposes also

$$e_2^k(y,u) = (I + hA)y - hb(y) - y^{k-1} - \mathcal{B}u,$$

where y^{k-1} denotes the state at the previous time slice.

The instantaneous optimal control problem for the semi-implicit time integration is given by

$$(\mathrm{P}^k) \qquad \text{minimize } J^k(y,u) \quad \text{s.t.} \quad e_1^k(y,u) = 0 \text{ in } V',$$

where $y^0 = \phi$. The initial value ϕ now is required to be an element of the space V. For given y^{k-1}, a pair (y^k, u^k) satisfies the subsidiary condition $e_1^k(y,u) = 0$ in V' if and only if

$$\nu h(y^k, \varphi)_V + (y^k, \varphi)_H = (y^{k-1}, \varphi)_H + \langle \mathcal{B}u^k + hb(y^{k-1}), \varphi \rangle_{V',V} \quad \forall \varphi \in V. \tag{3.2}$$

Since $\phi \in V$ holds, the right-hand side in this linear equation defines a bounded linear functional on V. Thus, for every $u^k \in \mathcal{U}$, (3.2) admits a unique solution $y^k \in V$ which satisfies the a priori estimate

$$|y^k|_V \le \frac{C}{\nu h} \left(|y^{k-1}|_H + h|y^{k-1}|^2{}_V + |u^k|_{\mathcal{U}} \right).$$

Since J^k is quadratic and e_1^k is linear, every problem (P^k), $k = 1, \dots, m$, admits a unique solution $(y_*^k, u_*^k) \in V \times \mathcal{U}$ which in fact defines a minimum for (P^k). Furthermore, the unique Lagrange multiplier $\lambda_*^k \in V$ together with the solution (y_*^k, u_*^k) satisfies the first-order necessary optimality conditions (note that A is self-adjoint)

$$(I + hA)y = \mathcal{B}u + y^{k-1} + hb(y^{k-1}), \tag{3.3a}$$

$$(I + hA)\lambda = -(y - z^k), \tag{3.3b}$$

$$\alpha u - \mathcal{B}^\star \lambda = 0. \tag{3.3c}$$

The optimal control problem (P^k) is equivalent to the unconstrained minimization of the functional

$$\hat{J}^k(u) = J^k(y(u), u)$$

over $\mathcal{U}$, where for a control $u \in \mathcal{U}$ the state $y(u) \in V$ is given as the unique solution to (3.2). The gradient of $\hat{J}^k$ at u is given by

$$\nabla \hat{J}^k(u) = \alpha u - \mathcal{B}^\star \lambda,$$

where for given u the function λ is obtained by first solving the linear quasi-Stokes problem (3.3a) for the state y and then solving (3.3b) for λ.

Remark 3.1. If one uses implicit time integration in problem (P^k), i.e., in the subsidiary condition the operator e_1^k is replaced by e_2^k, the adjoint equation (3.3b) alters to

$$(I + hA)\lambda - b'(y)^\star \lambda = -(y - z^k). \tag{3.4}$$

Into this equation the state y enters in two different fashions: first as observation $-(y - z^k)$, and second as coefficient in the nonlinearity $b'(y)^\star$. Since the gradient $\nabla \hat{J}^k(u)$ at a control u depends on λ, in the present case it depends on the observation $y^k - z^k$ and also on the whole state y^k in terms of the derivative of b. The structure of (3.4) is retained also in the case of boundary observation, where the observation enters as boundary condition into the adjoint equation, but the whole state y^k again enters as a coefficient function. As a consequence, computation of gradient information for $\hat{J}^k$ in the present case cannot be based on observations alone.

On the other hand, the adjoint equation (3.3b) depends only on the observation $y^k - z^k$. Therefore, gradient information for the functional $\hat{J}^k$ is available utilizing the observations only. In the particular case of boundary observation no information of the state in the whole computational domain is needed at all.

3.1. The algorithm. Instead of solving (P^k) exactly, the instantaneous control strategy provides only approximate solutions to this problem through applying one step of the steepest descent method with stepsize $\rho > 0$. The control obtained in this way is used to steer the system to the next time slice. With the gradient of the functional $\hat{J}^k$ available, this procedure in algorithmic form can be formulated as follows.

ALGORITHM 1 (instantaneous control).

1. Set $y^0 = \phi$, $k = 0$ and $t^0 = 0$.
2. Given an initial control u_0^k, solve
$$(I + hA)y = y^k + hb(y^k) + \mathcal{B}u_0^k,$$
$$(I + hA)\lambda = -(y - z^k).$$
3. Set $\nabla \hat{J}(u_0^k) = \alpha u_0^k - \mathcal{B}^\star \lambda$.
4. Given $\rho > 0$, set $u^{k+1} = u_0^k - \rho \nabla \hat{J}(u_0^k)$.
5. Solve
$$(I + hA)y^{k+1} = y^k + hb(y^k) + \mathcal{B}u^{k+1}.$$
6. Set $t_{k+1} = t_k + h$, $k = k + 1$. If $t_k < T$ goto 2.

Here, $\hat{J} = \hat{J}^k$. The choice of the stepsize ρ in step 4 of the algorithm is crucial. Since (P^k) is quadratic with linear constraints, the optimal choice ρ^* can be computed exactly and in the present situation is given by

$$\rho^* = -\frac{(y(u) - z, y(d))_H + \alpha(u, d)_{\mathcal{U}}}{|y(d)|_H^2 + \alpha|d|_{\mathcal{U}}^2} = \frac{|d|_{\mathcal{U}}^2}{|y(d)|_H^2 + \alpha|d|_{\mathcal{U}}^2} \le \frac{1}{\alpha}, \tag{3.5}$$

where $d = -\nabla \hat{J}(u)$. The computation of ρ^* requires only the computation of the auxiliary function $y(d)$.

It is shown in [11, 14] that Algorithm 1 allows the interpretation as a nonlinear discrete-in-time suboptimal closed loop control method, which turns out to be the stable time discretization of some continuous closed loop controller. For the convenience of the reader these results for the Navier–Stokes system are summarized in the following subsection. To simplify the exposition from here onward, $\mathcal{U} = V'$ and, thus, $\mathcal{B} = I$ are assumed. This choice is justified by the fact that in many applications of distributed control applied to systems governed by parabolic equations, the operator $\mathcal{B}$ defined in (2.1) plays the role of an extension operator.

3.2. Feedback control laws. It is shown in [11, Theorem 5.4.1], [14, Theorem 3] that Algorithm 1 allows the following interpretation.

THEOREM 3.2.

(i) *For $u_0^k = 0$ Algorithm 1 is equivalent to the semi-implicit time discretization with discretization stepsize h,*

(3.6)
$$(I + hA)y^{k+1} = y^k + hb(y^k) - \rho BB(y^k - z^k) - h\rho BB(b(y^k) - Az^k), \quad y^0 = \phi,$$

of the dynamical system

$$\dot{y} + Ay - b(y) = K(y) \quad \text{in } L^2(V') \quad \text{and} \quad y(0) = \phi, \tag{3.7}$$

where

$$K(y) = -\frac{\rho}{h}BB(y - z) - \rho BB(b(y) - Az). \tag{3.8}$$

(ii) *Choosing the initial control u_0^k in Algorithm 1 as the solution of*

$$\left(I - \frac{\rho}{1 - \rho\alpha}BB\right)u_0^k = \frac{1}{1 - \rho\alpha}\left(z^{k+1} - z^k + Az^{k+1} - b(z^k) + \rho BB(b(z^k) - Az^k)\right),$$

the algorithm is equivalent to

$$\begin{aligned}(I + hA)w^{j+1} = w^j + h(b(y^j) - b(z^j)) - \rho BBw^j \\ -\rho hBB(b(y^j) - b(z^j)), \quad w^0 = \phi - z(0),\end{aligned} \tag{3.9}$$

where $w := y - z$. The related continuous system is given by

$$y_t + Ay - b(y) = K(y) \quad \text{in } L^2(V') \quad \text{and} \quad y(0) = \phi, \tag{3.10}$$

where

$$K(y) = -\frac{\rho}{h}BB(y - z) - \rho BB(b(y) - b(z)) + z_t + Az - b(z). \tag{3.11}$$

It is now clear that the nonlinear operators K defined in (3.8) and (3.11), respectively, can be interpreted as nonlinear closed loop control policies for the Navier–Stokes equations. In this context it is important to note that the discretization stepsize h and the descent parameter ρ in the gradient step of Algorithm 1 in the continuous case may now be regarded as parameters defining the controller.

The discrete counterpart to (3.11) will frequently be used in what follows. It is given by

$$K^D(y) = -\frac{\rho}{h}BB(y - z^j) - \rho BB(b(y) - b(z^j)) + \frac{z^{j+1} - z^j}{h} + Az^{j+1} - b(z^j). \tag{3.12}$$

Unless otherwise stipulated, the following assumption holds from here onward.

Assumption 3.3. $0 \neq \phi \in V$, and $z \in H^{2,1}(Q)$.

Note that this assumption on the desired state z in particular implies that $z(0)$ is meaningful. Moreover, $z(0) \in V$.

4. Existence, uniqueness, and stability of solutions. In this section existence, uniqueness, stability, and regularity of a solution to (3.10) are discussed. The boundary $\partial\Omega$ is assumed to be as smooth as required by the existence and regularity results for the Stokes and quasi-Stokes problems considered in the proofs of the theorems below; see [23, Chap. I, Prop. 2.3]. It follows from the a priori estimates to be derived that the stabilized system (3.10) admits a unique solution. Moreover,

the H- and V-norms of difference $w = y - z$ decay exponentially with rate $-\frac{\rho}{h}$. To achieve these results the range of ρ has to be adapted to the size of $|z|^2_{L^\infty(H)}$. For this purpose set $d_H := 2|\phi - z(0)|^2_H$ and let the range of ρ be implicitly defined by

$$0 < \rho \le \rho_1 := \min\left(\rho_0, \frac{\nu^2}{2\nu^2 + \exp\left(\frac{4+\rho}{\nu}|z|^2_{L^2(V)}\right) d_H + |z|^2_{L^\infty(H)}}\right). \tag{4.1}$$

THEOREM 4.1. *Let $0 < \rho \le \rho_1$ with ρ_1 from (4.1) and let $h > 0$ be fixed. Further, let $\phi \in H$ and $z \in W$. Then (3.10) for every $T > 0$ admits a unique solution $y \in W$, and the difference $w = y - z$ satisfies*

$$\begin{cases} |w|^2_H \le C\,(\nu, |z|_W)\, e^{-\frac{\rho}{h}t} & \forall\, t \in [0,T], \\ |w|^2_{L^\infty(H)} \le C\,(\nu, |z|_W), & \\ |w|^2_{L^2(V)} \le C\,(\nu, |z|_W), & \text{and} \\ |w_t|^2_{L^2(V')} \le C\,(\nu, |z|_W)\left\{1 + \frac{\rho}{h}\right\}, & \end{cases} \tag{4.2}$$

where $C(\nu, |z|_W)$ is a positive constant independent of ρ and h.

Proof. Existence of a solution can be proved using a Galerkin ansatz in combination with the estimates derived below; compare [11, 14]. Uniqueness also follows from these estimates. To prove (4.2) use w as test function in the variational formulation of (3.10). This leads to

$$\begin{aligned} &\frac{d}{dt}|w|^2_H + \nu|w|^2_V + \frac{\rho}{h}|Bw|^2_H \\ &\quad = \int_\Omega ((w\nabla)w + (z\nabla)w + (w\nabla)z)w\,dx \\ &\qquad -\rho\int_\Omega (B((w\nabla)w + (z\nabla)w + (w\nabla)z))Bw\,dx = (i) + \cdots + (vi). \end{aligned}$$

Since $(i) = (ii) = 0$, it remains to estimate the terms $(iii), (iv), (v)$, and (vi) in order to derive a differential inequality for $y - z$. To begin, estimate, using Young's inequality and (2.7),

$$\begin{aligned} (iii) &\le \frac{\nu}{2}|w|^2_V + \frac{2}{\nu}|z|^2_V|w|^2_H, \\ (iv) \le \rho h|B((w\nabla)w)|^2_H + \frac{\rho}{4h}|Bw|^2_H &\le \frac{\rho}{2\nu}|w|^2_H|w|^2_V + \frac{\rho}{4h}|Bw|^2_H, \\ (v) + (vi) &\le \rho h\left\{|B((w\nabla)z)|^2_H + |B((z\nabla)w)|^2_H\right\} + \frac{\rho}{4h}|Bw|^2_H \\ &\le \frac{\rho}{4h}|Bw|^2_H + \frac{\rho}{2\nu}\left\{|w|^2_V|z|^2_H + |z|^2_V|w|^2_H\right\}, \end{aligned}$$

and apply estimate (2.6). This leads to

$$\frac{1}{2}\frac{d}{dt}|w|^2_H + \left(\nu\left(\frac{1}{2} - \rho\right) - \frac{\rho}{2\nu}\left(|w|^2_H + |z|^2_{L^\infty(H)}\right)\right)|w|^2_V + \frac{\rho}{2h}|w|^2_H \le \frac{4+\rho}{2\nu}|z|^2_V|w|^2_H.$$

Since ρ satisfies (4.1), $\nu(\frac{1}{2} - \rho) - \frac{\rho}{2\nu}(\exp(\frac{4+\rho}{\nu}|z|^2_{L^2(V)})d_H + |z|^2_{L^\infty(H)}) > 0$. For ρ in this range standard arguments yield

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}|w|^2_H + \left(\nu\left(\frac{1}{2} - \rho\right) - \frac{\rho}{2\nu}\left(\exp\left(\frac{4+\rho}{\nu}|z|^2_{L^2(V)}\right) d_H + |z|^2_{L^\infty(H)}\right)\right) \\ &\qquad \times |w|^2_V + \frac{\rho}{2h}|w|^2_H \le \frac{4+\rho}{2\nu}|z|^2_V|w|^2_H. \end{aligned} \tag{4.3}$$

Since the right-hand side in (4.3) is integrable, a further Gronwall argument gives the desired result. Note that the estimate for w_t is a direct consequence of the second and third estimates in (4.2). □

A similar result holds for the decay w.r.t. the V-norm of w. Now let the range of ρ be implicitly defined by the relation

$$0 < \rho \le \rho_2 := \min\left(\rho_0, \frac{1}{2}\frac{\nu^2}{2\nu^2 + 3\exp\left(\frac{4+\rho}{\nu}|z|^2_{L^2(V)}\right) d_H + 3|z|^2_{L^\infty(H)}}\right). \tag{4.4}$$

THEOREM 4.2. *Let ρ satisfy* (4.4) *and let y be the unique solution of* (3.10). *Then $y \in H^{2,1}(Q)$, and $w = y - z$ satisfies*

$$\begin{aligned}
&|w|^2_V \le C\left(\nu, |z|_{H^{2,1}(Q)}\right) e^{-\frac{\rho}{h}t} && \forall\, t \in [0,T],\\
&|w|^2_{L^\infty(V)} \le C\left(\nu, |z|_{H^{2,1}(Q)}\right),\\
&|w|^2_{L^2(H^2(\Omega)^2 \cap V)} \le C\left(\nu, |z|_{H^{2,1}(Q)}\right), && \text{and}\\
&|w_t|^2_{L^2(H)} \le C\left(\nu, |z|_{H^{2,1}(Q)}\right)\left\{1 + \frac{\rho}{h}\right\},
\end{aligned}$$

where $C(\nu, |z|_{H^{2,1}(Q)})$ is a positive constant independent of ρ and h.

Proof. Use Sw as test function in (3.10). This leads to

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}|w|^2_V + \nu|Sw|^2_H + \int_\Omega ((y\nabla)y - (z\nabla)z)\, Sw\, dx\\
&\quad = -\frac{\rho}{h}\int_\Omega BwBSw\, dx - \rho\int_\Omega B((y\nabla)y - (z\nabla)z)BSw\, dx.
\end{aligned}$$

Relation (4.4) implies $1 - \rho - \frac{3\rho}{\nu^2}(\exp(\frac{4+\rho}{\nu}|z|^2_{L^2(V)})d_H + |z|^2_{L^\infty(H)}) \ge \frac{1}{2}$. Restricting ρ to this range, similar to the derivation of (4.3) after several applications of (2.2), (2.8), Lemma 2.2, and Young's inequality, one ends up with

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}|w|^2_V + \frac{\nu}{2}|Sw|^2_H + \frac{\rho}{h}\left(1 - \rho - \frac{3\rho}{\nu^2}\left(\exp\left(\frac{4+\rho}{\nu}|z|^2_{L^2(V)}\right) d_H + |z|^2_{L^\infty(H)}\right)\right)|w|^2_V\\
&\quad \le C_\nu\left\{|w|^2_H|w|^2_V + |z|_H|Sz|_H + |z|^2_V + |z|^4_V\right\}|w|^2_V.
\end{aligned} \tag{4.5}$$

Since the right-hand side in (4.5) is integrable, a Gronwall argument gives the desired result. □

4.1. Stability of discrete controllers. First the stability properties of the instantaneous control procedure (3.9) are investigated. As will be shown, stability for a certain parameter range of h and ρ can be ensured only by requiring additionally either largeness of the viscosity parameter ν or smallness of $\phi - z(0)$. As is shown in [14], these restrictions do not apply when the procedure is applied to the instationary Burgers equation. Secondly, a slightly modified version of the controller (3.12) is applied to the fully implicit Euler discretization of the Navier–Stokes system. It turns out that the resulting discrete-in-time system for $w = y - z$ is unconditionally stable. Note that fully implicit discretization of the state is a realistic situation, since the discrete controller is applied to stabilize a physical system which is described by the

Navier–Stokes equations. Therefore, the choice of the discretization procedure for the uncontrolled state need not be linked to the discrete controller.

Throughout this section it is assumed that the following assumption holds.

Assumption 4.3. In addition to Assumption 3.3, let $z \in C([0,T]; H^{1,\infty}(\Omega)^2 \cap V)$.

In a preparatory step, an inequality relating the H-norms of $w = y^{j+1} - z^{j+1}$ and $v = y^j - z^j$ is derived, and $z := z^j$, $j \in \mathbb{N}$. To begin, test (3.9) with w. This gives

$$\begin{aligned} &\frac{1}{2}|w|_H^2 - \frac{1}{2}|v|_H^2 + \frac{1}{2}|w-v|_H^2 + \nu h|w|_V^2 \\ &\quad = h\int_\Omega \left((v\nabla)v + (z^j\nabla)v + (v\nabla)z^j\right) w\,dx - \rho\int_\Omega BvBw\,dx \\ &\qquad -\rho h\int_\Omega B\left((v\nabla)v + (z^j\nabla)v + (v\nabla)z^j\right) Bw\,dx = (i) + \cdots + (vii). \end{aligned} \tag{4.6}$$

Estimating, using (2.2),

$$|B((v\nabla)v)|_H^2 \le \frac{1}{2\nu h}|v|_H^2|v|_V^2$$

from Lemma 2.1 and Young's inequality yields

$$\begin{aligned} (i) &\le \frac{\nu h}{4}|w|_V^2 + \frac{2h}{\nu}|v|_H^2|v|_V^2, \\ (ii) &\le \frac{\nu h}{4}|w|_V^2 + \frac{|z|_\infty^2 h}{\nu}|v|_H^2, \\ (iii) &\le \frac{h}{2}|w|_H^2 + \frac{|z|_{1,\infty}^2 h}{2}|v|_H^2, \\ (iv) &= -\rho\int_\Omega |Bw|^2 dx - \rho\int_\Omega B(v-w)Bw \le \frac{\rho}{2}(\rho-1)|Bw|_H^2 + \frac{1}{2}|v-w|_H^2, \\ (v) &= -\rho h\int_\Omega B\left((v\nabla)v\right)Bw\,dx \le \frac{3\rho h}{2\nu(1-\rho)}|v|_H^2|v|_V^2 + \frac{\rho}{12}(1-\rho)|Bw|_H^2, \end{aligned}$$

and finally

$$\begin{aligned} (vi) + (vii) &= -\rho h\int_\Omega B\left((v\nabla)z^j + (z^j\nabla)v\right)Bw\,dx \\ &\le \left\{\frac{3\rho h^2|z|_{1,\infty}^2}{1-\rho} + \frac{3\rho h|z|_\infty^2}{4(1-\rho)\nu}\right\}|v|_H^2 + \frac{\rho(1-\rho)}{6}|Bw|_H^2. \end{aligned}$$

Now introduce

$$c_1(\rho,h) := \frac{1}{2} + \frac{\rho}{4}(1-\rho) - \frac{h}{2}$$

and

$$\begin{aligned} c_2^j(\rho,h,z) := \frac{1}{2} &+ \left[\frac{2h}{\nu} + \frac{3h\rho}{2\nu(1-\rho)}\right]|v|_V^2 \\ &+ \underbrace{\left[\frac{|z|_{1,\infty}^2 h}{2} + \frac{|z|_\infty^2 h}{\nu} + \frac{3\rho h^2|z|_{1,\infty}^2}{1-\rho} + \frac{3\rho h|z|_\infty^2}{(1-\rho)4\nu}\right]}_{=:\tilde{c}_2(h,\rho,z)}. \end{aligned}$$

With this notation the estimates above and Lemma 2.1 together with (4.6) give

$$c_1(\rho,h)|w|_H^2 - c_2^j(\rho,h,z)|v|_H^2 + \frac{\nu h}{2}(1-\rho+\rho^2)|w|_V^2 \le 0. \tag{4.7}$$

THEOREM 4.4 (conditional H-norm stability of instantaneous control). *Let $w^j := y^j - z^j$, where y^j denotes the iterates obtained by* (3.9). *Then*

$$\forall \rho \in (0,1)\ \exists h^*(\rho),\ 0<\kappa<1\ \forall j \in \mathbb{N} : |w^{j+1}|_H^2 \le \kappa^j |w^0|_H^2,$$

provided $0 < h \le h^*(\rho)$ and

$$crit := \frac{4-\rho}{\nu^2(1-\rho)(1-\rho+\rho^2)}|w^0|_H^2$$

is sufficiently small.

Proof. Fixing $\rho \in (0,1)$, define

$$\hat{c}_2(\rho,h,z) := \frac{1}{2} + crit + \tilde{c}_2(\rho,h,z)$$

and argue by induction as follows.

1. Set $j=0$, and choose $h_0 = h_0(\rho)$ and *crit* so small that for all $0 < h \le h_0$
   (a) $\hat{c}_2(\rho,h,z) \le 1$ and $c_2^0(\rho,h,z) \le 1$,
   (b) $\frac{\hat{c}_2(\rho,h,z)}{c_1(\rho,h)} = \kappa_1 < 1$, and
   (c) $\frac{c_2^0(\rho,h,z)}{c_1(\rho,h)} = \kappa_2 < 1$

hold. This is possible, since for $\rho \in (0,1)$ the term $\frac{\rho}{4}(1-\rho)$ in the definition of $c_1(\rho,h)$ is positive. Define $\kappa := \max(\kappa_1,\kappa_2)$. Then (4.7) implies

$$|w^1|_H^2 \le \kappa |w^0|_H^2$$

and

$$|w^1|_V^2 \le \frac{2}{\nu h(1-\rho+\rho^2)}|w^0|_H^2 = \frac{2}{\nu h(1-\rho+\rho^2)}\kappa^0|w^0|_H^2.$$

2. Now assume that for $j \in \mathbb{N}$
   (a) $|w^j|_H^2 \le \kappa^j |w^0|_H^2$ and
   (b) $|w^j|_V^2 \le \frac{2}{\nu h(1-\rho+\rho^2)}\kappa^{j-1}|w^0|_H^2$

hold true.

3. Then conclude from (4.7) that

$$\begin{aligned} c_2^j(\rho,h,z) &= \frac{1}{2} + \left[\frac{2h}{\nu} + \frac{3h\rho}{2\nu(1-\rho)}\right]|w^j|_V^2 + \tilde{c}_2(\rho,h,z) \\ &\le \frac{1}{2} + \left[\frac{2h}{\nu} + \frac{3h\rho}{2\nu(1-\rho)}\right]\frac{2}{\nu h(1-\rho+\rho^2)}\kappa^{j-1}|w^0|_H^2 + \tilde{c}_2(\rho,h,z) \\ &\le \frac{1}{2} + crit + \tilde{c}_2(\rho,h,z) = \hat{c}_2(\rho,h,z). \end{aligned}$$

Thus, a further application of (4.7) implies

$$|w^{j+1}|_H^2 \le \frac{c_2^j(\rho,h,z)}{c_1(\rho,h)}|w^j|_H^2 \le \frac{\hat{c}_2(\rho,h,z)}{c_1(\rho,h)}|w^j|_H^2 \le \kappa^{j+1}|w^0|_H^2$$

and

$$|w^{j+1}|_V^2 \le \frac{2}{\nu h(1-\rho+\rho^2)}c_2^j(\rho,h,z)|w^j|_H^2 \le \frac{2}{\nu h(1-\rho+\rho^2)}\kappa^j|w^0|_H^2,$$

which completes the proof of Theorem 4.4. ∎

The last estimate of the previous proof also yields stability with respect to the V-norm.

THEOREM 4.5 (conditional V-norm stability of instantaneous control).

$$\forall \rho \in (0,1)\ \exists h^*(\rho),\ 0<\kappa<1\ \forall j\in\mathbb{N},\ 0<h\le h^*: |w^j|_V \le \frac{2}{\nu h(1-\rho+\rho^2)}\kappa^{j-1}|w^0|_H,$$

provided

$$crit := \frac{4-\rho}{\nu^2(1-\rho)(1-\rho+\rho^2)}|w^0|_H^2$$

is sufficiently small.

Remark 4.6. The smallness of *crit* in Theorems 4.4 and 4.5 is a condition either on the smallness of the initial difference between state and desired state or on the smallness of the Reynolds number of the fluid. It has to be required since there are no better estimates available for the term (i) in (4.6). The term (v) in (4.6) could be estimated in a slightly different way to obtain a ρ^2 in front of $|v|_V^2$ (see the proof of the next theorem) and therefore could be reduced by decreasing ρ. However, for (i) in (4.6) there is no further knob to fix its size. For the Burgers equation the situation is much more comfortable at this stage. Due to the continuous embedding $H^1 \hookrightarrow L^\infty$ and the well-known L^2-H^1 interpolation estimate for L^∞ functions in one spatial dimension, one has

$$h\int_\Omega vv'w\,dx \le \frac{\nu h}{4}|w|_V^2 + h^{1-2\alpha}|w|_H^2 + h^{1+2\alpha}\frac{1}{2\sqrt{\nu}}|w|_V^2|w|_H^2 \quad \forall\ \alpha\in(0,1).$$

Following the lines of the proof of Theorem 4.4 one can now conclude that the smallness requirement on *crit* may be dropped provided ρ is sufficiently small, since the power of h in the last addend on the right-hand side of this estimate is larger than one. For more details see [14].

Finally, a discrete-in-time control policy for the Navier–Stokes system is considered which is unconditionally stable. Let

(4.8)
$$K^D(y) = -\frac{\rho}{h}BB(y-z^j) - \rho BB(b(y)-b(z^j)) + \frac{z^{j+1}-z^j}{h} + Az^{j+1} - b(z^{j+1}),$$

and consider the following discretization of (3.10):

$$\frac{y^{j+1}-y^j}{h} + Ay^{j+1} - b(y^{j+1}) = K^D(y^j), \quad j=0,1,\dots \quad \text{and} \quad y^0=\phi. \tag{4.9}$$

There holds the following theorem.

THEOREM 4.7 (H- and V-norm stability of (4.8)). *Let $w^j := y^j - z^j$. There exists some $\rho^*\in(0,1)$ such that for every $0<\rho\le\rho^*$ there exist an $h^*(\rho)>0$ and a positive $\kappa<1$ such that for all $j\in\mathbb{N}$*

$$|w^j|_H^2 \le \kappa^j|w^0|_H^2 \quad \textit{and}$$

$$|w^j|_V^2 \le \frac{2}{\nu h(1-\frac{2}{3}\rho+\frac{2}{3}\rho^2)}\kappa^{j-1}|w^0|_H^2,$$

provided $0<h\le h^*(\rho)$.

Proof. During the proof again let $v = y^j - z^j$, $w = y^{j+1} - z^{j+1}$ and test (4.9) with w. This leads to

$$\begin{aligned}
&\frac{1}{2}|w|_H^2 - \frac{1}{2}|v|_H^2 + \frac{1}{2}|w-v|_H^2 + h\nu|w|_V^2 \\
&\quad = h\int_\Omega \left((w\nabla)w + (z^j\nabla)w + (w\nabla)z^j\right) w\,dx - \rho\int_\Omega BvBw\,dx \\
&\quad\quad -\rho h\int_\Omega B\left((v\nabla)v + (z^j\nabla)v + (v\nabla)z^j\right) Bw\,dx \\
&\quad = (i)' + (ii)' + (iii)' + (iv) + (v)' + (vi) + (vii),
\end{aligned} \tag{4.10}$$

where (iv), (vi), and (vii) are defined in (4.6). There holds $(i)' = (ii)' = 0$, and $(iii)'$ can be estimated as

$$(iii)' \le \frac{\nu h}{2}|w|_V^2 + \frac{h}{\nu}|z|_V^2|w|_H^2.$$

Utilizing the estimate $|BBw|_V \le |w|_V$, one obtains

$$(v)' \le \frac{\nu h}{3}|w|_V^2 + \frac{3\rho^2 h}{2\nu}|v|_H^2|v|_V^2.$$

The remaining addenda can be estimated as above. Now introduce

$$c_1(\rho,h,z) := \frac{1}{2} + \frac{\rho}{3}(1-\rho) - \frac{2h}{\nu}|z|_{L^\infty(V)}^2 - \frac{h}{2}$$

and

$$c_2^j(\rho,h,z) := \frac{1}{2} + \frac{3\rho^2 h}{2\nu}|v|_V^2 + \underbrace{\left[\frac{3\rho h^2|z|_{1,\infty}^2}{1-\rho} + \frac{3\rho h|z|_\infty^2}{4(1-\rho)\nu}\right]}_{=:\tilde{c}_2(h,\rho,z)}.$$

With this notation and the estimates above, one concludes from (4.10) that

$$c_1(\rho,h,z)|w|_H^2 - c_2^j(\rho,h,z)|v|_H^2 + \frac{\nu h}{2}\left(1 - \frac{2}{3}\rho + \frac{2}{3}\rho^2\right)|w|_V^2 \le 0. \tag{4.11}$$

Now define

$$\hat{c}_2(\rho,h,z) := \frac{1}{2} + \frac{3\rho^2}{\nu^2(1-\frac{2}{3}\rho+\frac{2}{3}\rho^2)}|w^0|_H^2 + \tilde{c}_2(\rho,h,z)$$

and proceed as follows.

1. Choose $\rho^* \in (0,1)$ such that

$$\frac{3\rho}{\nu^2(1-\frac{2}{3}\rho+\frac{2}{3}\rho^2)}|w^0|_H^2 \le \frac{1}{6}(1-\rho) \quad \forall\,\rho\in(0,\rho^*]. \tag{4.12}$$

2. Fix $\rho \in (0,\rho^*]$ and choose $h^* = h^*(\rho) > 0$ such that
   (a) $\tilde{c}_2(h,\rho,z) \le \frac{1}{4}$,
   (b) $\frac{3\rho^2 h}{2\nu}|w^0|_V^2 \le \frac{1}{4}$,
   (c) $\frac{h}{2} + \frac{2h}{\nu}|z|_{L^\infty(V)}^2 + \tilde{c}_2(h,\rho,z) < \frac{\rho}{6}(1-\rho)$, and
   (d) $\frac{c_2^0(\rho,h,z)}{c_1(\rho,h,z)} \le \kappa_1 < 1$

hold for all h in the interval $(0,h^*]$. Let $h \in (0,h^*]$.

3. Now conclude from (4.12) and $\rho \in (0,1)$ that

$$\frac{3\rho^2}{\nu^2(1-\frac{2}{3}\rho+\frac{2}{3}\rho^2)}|w^0|_H^2 \le \frac{\rho}{6}(1-\rho) < \frac{1}{4},$$

which together with (c) implies that

$$\hat{c}_2(\rho,h,z) < 1 \quad \text{and} \quad \frac{\hat{c}_2(\rho,h,z)}{c_1(\rho,h,z)} \le \kappa_2 < 1.$$

Furthermore, (a) and (b) give $c_2^0(\rho,h,z) \le 1$, so that with (4.11)

$$|w^1|_V^2 \le \frac{2c_2^0(\rho,h,z)}{\nu h(1-\frac{2}{3}\rho+\frac{2}{3}\rho^2)}|w^0|_H^2 \le \frac{2}{\nu h(1-\frac{2}{3}\rho+\frac{2}{3}\rho^2)}\kappa^0|w^0|^2.$$

4. Now assume that for $j \in \mathbb{N}$
   (a) $|w^j|_H^2 \le \kappa^j|w^0|_H^2$ and
   (b) $|w^j|_V^2 \le \frac{2}{\nu h(1-\frac{2}{3}\rho+\frac{2}{3}\rho^2)}\kappa^{j-1}|w^0|_H^2$

hold true, where $\kappa := \max(\kappa_1,\kappa_2) < 1$.

5. Then conclude from (b) that

$$\begin{aligned} c_2^j(\rho,h,z) &= \frac{1}{2} + \frac{3h\rho^2}{2\nu}|w^j|_V^2 + \tilde{c}_2(\rho,h,z) \\ &\le \frac{1}{2} + \frac{3h\rho^2}{2\nu}\frac{2}{\nu h(1-\frac{2}{3}\rho+\frac{2}{3}\rho^2)}\kappa^{j-1}|w^0|_H^2 + \tilde{c}_2(\rho,h,z) \\ &\le \hat{c}_2(\rho,h,z). \end{aligned}$$

Thus, utilizing (4.11) one more time gives

$$|w^{j+1}|_H^2 \le \frac{c_2^j(\rho,h,z)}{c_1(\rho,h,z)}|w^j|_H^2 \le \frac{\hat{c}_2(\rho,h,z)}{c_1(\rho,h,z)}|w^j|_H^2 \le \kappa^{j+1}|w^0|_H^2$$

and

$$|w^{j+1}|_V^2 \le \frac{2}{\nu h(1-\frac{2}{3}\rho+\frac{2}{3}\rho^2)}c_2^j(\rho,h,z)|w^j|_H^2 \le \frac{2}{\nu h(1-\frac{2}{3}\rho+\frac{2}{3}\rho^2)}\kappa^j|w^0|_H^2,$$

which completes the proof of Theorem 4.7. ∎

Remark 4.8.

1. In the discrete scheme (4.9) all local quantities are discretized implicitly, and nonlocal quantities explicitly.

2. The proof of Theorem 4.7 is constructive. Estimates of ρ^* and $h^*(\rho)$ therefore can be deduced from (4.12) and 2(a)–(d) of the proof of the theorem, respectively.

3. Note further that the conditions on ρ in (4.1), (4.4) are satisfied for the stepsize of (3.5), provided the parameter α is chosen sufficiently large.

5. Numerical validation. Here the results obtained in the previous sections are numerically validated. In order to value the performance of the feedback operators (3.11) and (3.12), the numerical example is taken from [16]. As is demonstrated below the instantaneous controller presented here steers the H-norm and the V-norm

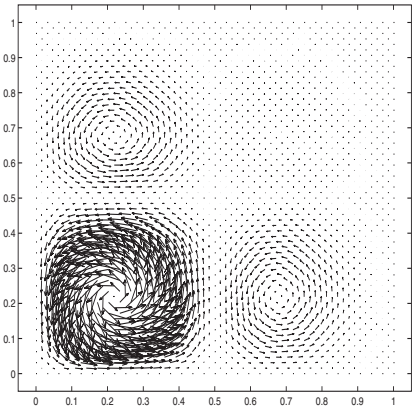


FIG. 1. *Desired flow at* $T = 2$.

of the difference $y - z$ to zero with exponential decay. This seems to be a more stable performance than that reported by Hou and Yan in [16] for their $(1, 1)$-rhc (i.e., control horizon length coincides with time stepsize). The instantaneous controls are compared to the optimal control, and it turns out that instantaneous controls give a much better reduction of the control gain but at significantly higher overall costs.

The control problem considered here is of tracking type and is given by (P) with cost functional

$$J(y,u) := \frac{1}{2}\int_Q |y-z|^2 dx\,dt + \frac{\alpha}{2}\int_Q |u|^2 dx\,dt$$

and control space $\mathcal{U} := L^2(\Omega)^2$, with $\mathcal{B}$ denoting the injection from U into V'. The initial value of the uncontrolled flow is chosen as

$$y(x,0) = e\begin{bmatrix}(\cos 2\pi x_1 - 1)\sin 2\pi x_2\\ -(\cos 2\pi x_2 - 1)\sin 2\pi x_1\end{bmatrix}$$

with e denoting the Euler number, and the desired state is time dependent and given by

$$z(t,x) = \begin{bmatrix}\varphi_{x_2}(t,x_1,x_2)\\ -\varphi_{x_1}(t,x_1,x_2)\end{bmatrix},$$

where φ is defined through the stream function

$$\varphi(t,x_1,x_2) = \theta(t,x_1)\theta(t,x_2)$$

with

$$\theta(t,y) = (1-y)^2(1-\cos 2k\pi t), \quad y \in [0,1].$$

For the results presented, $\alpha = 1.e-2$, $k = 1$, and the time interval is chosen as $[0, 2]$, i.e., $T = 2$. For the discretization in time, an equidistant grid with width $\delta t = 0.01$ is used, and for the spatial discretization, the Taylor–Hood finite element [15] is used on a grid containing 1024 triangles with 2113 velocity nodes and 545 pressure nodes. The number of unknowns in the discretized control problem therefore has the magnitude 1.65×10^6, including the primal, adjoint, and control variables.

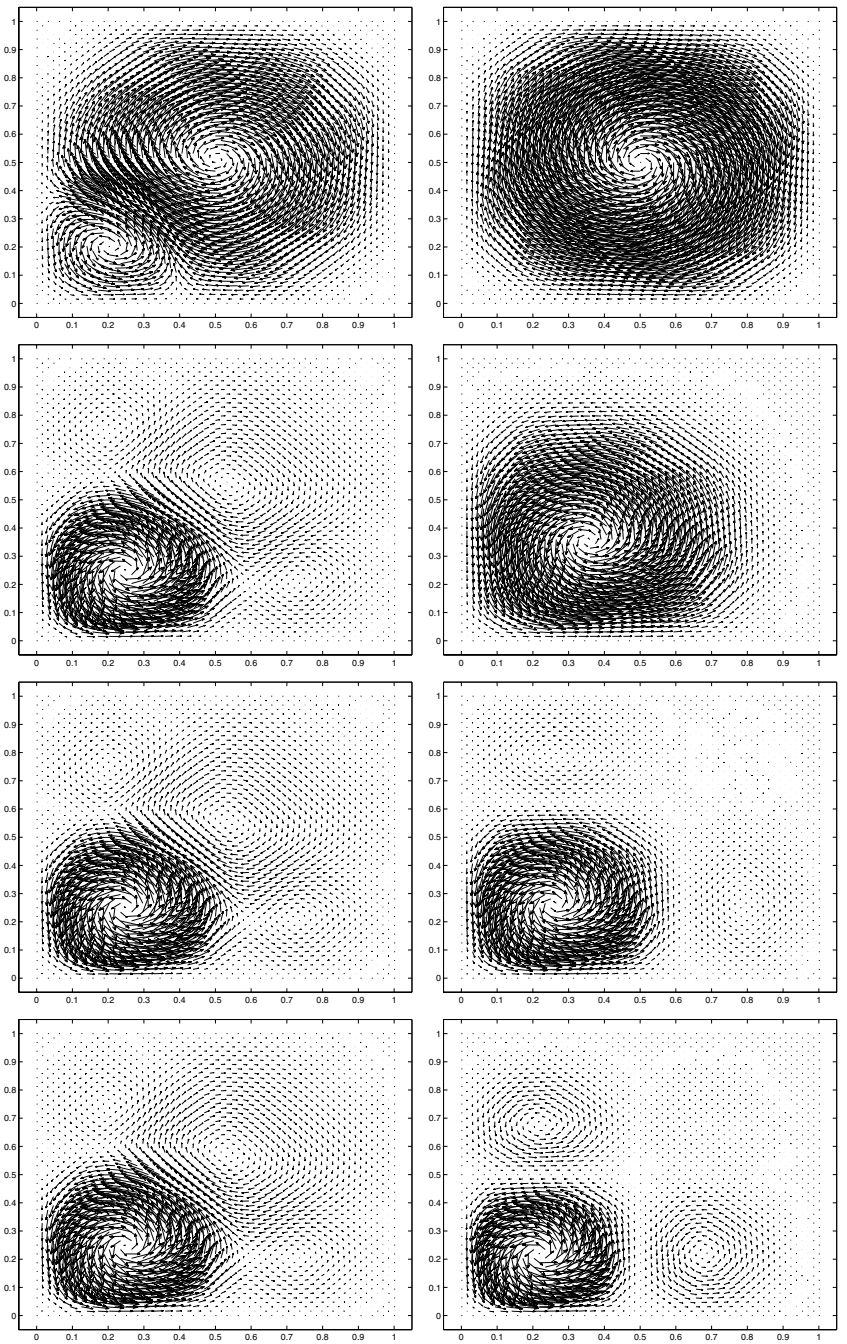


FIG. 2. *Optimally controlled (left) versus instantaneously controlled flows for* $t = 0.1, 1, 1.6, 2$.

In Figure 1 the desired flow at $T = 2$ is shown. It forms four cells with opposite flow directions near the cell borders.

In Figure 2 the evolution of the optimally controlled flow computed with Newton's method (see [11, 13] for computational details) and the instantaneously controlled flow

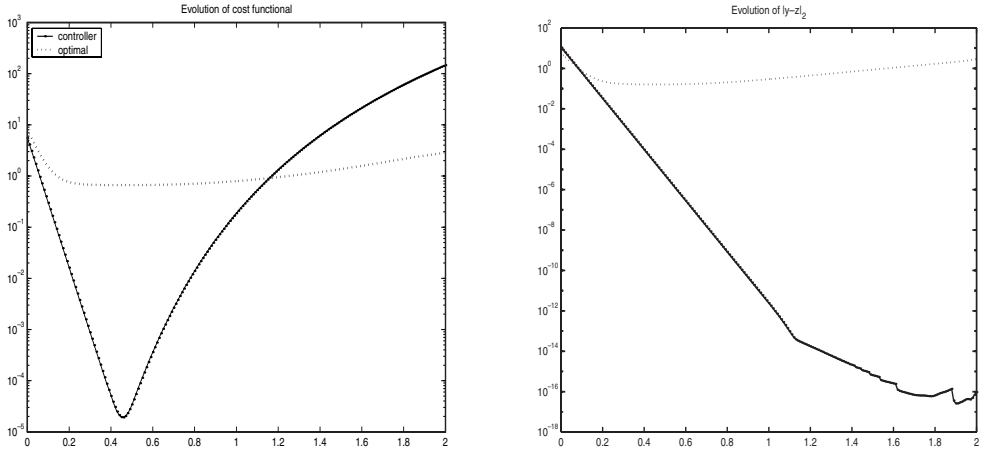


FIG. 3. *Evolution of cost (left) and control gain for* $h = 0.01$ *and* $\rho = 0.1$.

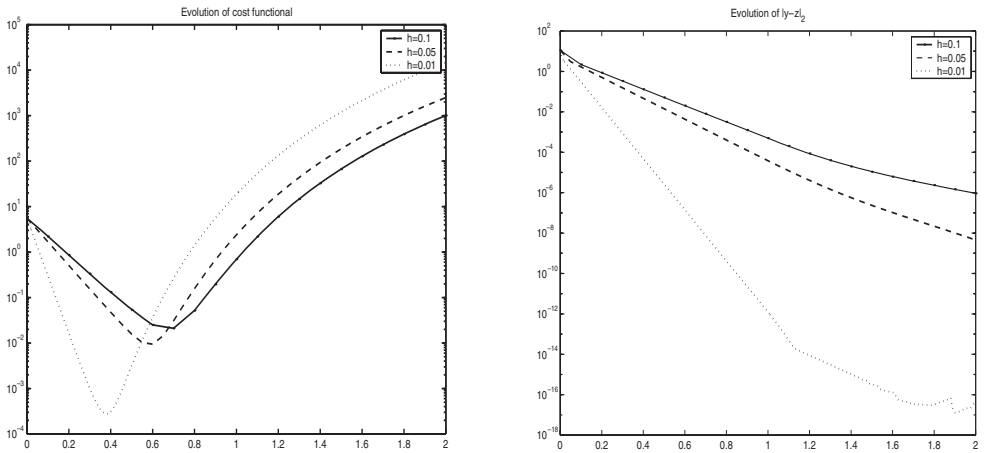


FIG. 4. *Evolution of cost (left) and control gain for* $\rho = 0.1$ *and* $h = 0.01, 0.05, 0.1$.

are illustrated at selected time instances. The costs are compared in Figure 3. For the instantaneous control strategy they become larger with increasing time. This is due to the increasing dynamics of the desired state. As is expected, the optimal control strategy equidistributes the costs over the time horizon, whereas instantaneous control at every time instance tries to match the desired state. This is also illustrated by the evolution in Figure 2.

For $\nu = 1/\mathrm{Re} = 1/10$ and $\gamma = 1.e-2$, the numerical computation of the optimal control takes about 45 minutes of cpu-time on a DEC-ALPHA^{TM} station 500. The instantaneous feedback controller takes about 2 minutes to compute a control function on the time horizon $[0, 2]$.

In Figure 4 the evolution of the L^2-cost for the instantaneous control law is shown for $\rho = 0.1$ and different values of h. In Figure 5, $h = 0.1$ is fixed, and the evolution of the control gain in the L^2- and H^1-norms for different values of ρ is shown. Exponential decay is observed, and thus the theoretical results of Theorems 4.1, 4.2, 4.4, and 4.5 are confirmed.

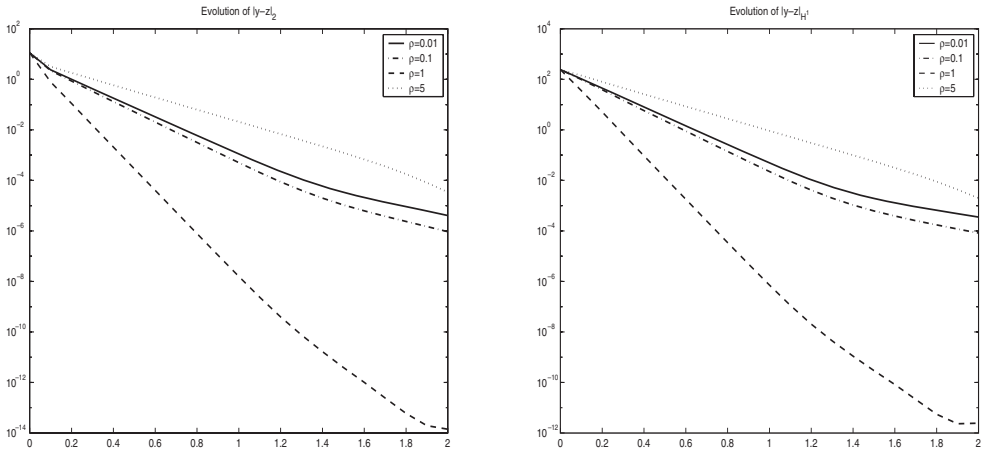


FIG. 5. *Evolution of control gain in L^2-norm (left) and H^1-norm for $h = 0.1$ and $\rho = 0.01, 0.1, 1, 5$.*

Acknowledgment. The author gratefully acknowledges the support of the Collaborative Research Center 609 (SFB 609) *Elektromagnetische Strömungsbeeinflussung in Metallurgie, Kristallzüchtung und Elektrochemie* funded by the German Research Foundation (DFG).

REFERENCES

[1] F. ABERGEL AND R. TEMAM, *On some control problems in fluid mechanics*, Theoret. Comput. Fluid Dyn., 1 (1990), pp. 303–325.

[2] M. BERGGREN, *Numerical solution of a flow-control problem: Vorticity reduction by dynamic boundary action*, SIAM J. Sci. Comput., 19 (1998), pp. 829–860.

[3] H. CHOI, *Suboptimal Control of Turbulent Flow Using Control Theory*, Lecture held at the Department of Mechanical Engineering, University of Tokyo, Tokyo, 1995.

[4] H. CHOI, M. HINZE, AND K. KUNISCH, *Instantaneous control of backward-facing-step flows*, Appl. Numer. Math., 31 (1999), pp. 133–158.

[5] H. CHOI, R. TEMAM, P. MOIN, AND J. KIM, *Feedback control for unsteady flow and its application to the stochastic Burgers equation*, J. Fluid Mech., 253 (1993), pp. 509–543.

[6] P. CONSTANTIN AND C. FOIAS, *Navier-Stokes Equations*, The University of Chicago Press, Chicago, 1988.

[7] C.E. GARCÍA, D.M. PRETT, AND M. MORARI, *Model predictive control: Theory and practice—a survey*, Automatica J. IFAC, 25 (1989), pp. 335–348.

[8] M.D. GUNZBURGER AND S. MANSERVISI, *Analysis and approximation for linear feedback control for tracking the velocity in Navier–Stokes flows*, Comput. Methods Appl. Mech. Engrg., 189 (2000), pp. 803–823.

[9] M.D. GUNZBURGER AND S. MANSERVISI, *Analysis and approximation of the velocity tracking problem for Navier–Stokes flows with distributed control*, SIAM J. Numer. Anal., 37 (2000), pp. 1481–1512.

[10] D.C. HILL, *Drag reduction strategies*, in CTR Annual Research Briefs, Center for Turbulence Research, Stanford University/NASA Ames Research Center, Stanford, CA, 1994, pp. 215–218.

[11] M. HINZE, *Optimal and Instantaneous Control of the Instationary Navier-Stokes Equations*, Habilitationsschrift, Fachbereich Mathematik, Technische Universität Berlin, 1999, revised version available from http://www.math.tu-dresden.de/∼hinze/publications.html.

[12] M. HINZE AND K. KUNISCH, *Control strategies for fluid flows—optimal versus suboptimal control*, in ENUMATH 97, H.G. Bock et al., eds., World Scientific, River Edge, NJ, 1998, pp. 351–358.

[13] M. HINZE AND K. KUNISCH, *Second order methods for optimal control of time-dependent fluid flow*, SIAM J. Control Optim., 40 (2001), pp. 925–946.

[14] M. Hinze and S. Volkwein, *Instantaneous control for the Burgers equation: Convergence analysis and numerical implementation*, Nonlinear Anal., 50 (2002), pp. 1–26.
[15] P. Hood and C. Taylor, *A numerical solution of the Navier-Stokes equations using the finite element technique*, Comput. & Fluids, 1 (1973), pp. 73–100.
[16] L.S. Hou and Y. Yan, *Dynamics and approximations of a velocity tracking problem for the Navier–Stokes flows with piecewise distributed controls*, SIAM J. Control Optim., 35 (1997), pp. 1847–1885.
[17] L.S. Hou and Y. Yan, *Dynamics for controlled Navier–Stokes systems with distributed controls*, SIAM J. Control Optim., 35 (1997), pp. 654–677.
[18] A. Kauffmann and K. Kunisch, *Optimal control of the solid fuel ignition model*, ESAIM Proc. 8, Soc. Math. Appl. Indus., Paris, 2000, pp. 65–76.
[19] C. Lee, J. Kim, and H. Choi, *Suboptimal control of turbulent channel flow for drag reduction*, J. Fluid Mech., 358 (1998), pp. 245–258.
[20] C. Min and H. Choi, *Suboptimal feedback control of vortex shedding at low Reynolds numbers*, J. Fluid Mech., 401 (1999), pp. 123–156.
[21] V. Nevistić and J.A. Primbs, *Finite receding horizon control: A general framework for stability and performance analysis*, preprint, Automatic Control Laboratory, ETH Zürich, Zürich, Switzerland, 1997.
[22] J.B. Rawlings and K.R. Muske, *The stability of constrained receding horizon control*, IEEE Trans. Automat. Control, 38 (1993), pp. 1512–1516.
[23] R. Temam, *Navier-Stokes Equations*, North-Holland, Amsterdam, 1979.
[24] R. Temam, *Infinite-Dimensional Dynamical Systems in Mechanics and Physics*, 2nd ed., Appl. Math. Sci. 68, Springer-Verlag, New York, 1997.
[25] F. Tröltzsch and A. Unger, *Fast solution of optimal control problems in selective cooling of steel*, ZAMM Z. Angew. Math. Mech., 81 (2001), pp. 447–456.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 2, pp. 584–617

MULTIRATE STABILIZATION OF LINEAR MULTIPLE SENSOR SYSTEMS VIA LIMITED CAPACITY COMMUNICATION CHANNELS*

ALEXEY S. MATVEEV† AND ANDREY V. SAVKIN‡

Abstract. The paper addresses a feedback stabilization problem involving bit-rate communication capacity constraints. A discrete-time partially observed linear system is studied. Unlike classic theory, the signals from multiple sensors are transmitted to the controller over separate finite capacity communication channels. The sensors do not have constant access to the channels, and the channels are not perfect: the messages incur time-varying transmission delays and may be corrupted or lost. However, we suppose that the time-average number of bits per sample period that can be successfully transmitted over the channel during a time interval converges to a certain limit as the length of the interval becomes large. Necessary and sufficient conditions for stabilizability are established. They give the tightest lower bounds on the channel capacities for which stabilization is possible. An algorithm for stabilization is also presented.

Key words. networked control systems, finite data rate, communication constraints, stabilizability

AMS subject classifications. 93D15, 93D20, 93B52

DOI. 10.1137/S0363012902419965

1. Introduction.

The standard assumption in classical control theory is that data transmission required by the algorithm can be performed with infinite precision. However, due to the growth in communication technology, it is becoming more common to employ digital finite capacity networks for the exchange of information between plant components. Examples concern complex dynamical processes like advanced aircraft, spacecraft, automotive, industrial and defense systems, arrays of microactuators, and power control in mobile communication. Bandwidth communication constraints are often major obstacles to control system design by means of classical theory. For instance, as was shown in [24], the design of control systems for platoons of underwater vehicles strongly highlights the need for control strategies that address explicitly the bandwidth limitation on communication between vehicles, which is severely restricted underwater. All these emerging applications motivate development of a new chapter of control theory that deals with networked systems and combines the control and communication issues, taking into account all the limitations on communication between sensors, controllers, and actuators.

Recently there was a good deal of research activity in this field. Starting with [4] and continuing with [1, 2, 3, 7, 8, 13, 15, 16, 17, 19, 22, 25, 27], various control schemes were proposed for stabilization via a limited capacity channel. The smallest data rate above in which stabilization is possible was derived in [19, 20, 25, 27] in various settings. However, up to now, only networks with the simplest topology, which involve only one "sensor-controller" and "controller-actuator" channel, and perfect,

*Received by the editors December 17, 2002; accepted for publication (in revised form) October 20, 2004; published electronically September 12, 2005. This work was supported by the Australian Research Council.

http://www.siam.org/journals/sicon/44-2/41996.html

†Department of Mathematics and Mechanics, Saint Petersburg University, Universitetskii pr. 28, Petrodvoretz, St. Petersburg, 198504, Russia (almat@am1540.spb.edu).

‡Corresponding author. School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney, 2052, Australia (a.savkin@unsw.edu.au).

i.e., noiseless and undelayed, communication were mainly considered. However, many modern control systems are implemented in a distributed fashion, which results in a less trivial topology with multiple sensors, controllers, and actuators communicating over a serial network. Various time delays in transmission are characteristic for such networks, and nonconstant access to the channel is typical [13].

In this paper, we also study a stabilization problem via quantized state feedback for a linear time-invariant partially observed system. However, we consider a multichannel communication between the multiple sensors and the controller, where each sensor is served by its own finite capacity channel and there is no information exchange between the sensors. There is also no feedback communication between the sensors and the controller, and the sensors have no access to the control. The objective is to establish first, the tightest lower bounds on the capacities of the channels for which the stabilization is possible and second, the rate of exponential stability that is achievable for given capacities obeying those bounds. To this end, we obtain necessary and sufficient conditions for stabilizability. In the particular case where the channels are perfect and the system is detectable via each sensor, these conditions formally come to those from [19, 20]. Thus we show that in this case, multiple noncommunicating sensors and channels with separate capacity constraints may be treated as a single sensor and a single channel with a united constraint, respectively. However, employing multiple sensors usually means that there are problems with detectability by means of a single sensor, and then the model with nondetecting sensors is often a good option.

Another crucial point is that we do not assume the channels to be perfect. Sensor signals may incur independent and time-varying delays and arrive at the controller out of order. There may be periods when the sensor is denied access to the channel. Transmitted data may be corrupted or even lost. However, we assume that the communication noise is compensated and so ultimately reveals itself only in the form of decay of the channel information capacity. For example, employing error correcting block codes [10, Chap. 12] means that the channel must be partly engaged in transmission of redundant check symbols, which decreases the average amount of carried primal messages from the sensors. We suppose that error correction is the function of the channel, i.e., the corresponding coder and decoder are given and considered as parts of what is termed "channel." The key assumption is that the time-average number of bits per sample period that can be successfully transmitted across the channel during a time interval converges to what we call the transmission capacity as the length of the interval becomes large. The stabilizability region is given in terms of these capacities. Note that bounded communication delays do not influence them and thus the region, though they affect the design of the stabilizing controller.

Stabilization with limited information feedback was studied in the presence of transmission delays in [27]. Unlike the current paper, the transmission time required to transfer one bit was assumed constant, the continuous time linear plant and the network with the simplest topology and constant access channels were considered, and conditions for nonasymptotic stability but a weaker property called containability were established.

Stabilization of multiple sensor systems via perfect (instantaneous and noiseless) channels was discussed in [25] under the assumption that the control is known at the sensor sites. In fact, a particular recurrent (i.e., changed according to a fixed rule as time progresses) stabilization scheme was examined. It is based on scaled quantization by means of special quantizers. Ideas coherent with Slepian–Wolf encoding of data

from correlated sources were also employed. It was established when the system can be stabilized via such a scheme. The answer is given in terms of the controller parameters called the rate vectors. They are tuples of naturals each being the number of the quantizer levels with respect to a certain state coordinate. These conditions can be reformulated in terms of the capacities of the channels. Then the criterion for stabilizability via the above scheme reduces to solvability of some linear system of inequalities in integers. The arguments from [25] also presuppose that the system is reducible to the real-diagonal form so that any "mode" is in a simple relation with any sensor. The latter means that the mode either does not affect the sensor outputs or can be completely determined from these outputs.[1]

In this paper, we show that there exist stabilizible systems, which, however, cannot be stabilized by means of the aforementioned control scheme. Moreover, insufficient are all schemes that merely have some features in common with that one (see section 9).[2] This stresses that stabilizability should be tested in the class of all controllers with a given information pattern. Such an analysis is offered in this paper. This gives rise to an additional trouble in the form of a gap between necessary and sufficient conditions, where all and specific controllers proposed in this paper are concerned, respectively. To fill this gap, not only time averaging but also convex duality techniques are employed. Contrary to [25], the final criterion is given in terms of only the plant and channels parameters. This criterion strictly improves that from [25].

We consider the case where the system is not necessarily reducible to a diagonal form. It is shown that nontrivial Jordan blocks may make it impossible to disintegrate the system into state-independent subsystems each in simple relations with the sensors. To treat this case, we employ sequential stabilization based on triangular decomposition into state-dependant subsystems. They are stabilized successively. In doing so, their interinfluence is interpreted as an exogenous disturbance, and ideas related to those from [6, 22] as well as [2, 19, 20, 25] are employed. Points of novelty concern an account for transmission delays and disturbances decaying at a known rate. Unlike the previous works, no other characteristics of the disturbance (e.g., an upper bound) are assumed to be known. Apart from state-dependency, the subsystems are also dependant via control. Since it is common, the control aimed to stabilize a particular subsystem may disturb the others. We offer a method to cope with this problem. Note that this issue was not addressed in [25]. Finally, contrary to [25], we consider the case when the sensors have no access to the control: there is no feedback communication between them and the decoder.

The paper is organized as follows. We first illustrate the problem statement and the main result by an example in section 2. The general problem statement is given in section 3. Section 4 offers basic definitions, assumptions, and notations. The main result is presented in section 5. Its proof is given in sections 7 and 8, where necessary and sufficient conditions for stabilizability are, respectively, justified. In section 6, the main result is applied to the example from section 2. The concluding section 9 comments on an important assumption imposed in this paper.

2. Example. We first illustrate the class of problems to be studied by an example.

[1] The arguments from [25] do not really require the diagonal form. However, the assumption that the system can be decomposed into independent subsystems each in simple relations with sensors seems to be crucial.

[2] For example, stabilization results from a control scheme, which is not recurrent but is cyclic (i.e., periodically varies in time).

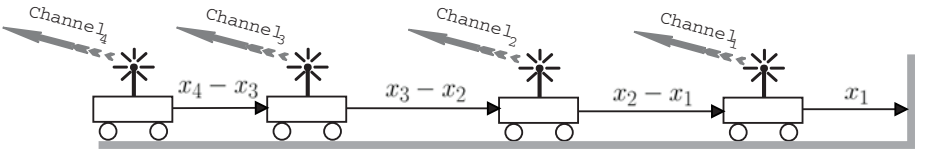


FIG. 2.1. *Platoon of autonomous vehicles.*

We consider a platoon composed of k vehicles moving along a line and enumerated from right to left. The dynamics of the platoon are uncoupled, and the vehicles are described by the equations

$$\dot{x}_i = v_i, \quad \dot{v}_i = u_i, \quad i = 1, \dots, k, \tag{2.1}$$

where x_i is the position of the ith vehicle, v_i is its velocity, and u_i is the control input. Each vehicle is equipped with a sensor giving the distance $y_i = x_i - x_{i-1}$ from it to the preceding one for $i \geq 2$ and the position $y_1 = x_1$ for $i = 1$. It is also endowed with a digital communication channel over which the measurement y_i is sent to the central controller. To this end, the sensor signals are sampled with a period $\Delta > 0$. This channel is delayed, nonstationary, and lossy and transmits on average $\mathfrak{c}_i > 0$ bits per sample period. Employing the data that arrives over all channels, the central controller produces the control inputs for all vehicles at the sample times. The objective is to stabilize the platoon motion about a given constant-velocity trajectory: $v_i = v_i^0, x_i(t) = x_i^0 + v_i^0 t \ \forall i$. This situation is illustrated in Figure 2.1 for $k = 4$. In this context, we pose the following questions:

1. *What is the minimum rate of the information transmission for which stabilization is possible?*
2. *Which rate of stability can be achieved for channels with given capacities* $\mathfrak{c}_j$ *and sample period* Δ*?*

More precisely, we are interested in the rate μ at which the platoon is able to approach the following desired trajectory:

$$|v_i(t) - v_i^0| \leq K_{v,i}\mu^{t/\Delta}, \qquad |x_i(t) - x_i^0 - v_i^0 t| \leq K_{x,i}\mu^{t/\Delta}.$$

As will be shown in section 6, stabilization of the platoon is possible for any capacities $\mathfrak{c}_i > 0$ and at any rate $\mu > \mu^0 := \sqrt{2}^{\,-\mathfrak{c}_{\min}}$, where $\mathfrak{c}_{\min} := \min_{i=1,\dots,k} \mathfrak{c}_i$. At the same time, no rate $\mu < \mu^0$ is achievable.

Now consider another situation where the sensor system accommodated by each vehicle is able to give the distances to $l < k$ vehicles to the right, as well as to l vehicles to the left. Then the platoon motion remains stabilizable for any capacities

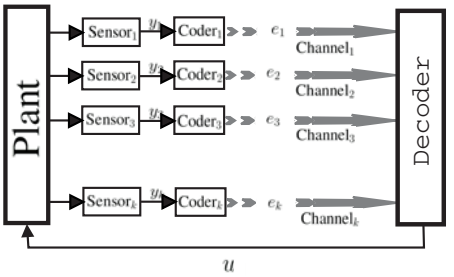


FIG. 3.1. *Feedback control by means of communication channels.*

$\mathfrak{c}_i$. However, the above threshold stability rate μ^0 is changed: $\mu^0 = \sqrt{2}^{\,-\mathfrak{c}_{k,l}}$. Here

$$\mathfrak{c}_{k,l} := \min\left\{\mathfrak{c}_{k,l}^{(1)}, \mathfrak{c}_{k,l}^{(2)}, \mathfrak{c}_{k,l}^{(3)}\right\} \quad \text{if} \quad 2l \geq k, \quad \text{and} \quad \mathfrak{c}_{k,l} := \min\left\{\mathfrak{c}_{k,l}^{(4)}, \mathfrak{c}_{k,l}^{(5)}, \mathfrak{c}_{k,l}^{(6)}\right\} \quad \text{if} \quad 2l < k, \quad \text{where} \tag{2.2}$$

$$\mathfrak{c}_{k,l}^{(1)} := \min_{i=1,\dots,k-l} \frac{1}{i}\sum_{j=1}^{i} \mathfrak{c}_j, \quad \mathfrak{c}_{k,l}^{(2)} := \frac{1}{k}\sum_{j=1}^{k} \mathfrak{c}_j, \quad \mathfrak{c}_{k,l}^{(3)} := \min_{i=l+1,\dots,k-1} \frac{1}{k-i}\sum_{j=i+1}^{k} \mathfrak{c}_j,$$

$$\mathfrak{c}_{k,l}^{(4)} := \min_{i=1,\dots,l+1} \frac{1}{i}\sum_{j=1}^{i} \mathfrak{c}_j, \quad \mathfrak{c}_{k,l}^{(5)} := \min_{i=l+2,\dots,k-l} \mathfrak{c}_i, \quad \mathfrak{c}_{k,l}^{(6)} := \min_{i=k-l,\dots,k-1} \frac{1}{k-i}\sum_{j=i+1}^{k} \mathfrak{c}_j.$$

The objective of this paper is to develop a general theory that enables one to obtain results like these.

3. General problem statement. We consider linear discrete-time multiple sensor systems of the form

$$\boldsymbol{x}(t+1) = A\boldsymbol{x}(t) + B\boldsymbol{u}(t); \qquad \boldsymbol{x}(0) = \boldsymbol{x}_0; \tag{3.1}$$

$$\boldsymbol{y}_j(t) = C_j\boldsymbol{x}(t), \qquad j = 1, \dots, k. \tag{3.2}$$

Here $\boldsymbol{x} \in \mathbb{R}^n$ is the state, $\boldsymbol{u} \in \mathbb{R}^{n_u}$ is the control, and $\boldsymbol{y}_j \in \mathbb{R}^{n_{y,j}}$ is the output of the jth sensor. The system is unstable: there is an eigenvalue λ of the matrix A with $|\lambda| \geq 1$. The objective is to stabilize the plant: $\boldsymbol{x}(t) \xrightarrow{t\to\infty} 0$.

We consider a remote control setup. Each sensor is served by its own communication channel capable of transmitting signals from a finite *alphabet* $\mathfrak{E}_j$. Over this channel, the jth *coder* sends a message $e_j(t) \in \mathfrak{E}_j$ based on the prior measurements

$$e_j(t) = \mathcal{E}_j[t, \boldsymbol{y}_j(0), \dots, \boldsymbol{y}_j(t)]. \tag{3.3}$$

On the basis of the data $\overline{\boldsymbol{e}}(t)$ received over all channels up to the current time t, the *decoder* selects the control

$$\boldsymbol{u}(t) = \mathcal{U}\left[t, \overline{\boldsymbol{e}}(t)\right]. \tag{3.4}$$

In the situation illustrated in Figure 3.1, the *networked controller* is constituted by the decoder and the set of coders

$$\mathcal{NC} := \left[\mathcal{E}_1(\cdot), \dots, \mathcal{E}_k(\cdot), \mathcal{U}(\cdot)\right]. \tag{3.5}$$

Transmitted messages incur delays and may be lost: the message sent at time t arrives at the decoder at the discrete time $t+\tau_j(t) \geq t$, where $\tau_j(t) := \infty$ if it is lost. So the data available to the decoder at time t is

$$\overline{\boldsymbol{e}}(t) := \big[\overline{\boldsymbol{e}}_1(t), \dots, \overline{\boldsymbol{e}}_k(t)\big], \quad \text{where} \quad \overline{\boldsymbol{e}}_j(t) := \big[e_j(\theta_1), \dots, e_j(\theta_{\sigma_j})\big] \tag{3.6}$$

is the data that arrived via the jth channel by the time t: $\{\theta_1 < \theta_2 < \dots < \theta_{\sigma_j}\} = \{\theta = 0, 1, \dots : \theta + \tau_j(\theta) \leq t\}$.

The main question to be considered is what is the minimum rate of the information exchange in the system for which stabilization is possible? In other words, we look for necessary and sufficient conditions for stabilizability expressed in terms of the channels *transmission capacities* $\mathfrak{c}_1, \dots, \mathfrak{c}_k$, along with the plant-sensors parameters A, B, C_j. Roughly speaking, such a capacity is the average number of bits transmitted over the channel during the sample period, despite the losses and delays.

4. Definitions, notations, and assumptions. It should be remarked that there may be a difference between the number of bits that happen to reach the decoder thanks to occasional favorable circumstances and the number of bits that can be successfully transmitted under any circumstances. In fact, these numbers give rise to two concepts of capacity. The first and second of these are concerned with the necessary and sufficient conditions for stabilizability, respectively. To simplify matters, we postulate that these capacities coincide: the discrepancy between those numbers is considerably less than the time of a long experiment.

We also consider the case where there is an uncertainty about the channel. Specifically, its *regime of operation* given by the distribution of integer transmission delays $\tau_j(t)$ over time t may not be known in advance. However, we suppose that it satisfies certain assumptions, and the designer of the controller is aware of some lower and upper bounds for the number of bits transmitted across the channel during a time interval of a given duration.

To understand the details, we start with the following.

DEFINITION 4.1. *We say that a message is* transmitted within a time interval $[t_0 : t_1]$ *if it departs and arrives at times* t *and* $t+\tau_j(t)$ *from this interval:* $t, t+\tau_j(t) \in [t_0 : t_1]$.

The *length* or *duration* of a discrete time interval $[t_0 : t_1]$ is defined to be $t_1 - t_0$.

ASSUMPTION 4.1. *For each channel, there exist two integer functions* $b_j^-(r)$ *and* $b_j^+(r)$ *of time* r *such that*

1. *no more than* $b_j^+(r)$ *bits are brought by the transmissions that occur within any time interval of length* r*;*
2. *given a time interval of duration* r*, there exists a way to transmit without losses and errors no less than* $b_j^-(r)$ *bits of information within this interval;*
3. *as the length* r *of the time interval increases, the averaged numbers* $b_j^+(r)/r$ *and* $b_j^-(r)/r$ *converge to a common limit* $\mathfrak{c}_j$ *called the* transmission capacity *of the channel*

$$\mathfrak{c}_j = \lim_{r\to\infty} \frac{b_j^-(r)}{r} = \lim_{r\to\infty} \frac{b_j^+(r)}{r}. \tag{4.1}$$

Claim 1 means that $p_j(t_0, t_1) \cdot \log_2 N_j \leq b_j^+(t_1 - t_0)$. Here $p_j(t_0, t_1)$ and N_j denote the numbers of messages transmitted in fact during the time interval $[t_0 : t_1]$ and the size of the channel alphabet, respectively. Note that each message from this alphabet carries $\log_2 N_j$ bits. In claim 2, the "way" is constituted by encoding and

decoding rules. The former translates b-bit words $\boldsymbol{\beta} = (\beta_1, \dots, \beta_b), \beta_i = 0, 1$, into sequences of messages $e \in \mathfrak{E}_j$ sent consecutively during the interval at hand. The decoder transforms the sequence of messages that arrived within this interval into a b-bit word $\boldsymbol{\beta}'$. The overall transmission must be errorless: $\boldsymbol{\beta}' = \boldsymbol{\beta}$.

We suppose that the designer of the controller is aware of these rules, along with the functions $b_j^-(r), b_j^+(r)$. A regime of the channel operation $\{\tau_j(\cdot)\}$ compatible with these data is said to be *possible*.

Now we offer two examples of channels satisfying Assumption 4.1. Other such examples can be found in [18].

Noiseless instantaneous channel. The channel is constantly accessible; any transmission is successful and instantaneous $\tau_j(t) = 0$. Then $b_j^-(r) = \lfloor (r+1) \cdot \log_2 N_j \rfloor, b_j^+(r) = \lceil (r+1) \cdot \log_2 N_j \rceil$, and $\mathfrak{c}_j = \log_2 N_j$.

Noiseless delayed channel. Now suppose that in the previous example, the transmission time is nonzero, is not known in advance, but is bounded $0 \le \tau_j(t) \le \tau_j^+$ by a known constant. The messages do not overtake each other. Then $b_j^-(r) = \lfloor (r - \tau_j^+ + 1) \cdot \log_2 N_j \rfloor$ if $r \ge \tau_j^+$ and $b_j^-(r) = 0$, otherwise $b_j^+(r) = \lceil (r+1) \cdot \log_2 N_j \rceil$ and $\mathfrak{c}_j = \log_2 N_j$.

Now we introduce two concepts of stabilizability: weak and strong ones.

DEFINITION 4.2. *The system is said to be* stabilizable *if some networked controller makes all trajectories converging to zero* $\boldsymbol{x}(t) \xrightarrow{t \to \infty} 0$ *for at least one possible regime of channels operation.*

DEFINITION 4.3. *We say that a networked controller* uniformly exponentially stabilizes *the system* at the rate $\mu \in (0, 1)$ *if the corresponding trajectories obey the inequalities*

$$|\boldsymbol{x}(t)| \le K_x \mu^t, \quad |\boldsymbol{u}(t)| \le K_u \mu^t \qquad \forall t = 0, 1, 2, \dots \tag{4.2}$$

whenever $|\boldsymbol{x}_0| \le K_0$. *This must be true irrespective of the regime of channels operation, provided it is possible. The constants* K_x *and* K_u *may depend on* K_0 *but must not depend on time* t *and this regime.*

DEFINITION 4.4. *The system is said to be* uniformly exponentially stabilizable at the rate $\mu \in (0, 1)$ *if there exists a networked controller that uniformly exponentially stabilizes the system at this rate.*

Note that this controller may depend on μ, along with $b_j^+(\cdot), b_j^-(\cdot)$ and A, B, C_j.

DEFINITION 4.5. *The system uniformly exponentially stabilizable at some rate* $\mu \in (0, 1)$ *is said to be* uniformly exponentially stabilizable. *The infimum value of* μ *is called the* rate of exponential stabilizability.

4.1. General notations. $\mathbf{mes}$—the Lebesgue measure; B_x^d—the open ball centered at x with the radius d; $\det \mathcal{A}$—the determinant of the operator $\mathcal{A}$ in a finite dimensional linear space; $\sigma(\mathcal{A})$—the spectrum of $\mathcal{A}$; $\sigma^+(\mathcal{A}) := \{\lambda \in \sigma(\mathcal{A}) : |\lambda| \ge 1\}$; $\sigma^-(\mathcal{A}) := \sigma(\mathcal{A}) \setminus \sigma^+(\mathcal{A})$; $M_\sigma = M_\sigma(\mathcal{A})$—the invariant subspace of $\mathcal{A}$ related to the spectrum set $\sigma \subset \sigma(\mathcal{A})$; $M_{st}(\mathcal{A}) := M_{\sigma^-(\mathcal{A})}$; $M_{unst}(\mathcal{A}) := M_{\sigma^+(\mathcal{A})}$; $\mathcal{A}|_L$—the operator $\mathcal{A}$ acting in its invariant subspace L; $\lceil z \rceil := \min\{i = 0, \pm 1, \pm 2, \dots : i \ge z\}, \lfloor z \rfloor := \max\{i = 0, \pm 1, \pm 2, \dots : i \le z\}$.

4.2. Assumptions about the system (3.1), (3.2).

ASSUMPTION 4.2. *The pair* (A, B) *is stabilizable.*

The next assumption concerns the subspaces that are not observed and detected,

respectively, by a given sensor:

(4.3)
$$L_j^{-obs} := \{\boldsymbol{x} \in \mathbb{R}^n : C_j A^\nu \boldsymbol{x} = 0 \quad \forall \nu = 0, \dots, n-1\}, \qquad L_j^- := M_{unst}(A) \cap L_j^{-obs}.$$

To state this assumption, we introduce the following.

DEFINITION 4.6. *A spectral set $\sigma \subset \sigma(A)$ is said to be* elementary *if it consists of either one real eigenvalue or a couple of conjugate complex ones.*

ASSUMPTION 4.3. *Let $\sigma \subset \sigma^+(A)$ be any elementary set that gives rise to more than one real Jordan block. Consider the subspace $L_j^- \cap M_\sigma$ of states $\boldsymbol{x} \in M_\sigma$ ignored by the jth sensor. Then the variety of all these subspaces $j = 1, \dots, k$ has the atomic structure: the space M_σ can be decomposed into a direct sum of "atom" subspaces M_σ^i, $i = 1, \dots, m_\sigma$, so that any $L_j^- \cap M_\sigma$ is the sum of several "atoms" (the sum over the empty set is $\{0\}$):*

$$L_j^-(\sigma) := L_j^- \cap M_\sigma = \oplus_{i \in I(j)} M_\sigma^i, \qquad \text{where} \quad I(j) \subset [1 : m_\sigma]. \tag{4.4}$$

If the set σ gives rise to only one real Jordan block, this assumption is necessarily true (see Lemma 8.2).

Assumption 4.3 is technical, imposed to meet the paper length limitation, and will be commented on in section 9. A typical example of the situation forbidden by this assumption is as follows:

$$\boldsymbol{x}(t+1) = \lambda \boldsymbol{x}(t) + \boldsymbol{u}(t) \in \mathbb{R}^2, \quad y_1(t) = x_1(t), \quad y_2(t) = x_2(t), \quad y_3(t) = x_1(t) - x_2(t), \quad \lambda > 1, \tag{4.5}$$

where $\boldsymbol{x} = (x_1, x_2)$. Indeed, here the invariant subspaces nondetectable by the sensors are $L_1^- = \{\boldsymbol{x} : x_1 = 0\}, L_2^- = \{\boldsymbol{x} : x_2 = 0\}$, and $L_3^- = \{\boldsymbol{x} : x_1 = x_2\}$. It is easy to see that $\mathbb{R}^2$ cannot be decomposed into a direct sum of "atom" subspaces M^i so that its special partial sums give each space L_i^-, as is required by Assumption 4.3. Anyhow, this assumption holds if any eigenvalue $\lambda : |\lambda| \geq 1$ gives rise to only one real Jordan block, or each of the sensors detects or ignores any of the above subspaces M_σ completely: either $M_\sigma \cap L_j^- = \{0\}$ or $M_\sigma \subset L_j^-$.

5. The main result. For any group of sensors $J \subset [1 : k]$, we introduce the space of states

$$L(J) := \bigcap_{j \in J} L_j^- \tag{5.1}$$

nondetectable by this group. (We recall that L_j^- is given by (4.3).) For consistency, we assign $M_{unst}(A)$ to the empty group. Gathering $L = L(J)$ except for $L = \{0\}$ over all groups of sensors J gives rise to a set $\mathfrak{L} = \{L\}$. Its size may be less than the number of all such groups since different groups may produce a common space $L(J)$.

Now we are in a position to state the main result of the paper.

THEOREM 5.1. *Suppose that Assumptions* 4.1–4.3 *hold. Then the following two statements are equivalent:*

1. *The system* (3.1), (3.2) *is uniformly exponentially stabilizable (see Definition* 4.5*);*

2. *For every subspace* (5.1) $L \in \mathfrak{L}$ *constituted by all states nondetectable by a certain group of sensors,*

(5.2)
$$\log_2 |\det A|_L| < \sum_{j \notin J(L)} \mathfrak{c}_j, \ \textit{where} \quad J(L) := \{j = 1, \dots, k : C_j x = 0 \quad \forall x \in L\}.$$

Here the sum is over the sensors that do not completely ignore the subspace L *at hand, and* $\mathfrak{c}_j$ *is the transmission capacity* (4.1) *of the* j*th channel.*

If the equivalent claims 1 *and* 2 *are true, the rate* μ^0 *of exponential stabilizability of the system is given by*

$$\log_2 \mu^0 = \max_{L \in \mathfrak{L}} \frac{1}{\mathbf{dim} L} \left(\log_2 |\det A|_L| - \sum_{j \notin J(L)} \mathfrak{c}_j \right). \tag{5.3}$$

A stabilizing networked controller will be constructed in section 8. We note that it uses only a finite and fixed number of recent observations $\boldsymbol{y}_j$ and messages e_j in (3.3) and (3.4), respectively.

As was shown in [20], the quantity $\log_2 |\det A|_L|$ from (5.2) represents the unit time increment of the number of bits required to describe to a particular accuracy the state of the open-loop system (3.1) considered on the invariant subspace L. More precisely, if the initial state $\boldsymbol{x}_0$ is randomly distributed over the subspace L in accordance with a certain probability density and $\boldsymbol{u}(t) \equiv 0$, the differential entropy $H[\boldsymbol{x}(t)]$ (see, e.g., [5] for the definition) of the state $\boldsymbol{x}(t)$ evolves as follows: $H[\boldsymbol{x}(t+1)] = H[\boldsymbol{x}(t)] + \log_2 |\det A|_L|$. At the same time, the right-hand side of (5.2) can be interpreted as the joint capacity of all channels except for those carrying no information about the state $\boldsymbol{x} \in L$. (The latter serve the sensors that completely ignore such states $C_j \boldsymbol{x} = 0 \ \forall \boldsymbol{x} \in L$.) Thus the condition (5.2) means that the amount of information concerning the state $\boldsymbol{x} \in L$ that the decoder may receive over all channels for the unit time exceeds the unit time growth of the number of bits required to describe the state to a particular accuracy. It must be noted here that some of the bits counted on the right-hand side in (5.2) characterize the state $\boldsymbol{x} \in L$ only partly. They correspond to any sensor whose outputs are not sufficient to reconstruct the entire state $\boldsymbol{x} \in L$. Moreover all the sensors may be of such a kind. Nevertheless, when inequalities (5.2) are taken for all subspaces L (or in other words, groups of sensors), they constitute a sufficient and necessary criterion for stabilizability.

Remark 5.1. The conditions (5.2) imply that the system is detectable via the entire set of the sensors.

Indeed, otherwise (5.2) fails to be true for $L := \bigcap_{j=1}^{k} L_j^- \neq \{0\}$ since the sum over the empty set is defined to be 0.

Remark 5.2. If the system is detectable by each sensor $L_j^- = \{0\} \ \forall j$, the set $\mathfrak{L}$ contains only one space $M_{unst} := M_{unst}(A)$ and so the condition 2 reduces to only one inequality,

$$\log_2 |A|_{M_{unst}}| < \mathfrak{c} := \sum_{j=1}^{k} \mathfrak{c}_j.$$

The sum $\mathfrak{c}$ can be interpreted as the capacity of the channel composed of all the channels at hand. So the inequality is in a harmony with those from [11, 14, 19, 20, 25] concerning the case of one channel.

In general, the number of inequalities (5.2) does not exceed 2^k (the total number of all groups of sensors). It also does not exceed the number N of "unstable" invariant subspaces. (If any eigenvalue $\lambda : |\lambda| \geq 1$ gives rise to only one Jordan block, then $N \leq 2^n - 1$.) Generally speaking, relations (5.2) are not independent. However, revealing "superfluous" inequalities is usually a much harder task than direct verification of the entire inequality set (5.2).

Remark 5.3. Whenever the system is stabilizable in the sense of Definition 4.2, the set of nonstrict inequalities (5.2) holds. Moreover, if the second limit in (4.1) is approached quickly enough, e.g.,

$$\left| \frac{b_j^+(r)}{r} - \mathfrak{c}_j \right| \leq \frac{\text{const}}{r},$$

the strict inequalities (5.2) are necessary for the weak stabilizability introduced by Definition 4.2. Then Theorem 5.1 ensures that this stabilizability implies the strong one described by Definitions 4.4 and 4.5.

The necessity of inequalities (5.2) can be complemented by the following fact. Whenever one of the inequalities strictly violates (i.e., $>$ occurs instead of $<$ in (5.2)), the state $\boldsymbol{x}(t)$ exponentially diverges from zero for almost all (with respect to the Lebesgue measure) initial states $\boldsymbol{x}_0 \in \mathbb{R}^n$, irrespective of which networked controller is employed.

We do not prove these remarks since they are not utilized further.

Remark 5.4. The implication $1 \Rightarrow 2$ remains true even if Assumption 4.3 is dropped.

This easily follows from the proof of this implication presented in section 7.

The proof of Theorem 5.1 is given in sections 7 and 8. In section 7, we prove its necessity part $1 \Rightarrow 2$. The converse $1 \Leftarrow 2$ is established in section 8, where formula (5.3) is also justified.

6. Application of Theorem 5.1 to the example from section 2. In this section, we justify the statements from section 2. We recall that they concern a platoon of k vehicles described by (2.1). Each of them is equipped with a sensor giving the distance $y_i = x_i - x_{i-1}$ from it to the preceding vehicle for $i \geq 2$ and the position $y_1 = x_1$ for $i = 1$. It is also served by a communication channel with the transmission capacity $\mathfrak{c}_i > 0$ carrying signals to the controller with the sample period $\Delta > 0$. The objective is to stabilize the platoon motion about a given constant-velocity trajectory: $v_i = v_i^0, x_i(t) = x_i^0 + v_i^0 t \ \forall i$.

The substitution of the variables $v_i := v_i - v_i^0, x_i := x_i - x_i^0 - v_i^0 t$ keeps the dynamics equations unchanged and shapes the control goal into $x_i = 0, v_i = 0$. To put the problem into the framework adopted in this paper, we consider the trajectory only at sampling times: $x_i(\tau) := x_i(\tau\Delta), v_i(\tau) := v_i(\tau\Delta), u_i(\tau) := u_i(\tau\Delta + 0)$. Then

$$(6.1)\quad x_i(\tau+1) = x_i(\tau) + \Delta \cdot v_i(\tau) + \frac{\Delta^2}{2} u_i(\tau), \quad v_i(\tau+1) = v_i(\tau) + \Delta \cdot u_i(\tau),$$
$$y_i(\tau) = x_i(\tau) - x_{i-1}(\tau),$$

where $x_0 := 0$. Now $A = \mathbf{diag}(A_1, \dots, A_k)$, $A_i = \mathcal{A} = \left(\begin{smallmatrix} 1 & \Delta \\ 0 & 1 \end{smallmatrix}\right)$. So the system has only one eigenvalue 1, which gives rise to k Jordan blocks. The nonobservable and nondetectable subspaces (4.3) are identical and equal,

$$L_j^- = \{\mathfrak{x} := (z_1, w_1, \dots, z_k, w_k) : z_j = 0, w_j = 0\}, \text{ where } z_i := x_i - x_{i-1}, w_i := v_i - v_{i-1}$$

for $i \geq 2$ and $z_1 := x_1, w_1 := v_1$. Assumption 4.3 holds with $M^i_\sigma := \{\mathfrak{x} : z_j = 0, w_j = 0\ \forall j \neq i\}, i = 1, \dots, k$; Assumption 4.2 is immediate from (6.1). Since $|\det A|_L| = 1$ for any invariant subspace L, Theorem 5.1 guarantees that *the platoon is uniformly exponentially stabilizable under arbitrary transmission capacities* $\mathfrak{c}_i > 0$.

To determine the rate of stabilizability μ^0, note that the states nondetectable by (maybe, empty) group $\mathcal{J} \subset [1:k]$ of sensors constitute the subspace $L(\mathcal{J}) := \{\mathfrak{x} : z_j = 0, w_j = 0\ \forall j \in \mathcal{J}\}$. Since $\mathbf{dim} L(\mathcal{J}) = 2(k - |\mathcal{J}|)$, where $|\mathcal{J}|$ is the size of $\mathcal{J}$, relation (5.3) shapes into

$$\log_2 \mu^0 = \max_{\mathcal{J}} \frac{1}{2(k-|\mathcal{J}|)} \left(-\sum_{j \notin \mathcal{J}} \mathfrak{c}_j \right) = -\frac{1}{2} \mathfrak{c}_{\min}, \quad \text{where} \quad \mathfrak{c}_{\min} := \min_{j=1,\dots,k} \mathfrak{c}_j. \tag{6.2}$$

Thus *the rate of the platoon exponential stabilizability equals* $\sqrt{2}^{\,-\mathfrak{c}_{\min}}$ *per sample period*.

Now consider the situation where the sensor system accommodated by each vehicle gives the distances to $l < k$ vehicles to the right, as well as to l vehicles to the left. (We suppose that there is an imaginary vehicle numbered by 0 and staying at the origin.) Then

$$L^-_j = \Big\{ \mathfrak{x} : z_i = 0, w_i = 0 \quad \forall i = \max\{j - l + 1, 1\}, \dots, \min\{j + l, k\} \Big\}.$$

Assumption 4.3 remains true with the same subspaces M^i_σ, and the platoon evidently remains uniformly exponentially stabilizable. What can be said about the rate of stabilizability? Now the collection $\mathfrak{L}$ consists of spaces $L(\mathcal{J})$ related to sets $\mathcal{J}$ which along with any element $j \in \mathcal{J}$, contain a certain interval of the form $[i - l + 1 : i + l] \cap [1:k] \ni j, i = 1, \dots, k$. Furthermore, such sets are said to be *wide*. To proceed, we consider separately two cases.

1. Let $2l \geq k$. Then any two such intervals contain a common point. It follows that apart from $\mathcal{J} = \emptyset$, the wide sets $\mathcal{J}$ are intervals of the form $[1:i], i \geq l+1$, or $[i:k], i \leq k-l+1$. By retracing (6.2), we see that $\mu^0 = \sqrt{2}^{\,-\mathfrak{c}_{k,l}}$, where $\mathfrak{c}_{k,l}$ is given by (2.2).

2. Now let $2l < k$. Then the sets

$$[1:i-1] \cup [i+1:k], i = l+2, \dots, k-l, \quad [i:k], i = 2, \dots, l+2, \quad [1:i], i = k-l, \dots, k-1,$$

are wide. By restricting the maximum in (6.2) to only these sets $\mathcal{J}$, we see that $\mu^0 \geq \sqrt{2}^{\,-\mathfrak{c}_{k,l}}$, where $\mathfrak{c}_{k,l}$ is given by (2.2). In fact, $\mu^0 = \sqrt{2}^{\,-\mathfrak{c}_{k,l}}$. To prove this, it suffices to show that

$$\frac{1}{k - |\mathcal{J}|} \sum_{j \notin \mathcal{J}} \mathfrak{c}_j \geq \mathfrak{c}_{k,l} \tag{6.3}$$

for any wide set $\mathcal{J}$. To this end, we put $i_- := \min\{j : j \in \mathcal{J}\}$ and $i_+ := \max\{j : j \in \mathcal{J}\}$. Then $i_- \leq l+1 \Rightarrow [i_- : l+1] \subset \mathcal{J}$ and $i_+ \geq k-l+1 \Rightarrow [k-l+1 : i_+] \subset \mathcal{J}$ by the definition of the wide set. Hence $\{j : j \notin \mathcal{J}\} = \mathcal{J}_1 \cup \dots \cup \mathcal{J}_s$, where the sets $\mathcal{J}_\nu$ are pairwise disjoint and any of them either contains only one index, $\mathcal{J}_\nu = \{i\}, i = l+2, \dots, k-l$, or $\mathcal{J}_\nu = [1:i], i = 1, \dots, l+1$, or $\mathcal{J}_\nu = [i:k], i = k-l+1, \dots, k$. By (2.2), $1/|\mathcal{J}_\nu| \sum_{j \in \mathcal{J}_\nu} \mathfrak{c}_j \geq \mathfrak{c}_{k,l}$. This implies (6.3) as follows:

$$\frac{1}{k - |\mathcal{J}|} \sum_{j \notin \mathcal{J}} \mathfrak{c}_j = \frac{1}{k - |\mathcal{J}|} \sum_{\nu=1}^{s} |\mathcal{J}_\nu| \frac{1}{|\mathcal{J}_\nu|} \sum_{j \in \mathcal{J}_\nu} \mathfrak{c}_j \geq \mathfrak{c}_{k,l}.$$

7. Inequalities (5.2) as necessary conditions for stabilizability. The current and next sections contain the proof of Theorem 5.1. For technical reasons which will become clear soon, we extend the class of systems

$$\boldsymbol{x}(t+1) = A\boldsymbol{x}(t) + \mathcal{B}[t, \boldsymbol{u}(0), \dots, \boldsymbol{u}(t)], \tag{7.1}$$

where $\mathcal{B}(\cdot)$ is a given function. We also generalize on them a fact well known for the systems (3.1) (see [11, 20, 25]).

LEMMA 7.1. *Suppose that there is only one channel $k = 1$. Then $\log_2 |\det A| < \mathfrak{c} := \mathfrak{c}_1$ whenever the system (7.1) is uniformly exponentially stabilizable.*

Proof. Let some networked controller uniformly exponentially stabilize the system. Then

$$|\boldsymbol{x}(t)| \le K\mu^t \quad \forall t = 01, 2, \dots \quad \text{whenever} \quad |\boldsymbol{x}_0| < 1, \quad \text{where} \quad \mu \in (0, 1). \tag{7.2}$$

Thanks to 1 of Assumption 4.1, no more than $b_1^+(t)$ bits arrive at the decoder by time t. So in the formula

$$\boldsymbol{x}(t) = A^t\boldsymbol{x}_0 - \boldsymbol{\omega}(t), \quad \boldsymbol{\omega}(t) := -\sum_{\nu=0}^{t-1} A^{t-1-\nu}\mathcal{B}\big[\nu, \boldsymbol{u}(0), \dots, \boldsymbol{u}(\nu)\big] \tag{7.3}$$

$\overset{(3.4),(3.6)}{=\!=} \mathcal{V}[t, \overline{\boldsymbol{e}}(t)]$, the tuple $\overline{\boldsymbol{e}}(t)$ and the vector $\boldsymbol{\omega}(t)$ take values from some sets each containing no more than $2^{b_1^+(t)}$ elements and not depending on the initial state $\boldsymbol{x}_0$. So (7.2) means that the image $A^tB_0^1$ of the unit ball can be covered by the union of no more than $2^{b_1^+(t)}$ balls of the form $B^{K\mu^t}_{\boldsymbol{\omega}(t)}$. Hence

$$|\det A|^t \,\mathbf{mes}\, B_0^1 = \mathbf{mes}\,\big[A^tB_0^1\big] \le 2^{b_1^+(t)}\,\mathbf{mes}\, B_0^{K\mu^t} = 2^{b_1^+(t)}\mu^{nt}K^n\,\mathbf{mes}\, B_0^1$$
$$\Downarrow$$
$$\log_2|\det A| \le \frac{b_1^+(t)}{t} + n\log_2\mu + \frac{n\log_2 K}{t}.$$

Letting $t \to \infty$ and taking into account 3 of Assumption 4.1 and the inequality $\mu < 1$ completes the proof. □

Lemma 7.1 evidently remains true if the regime of channel operation (given by $\tau_1(\cdot)$) is known in advance.

To prove (5.2), we revert to the system (3.1) and pick a subspace $L \in \mathfrak{L}$ constituted by the states not detectable by a certain group of sensors. Then we restrict ourselves to trajectories $\{\boldsymbol{x}(t)\}$ starting at $\boldsymbol{x}_0 \in L$ and apply Lemma 7.1 to them. More precisely, we take into account that $\{\boldsymbol{x}(t)\}$ may leave L due to controls, and consider $\boldsymbol{x}_L(t) := \pi\boldsymbol{x}(t)$. Here π is an arbitrarily chosen projector from $\mathbb{R}^n$ onto L. By applying formula (7.3) to the system (3.1), it is easy to check that the evolution of $\boldsymbol{x}_L$ is governed by the equations of the form (7.1),

$$\boldsymbol{x}_L(t+1) = A|_L\boldsymbol{x}_L(t) + \pi B\boldsymbol{u}(t) + \sum_{i=0}^{t-1}(\pi A - A\pi)A^{t-1-i}B\boldsymbol{u}(i),\ \boldsymbol{x}_L(t) \in L,\ \boldsymbol{x}_L(0) = \boldsymbol{x}_0. \tag{7.4}$$

(The first equation simplifies if $\pi A = A\pi$. However, such a projector π exists if and only if there exists an A-invariant subspace that is complementary to L, which is not

true in general.) We interpret (7.4) as equations of an auxiliary system and equip it with the sensor

$$\boldsymbol{y}_L(t) = C_{J^c}\boldsymbol{x}_L(t). \tag{7.5}$$

Here $J^c := \{j : j \notin J\}$ is the complement to the set $J = \{j : C_j\boldsymbol{x} = 0\ \forall \boldsymbol{x} \in L\}$ of sensors ignoring the subspace L. Furthermore, for $I \subset [1:k]$, the symbol C_I denotes the block matrix that results from arranging the blocks C_j with $j \in I$ into a column. (For any entities v_j enumerated by $j \in [1:k]$, the symbol v_I is defined likewise.) We also suppose that all channels with $j \notin J$ are commissioned to transmit $\boldsymbol{y}_L$.

The sum on the right-hand side of (5.2) equals the capacity of the union of these channels. Hence (5.2) follows from Lemma 7.1 applied to the system (7.4), (7.5). To complete the proof, it suffices to show that this system is stabilizable whenever the original one (3.1), (3.2) can be stabilized. In doing so, one must cope with the fact that the trajectory of the original closed-loop system (3.1)–(3.4) may leave the subspace L. So the observations (3.2) and (7.5) may differ. Moreover, the sensors omitted in (7.5) may see the state $\boldsymbol{x}(t)$ for $t \geq 1$. It should be shown that they are yet useless and can be dropped.

LEMMA 7.2. *Let the system* (3.1), (3.2) *be exponentially stabilized by some networked controller, and the regime of channels operation be known in advance. Then the system* (7.4), (7.5) *is also exponentially stabilizable.*

Proof. We first show that for $\boldsymbol{x}_0 \in L$, the process in the original closed-loop system obeys the relations

$$\begin{aligned} e_J(t) = \mathcal{E}'_j\big[t, \overline{e}_{J^c}(t-1)\big], \quad \boldsymbol{y}_J(t) = \mathcal{Y}'\big[t, \overline{e}_{J^c}(t-1)\big], \\ \boldsymbol{y}_{J^c}(t) = \boldsymbol{y}_L(t) + \mathcal{Y}''\big[t, \overline{e}_{J^c}(t-1)\big]. \end{aligned} \tag{7.6}$$

We recall that the data $\overline{e}_j(t)$ that arrived via the jth channel by time t is given by (3.6). The observation $\boldsymbol{y}_L(t)$ is defined by (7.4) and (7.5) for the sequence of controls $\boldsymbol{u}(t)$ identical to that driving the original system.

For $t = 0$, we have $\boldsymbol{x}(0) \in L \Rightarrow \boldsymbol{y}_J(0) = 0$ and $\boldsymbol{y}_{J^c}(0) = \boldsymbol{y}_L(0)$, and (7.6) with $t = 0$ follows from (3.3). Now suppose that (7.6) with $t := \theta$ holds for all $\theta \leq t$. Then $\overline{e}_J(\theta) = \overline{\mathcal{E}}'\big[\theta, \overline{e}_{J^c}(t-1)\big]$ and so

$$\overline{e}(\theta) = \big[\overline{e}_J(\theta), \overline{e}_{J^c}(\theta)\big] = \overline{\mathcal{E}}\big[\theta, \overline{e}_{J^c}(t)\big] \overset{(3.4)}{\Longrightarrow} \boldsymbol{u}(\theta) = \mathcal{U}'\big[\theta, \overline{e}_{J^c}(t)\big], \qquad \theta \leq t. \tag{7.7}$$

Now we invoke (7.3) and note that $\boldsymbol{x}_0 \in L \Rightarrow A^{t+1}\boldsymbol{x}_0 \in L \Rightarrow C_J A^{t+1}\boldsymbol{x}_0 = 0and(I - \pi)A^{t+1}\boldsymbol{x}_0 = 0$:

$$\begin{aligned} \boldsymbol{y}_J(t+1) &= \underbrace{C_J A^{t+1}\boldsymbol{x}_0}_{=0} + \sum_{\theta=0}^{t} C_J A^{t-\theta} B\boldsymbol{u}(\theta) =: \mathcal{Y}'\big[t+1, \overline{e}_{J^c}(t)\big], \\ \boldsymbol{y}_{J^c}(t+1) - \boldsymbol{y}_L(t+1) &= C_{J^c}\big[\boldsymbol{x}(t+1) - \pi\boldsymbol{x}(t+1)\big] = \underbrace{C_{J^c}(I-\pi)A^{t+1}\boldsymbol{x}_0}_{=0} \\ &\quad + C_{J^c}\sum_{\theta=0}^{t}(I-\pi)A^{t-\theta}B\boldsymbol{u}(\theta) =: \mathcal{Y}''\big[t+1, \overline{e}_{J^c}(t)\big], \end{aligned}$$

i.e., the last two relations from (7.6) do hold with $t := t+1$. Then the first relation follows from (3.3).

It follows from (3.3) and (7.6) that the signal $e_{J^c}(t)$ is determined by the prior measurements from (7.5),

$$e_{J^c}(t) = \mathcal{E}_L[t, \boldsymbol{y}_L(0), \dots, \boldsymbol{y}_L(t)].$$

Now we interpret this as the equation of the coder, and the last relation from (7.7) (where $\theta := t$) as that of the decoder for the system (7.4), (7.5). By the foregoing, this coder-decoder pair generates the trajectory $\pi\boldsymbol{x}(t), \boldsymbol{u}(t), t = 0, 1, \dots$, where $\boldsymbol{x}(t), \boldsymbol{u}(t)$ is the trajectory of the original closed-loop system. So the inequalities (4.2) are inherited by the system (7.4), (7.5), which completes the proof. □

As was remarked, the part 1 ⇒ 2 of Theorem 5.1 is immediate from Lemmas 7.1 and 7.2.

8. Inequalities (5.2) as sufficient conditions for stabilizability. From now on, we suppose that 2 of Theorem 5.1 holds. We also assume, until otherwise stated, that *the system* (3.1) *has no stable modes*. In the general case, a stabilizing controller will be obtained by applying that presented below to the unstable part of the system.

To stabilize the system, we employ the scaled quantization scheme (see, e.g., [2, 12, 14, 19, 20, 22, 23, 25]). It was mainly developed for only one channel and is stated briefly as follows. Both coder and decoder compute a common upper bound δ of the current state norm $|\boldsymbol{x}| := \max_i |x_i|$. They are also given a partition of the unit ball into m balls (cubes) with small radii $\le \sigma(m)$. The number m matches the channel capacity so that the serial number of the cube can be communicated to the decoder. The coder determines the current state from the observations and notifies the decoder which cube contains this state divided by δ. Since the decoder knows δ, it thus becomes aware of a ball B with the radius $\le \delta\sigma(m)$ containing the current state. Then it selects a control that drives the system from the center of this ball to zero. The ball itself is expanded due to the unstable dynamics of the system and transformed into a set $D_+(B)$ centered about zero: $D_+(B) \subset B_0^\rho$. So the radius ρ can be taken as a new upper bound δ. Here $\rho \le \delta\sigma(m)\alpha$, where α characterizes the expansion rate of the system. If $\sigma(m)\alpha < 1$, the bound δ is thus improved $\delta := \delta\sigma(m)\alpha < \delta$ and by continuing likewise, driven to zero $\delta \to 0$, along with the state $\boldsymbol{x}$.

In the context of this paper, a problem with the above scheme is that no coder may be aware of the entire state $\boldsymbol{x}$. So a natural idea [25] is to disintegrate the system (3.1) into subsystems each observable by some sensor. Then each subsystem can be stabilized by following the above lines, provided the stability condition $\sigma(m)\alpha < 1$ holds for it. In fact, this condition means that there is a way to communicate a sufficiently large amount of information from the subsystem to the decoder: the smaller the radius $\sigma(m)$, the larger the size m of the partition, and so the larger the number of bits required to describe which of m cubes contains the state.

No channel in itself may meet the above stability condition. At the same time, this condition may be met if several channels are commissioned to transmit information about a given subsystem. Then each channel may carry only a part of this information, whereas the decoder assembles these parts and thus gets the entire message. Certainly, these channels must be chosen among those serving the sensors that observe the subsystem at hand.

Since a given sensor may observe several subsystems, the above scheme means that each channel must transmit a set of messages each concerning a particular subsystem. This gives rise to the question: is it possible to distribute the required information about any particular subsystem over parallel channels in such a way that the total amount of information carried via every channel meets its capacity? It will be shown

via convex duality arguments that the answer is in the affirmative whenever 2 of Theorem 5.1 holds.

Another problem is how to use a sensor observing the subsystem only partly. Certainly, this problem does not hold if there are no such sensors and subsystems: the state of each subsystem either is completely determined from or does not affect the outputs of any given sensor. Decomposition into a set of such subsystems is possible. However, in general, these subsystems are dependant. First, the control is common. Second, Jordan blocks may entail an unavoidable interinfluence between the states of subsystems. Now we illustrate this by an example.

Example. Consider the system whose dynamical matrix is a standard Jordan block,

(8.1)
$$\begin{array}{rcccccccc} x_1(t+1) & = & \lambda x_1(t) & + & 0 & + & b_1^* u(t) & & y_1(t)=x_1(t) \\ x_2(t+1) & = & \lambda x_2(t) & + & x_1(t) & + & b_2^* u(t) & & y_2(t)=x_2(t) \\ & & & & \vdots & & & , & \vdots \\ x_d(t+1) & = & \lambda x_d(t) & + & x_{d-1}(t) & + & b_d^* u(t) & & y_d(t)=x_d(t) \end{array} \quad , \qquad |\lambda|>1.$$

The unobservable subspaces of the sensors are, respectively,

$$L_1^- := \{\boldsymbol{x} : x_1 = 0\}, \quad L_2^- = \{\boldsymbol{x} : x_1 = x_2 = 0\}, \dots, L_{d-1}^- = \{\boldsymbol{x} : x_1 = \dots = x_{d-1} = 0\},$$
$$L_d^- = \{0\}.$$

There are no other invariant subspaces (see Lemma 8.2 below), so this system cannot be decomposed into subsystems with independent open-loop dynamics. At the same time, any sensor except for the last one observes the state $\boldsymbol{x}$ only partly. So to exclude such a partial vision, disintegration into state-dependant subsystems is unavoidable.

To deal with them, we employ sequential stabilization. We define the sth subsystem as that described in the sth row from (8.1). Its state is x_s. Then we stabilize the first subsystem, which is independent of the other ones. This makes x_1 exponentially decaying. In the equations of the second subsystem, we interpret x_1 as an exogenous disturbance. By constructing a device stabilizing this subsystem under any exponentially vanishing disturbances, we make x_2 exponentially decaying. The entire system is stabilized by continuing likewise.

These arguments, however, do not take into account that the control affects all subsystems and some of them may be unstabilizable (though the entire system is controllable). For example, the subsystems with $s \geq 2$ are unstabilizable if $b_1 = 1, b_2 = \dots b_d = 0$ in (8.1). These obstacles can be easily overcome via increasing the sample period.

Indeed, let us pick $r = 1, 2, \dots$. The state $\boldsymbol{x}_i := \boldsymbol{x}(ir)$ evolves as follows:

$$\boldsymbol{x}_{i+1} = A^r \boldsymbol{x}_i + \mathfrak{B}\boldsymbol{U}^{\,i}, \tag{8.2}$$

(8.3)
$$\text{where} \quad \boldsymbol{U}^{\,i} := \big[\boldsymbol{u}(ir), \boldsymbol{u}(ir+1), \dots, \boldsymbol{u}(ir+r-1)\big] \quad \text{and} \quad \mathfrak{B}\boldsymbol{U} := \sum_{j=0}^{r-1} A^{r-1-j} B \boldsymbol{u}_j$$

is the state to which the control program $\boldsymbol{U} = \big[\boldsymbol{u}_0, \dots, \boldsymbol{u}_{r-1}\big]$ drives the system at time $t = r$ from $\boldsymbol{x}(0) = 0$. Since the system (3.1) with no stable modes is controllable

by Assumption 4.2, the operator $\mathfrak{B}$ is onto if $r > n$. Related to the decomposition of the system $\boldsymbol{x} = (\boldsymbol{x}^1, \dots, \boldsymbol{x}^d)$ is a block partition $\mathfrak{B}\boldsymbol{U} = \left[\mathfrak{B}_1\boldsymbol{U}, \dots, \mathfrak{B}_d\boldsymbol{U}\right]$. Since all operators $\mathfrak{B}_s$ have full rank, any subsystem is controllable. Moreover, any control action $\boldsymbol{y} = \mathfrak{B}_s\boldsymbol{U}$ in the sth subsystem can be implemented by a control $\boldsymbol{U}$ that does not disturb the other subsystems, $\mathfrak{B}_j\boldsymbol{U} = 0\ \forall j \neq s$.

Summarizing, we adopt the following plan of proving the sufficiency part of Theorem 5.1.

1. We decompose the system so that, for any given sensor, the state of each subsystem either does not affect or is determined from the sensor outputs and second, the decomposition is triangular. The latter permits us to employ the sequential stabilization approach.
2. We increase the sample period and, for each subsystem, offer a class of networked controllers stabilizing it under any exponentially decaying disturbance. In doing so, we assume that the coder is aware of the current state at any sample time $t = ir$, and there is a way to transmit as much information as desired from the coder to the decoder. Then we find the controller for which the information traffic is minimal.
3. We show that if all subsystems are equipped with the above controllers, the entire system is stabilized.
4. We obtain conditions under which the entire set of these controllers can be implemented on the basis of real channels and sensors. These conditions are not constructive and require that a linear system of inequalities be solvable in integers.
5. By employing convex duality arguments, we show that these conditions are equivalent to (5.2).

8.1. Decomposition of the system. Now we perform step 1 of the plan of proving the sufficiency part of Theorem 5.1.

PROPOSITION 8.1. *Suppose that the system* (3.1) *has no stable modes. Then after a proper one-to-one linear transformation and partition of the state*

$$\boldsymbol{x} = (\boldsymbol{x}^1, \dots, \boldsymbol{x}^d) \tag{8.4}$$

into several blocks $\boldsymbol{x}^s \in \mathbb{R}^{n_s}$ *interpreted as the states of subsystems, the following statements hold:*

1. *The unobservable subspace* (4.3) $L_j^{-obs} = L_j^-$ *of any sensor is composed of several blocks,*

$$L_j^- = \{\boldsymbol{x} : \boldsymbol{x}^s = 0\ \forall s \in O_j\}, \qquad \text{where} \quad O_j \subset [1:d].$$

2. *The block representation of the dynamics equations* (3.1) *is lower triangular,*

$$\boldsymbol{x}^s(t+1) = \sum_{i=1}^{s} A_{si}\boldsymbol{x}^i(t) + B_s\boldsymbol{u}(t), \qquad s = 1, \dots, d. \tag{8.5}$$

By 1, the states $\boldsymbol{x}^s$ of subsystems $s \notin O_j$ do not affect the outputs of the jth sensor, whereas the states $\boldsymbol{x}^s$ with $s \in O_j$ are uniquely determined from these outputs.

To prove the proposition, we start with two technical facts.

LEMMA 8.2. *Assumption* 4.3 *holds for any elementary spectral set* σ.

Proof. It suffices to prove the lemma assuming that the set σ gives rise to only one real Jordan block. We put $M := M_\sigma, A_\sigma := A|_M$ and note that $\det[\lambda I - A_\sigma] = \varphi(\lambda)^p$, where the polynomial φ is irreducible over the field of real numbers. By employing the basis in M reducing A_σ to the real Jordan form, it is easy to see that the formula $\mathcal{L}(\nu) := \ker\left[\varphi(A_\sigma)\right]^\nu$ produces $(p+1)$ distinct $\{0\} = \mathcal{L}(0) \subset \mathcal{L}(1) \subset \cdots \subset \mathcal{L}(p) = M$ invariant subspaces and $\mathbf{dim}\mathcal{L}(\nu) = \nu \deg\varphi$. We are going to show that there are no other invariant subspaces.

Indeed let L be such a subspace and ψ be the minimal annihilating polynomial of L. Then ψ is a divisor of φ^p and so $\psi = \varphi^\nu, \nu = 0, \dots, p$. Hence $L \subset \ker\left[\varphi(A_\sigma)\right]^\nu = \mathcal{L}(\nu)$. At the same time, Theorem 2 of [9, p. 180] implies that $\mathbf{dim}L = \deg\psi$. Thus $\mathbf{dim}L = \nu, \deg\varphi = \mathbf{dim}\mathcal{L}(\nu)$, and so $L = \mathcal{L}(\nu)$.

As a result, we see that all invariant subspaces $L_j^- \cap M_\sigma$ are among $\mathcal{L}(0), \dots, \mathcal{L}(p)$. It remains to pick $M_\sigma^1 := \mathcal{L}(1), m_\sigma := p$, and for $i = 2, \dots, p$, choose M_σ^i so that $\mathcal{L}(i-1) \oplus M_\sigma^i = \mathcal{L}(i)$. □

The next lemma plays the key role in the proof of Proposition 8.1.

LEMMA 8.3. *In Assumption* 4.3*, the atoms $M_\sigma^i, i = 1, \dots, m_\sigma$, can be chosen so that all partial direct sums $M_\sigma^1 \oplus \cdots \oplus M_\sigma^i, i = 1, \dots, m_\sigma$, are A-invariant.*

Proof. We consider the set of atoms with the minimal size m_σ. We also introduce the undetectable subspaces $L_j := L_j^- \cap M_\sigma$ of M_σ, then form all their intersections $L^\cap = L_{j^1} \cap \cdots \cap L_{j^p}$, and then all algebraic sums (not necessarily direct) of such intersections $L^\Sigma = L_{i^1}^\cap + \cdots + L_{i^r}^\cap$. Here p, r and the subspaces $L_{j^\nu}, L_{i^\mu}^\cap$ are chosen arbitrarily. Let $\mathfrak{M}$ denote the set of all L^Σ. It is clear that (1) any space $L \in \mathfrak{M}$ is invariant and decomposed into a direct sum of several atoms, (2) $L \in \mathfrak{M} \Rightarrow L \cap L_j \in \mathfrak{M} \forall j$, (3) $M_\sigma \in \mathfrak{M}$, (4) $L', L'' \in \mathfrak{M} \Rightarrow L' + L'' \in \mathfrak{M}$, and (5) the set $\mathfrak{M}$ is finite. Now we pick a minimal element $L_{\min}$ among $L \in \mathfrak{M}, L \neq \{0\}$, i.e., such that $L \subset L_{\min}$ and $L \in \mathfrak{M}$ and $L \neq \{0\} \Rightarrow L = L_{\min}$. By trying here $L := L_{\min} \cap L_j$, we see that either $L_{\min} \subset L_j$ or $L_{\min} \cap L_j = \{0\}$. Hence any L_j contains either all atoms constituting $L_{\min}$ or none of them. So these atoms can be replaced by their sum in Assumption 4.3. Since the number of all atoms is minimal, the only concern is $L_{\min} = M_\sigma^\nu$. By permuting the atoms, we set $\nu = 1$. Then the claim of the lemma holds for $i = 1$ by (1).

Now let $L_{\min}$ denote a minimal element among $L \in \mathfrak{M}, L \supset M_\sigma^1, L \neq M_\sigma^1$. By (2) and (4), $L := M_\sigma^1 + L_j \cap L_{\min} \in \mathfrak{M}$. So the minimum property yields that either $L = M_\sigma^1$ or $L_{\min} \subset L$. In terms of the decomposition from (1) $L_{\min} = M_\sigma^1 \oplus M_{\min} \in \mathfrak{M}$ (where $M_{\min}$ is the sum of several atoms), this means that either $M_{\min} \cap L_j = \{0\}$ or $M_{\min} \subset L_j$. Like above, this implies that $M_{\min}$ consists of only one atom M_σ^ν. By permuting the atoms, we set $\nu = 2$ and make the claim of the lemma true for $i = 2$ by (1). The proof is completed by continuing likewise. □

Proof of Proposition 8.1. We decompose the spectrum $\sigma(A) = \sigma^1 \cup \cdots \cup \sigma^p$ into the union of disjoint elementary sets. Then $\mathbb{R}^n = M_{\sigma^1} \oplus \cdots \oplus M_{\sigma^p}$, and any invariant subspace L_j^- is decomposed $L_j^- = L_j^-(1) \oplus \cdots \oplus L_j^-(p)$ into invariant subspaces $L_j^-(\nu) := L_j^- \cap M_{\sigma^\nu}$. So it suffices to show that, for any ν, there exist linear coordinates in M_{σ^ν} and their block partition for which any subspace $L_j^-(\nu), j = 1, \dots, k$, is the direct sum of several "blocks" and the operator $A|_{M_{\sigma^\nu}}$ has a lower triangular form with respect to this partition. These blocks $\boldsymbol{z}^i$ are in fact given by Lemma 8.3: $\boldsymbol{z}^i \in M_{\sigma^\nu}^{m_{\sigma^\nu}-i+1}$. More precisely, it suffices to pick a basis in each subspace $M_{\sigma^\nu}^i, i = 1, \dots, m_{\sigma^\nu}$, unite them to produce a basis in M_{σ^ν}, and then consider the coordinates with respect to this basis and their partition that corresponds to the partition $\boldsymbol{z} = \boldsymbol{z}^1 + \cdots + \boldsymbol{z}^{m_{\sigma^\nu}}$ of $\boldsymbol{z}$ into $\boldsymbol{z}^i \in M_{\sigma^\nu}^{m_{\sigma^\nu}-i+1}$.

8.2. Separate stabilization of subsystems. In this subsection, we pick an integer parameter r and focus attention only on the states at times $\tau_i = i \cdot r$. The evolution of these states is given by (8.2), which evidently inherits the lower triangular structure from (8.5),

$$\boldsymbol{x}^s_{i+1} = \sum_{\nu=1}^{s} A^{(r)}_{s\nu} \boldsymbol{x}^\nu_i + \mathfrak{B}_s \boldsymbol{U}_i \tag{8.6}$$

for $s = 1, \dots, d$. Here $\boldsymbol{x}^s_i := \boldsymbol{x}^s(\tau_i), \boldsymbol{U}_i = \big[\boldsymbol{u}(\tau_i), \dots, \boldsymbol{u}(\tau_i + r - 1)\big]$, and the diagonal coefficients from (8.6) are the rth powers of matching coefficients from (8.5), $A^{(r)}_{ss} = A^r_{ss}$. The *sth subsystem* is described by the following equations:

$$\boldsymbol{x}^s_{i+1} = A^r_{ss} \boldsymbol{x}^s_i + \mathfrak{B}_s \boldsymbol{U}_i + \boldsymbol{\xi}_{s,i}, \qquad i = 0, 1, \dots . \tag{8.7}$$

Here in accordance with (8.6),

$$\boldsymbol{\xi}_{s,i} = 0 \quad \text{for} \quad s = 1 \quad \text{and} \qquad \boldsymbol{\xi}_{s,i}(t) = \sum_{\nu=1}^{s-1} A^{(r)}_{s\nu} \boldsymbol{x}^\nu_i \quad \text{otherwise.} \tag{8.8}$$

In this subsection, we ignore this rule and interpret $\boldsymbol{\xi}_{s,i}$ as an exogenous disturbance. This permits us to study each subsystem independently of the others. We also suppose that the disturbance decays at a known rate ρ_ξ,

$$|\boldsymbol{\xi}_{s,i}| \le K_\xi \rho^i_\xi, \qquad \rho_\xi \in [0, 1), \quad i = 0, 1, \dots, \tag{8.9}$$

whereas K_ξ is unknown, and offer a controller that stabilizes the sth subsystem under all such disturbances. In doing so, we assume that the current state $\boldsymbol{x}^s_i$ is measured on-line. The proposed controller uses only finitely many bits of information about $\boldsymbol{x}^s_i$.

It will be mainly based on the ideas from [2, 12, 14, 19, 20, 22, 23, 25]. A novelty concerns two points. First, we consider the case where the transmission of the above bits to the decoder takes some time (specifically, r units of time). This implies complements to the stabilization scheme, e.g., the need to quantize not the current state but the state prognosis. Second, we take into account exogenous disturbances decaying at a known rate.

We start with introducing components of which the coder and decoder will be assembled.

Quantizer. An m-level *quantizer* $\mathfrak{Q}^s$ in $\mathbb{R}^{n_s}$ is a partition of the unit ball $B^1_0 \subset \mathbb{R}^{n_s}$ with respect to some norm $|\cdot|$ into m disjoint sets $Q_1, \dots, Q_m$ each equipped with a *centroid* $\boldsymbol{q}^{Q_i} \in Q_i$. Such a quantizer associates any vector $\boldsymbol{x}^s \in Q_i$ with its *quantized value* $\mathfrak{Q}^s(\boldsymbol{x}^s) := \boldsymbol{q}^{Q_i}$ and any vector $\boldsymbol{x}^s \notin B^1_0$ with an alarm symbol ✠.

Definition 8.4. *The quantizer $\mathfrak{Q}^s$ is said to be* r-contracted *(for the sth subsystem) if*

$$A^r_{ss}\big(Q - \boldsymbol{q}^Q\big) \subset \rho_{\mathfrak{Q}^s} B^1_0 \quad \forall Q = Q_i,\ i = 1, \dots, m, \qquad \textit{where} \quad \rho_{\mathfrak{Q}^s} \in (0, 1). \tag{8.10}$$

Deadbeat stabilizer. This is a linear transformation $\mathfrak{S}$ of an initial state $\boldsymbol{x}^s_0$ into a control program $\boldsymbol{U}$ that drives the unperturbed $\boldsymbol{\xi}_{s,i} \equiv 0$ subsystem (8.7) to zero,

$$0 = \boldsymbol{x}^s_1 (= \boldsymbol{x}^s(r)) = A^r_{ss} \boldsymbol{x}^s_0 + \mathfrak{B}_s \boldsymbol{U} \quad \text{for} \quad \boldsymbol{U} := \mathfrak{S} \boldsymbol{x}^s_0 \quad \text{and any} \quad \boldsymbol{x}^s_0. \tag{8.11}$$

It is supposed that a contracted quantizer and deadbeat stabilizer are given. Furthermore, we show that they do exist.

Parameters. Apart from r, the controller employs two more parameters ρ and γ. They are chosen so that

$$r > n, \qquad \gamma > \|A_{ss}\|^r, \qquad \text{and} \qquad 1 > \rho > \max\{\rho_\xi, \rho_{\mathfrak{O}^s}\}, \tag{8.12}$$

where A_{ss}, ρ_ξ, and $\rho_{\mathfrak{O}^s}$ are taken from (8.5), (8.9), and (8.10), respectively.

Both coder and decoder compute controls $\boldsymbol{U}^{\,c}_i$, $\boldsymbol{U}^{\,d}_i$ and upper bounds δ^c_i, δ^d_i for the state norm $|\boldsymbol{x}^s_i|$, respectively. Actually, acting upon the plant is the control $\boldsymbol{U}^{\,d}_i$. The initial bounds are common: $\delta^c_0 = \delta^d_0 = \delta_0$. (The inequality $\delta_0 \geq |x^s_0|$ may be violated.) At any time $\tau_i = ir$, the coder selects a finite-bit message based on $\boldsymbol{x}^s_i$. This message is sent to the decoder and arrives by time τ_{i+1}. Specifically, the coder and decoder act as follows.

The *s*th coder (at the times $t = \tau_i, i = 1, 2, \dots$).

c.1. Proceeding from the current state $\boldsymbol{x}^s_i$, computes the state prognosis for the time $t = \tau_{i+1}$,

$$\widehat{\boldsymbol{x}}^s_{i+1} := A^r_{ss}\boldsymbol{x}^s_i + \mathfrak{B}_s\boldsymbol{U}^{\,c}_i; \tag{8.13}$$

c.2. Employs the r-contracted quantizer $\mathfrak{Q}^s$ to compute the quantized value $\boldsymbol{q}_i$ of the scaled state at $t = \tau_{i+1}$,

$$\boldsymbol{\varepsilon}_i := \big[\delta^c_i\big]^{-1}\widehat{\boldsymbol{x}}^s_{i+1}, \qquad \boldsymbol{q}_i := \mathfrak{Q}^s\big[\boldsymbol{\varepsilon}_i\big]; \tag{8.14}$$

c.3. Encodes this quantized value $\boldsymbol{q}_i$ for transmission and sends it to the decoder;

c.4. Computes the next control program by means of the deadbeat stabilizer $\mathfrak{S}$ and corrects the upper bound,

$$\boldsymbol{U}^{\,c}_{i+1} := \mathfrak{S}\big[\delta^c_i \overset{\star}{\boldsymbol{q}}_i\big], \qquad \delta^c_{i+1} := \delta^c_i \times \langle \boldsymbol{q}_i\rangle_{\rho,\gamma}, \qquad \text{where} \tag{8.15}$$

$$\overset{\star}{\boldsymbol{q}} := \begin{cases} \boldsymbol{q} & \text{if } \boldsymbol{q} \neq \maltese, \\ 0 & \text{otherwise,} \end{cases} \qquad \langle q\rangle_{\rho,\gamma} := \begin{cases} \rho & \text{if } q \neq \maltese, \\ \gamma & \text{otherwise.} \end{cases} \tag{8.16}$$

The *s*th decoder (at the times $t = \tau_i, i = 2, 3, \dots$).

d.1. Decodes the newly received data and thus acquires $\boldsymbol{q}_{i-1}$;

d.2. Computes the current control program and corrects the upper bound,

$$\boldsymbol{U}^{\,d}_i := \mathfrak{S}\big[\delta^d_i \overset{\star}{\boldsymbol{q}}_{i-1}\big], \qquad \delta^d_{i+1} := \delta^d_i \times \big\langle \boldsymbol{q}_{i-1}\big\rangle_{\rho,\gamma}. \tag{8.17}$$

For definiteness, the initial control programs $\boldsymbol{U}^{\,c}_0, \boldsymbol{U}^{\,d}_0, \boldsymbol{U}^{\,c}_1, \boldsymbol{U}^{\,d}_1$ are taken to be zero.

We introduced separate controls $\boldsymbol{U}^{\,c}_i, \boldsymbol{U}^{\,d}_i$ and bounds δ^c_i, δ^d_i to stress that the coder and decoder compute them independently. However, it easily follows from (8.15), (8.17) and induction on i that they in fact coincide,

$$\delta^d_i = \delta^c_{i-1}, \qquad \boldsymbol{U}^{\,d}_i = \boldsymbol{U}^{\,c}_i, \qquad i = 1, 2, \dots . \tag{8.18}$$

The second relation implies that the error in the state prognosis (8.13) is equal to the disturbance from (8.7),

$$\widehat{\boldsymbol{x}}^s_{i+1} = \boldsymbol{x}^s_{i+1} - \boldsymbol{\xi}_{s,i}. \tag{8.19}$$

Stabilizing properties of the offered controller are revealed by the following main result of the subsection.

PROPOSITION 8.5. *Suppose that the disturbance in the sth subsystem* (8.7) *satisfies* (8.9) *and that* (8.12) *holds. Then the above coder-decoder pair uniformly exponentially stabilizes this subsystem,*

$$|\boldsymbol{x}_i^s| \le \mathcal{K}_x\rho^i, \quad |\boldsymbol{U}_i^d| \le \mathcal{K}_u\rho^i \quad \forall i = 0,1,2,\dots \qquad \text{whenever} \quad |\boldsymbol{x}_0^s| \le K_0. \tag{8.20}$$

Here ρ is the parameter of the controller and the constants $\mathcal{K}_x, \mathcal{K}_u$ may depend on K_ξ from (8.9) *and K_0.*

The proof of this proposition consists of several lemmas. We start with rough estimates of concerned variables.

LEMMA 8.6. *The following inequalities hold for all $i = 1,2,\dots, h = 0,1,\dots,$ and $p \ge h$:*

$$\begin{aligned} &\delta_0\rho^{i-1} \le \delta_i^c \le \delta_0\gamma^{i-1}, \quad |\boldsymbol{U}_i^d| \le \|\mathfrak{S}\|\delta_{i-1}^c, \\ &|\boldsymbol{x}_p^s| \le \|A_{ss}\|^{r(p-h)}\Big[|\boldsymbol{x}_h^s| + K'_\xi\rho_\xi^h\Big] + K_\gamma\gamma^p|\mathfrak{I}(p,h)|, \end{aligned} \tag{8.21}$$

where $|\mathfrak{I}(p,h)|$ is the size of the set $\mathfrak{I}(p,h) := \{j = h,\dots,p-1 : j \ge 2 \;\&\; \boldsymbol{q}_{j-1} \ne \maltese\}$, $K'_\xi := K_\xi/(\|A_{ss}\|^r - 1)$, and the constant K_γ does not depend on $\boldsymbol{x}_0, i, h, p, K_\xi$.

Proof. The first formula is immediate from (8.15) and (8.16) since $\rho < 1 < \gamma$ by (8.12). The second one results from (8.17) and (8.18) since $|\overset{\star}{\boldsymbol{q}}| \le 1$ due to (8.16). To prove the last formula, we first note that $\boldsymbol{U}_j^d = 0\,\forall j \notin \mathfrak{I}(p,h), h \le j \le p-1$, by (8.17) and (8.16). Hence

$$\begin{aligned} |x_p^s| &\overset{(8.7)}{=\!=} \left|A_{ss}^{r(p-h)}x_h^s + \sum_{j=h}^{p-1} A_{ss}^{r(p-1-j)}\Big[\mathfrak{B}_s\boldsymbol{U}_j^d + \boldsymbol{\xi}_{s,j}\Big]\right| \\ &\overset{(8.9)}{\le} \|A_{ss}\|^{r(p-h)}|x_h^s| + \|\mathfrak{B}_s\|\|\mathfrak{S}\|\delta_0 \sum_{j\in\mathfrak{I}(p,h)} \underbrace{\|A_{ss}\|^{r(p-1-j)}}_{\le\gamma^{p-1-j}\text{by (8.12)}}\gamma^{j-2} \\ &\quad + K_\xi\sum_{j=h}^{p-1}\|A_{ss}\|^{r(p-1-j)} \underbrace{\rho_\xi^j}_{\le\rho_\xi^h\text{by (8.9)}} \\ &\le \|A_{ss}\|^{r(p-h)}|x_h^s| + \|\mathfrak{B}_s\|\|\mathfrak{S}\|\delta_0|\mathfrak{I}(p,h)|\gamma^{p-3} + K_\xi\rho_\xi^h\frac{\|A_{ss}\|^{r(p-h)}-1}{\|A_{ss}\|^r - 1}, \end{aligned}$$

which yields the last formula from (8.21). ∎

To prove stability, it suffices to show that δ_i^c are true bounds for the state prognosis $|\widehat{\boldsymbol{x}}_{i+1}^s| \le \delta_i^c$ for all large i. Indeed, then $|\boldsymbol{\varepsilon}_i| \le 1\ \forall i \approx \infty$ by (8.14). Hence (8.15) and (8.16) ensure that the bound δ_i^c and thus $\widehat{\boldsymbol{x}}_{i+1}^s$ decay exponentially $\delta_{i+1}^c = \rho\delta_i^c$ for $i \approx \infty$. Then so does the state $\boldsymbol{x}_{i+1}^s$ thanks to (8.9) and (8.19), i.e., the system is stable. We start by showing that even if the bound δ_i^c is incorrect for some i, it becomes true later.

LEMMA 8.7. *For any K_0, K_ξ, and i_0, there exists an integer $p_0 \ge i_0$ such that the bound δ_i^c is correct $|\widehat{\boldsymbol{x}}_{i+1}^s| \le \delta_i^c$ for at least one index $i \in [i_0 : p_0]$ whenever $|\boldsymbol{x}_0^s| \le K_0$ and* (8.9) *holds.*

Proof. The letter $\mathcal{K}$ (with possible indices) is used to denote a constant that depends on K_0, K_ξ but not $\boldsymbol{x}_0^s$ and $\boldsymbol{\xi}_{s,i}$. By putting $h := 0$ and the estimate $|\mathfrak{I}(p,h)| \le p - h$ into the last inequality from (8.21), we see that

$$|\boldsymbol{x}_{p+1}^s| \le \mathcal{K}^{(p)} \quad \text{whenever } |\boldsymbol{x}_0^s| \le K_0 \text{ and (8.9) holds.} \tag{8.22}$$

Now suppose that the bound δ_i^c is incorrect for all i from some interval $[i_0 : i_1]$ with the left end i_0. To estimate i_1, we note that $\mathfrak{I}(p,i) = \emptyset$ for $i := i_0 + 1$ and $p := i_1 + 1$ due to (8.14). So the last inequality from (8.21) yields

$$\begin{aligned}
|\boldsymbol{x}_{i_1+1}^s| &\le \|A_{ss}\|^{r(i_1-i_0)}\Big[\mathcal{K}^{(i_0+1)} + K'_\xi \rho_\xi^{i_0+1}\Big]; |\widehat{\boldsymbol{x}}_{i_1+1}^s| \overset{(8.19)}{=\!=} |\boldsymbol{x}_{i_1+1}^s - \boldsymbol{\xi}_{s,i_1}| \\
&\overset{(8.9)}{\le} \|A_{ss}\|^{r(i_1-i_0)}\Big[\mathcal{K}^{(i_0+1)} + K'_\xi \rho_\xi^{i_0+1}\Big] + K_\xi \rho_\xi^{i_1} \\
&\overset{\rho_\xi<1}{\le} \|A_{ss}\|^{r(i_1-i_0)}\Big[\mathcal{K}^{(i_0+1)} + K'_\xi\Big] + K_\xi \overset{1\le\|A_{ss}\|}{\le} \big(\mathcal{K}^{(i_0+1)} + K'_\xi + K_\xi\big)\|A_{ss}\|^{r(i_1-i_0)}.
\end{aligned}$$

At the same time, (8.15) and (8.16) entail that $\delta_{i+1}^c = \gamma \delta_i^c$ for $i \in [i_0 : i_1]$. So $\delta_{i_1}^c = \gamma^{i_1-i_0}\delta_{i_0}^c \ge \gamma^{i_1-i_0}\rho^{i_0-1}\delta_0$, where the last inequality is based on (8.21). Since the bound $\delta_{i_1}^c$ is incorrect, it follows that

$$1 \le \frac{|\widehat{\boldsymbol{x}}_{i_1+1}^s|}{\delta_{i_1}^c} \le \left(\frac{\|A_{ss}\|^r}{\gamma}\right)^{i_1-i_0} \frac{\mathcal{K}^{(i_0+1)} + K'_\xi + K_\xi}{\rho^{i_0-1}\delta_0}.$$

By invoking the second relation from (8.12), we conclude that

$$i_1 \le i_0 + \nu, \quad \text{where} \quad \nu := \left\lfloor \frac{\log_2(\mathcal{K}^{(i_0+1)} + K'_\xi + K_\xi) - (i_0-1)\log_2\rho - \log_2\delta_0}{\log_2\gamma - r\log_2\|A_{ss}\|} \right\rfloor.$$

So one may pick $p_0 := i_0 + 1 + \max\{\nu, 0\}$. The claim of the lemma remains true with the same p_0 if the interval $[i_0 : i_1]$ does not exist, because the bound $\delta_{i_0}^c$ is correct. ∎

The next lemma in fact completes the proof of Proposition 8.5.

LEMMA 8.8. *Suppose that $|\boldsymbol{x}_0^s| \le K_0$ and (8.9) holds. Whenever the bound δ_i^c becomes correct $|\widehat{\boldsymbol{x}}_{i+1}^s| \le \delta_i^c$, it is kept true afterwards, provided that $i \ge i_0$. Here i_0 is taken so that*

$$\frac{\rho_{\mathfrak{O}^s}}{\rho} + \frac{\|A_{ss}\|^r K_\xi}{\delta_0}\left(\frac{\rho_\xi}{\rho}\right)^i < 1 \qquad \forall i \ge i_0. \tag{8.23}$$

Remark 8.1. Such an i_0 exists due to the last inequality from (8.12).

Proof of Lemma 8.8. By (8.14), $|\boldsymbol{\varepsilon}_i| \le 1$. So (8.14), (8.15), and (8.16) imply that

$$\boldsymbol{\varepsilon}_i \in Q, \quad \boldsymbol{q}_i = \boldsymbol{q}^Q \quad \text{for some} \quad Q \in \{Q_1, \dots, Q_m\}; \quad \boldsymbol{U}_{i+1}^c = \mathfrak{S}\big[\delta_i^c \boldsymbol{q}_i\big], \quad \delta_{i+1}^c = \rho\delta_i^c, \tag{8.24}$$

where Q_j are the level sets of the quantizer. By (8.11), the third relation yields $\delta_i^c A_{ss}^r \boldsymbol{q}_i + \mathfrak{B}_s \boldsymbol{U}_{i+1}^c = 0$. Hence

$$\begin{aligned}
\big(\delta_{i+1}^c\big)^{-1}\big|\widehat{\boldsymbol{x}}_{i+2}^s\big| &\overset{(8.13)}{=\!=} \big(\delta_{i+1}^c\big)^{-1}\big|A_{ss}^r \boldsymbol{x}_{i+1}^s + \mathfrak{B}_s \boldsymbol{U}_{i+1}^c\big| = \big(\delta_{i+1}^c\big)^{-1}\big|A_{ss}^r \boldsymbol{x}_{i+1}^s - \delta_i^c A_{ss}^r \boldsymbol{q}_i\big| \\
&\overset{(8.19)}{=\!=} \big(\delta_{i+1}^c\big)^{-1}\big|A_{ss}^r\big[\widehat{\boldsymbol{x}}_{i+1}^s + \boldsymbol{\xi}_{s,i}\big] - \delta_i^c A_{ss}^r \boldsymbol{q}_i\big| \overset{(8.14)}{\le} \frac{\delta_i^c}{\delta_{i+1}^c}\big|A_{ss}^r\big[\boldsymbol{\varepsilon}_i - \boldsymbol{q}_i\big]\big| + \big(\delta_{i+1}^c\big)^{-1}\big|A_{ss}^r \boldsymbol{\xi}_{s,i}\big|.
\end{aligned}$$

It follows from (8.10) and the first two relations in (8.24) that $\left|A_{ss}^r\left[\boldsymbol{\varepsilon}_i - \boldsymbol{q}_i\right]\right| \le \rho_{\mathfrak{O}^s}$. We proceed by invoking (8.9) and the last relation from (8.24), along with the first inequality from (8.21),

$$\left(\delta_{i+1}^c\right)^{-1}\left|\widehat{\boldsymbol{x}}_{i+2}^s\right| \le \frac{\rho_{\mathfrak{O}^s}}{\rho} + \|A_{ss}\|^r K_\xi \frac{\rho_\xi^i}{\delta_{i+1}^c} \le \frac{\rho_{\mathfrak{O}^s}}{\rho} + \frac{\|A_{ss}\|^r K_\xi}{\delta_0}\left(\frac{\rho_\xi}{\rho}\right)^i \overset{(8.23)}{<} 1.$$

Thus the bound δ_{i+1}^c is true, which completes the proof. □

Proof of Proposition 8.5. Consider the number p_0 from Lemma 8.7, where i_0 is taken from Lemma 8.8. By these lemmas, the bound δ_i^c is true $|\widehat{\boldsymbol{x}}_{i+1}^c| \le \delta_i^c \forall i \ge p_0$. Then $\delta_{i+1}^c = \rho\delta_i^c\ \forall i \ge p_0$ thanks to (8.14), (8.15), and (8.16). With regard to the first relation from (8.21), we see that $\delta_i^c \le \delta_0\gamma^i\ i \le p_0, \delta_i^c = \delta_{p_0}^c\rho^{i-p_0} \le \delta_0\big(\gamma/\rho\big)^{p_0}\rho^i\ i \ge p_0$ and so $\delta_i^c \le \mathcal{K}_\delta\rho^i\ \forall i$, where $\mathcal{K}_\delta := \delta_0\big(\gamma/\rho\big)^{p_0}$. This and the second formula from (8.21) give the second inequality in (8.20). To prove the first one, we note that

$$\left|\boldsymbol{x}_{i+1}^c\right| \overset{(8.19)}{=\!=} \left|\widehat{\boldsymbol{x}}_{i+1}^c + \boldsymbol{\xi}_{s,i}\right| \overset{(8.9)}{\le} \delta_i^c + K_\xi\rho_\xi^i \overset{(8.12)}{\le} \mathcal{K}_x^0\rho^{i+1} \quad \forall i \ge p_0,$$
$$\text{where} \quad \mathcal{K}_x^0 := \rho^{-1}\left(\mathcal{K}_\delta + K_\xi\right).$$

For $i \le p_0 + 1$, inequality (8.22) yields $|\boldsymbol{x}_i^c| \le \mathcal{K}^{(i)} \le \mathcal{K}' := \max_{j=0,\dots,p_0+1}\mathcal{K}^{(j)}$. Thus the first inequality in (8.20) does hold with $K_x := \max\{\mathcal{K}_x^0; \mathcal{K}'\rho^{-p_0-1}\}$. □

8.3. Existence of a contracted quantizer and deadbeat stabilizer. To implement the proposed scheme, the state $\boldsymbol{x}_i^s$ should be determined at a certain site, which may be only the site of a sensor. Then the quantized value $\boldsymbol{q}_i = \mathfrak{O}^s(\boldsymbol{x}_i^s)$ should be communicated for r units of time to the decoder. Since the capacities of the channels serving the sensors are limited, we are interested in minimizing the number b of communicated bits. This number is that $b = \lceil\log(m+1)\rceil$ required to describe a quantized value (including the alarm one ✠) for the m-level quantizer $\mathfrak{O}^s$. Thus finding an r-contracted quantizer with the minimum number of levels m is of interest.

Now we establish tight lower and upper bounds for this number.

LEMMA 8.9. *For any m-level r-contracted* (8.10) *quantizer, $m > |\det A_{ss}|^r$. This inequality is sufficient up to a polynomial factor. In other words, there is a polynomial $\varphi_s(\cdot)$ (depending on A_{ss}) such that for any $r = 1, 2, \dots$ there exists an r-contracted quantizer $\mathfrak{O}^s$ with the number of levels*

$$m_s \le \varphi_s(r)|\det A_{ss}|^r. \tag{8.25}$$

Proof. Necessity. Due to (8.10), $|\det A_{ss}|^r\,\mathbf{mes}\,(Q_i) = |\det A_{ss}|^r\,\mathbf{mes}\,[Q_i - q^{Q_i}] = \mathbf{mes}\left[A_{ss}^r(Q_i - q^{Q_i})\right] \le \rho_{\mathfrak{O}^s}^{n_s}\,\mathbf{mes}\,B_0^1 < \mathbf{mes}\,B_0^1$. Summing over i gives $m > |\det A_{ss}|^r$.

Sufficiency. Note first that whenever the claim is true for two matrices A'_{ss} and A''_{ss}, it is also true for the block matrix $\left(\begin{smallmatrix} A'_{ss} & 0 \\ 0 & A''_{ss}\end{smallmatrix}\right)$. By employing the canonical Jordan form of A_{ss}, this reduces the proof to the case where the matrix is a real Jordan block. Let n_s denote its size, λ its eigenvalue, and $\omega := |\lambda|$. As follows from, e.g., [26, Lemma 3.1, p. 64], $\Xi(r) := \omega^{-r}\varphi(r)^{-1}A_{ss}^r \to 0$ as $r \to \infty$ for some polynomial $\varphi(\cdot)$. So $\|\Xi(r)\| < \rho < 1$ for $r \approx \infty$. Here $\|\cdot\|$ is the operator norm associated with the norm $|\boldsymbol{z}| := \max_i |z_i|$ in $\mathbb{R}^{n_s} = \{\boldsymbol{z} = (z_1, \dots, z_{n_s})\}$. Multiplying the polynomial $\varphi(r)$ by a sufficiently large scalar factor makes the inequality $\|\Xi(r)\| < \rho$ true for all r. Now consider the uniform quantizer $\mathfrak{O}^s$ partitioning the unit ball B_0^1 into $m_s := k^{n_s}$

balls Q_i of radius $\frac{1}{k}$, where $k := \lceil \omega^r \varphi(r) \rceil$. The centroid q^{Q_i} is the center of the ball Q_i. Then

$$\|\Xi(r)\| < \rho \Rightarrow \Xi(r)[Q_i - q^{Q_i}] \subset \rho[Q_i - q^{Q_i}] = \frac{\rho}{k} B_0^1 \Rightarrow A^r[Q_i - q^{Q_i}] \subset \rho \frac{\omega^r \varphi(r)}{k} B_0^1 \subset \rho B_0^1.$$

Thus the quantizer is r-contracted. It remains to note that

$$m_s = k^{n_s} \le [\omega^r \varphi(r) + 1]^{n_s} \le 2^{n_s - 1} \Big([\omega^r \varphi(r)]^{n_s} + 1 \Big) = 2^{n_s - 1} \Big(|\det A_{ss}|^r \varphi(r)^{n_s} + 1 \Big)$$
$$\overset{1 \le |\det A_{ss}|}{\le} 2^{n_s - 1} |\det A_{ss}|^r \Big[\varphi(r)^{n_s} + 1 \Big].$$

Thus (8.25) does hold with $\varphi_s(r) := 2^{n_s - 1}[\varphi(r)^{n_s} + 1]$. □

When all subsystems are equipped with the proposed controllers, the stability of a particular subsystem may be violated by the controllers serving the other subsystems since the control is common. To avoid this, it suffices to choose the controls in such a way that they influence only the subsystem for which they are intended. For the basic (unit) sample period, this may be impossible. Now we show that this can be done if the sample period r is properly increased: the controls generated for a given subsystem do not affect the other ones at times $t = i \cdot r, i = 0, 1, \dots$.

Common controls give rise to another trouble. As was remarked, the sth coder will be implemented at the sites of sensors observing the sth subsystem. To compute the states $\boldsymbol{x}_i^s = \boldsymbol{x}^s(\tau_i)$ utilized by this coder, not only the observations but also controls must be known at these sites. However, the sth coder may know the control only partly. It is aware of its own summand in the overall control, which is the sum of the controls generated for all subsystems. At the same time, it is not able to determine the summands based on the modes $\boldsymbol{x}^j$ invisible at its site. To overcome this obstacle, it suffices to note that the controls must be known for only n times t preceding τ_i. So it suffices to ensure that all controllers produce zero controls at these times.

Now we show that deadbeat stabilizers with the above properties do exist.

LEMMA 8.10. *Whenever $r > n$, a deadbeat stabilizer for the sth subsystem exists. Moreover, it can be chosen so that it generates control programs $\boldsymbol{U} = (\boldsymbol{u}_0, \dots, \boldsymbol{u}_{r-1})$ vanishing since the time $t = n$, i.e., $\boldsymbol{u}_n = \dots = \boldsymbol{u}_{r-1} = 0$, and does not disturb the other subsystems, i.e., $\mathfrak{B}_j \boldsymbol{U} = 0$ for $j \ne s$, $\boldsymbol{U} = \mathfrak{S}\boldsymbol{x}^s$, and any $\boldsymbol{x}^s$.*

Proof. By (8.11), a deadbeat stabilizer is the right inverse to the operator $\mathfrak{D}\boldsymbol{U} := -A_{ss}^{-r} \mathfrak{B}_s \boldsymbol{U}$. In (8.2), $\mathfrak{B}\boldsymbol{U}$ is the state to which the control program $\boldsymbol{U}$ drives the system (3.1) at time $t = r$ from $\boldsymbol{x}(0) = 0$. By assumption, this system has no stable modes. So it is controllable thanks to Assumption 4.2. It follows that the operator $\mathfrak{B}$ is onto. Moreover, $\mathfrak{B}|_M$ is onto, where $M := \{\boldsymbol{U} : \boldsymbol{u}_n = \dots = \boldsymbol{u}_{r-1} = 0\}$. Indeed for any $\boldsymbol{x}$, it suffices to pick the control program $\boldsymbol{u}_0, \dots, \boldsymbol{u}_{n-1}$ that drives the system from 0 at $t = 0$ to $A^{n-r}\boldsymbol{x}$ at $t = n$ and extend it by zeros to form $\boldsymbol{U} \in M$. Then evidently $\mathfrak{B}\boldsymbol{U} = \boldsymbol{x}$. Now consider $\boldsymbol{x}$ such that in (8.4) all blocks are zeros except for $\boldsymbol{x}^s \in \mathbb{R}^{n_s}$. Since this block can be chosen arbitrarily, it follows that the operator $\mathfrak{B}_s$ maps $L := \{\boldsymbol{U} \in M : \mathfrak{B}_j \boldsymbol{U} = 0 \, \forall j \ne s\}$ onto $\mathbb{R}^{n_s}$. So evidently does $\mathfrak{D}$. It remains to define $\mathfrak{S}$ as the right inverse to $\mathfrak{D}|L$. □

8.4. Stabilization of the entire system. Now we study the set of all subsystems in their actual connection. So the disturbance in (8.7) is given by (8.8). We pick $r > n$ and suppose that the following claims hold for any s:

A1. *The block $\boldsymbol{x}^s(\tau_i)$ of the state can be determined at any time $\tau_i = i \cdot r$ at a certain site* (called the sth site)*;*

A2. *The sth coder from subsection* 8.2 *is implemented at this site;*

A3. *There is a way to communicate the quantized value* $\boldsymbol{q}_i^s$ *generated by the sth coder at the step* **c.3** *to the decoder site during the time interval* $[\tau_i : \tau_{i+1})$.

Note that the sth coder is driven by only the sequence of states $\boldsymbol{x}_i^s = \boldsymbol{x}^s(\tau_i)$. So A1 makes A2 possible. In A3, considered is the site where the actual decoder (see Figure 3.1) must be situated. The sth decoder from subsection 8.2 is implemented at this site for all s. This is possible since it is driven only by the sequence of quantized values $\boldsymbol{q}_i^s, i = 0, 1, \dots$. Each decoder produces its own sequence of controls $\boldsymbol{U}_i^d = \big[\boldsymbol{u}^s(ir), \boldsymbol{u}^s(ir+1), \dots, \boldsymbol{u}^s(ir+r-1)\big]$. These sequences are summed over all decoders to produce the control sequence acting upon the plant:

$$\boldsymbol{u}(t) := \boldsymbol{u}^1(t) + \boldsymbol{u}^2(t) + \cdots + \boldsymbol{u}^d(t).$$

To complete the description of the coders, a quantizer, deadbeat stabilizer, and parameters r, γ, ρ should be chosen for each coder. The parameter $r > n$ is already picked. For any subsystem s, the quantizer and stabilizer are taken from Lemmas 8.9 and 8.10, respectively. The parameter $\gamma = \gamma_s$ is chosen to satisfy the second relation from (8.12). As for the third relation, it is indefinite under the circumstances since ρ_ξ from (8.9) is not given. So now we pick the parameter $\rho = \rho_s$ in another way. It is chosen successively for $s = 1, 2, \dots, d$ and so that

(8.26)
$$\rho_1 > \rho_{\mathfrak{Q}^1}, \quad \rho_2 > \max\{\rho_{\mathfrak{Q}^2}; \rho_1\}, \quad \rho_3 > \max\{\rho_{\mathfrak{Q}^3}; \rho_2\}, \dots, 1 > \rho_d > \max\{\rho_{\mathfrak{Q}^d}; \rho_{d-1}\},$$

where $\rho_{\mathfrak{Q}^s}$ is taken from (8.10). Stabilizing properties of this control scheme are described by the following.

PROPOSITION 8.11. *Suppose that assumptions* A1–A3 *hold. Then the proposed networked controller uniformly exponentially stabilizes the entire system* (3.1) *at the rate* $\mu = \rho_d^{1/r}$.

We preface the proof of this proposition with a simple technical fact.

LEMMA 8.12. *Suppose that a trajectory of the system* (3.1) *satisfies the estimates*

$$|\boldsymbol{x}_i| \le \mathcal{K}_x \rho^i, \quad |\boldsymbol{U}_i| \le \mathcal{K}_u \rho^i, \qquad i = 0, 1, 2, \dots, \tag{8.27}$$

where $\boldsymbol{x}_i := \boldsymbol{x}(\tau_i), \rho \in [0, 1), \tau_i := i \cdot r$, *and* $\boldsymbol{U}_i := [\boldsymbol{u}(\tau_i), \boldsymbol{u}(\tau_i + 1), \dots, \boldsymbol{u}(\tau_i + r - 1)]$. *Then* (4.2) *holds, where* $\mu := \rho^{1/r}$ *and the constants* K_x, K_u *are determined by* $\mathcal{K}_x, \mathcal{K}_u, \rho$ *(for a given system).*

Proof. Whenever $t \in [\tau_i : \tau_{i+1})$, we have $\rho^i = \mu^{\tau_i} = \mu^{\tau_i - t}\mu^t \le \mu^{-r}\mu^t = \rho^{-1}\mu^t$. So $|\boldsymbol{u}(t)| \le |\boldsymbol{U}_i| \le \mathcal{K}_u\rho^i \le \mathcal{K}_u\rho^{-1}\mu^t$, i.e., the second inequality from (4.2) does hold. We denote $\varkappa := 1 + \|A\|$. Then

$$\begin{aligned}
|\boldsymbol{x}(t)| = \left| A^{t-\tau_i}\boldsymbol{x}(\tau_i) + \sum_{j=\tau_i}^{t-1} A^{t-1-j} B\boldsymbol{u}(j) \right| &\le \|A\|^{t-\tau_i}|\boldsymbol{x}(\tau_i)| + \sum_{j=\tau_i}^{t-1} \|A\|^{t-1-j}\|B\||\boldsymbol{u}(j)| \\
&\le \left[\|A\|^{t-\tau_i} + \|B\| \sum_{j=\tau_i}^{t-1} \|A\|^{t-1-j} \right] \times [|\boldsymbol{x}(\tau_i)| + |\boldsymbol{U}_i|] \\
&\le \left[\varkappa^r + \|B\| \sum_{j=\tau_i}^{\tau_{i+1}-1} \varkappa^{\tau_{i+1}-1-j} \right] [\mathcal{K}_x + \mathcal{K}_u]\, \rho^i,
\end{aligned}$$

where $\rho^i \le \rho^{-1}\mu^t$. By the index substitution $j := \tau_i + \nu$ in the last sum, we see that (4.2) is true. □

Proof of Proposition 8.11. Suppose that $|\boldsymbol{x}(0)| \le K_0$, where K_0 is given. The controls $\boldsymbol{u}^s$ with $s \ge 2$ do not disturb the first block $\boldsymbol{x}_i^1 := \boldsymbol{x}^1(\tau_i)$ of the state at times $\tau_i = i \cdot r$ by the choice of the deadbeat stabilizers. So this block $\boldsymbol{x}_i^1, i = 0, 1, \dots,$ evolves just as in the first subsystem (8.7) driven by the first coder and decoder and perturbed by the noise $\boldsymbol{\xi}_{1,i}$, which is zero by (8.8). Then Proposition 8.5 and the first inequality from (8.26) imply that the first subsystem $s = 1$ is uniformly exponentially stabilized (8.20) at the rate $\rho := \rho_1$. This and (8.8) imply that the noise $\boldsymbol{\xi}_{2,i}$ in the second subsystem (8.7) (where $s = 2$) exponentially decays (8.9) at the rate ρ_1. Now we retrace the above arguments with respect to this subsystem and employ the second relation from (8.26). As a result, we establish that this subsystem is stabilized at the rate ρ_2, i.e., (8.20) holds for $s = 2$ and $\rho := \rho_2$. By continuing likewise, we see that for any s, inequalities (8.20) are true with $\rho := \rho_s$ and proper constants $\mathcal{K}_x, \mathcal{K}_u$ (depending on s) whenever $|\boldsymbol{x}(0)| \le K_0$. Since $\rho_d \ge \rho_s \, \forall s$ by (8.26), it follows that (8.27) holds with $\rho := \rho_d$ and some constants $\mathcal{K}_x, \mathcal{K}_u$ depending on K_0. Lemma 8.12 and Definition 4.3 complete the proof. □

8.5. Analysis of Assumptions A1 and A3. Our next goal is to show that these assumptions stated in the previous subsection are satisfied whenever $r > 2n$ and (5.2) holds. This in fact will complete the proof of Theorem 5.1. In this subsection, we perform the first step to this end.

We start with assumption A1. By 1 of Proposition 8.1, the unobservable subspace (4.3) $L_j^{-obs} = L_j^-$ of the jth sensor is composed of several blocks $\boldsymbol{x}^s, s \notin O_j$, of the state (8.4). These blocks do not affect its outputs y_j, whereas all other blocks $\boldsymbol{x}^s, s \in O_j$, can be determined from these outputs.

LEMMA 8.13. *Whenever $r > 2n$, assumption* A1 *holds. For any s, the site of any sensor j with $O_j \ni s$ can be taken as the sth site in* A1.

Proof. We recall that the deadbeat stabilizers were taken from Lemma 8.10. So they produce control programs $\boldsymbol{U} = (\boldsymbol{u}_0, \dots, \boldsymbol{u}_{r-1})$ with zeros $\boldsymbol{u}_i = 0$ at any place $i \ge n$. For $r > 2n$, this means that the corresponding control sequence $u(t), t = 0, 1, \dots,$ vanishes $u(t) = 0$ for at least n times t preceding each $\tau_i = i \cdot r, i = 0, 1, \dots$. It remains to invoke the remarks prefacing Lemma 8.10. □

Now we turn to an analysis of A3. We recall that in A3, the value $\boldsymbol{q}_i^s$ is given by an m_s level quantizer $\mathfrak{Q}^s$. Description of such a value (which may equal ✠) requires $b_s = \lceil \log_2(m_s + 1) \rceil$ bits. This number may exceed the capacity of the channel that serves any particular sensor j observing the block $\boldsymbol{x}^s$. So we employ all such channels. Specifically, the following scheme of transmission $\boldsymbol{q}_i^s$ to the decoder site is used for each subsystem $s = 1, \dots, d$:

- T1. The sth coder is implemented at the sites of all sensors j observing the state $\boldsymbol{x}^s$, i.e., such that $s \in O_j$;
- T2. By employing a common encoding rule, the value $\boldsymbol{q}_i^s$ produced at each of these sites is then transformed into a b_s-bit sequence $\boldsymbol{\beta}_i^s = (\beta_1, \beta_2, \dots, \beta_{b_s})$ of binary digits $\beta_\nu = 0, 1$;
- T3. By applying a common rule, this sequence $\boldsymbol{\beta}$ is split into several subsequences $\boldsymbol{\beta}_i^{s,j}$ each associated with one of the concerned sensors j, i.e., such that $s \in O_j$;
- T4. Each of these sensors j sends only its own subsequence $\boldsymbol{\beta}_i^{s,j}$ over the attached channel to the decoder site;
- T5. At the decoder site, the required value $\boldsymbol{q}_i^s$ is reconstructed by reversing the rules from T2 and T3.

We assume that the rules from T2 and T3 do not change as i progresses and are known at the decoder site. Furthermore, the rule from T2 is lossless: the value $\boldsymbol{q}_i^s$ can be reconstructed from $\boldsymbol{\beta}$. This makes T5 possible.

The above scheme means that several binary words $\boldsymbol{\beta}_i^{s,j}, s \in O_j$, must be transmitted over the common jth channel during any time interval $[\tau_i : \tau_{i+1})$ of duration $r-1$. By 2 of Assumption 4.1, this is possible if the total length of these words does not exceed $b_j^-(r-1)$. Now we denote by b_{sj} the number of bits in $\boldsymbol{\beta}_i^{s,j}$ whenever $s \in O_j$ and put $b_{sj} := 0$ otherwise. Summarizing, we arrive at the following lemma.

LEMMA 8.14. *Assumption* A3 *is satisfied whenever there exist nonnegative integer numbers $b_{sj}, s = 1, \dots, d$, $j = 1, \dots, k$, such that the following relations hold:*

$$\sum_{j=1}^{k} b_{sj} = b_s = \lceil \log_2(m_s + 1) \rceil \ \forall s, \quad \sum_{s=1}^{d} b_{sj} \le b_j^-(r-1) \ \forall j, \quad \textit{and} \quad b_{sj} = 0 \ \textit{whenever} \quad s \notin O_j. \tag{8.28}$$

Here k and d are the numbers of sensors and subsystems, respectively, m_s is the number of levels for the r-contracted quantizer taken from Lemma 8.9*, whereas $b_j^-(\cdot)$ and O_j are taken from* 2 *of Assumption* 4.1 *and* 1 *of Proposition* 8.1*, respectively.*

8.6. Inconstructive sufficient conditions for stabilizability. These conditions are immediate from Proposition 8.11 combined with Lemmas 8.13 and 8.14.

PROPOSITION 8.15. *Suppose that the following system of relations*

$$\log_2 |\det A_{ss}| < \sum_{j=1}^{k} \alpha_{sj} \ \forall s, \quad \sum_{s=1}^{d} \alpha_{sj} < \mathfrak{c}_j \ \forall j, \quad \alpha_{sj} \ge 0 \ \forall s, j, \quad \alpha_{sj} = 0 \ \textit{whenever} \ s \notin O_j \tag{8.29}$$

is solvable in real numbers α_{sj}. Here A_{ss} is taken from 2 *of Proposition* 8.1 *and $\mathfrak{c}_j$ is the transmission capacity* (4.1) *of the jth channel. Then the system* (3.1), (3.2) *is uniformly exponentially stabilizable.*

Proof. It suffices to show that for all large r, the system (8.28) is solvable in nonnegative integers b_{sj}. Indeed such an r can be clearly chosen so that $r > 2n$. Then Lemmas 8.13 and 8.14 ensure that assumptions A1 and A3 from subsection 8.4 hold, whereas A2 $\Leftarrow$ A1. In A2, the parameters of the sth coder are chosen as was indicated in that subsection. Proposition 8.11 completes the proof. □

We note first that in (8.28), the first relation can be replaced by the inequality

$$\sum_{j=1}^{k} b_{sj} \ge \lceil \log_2(m_s + 1) \rceil. \tag{8.30}$$

Indeed, if after this the system is solvable, than a solution for the original relation can be obtained by properly decreasing the nonnegative integers b_{sj}. Specifically, they are decreased to satisfy the first relation from (8.28), which may only enhance the second relation and keep the third relation true.

We are going to show that a solution is given by $b_{sj} := \lfloor r \cdot \alpha_{sj} \rfloor$, provided $r \approx \infty$.

Indeed the third relation in (8.28) follows from the last one in (8.29). Furthermore,

$$\frac{1}{r}\sum_{j=1}^{k} b_{sj} \xrightarrow{r\to\infty} \sum_{j=1}^{k} \alpha_{sj} \overset{(8.29)}{=\!=} \log_2|\det A_{ss}| + \varkappa_s, \text{ where } \varkappa_s > 0,$$

$$\frac{1}{r}\lceil\log_2(m_s+1)\rceil \le \frac{1}{r}\left[\log_2(m_s+1)+1\right]$$

$$\overset{(8.25)}{\le} \frac{1}{r}\left[\log_2\left(\varphi_s(r)|\det A_{ss}|^r+1\right)+1\right] \xrightarrow{r\to\infty} \log_2|\det A_{ss}|.$$

It follows that (8.30) does hold for all $r \approx \infty$. Likewise,

$$\frac{1}{r}\sum_{s=1}^{d} b_{sj} \xrightarrow{r\to\infty} \sum_{s=1}^{d} \alpha_{sj} \overset{(8.29)}{=\!=} \mathfrak{c}_j - \eta_j, \text{ where } \eta_j > 0, \qquad \frac{b_j^-(r-1)}{r} \xrightarrow{(4.1)} \mathfrak{c}_j \text{ as } r\to\infty.$$

Thus the second relation from (8.28) is also true for all $r \approx \infty$. □

8.7. Convex duality and a criterion for the system (8.29) to be solvable. Now we in fact perform the final step in the proof of the sufficiency part of Theorem 5.1 by justifying the following claim.

PROPOSITION 8.16. *The system* (8.29) *is solvable in real numbers* α_{sj} *if and only if* 2 *of Theorem* 5.1 *holds.*

Then by invoking Proposition 8.15, we arrive at the following corollary.

COROLLARY 8.1. *Whenever the system* (3.1) *has no stable modes,* 2 *of Theorem* 5.1 *implies* 1.

We preface the proof of Proposition 8.16 with a useful reformulation of 2 from Theorem 5.1 in terms of the decomposition from Proposition 8.1.

LEMMA 8.17. *Along with the sets* O_j *from* 1 *of Proposition* 8.1, *consider all their unions* $O = \bigcup_{j\in J} O_j$, *where* J *ranges over all groups of sensors. (The union of the empty group of sets* O_j *is included and interpreted as the empty set.) Then* 2 *of Theorem* 5.1 *is true if and only if for any such a union* $O \neq [1:d]$,

$$\sum_{s\notin O} \log_2|\det A_{ss}| < \sum_{j:O_j\not\subset O} \mathfrak{c}_j. \tag{8.31}$$

Proof. Due to 1 of Proposition 8.1, the sets (5.1) $L = \bigcap_{j\in J} L_j^-$ have the form

$$L = \{\boldsymbol{x} : \boldsymbol{x}^s = 0 \ \forall s \in O\}, \qquad \text{where} \quad O = \bigcup_{j\in J} O_j.$$

So (8.5) implies $\det A_L = \prod_{s\notin O} \det A_{ss}$. Hence the left-hand sides in (5.2) and (8.31) coincide. It remains to note that so do the right-hand ones since in (5.2) $J(L) = \{j : O_j \subset O\}$ owing to (4.3), (5.1), and 1 of Proposition 8.1.

Proof of Proposition 8.16. *Necessity.* Suppose that (8.29) has a solution α_{sj}. Then

$$\sum_{s\notin O} \log_2|\det A_{ss}| \overset{(8.29)}{<} \sum_{s\notin O}\sum_{j=1}^{k} \alpha_{sj} = \sum_{j=1}^{k}\sum_{s\notin O} \alpha_{sj} \overset{(8.29)}{=\!=} \sum_{j:O_j\not\subset O}\sum_{s\notin O} \alpha_{sj} \le \sum_{j:O_j\not\subset O}\sum_{s=1}^{d} \alpha_{sj}$$

$$\overset{(8.29)}{<} \sum_{j:O_j\not\subset O} \mathfrak{c}_j,$$

i.e., (8.31) holds. By Lemma 8.17, so does (5.2).

Sufficiency. Now suppose that 2 of Theorem 5.1 is true. By Lemma 8.17, this means that (8.31) holds for the union O of any sets O_j, provided $O \neq [1:d]$. It should be shown that (8.29) is solvable in real numbers α_{sj}.

Suppose the contrary. Then the following convex polyhedra in the space of matrices $\alpha = (\alpha_{sj})$ are disjoint:

$$C_1 := \left\{ \alpha : \log_2 |\det A_{ss}| < \sum_{j=1}^{k} \alpha_{sj} \ \forall s \right\},$$

$$C_2 := \left\{ \alpha : \sum_{s=1}^{d} \alpha_{sj} < \mathfrak{c}_j \ \forall j, \alpha_{sj} \geq 0 \ \forall s, j, \alpha_{sj} = 0 \text{ if } s \notin O_j \right\}.$$

Hence they are separated by a hyperplane: there exists a nonzero matrix $\gamma = (\gamma_{sj})$ such that

$$\inf_{\alpha \in C_1} \sum_{s,j} \gamma_{sj} \alpha_{sj} \geq \sup_{\alpha \in C_2} \sum_{s,j} \gamma_{sj} \alpha_{sj}. \tag{8.32}$$

The definition of C_1 implies that

$$\inf_{\alpha \in C_1} \sum_{s,j} \gamma_{sj} \alpha_{sj} = \sum_{s=1}^{d} \inf_{(\alpha_j): \sum_{j=1}^{k} \alpha_j > \log_2 |\det A_{ss}|} \sum_{j=1}^{k} \gamma_{sj} \alpha_j.$$

The infimum on the right is that of a linear functional over a half-space of (α_j) bounded by a hyperplane with the normal $(1, \dots, 1)$. This infimum is finite only if the functional is generated by a vector colinear with the normal. So $\gamma_{sj} = \theta_s \ \forall j$ for some $\theta_s \geq 0$ and $\sum_s \theta_s > 0$. It follows that

$$\inf_{\alpha \in C_1} \sum_{s,j} \gamma_{sj} \alpha_{sj} = \sum_{s=1}^{d} \theta_s \log_2 |\det A_{ss}|.$$

At the same time, the definition of C_2 implies that

$$\sup_{\alpha \in C_2} \sum_{s,j} \gamma_{sj} \alpha_{sj} = \sup_{\substack{\alpha_{sj} \geq 0, \sum_s \alpha_{sj} < \mathfrak{c}_j, \\ s \notin O_j \Rightarrow \alpha_{sj} = 0}} \sum_{sj} \theta_s \alpha_{sj} = \sum_{j=1}^{k} \max_{\alpha_s \geq 0, \sum_s \alpha_s \leq \mathfrak{c}_j} \sum_{s \in O_j} \theta_s \alpha_s$$

$$= \sum_{j=1}^{k} \mathfrak{c}_j \max_{s \in O_j} \theta_s.$$

By (8.32), the cone $K := \{\theta = (\theta_1, \dots, \theta_d) \in \mathbb{R}^d : \theta_s \geq 0\}$ contains a nonzero solution of the inequality

$$\sum_{s=1}^{d} \theta_s \log_2 |\det A_{ss}| \geq \sum_{j=1}^{k} \mathfrak{c}_j \max_{s \in O_j} \theta_s. \tag{8.33}$$

This cone can be partitioned into a finite number of convex polyhedral subcones such that the right-hand side of (8.33) is linear on any subcone. It follows that (8.33) must

be satisfied on some extreme ray of some subcone. Any of them is bounded by a finite number of hyperplanes, each described by an equation of the form either $\theta_\nu = 0$ or $\theta_\mu = \theta_\nu$, where $\nu \neq \mu$ and $\nu, \mu \in O_j$ for some j. This implies [21, p. 104] that the extreme ray is described by a finite system of such equations, which determines its solution uniquely up to multiplication by a scalar. It is easy to see that the solution of such a system looks as follows: $\theta_s = \theta$ whenever $s \notin \mathcal{O}$, and $\theta_s = 0$ otherwise. Here $\mathcal{O} \subset [1:d]$ is some set, $\mathcal{O} \neq [1:d]$. For vectors on the above extreme ray, we have $\theta > 0$, and (8.33) shapes into

$$\sum_{s \notin \mathcal{O}} \log_2 |\det A_{ss}| \geq \sum_{j: O_j \not\subset \mathcal{O}} \mathfrak{c}_j.$$

Changing $\mathcal{O} := O := \bigcup_{j: O_j \subset \mathcal{O}} O_j$ does not alter the right-hand side, possibly increases the left-hand one, and thus keeps the inequality true, in violation of (8.31). The contradiction obtained proves that the system (8.29) is solvable in real numbers α_{sj}. □

8.8. Proof of Theorem 5.1. It was shown in section 7 that $1 \Rightarrow 2$. The converse $2 \Rightarrow 1$ is given by Corollary 8.1 if the system (3.1) has no stable modes. Thus it remains to drop the last requirement and prove (5.3).

To achieve the first objective, we consider the system (3.1) with both unstable and stable modes that satisfies 2. It is clear that it suffices to stabilize only its unstable part,

(8.34)
$$x_+(t+1) = A_+ x_+(t) + \pi_+ B u(t), \qquad x_+(0) := \pi_+ x_0 \in L_+, \qquad y_+(t) = C x_+(t).$$

Here $L_+ := M_{unst}(A)$ and $L_- := M_{st}(A)$ are the invariant subspaces of A related to the unstable and stable parts of its spectrum, π_+ and π_- are the projectors onto L_+ parallel to L_- and vice versa, respectively, and $A_\pm := A|_{L_\pm}$. Thanks to the second relation from (4.3), 2 still holds for the system (8.34). By the foregoing, this system can be uniformly exponentially stabilized by some networked controller. While constructing it, we employ the parameter $r > 2n$. Now we apply this controller to the primal system (3.1). In doing so, the proof of possibility of T1 from subsection 8.5 must be revisited. Indeed the sth coder can be implemented at the jth sensor site (where $s \in O_j$) only if $\boldsymbol{x}^s(\tau_i), \tau_i := i \cdot r$ can be determined there. Formerly this was done on the base of the past measurements from (8.34). Now we must employ the observations from (3.1). This is possible due to (4.3) and 1 of Proposition 8.1 since the dynamics of the system (3.1) is free $u(t) = 0$ at least n time steps before τ_i.

By Definition 4.3, there exists $\mu \in [0, 1)$ such that whenever a constant K_0 is given and $|\boldsymbol{x}_0| \leq K_0$,

$$|\pi_+ \boldsymbol{x}(t)| \leq K_x^+ \mu^t, \quad |\boldsymbol{u}(t)| \leq K_u \mu^t \qquad \forall t = 0, 1, 2, \dots .$$

The evolution of $\boldsymbol{x}_-(t) := \pi_- \boldsymbol{x}(t)$ is described by the first two equations from (8.34), where the index $_+$ is switched to $_-$. Since the operator A_- is stable and the controls $\boldsymbol{u}(t)$ exponentially decay, so do the states $|\boldsymbol{x}_-(t)| \leq K_x^- \rho^t$. Here $\rho \in (0, 1)$ does not depend on K_0. Since $|\boldsymbol{x}(t)| = |\boldsymbol{x}_-(t) + \boldsymbol{x}_+(t)| \leq |\boldsymbol{x}_-(t)| + |\boldsymbol{x}_+(t)|$, increasing $\mu := \max\{\mu, \rho\}$ yields (4.2). Definitions 4.3 and 4.4 complete the proof of 1 from Theorem 5.1.

It remains to justify (5.3). To this end, we note that the transformation $z(t) := \mu^{-t} x(t), v(t) := \mu^{-t} u(t)$ establishes a one-to-one correspondence between the trajectories $\{x(t), u(t)\}$ and $\{z(t), v(t)\}$ of the systems given by (3.1) and the equation

$z(t+1) = \mu^{-1}Az(t) + \mu^{-1}Bv(t)$, respectively. We equip the latter with the sensors $\widetilde{y}_j = C_j z, j = 1, \dots, k$. It easily follows from Definitions 4.3 and 4.4 that the initial system is uniformly exponentially stabilizable at a rate $\mu' \in (0, \mu)$ if and only if the second system is uniformly exponentially stabilizable. By applying the $1 \Leftrightarrow 2$ part of Theorem 5.1 to it, we get

$$-\mathbf{dim}L \cdot \log_2 \mu + \log_2 |\det A|_L|$$

$$< \sum_{j \not\in J(L)} \mathfrak{c}_j \ \forall L \in \mathfrak{L} \Rightarrow \log_2 \mu > \max_{L \in \in \mathfrak{L}} \frac{1}{\mathbf{dim}L} \left(\log_2 |\det A|_L| - \sum_{j \not\in J(L)} \mathfrak{c}_j \right).$$

To arrive at (5.3), it remains to note that the rate of exponential stabilizability μ^0 is the infimum of all such μ.

9. Comments on Assumption 4.3. Now we explain why it has such a big impact on the controller design. We also briefly discuss ideas underlying such a design in the case where this assumption does not hold.

To start with, we illuminate the role of Assumption 4.3. A "subsystem" (arising from (8.4)) is said to be in a *simple relation* with the jth sensor if it either does affect this sensor or its state can be uniquely determined from the sensor outputs. The simplest case in stabilization of a multiple sensor system is where the system can be decomposed into independent subsystems each in a simple relation with any sensor. In general, this is impossible. As was shown, a nontrivial Jordan block may form a barrier to decomposition into independent subsystems. The example (4.5) proves that it may be still worse: the system cannot be disintegrated into (even dependent) subsystems each in simple relations with sensors. Assumption 4.3 in fact describes when this worst case does not occur.

So if this assumption is violated, unavoidable is the situation where some sensor partly observes some subsystem: its state cannot be determined on the site of this sensor though the sensor signals contain information about this state. Then an additional problem arises: how to utilize this information in the coding-decoding scheme for stabilization purposes? As will be shown, the answer requires the revision of some basic principles on which the design of such schemes was based up to now. In this paper, this is omitted due to space limitations.

To come into details, we pick natural $\mathfrak{c}$ and real $\lambda \approx \sqrt{2}^{3\mathfrak{c}}, \lambda < \sqrt{2}^{3\mathfrak{c}}$ numbers and revert to the example (4.5),

$$\boldsymbol{x}(t+1) = \lambda \boldsymbol{x}(t) + \boldsymbol{u}(t) \in \mathbb{R}^2, \quad y_1(t) = x_1(t), \quad y_2(t) = x_2(t), \quad y_3(t) = x_1(t) - x_2(t), \quad t = 0, 1, \dots, \tag{9.1}$$

where $\boldsymbol{x} = (x_1, x_2)$. Any of three channels transmits $\mathfrak{c}$ bits per unit time without delays and losses, i.e., $\mathfrak{c}_1 = \mathfrak{c}_2 = \mathfrak{c}_3 = \mathfrak{c}$. The necessary conditions for stabilizability (5.2) take the form $\lambda < \sqrt{2}^{3\mathfrak{c}}$ and are satisfied.

Assumption 4.3 clearly does not hold: one of the sensors observes a certain subsystem only partly for any decomposition of the system (9.1).For example, consider the natural decomposition $\boldsymbol{x} = (x_1, x_2)$, where x_1 and x_2 are interpreted as the states of the subsystems. They are in simple relations with the first and second sensors. However, they are not in such relations with the third one. Indeed the state x_i influences its outputs $y_3 = x_1 - x_2$ but cannot be determined on the basis of them. Moreover,

the only linear coordinate (i.e., function) of the state that can be determined on the site of the third sensor is its output y_3 (up to a scalar factor). Likewise, the first and second sensors permit us to find only x_1 and x_2, respectively. This conclusion holds for any decomposition.

Now we are going to show that though *the system* (9.1) *is stabilizable, it cannot be stabilized by a controller with the following features*, which are characteristic for the most known relevant controllers:

1. Not only the "mode" y_i but also its upper (maybe incorrect) bound δ_i is determined at the sensor site;
2. The state $\boldsymbol{x}$ and these bounds in fact constitute the state of the closed-loop system, which is time-invariant;
3. At the decoder site, the information about the "mode" $y_i(t)$ comes to its quantized scaled value $e_i(t) = \mathfrak{Q}_i[\delta_i(t)^{-1}y_i(t)]$ given by a static quantizer $\mathfrak{Q}_i(\cdot)$ with convex level sets and the number m_i of levels matching $m_i \le 2^{\mathfrak{c}} - 1$ the channel capacity;
4. The next bound $\delta_i(t+1)$ is determined from $\delta_i(t)$ and the knowledge of whether $e_i(t) = \maltese$ or not;
5. Whenever all bounds are true $\delta_i(t) \ge |y_i(t)|\,\forall i$, they remain true afterwards.

It is clear that these features mainly concern the coding algorithm.

Remark 9.1. In [25], a specific class of networked controllers was proposed for stabilization of multiple sensor systems. Since those controllers satisfy 1–5, they are unable to stabilize some stabilizable systems, e.g., the system (9.1). In [25], it was established when a system can yet be stabilized by a controller from this class. The corresponding results are formulated as criteria for "stabilizability" with no reference to the class. This does not seem a good idea because then some actually stabilizable systems are in fact classified as "unstabilizable."

LEMMA 9.1. *Let* $\mathfrak{c} \ge 2$ *and a networked controller satisfying* 1–5 *be given. Then the closed-loop system* (9.1) *is neither stable nor dissipative:* $\limsup_{t\to\infty}\sup_{\boldsymbol{x}_0\in B_0^\delta}|\boldsymbol{x}(t)| = \infty$ *for all* $\delta > 0$ *and initial bounds* $\delta_i^0 > 0$.

Proof. By 2 and 4, $\delta_i(t+1) = \mathcal{D}_i[\delta_i(t)]$ whenever $e_i(t) \ne \maltese$. We are going to estimate $\mathcal{D}_i(\cdot)$ from below. Due to 3, any quantizer $\mathfrak{Q}_i$ is related to a partition of the interval $[-1,1]$ into m_i subintervals (level sets) $\Delta_1^{(i)},\dots,\Delta_{m_i}^{(i)}$, $\Delta_j^{(i)} = [\alpha_j^{(i)},\beta_j^{(i)}]$. Since $m_i \le 2^{\mathfrak{c}} - 1$, one of them has the length $\beta_{j^i}^{(i)} - \alpha_{j^i}^{(i)} \ge 2\cdot 2^{-\mathfrak{c}}$. Now we pick $\delta > 0$, set the initial bounds $\delta_1(0) = \delta_2(0) = \delta, \delta_3(0) = 2\delta$, and note that all initial states from the segment $S := \{\boldsymbol{x}_0 = (\delta\alpha_{j^1}^{(1)} + \theta, \delta\alpha_{j^2}^{(2)} + \theta) : 0 \le \theta \le 2\delta 2^{-\mathfrak{c}}\}$ give rise to common outputs for each quantizer $i = 1,2,3$ at $t = 0$. So they give rise to a common control $\boldsymbol{u}(0) = (u_1, u_2)$. For all these states, the above initial bounds are correct. Then 5 ensures that for $i = 1, 2$

$$\delta_i(1) = \mathcal{D}_i[\delta_i(0)] = \mathcal{D}_i[\delta] \ge |y_i(1)| = |\lambda x_i(0) + u_i|.$$

Here $\lambda x_i(0) + u_i$ runs over an interval of length $\ge 2\lambda\delta 2^{-\mathfrak{c}}$ as $\boldsymbol{x}_0$ ranges over S. Hence

$$\mathcal{D}_i(\delta) \ge \lambda\delta 2^{-\mathfrak{c}} \tag{9.2}$$

for $i = 1, 2$. This inequality is extended on $i = 3$ by putting $\delta_1(0) = \delta_3(0) = \delta, \delta_2(0) = 2\delta, S := \{\boldsymbol{x}_0 = (\delta\alpha_{j^1}^{(1)} + \theta, \delta\alpha_{j^1}^{(1)} - \delta\alpha_{j^3}^{(3)}) : 0 \le \theta \le 2\delta 2^{-\mathfrak{c}}\}$ and retracing the above arguments.

Now we suppose that for some $\delta_i^0 > 0$ and $\delta > 0$ the conclusion of the lemma violates $c := \sup|\boldsymbol{x}(t)| < \infty$, where sup is over $\boldsymbol{x}_0 \in B_0^\delta$ and all t. By decreasing δ, one

can ensure that $\delta < \min\{\delta_1^0, \delta_2^0, \frac{1}{2}\delta_3^0\}$. Then for all initial states $\boldsymbol{x}_0 \in B_0^\delta$, the bounds δ_i are correct for $t = 0$. Thanks to 4 and 5, they remain correct for all t and common for all $\boldsymbol{x}_0 \in B_0^\delta$. Then (9.2) yields $\delta_i(t) \geq \delta \left(\frac{\lambda}{2^{\mathfrak{c}}}\right)^t$. Here $\lambda \approx \sqrt{2}^{3\mathfrak{c}}$. So $\delta_i(t) \to \infty$ as $t \to \infty$. As a result, the interval $[-2c, 2c]$ is covered by at most two intervals $\delta_i(t)\Delta_j^{(i)}$ for each $i = 1, 2, 3$, provided t is large enough. Since $|\boldsymbol{x}(t)| \leq c \Rightarrow |y_i(t)| \leq 2c, i = 1, 2, 3$, this and 3 mean that for $\boldsymbol{x}_0 \in B_0^\delta$ each $\mathfrak{Q}_i$ in fact acts as a binary quantizer at any large time. Thus the decoder receives in fact no more than one bit of information about processes with $\boldsymbol{x}_0 \in B_0^\delta$ via each channel. By treating three channels as one and retracing the arguments from the proof of Lemma 7.1 (see also [20, 25, 11]), we arrive at the necessary conditions for dissipativity $\lambda^2 \leq 2^3 \Leftrightarrow \lambda \leq 2^{3/2}$. At the same time, $\mathfrak{c} \geq 2$ and $\lambda \approx \sqrt{2}^{3\mathfrak{c}} > 2^{3/2}$. The contradiction obtained proves the lemma. □

Now we show that nevertheless *the system* (9.1) *is stabilizable*. The stabilizing controller lacks properties 2 and 3. It employs a 2-periodic quantization scheme applied to not only scaled but also shifted observations. We give only a sketch of the proof and focus on the main part of the stabilization process. It starts when a correct upper bound of the state norm is found via successive multiplying by a sufficiently large factor (see, e.g., subsection 8.2).

Each coder computes a number $\delta_i(t)$, and the decoder duplicates these computations. The meaning of the numbers and actions of the controller are different for odd and even times t. To explain them, we put $N := 2^{\mathfrak{c}}$.

At *odd* times, $\delta_1(t) = \delta_2(t), \delta_3(t) = \delta_1(t) + \delta_2(t)$, $\boldsymbol{x}(t) \in M(t) := \{\boldsymbol{x} : -\delta_i(t) \leq y_i < \delta_i(t), i = 1, 2\}$ and

1. For $i = 1, 2$, the ith coder determines which interval $[\mu_{j'}^{(i)}, \mu_{j'+1}^{(i)})$, $\mu_{j'}^{(i)} := j'\frac{2\delta_i(t)}{N}, j' = 0, \ldots, N-1$, contains $\overline{y}_i(t) := y_i(t) + \delta_i(t)$, and notifies the decoder about its serial number $j' = j^{(i)}$;
2. For $i = 3$, the coder (a) finds which interval $[\omega_j, \omega_{j+1}), \omega_j := j\frac{\delta_i(t)}{N}$ contains $\overline{y}_i(t)$, (b) notifies the decoder which subinterval $[\omega_j^{(\nu)}, \omega_j^{(\nu+1)})$, $\omega_j^{(\nu)} := \omega_j + \nu\frac{\omega_{j+1}-\omega_j}{N-2}$, $\nu = 0, \ldots, N-3$, contains $\overline{y}_i(t)$, and (c) uses the remaining bit to make the decoder aware of whether j is odd or even.

It is easy to see that $(j^{(1)} + j_*^{(2)})\frac{\delta_3(t)}{N} < \overline{y}_3(t) < (j^{(1)} + j_*^{(2)} + 2)\frac{\delta_3(t)}{N}$, where $j_*^{(2)} := N - j^{(2)} - 1$. So either $j = j^{(1)} + j_*^{(2)}$ or $j = j^{(1)} + j_*^{(2)} + 1$. Hence j can be found from $j^{(1)}, j^{(2)}$ and the information from (c).

3. From $j^{(1)}$, the decoder finds the strip $\{\boldsymbol{x} : y_1 \in -\delta_1(t) + [\mu_{j^{(1)}}^{(1)}, \mu_{j^{(1)}+1}^{(1)})\}$ that contains $\boldsymbol{x}(t)$. By reconstructing j and using ν, it finds another such strip $\{\boldsymbol{x} : y_3 \in -\delta_3(t) + [\omega_j^{(\nu)}, \omega_j^{(\nu+1)})\}$. Then the decoder selects a control driving the system from the center of the intersection of the strips to zero.

As a result, $\boldsymbol{x}(t+1) \in M(t+1) := \{\boldsymbol{x} : -\boldsymbol{\varepsilon}' \leq y_1 < \boldsymbol{\varepsilon}', -\boldsymbol{\varepsilon}'' \leq y_3 < \boldsymbol{\varepsilon}''\}$, where $\boldsymbol{\varepsilon}' := \lambda\delta_1(t)N^{-1} \approx \delta_1(t)2^{3/2\mathfrak{c}}2^{-\mathfrak{c}} = \delta_1(t)2^{\mathfrak{c}/2} > \delta_1(t)$ and $\boldsymbol{\varepsilon}'' := \lambda\delta_1(t)[N(N-2)]^{-1} \approx \delta_1(t)2^{\mathfrak{c}/2}(2^{\mathfrak{c}} - 1)^{-1} \leq 2\delta_1(t)2^{-\mathfrak{c}/2} < \delta_3(t)$ for $\mathfrak{c} \geq 2$. Thus, for one step, the domain $M(t) \ni \boldsymbol{x}(t)$ is stretched in one direction and tightened in the other.

4. The ith coder defines the next number δ_i as the bound for y_i when $\boldsymbol{x} \in M(t+1)$, i.e., $\delta_1(t+1) := \lambda N^{-1}\delta_i(t), \delta_3(t+1) := \frac{1}{2}\lambda\delta_3(t)[N(N-2)]^{-1}$, and $\delta_2(t+1) := \lambda\delta_2(t)N^{-1}[1 + (N-2)^{-1}]$.

At *even* times, $\delta_3(t) = \frac{\delta_1(t)}{N-2}, \delta_2(t) = \delta_1(t) + \delta_3(t), \boldsymbol{x}(t) \in M(t) = \{\boldsymbol{x} : -\delta_i(t) \leq y_i < \delta_i(t), i = 1, 3\}$, and

5. The operation 1 is carried out by the first and third coders $i = 1, 3$, and 2 is

done by the second one $i = 2$ with ω_j altered: $\omega_j := j\frac{4\delta_2(t)}{N}$.

The definitions of $j^{(1)}, j^{(3)}$ and j from 1 and 2, respectively, imply that $j^{(1)}, j^{(3)} \in [0 : N-1]$ and

$$-\delta_2(t) + \tfrac{2\delta_2(t)}{N}\left[j^{(1)} + \tfrac{N-1-j^{(3)}-j^{(1)}}{N-1}\right] < y_2 < -\delta_2(t) + \tfrac{2\delta_2(t)}{N}\left[j^{(1)} + \tfrac{N-1-j^{(3)}-j^{(1)}}{N-1}\right] + \tfrac{2\delta_2(t)}{N} \quad \Downarrow$$

$$-\delta_2(t) + \tfrac{2\delta_2(t)}{N}(j^{(1)} - 1) < y_2 < -\delta_2(t) + \tfrac{2\delta_2(t)}{N}(j^{(1)} + 2),$$

whereas $-\delta_2(t) + \frac{4\delta_2(t)}{N} j \le y_2 < -\delta_2(t) + \frac{4\delta_2(t)}{N}(j+1)$. It follows that either $j = j^{(1)}/2$ or $j = j^{(1)}/2 - 1$ if $j^{(1)}$ is even, and either $j = (j^{(1)} - 1)/2$ or $j = (j^{(1)} - 1)/2 + 1$ if $j^{(1)}$ is odd. Thus the number j can be found by the decoder on the basis of $j^{(1)}$ and the information from 2.c).

6. The decoder finds two strips containing $\boldsymbol{x}(t)$. The first of them is $\{\boldsymbol{x} : y_3 \in -\delta_3(t) + [\mu^{(3)}_{j^{(3)}}, \mu^{(3)}_{j^{(3)}+3})\}$, and the second one is $\{\boldsymbol{x} : y_2 \in -\delta_2(t) + [\omega^{(\nu)}_j, \omega^{(\nu+1)}_j)\}$. Then the decoder selects a control driving the system from the center of the intersection of the strips to zero.

As a result, $\boldsymbol{x}(t+1) \in M'(t+1) := \{\boldsymbol{x} : -\boldsymbol{\varepsilon}' \le y_3 < \boldsymbol{\varepsilon}', -\boldsymbol{\varepsilon}'' \le y_2 < \boldsymbol{\varepsilon}''\}$, where $\boldsymbol{\varepsilon}' := \lambda\delta_3(t)N^{-1}$ and $\boldsymbol{\varepsilon}'' := 1/2[\omega^{(\nu+1)}_j - \omega^{(\nu)}_j] = 2\lambda\delta_2(t)[N(N-2)]^{-1} = 2\lambda\delta_3(t)(N-1)[N(N-2)]^{-1} \le 3\lambda\delta_3(t)N^{-1}$ (since $N = 2^{\mathfrak{c}}$ and $\mathfrak{c} \ge 2$). The set $M'(t+1)$ is covered by the square $M(t+1) := \{\boldsymbol{x} : |x_1|, |x_2| \le 3\lambda\delta_3(t)N^{-1}\}$.

7. The numbers δ_i are updated so that $\delta_1 = \delta_2$ become the half length of the edge of $M(t+1)$ and $\delta_3 = \delta_1 + \delta_2$, i.e., $\delta_1(t+1) := 3\lambda\delta_1(t)[N(N-2)]^{-1}, \delta_2(t+1) := 3\lambda\delta_2(t)[N(N-1)]^{-1}, \delta_3(t+1) := 6\lambda\delta_3(t)N^{-1}$.

Now observe that for two steps, the square $M(t) \ni \boldsymbol{x}(t)$ with the edge $2\delta_1(t)$ (where t is odd) is transformed into the square $M(t+2)$ with the edge $2\delta_1(t) \times \frac{3\lambda^2}{N^2(N-2)}$. So if $\frac{3\lambda^2}{N^2(N-2)} < 1 \Leftrightarrow \lambda < 2^{3/2\mathfrak{c}}\sqrt{1/3(1-2^{1-\mathfrak{c}})}$, the system is stabilized. The last inequality is a bit worse than the necessary condition for stabilizability $\lambda < 2^{3/2\mathfrak{c}}$. This gap can be discarded by increasing the sample period to r time units. Indeed, this "transforms" λ into λ^r, N into N^r, and the sufficient condition $\lambda < 2^{3/2\mathfrak{c}}\sqrt{1/3(1-2^{1-\mathfrak{c}})}$ for stabilizability into $\lambda^r < 2^{3/2r\mathfrak{c}}\sqrt{1/3(1-2^{1-r\mathfrak{c}})} \Leftrightarrow \lambda < 2^{3/2\mathfrak{c}}[1/3(1-2^{1-r\mathfrak{c}})]^{2/r}$. The latter reduces to $\lambda < 2^{3/2\mathfrak{c}}$ as $r \to \infty$.

Thus even for a very simple system, violation of Assumption 4.3 complicates the coding-decoding scheme. We consider the study of the general case where this assumption does not hold as a topic of separate research.

REFERENCES

[1] J. Baillieul, *Feedback designs for controlling device arrays with communication channel bandwidth constraints*, in ARO Workshop on Smart Structures, Univeristy Park, Pennsylvania, 1999.

[2] R. W. Brockett and D. Liberzon, *Quantized feedback stabilization of linear systems*, IEEE Trans. Automat. Control, 45 (2000), pp. 1279–1289.

[3] S. Dasgupta, *Control over bandlimited communication channels: Limitations on stabilizability*, in Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, 2003, pp. 176–181.

[4] D. F. Delchamps, *Stabilizing a linear system with quantized state feedback*, IEEE Trans. Automat. Control, 35 (1990), pp. 916–924.

[5] A. Dembo, T. M. Cover, and J. A. Thomas, *Information-theoretic inequalities*, IEEE Trans. Inform. Theory, 37 (1991), pp. 1501–1518.

[6] N. G. Dokuchaev and A. V. Savkin, *A new class of hybrid dynamical systems: State estimators with bit-rate constraints*, Internat. J. Hybrid Systems, 1 (2001), pp. 33–50.
[7] N. Elia, *Design of hybrid systems with guaranteed performance*, in Proceedings of the 39th IEEE Conference on Decision and Control, Sydney, Australia, 2000, pp. 421–425.
[8] N. Elia and S. Mitter, *Stabilization of linear systems with limited information*, IEEE Trans. Automat. Control, 46 (2001), pp. 1384–1400.
[9] F. R. Gantmacher, *The Theory of Matrices*, Vol. 1, Chelsea, New York, 1959.
[10] S. Guiasu, *Information Theory with Applications*, McGraw–Hill, New York, 1977.
[11] J. Hespanha, A. Ortega, and L. Vasudevan, *Towards the control of linear systems with minimum bit-rate*, in Proceedings of the 15th International Symposium on Mathematical Theory of Networks and Systems, Notre Dame, IN, 2002; available online at http://www.ece.ucsb.edu/hespanha/published.html.
[12] H. Ishii and T. Basar, *Remote control of LTI systems over networks with state quantization*, in Proceedings of the 41st IEEE Conference on Decision and Control, Las Vegas, 2002, pp. 830–835.
[13] H. Ishii and B. Francis, *Limited data rate in control systems with networks*, in Lecture Notes in Control and Inform. Sci., M. Thoma and M. Morari, eds., Springer, Berlin, 2002.
[14] R. Jain, T. Simsek, and P. Varaiya, *Control under communication constaints*, in Proceedings of the 41st IEEE Conference on Decision and Control, Las Vegas, 2002, pp. 3209–3216.
[15] D. Liberzon, *A note on stabilization of linear systems using coding and limited communication*, in Proceedings of the 41st IEEE Conference on Decision and Control, Las Vegas, 2002, pp. 836–841.
[16] D. Liberzon, *Stabilizing a nonlinear system with limited information feedback*, in Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, 2003, pp. 182–186.
[17] A. S. Matveev and A. V. Savkin, *The problem of LQG optimal control via a limited capacity communication channel*, Systems Control Lett., 53 (2004), pp. 51–64.
[18] A. S. Matveev and A. V. Savkin, *Stabilization of multisensor networked control systems with communication constraints*, in Proceedings of 5th Asian Control Conference, Melbourne, Australia, 2004, pp. 1916–1923.
[19] G. N. Nair and R. J. Evans, *Stabilization with data-rate-limited feedback: Tightest attainable bounds*, Systems Control Lett., 41 (2000), pp. 49–56.
[20] G. N. Nair and R. J. Evans, *Mean square stabilisability of stochastic linear systems with data rate constraints*, in Proceedings of 41st IEEE Conference on Decision and Control, Las Vegas, 2002, pp. 1632–1637.
[21] M. J. Panik, *Fundamentals of Convex Analysis: Duality, Separation, Representation, and Resolution*, Kluwer Academic Publishers, Boston, 1993.
[22] I. R. Petersen and A. V. Savkin, *Multi-rate stabilization of multivariable discrete-time linear systems via a limited capacity communication channel*, in Proceedings of the 40th IEEE Conference on Decision and Control, Orlando, FL, 2001, pp. 304–309.
[23] A. V. Savkin and I. R. Petersen, *Set-valued state estimation via a limited capacity communication channel*, IEEE Trans. Automat. Control, 48 (2003), pp. 676–680.
[24] D. J. Stiwell and B. E. Bishop, *Platoons of underwater vehicles*, IEEE Control Systems Magazine, 20 (2000), pp. 45–52.
[25] S. C. Tatikonda, *Control Under Communication Constraints*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 2000.
[26] R. Varga, *Iterative Matrix Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1965.
[27] W. S. Wong and R. W. Brockett, *Systems with finite communication bandwidth constraints—*II: *Stabilization with limited information feedback*, IEEE Trans. Automat. Control, 44 (1999), pp. 1049–1053.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 2, pp. 618–645

2-REGULARIZED NYQUIST CRITERION IN LINEAR CONTINUOUS-TIME PERIODIC SYSTEMS AND ITS IMPLEMENTATION*

JUN ZHOU† AND TOMOMICHI HAGIWARA†

Abstract. First, by using the 2-regularized determinant technique for Hilbert–Schmidt operators, the computation formula, infinite-product convergence, and analyticity of the 2-regularized determinant of the modified harmonic state operator in finite-dimensional linear continuous-time periodic (FDLCP) systems are derived in this paper. Second, based on these results, a 2-regularized Nyquist criterion is established for asymptotic stability analysis of a class of FDLCP systems for the first time. Third, a numeric implementation algorithm for the 2-regularized Nyquist criterion is also proposed via the staircase truncation on the harmonic transfer operator of the FDLCP system concerned. Finally, to illustrate the results of this paper, asymptotic stability of the lossy Mathieu differential equation is investigated.

Key words. continuous-time periodic system, Nyquist criterion, determinant of Hilbert–Schmidt operators, implementation, convergence

AMS subject classifications. 43A32, 65N12, 65F40, 93D20

DOI. 10.1137/S0363012902416900

1. Introduction.

Stability analysis is a difficult topic in finite-dimensional linear continuous-time periodic (FDLCP) systems [3], [8], [14], [19], [27], which are encountered in many engineering applications. For instance, the flapping dynamics of helicopter rotors [7], [26] and rolling motion of ships in waves [1] can be related to FDLCP models. Other examples include robot arms moving along periodic trajectories and electromechanical oscillation in AC generators [22]. Different types of (closed-loop) stability of FDLCP systems are discussed via various methods in the literature. Absolute stability of FDLCP systems with nonlinearities satisfying integral quadratic constraints is dealt with in [16] and [28] by the cutting plane algorithm and the Hamiltonian approach, respectively, while input/output stability and Youla-style parameterization of stabilizing controllers are discussed in [6] via the graph representation theory. As for asymptotic stability analysis of FDLCP systems, the Floquet theorem [18] completes the task by testing the eigenvalues of its monodromy matrix that is hard to find. Asymptotic stability has also been examined by a Lyapunov method [4] and the harmonic analysis [31]. Perturbation methods to study stability in FDLCP systems can be found in [20]. Nyquist-type stability criteria have also been considered in the FDLCP cases; for example, two generalized Nyquist criteria are suggested in [15] and [26]. The former is an integral-operator-based Nyquist criterion, while the latter is given in terms of the Hill-determinant of the infinite-dimensional harmonic transfer operator of an FDLCP system. However, due to the infinite-dimensionality and various convergence issues in the Hill-determinant, the validity of the latter generalized Nyquist criterion [26] remains as an open problem in general situations. Similar comments also apply to the former criterion. As for integral-operator modeling of periodic systems, we refer the readers to [2] for a general idea.

*Received by the editors November 1, 2002; accepted for publication (in revised form) March 10, 2005; published electronically September 12, 2005.
http://www.siam.org/journals/sicon/44-2/41690.html
†Department of Electrical Engineering, Kyoto University, Kyotodaigaku-Katsura, Nishikyo-ku, Kyoto 615-8510, Japan (zhouj@kuee.kyoto-u.ac.jp, hagiwara@kuee.kyoto-u.ac.jp).

A crucial observation for establishing Nyquist-type stability criteria in more general FDLCP systems is that the harmonic transfer operators of FDLCP systems are Hilbert–Schmidt operators under mild assumptions. Hence the validity of the 2-regularized determinant on the harmonic transfer operators [5] can be justified. It should be pointed out that the harmonic transfer operators do not belong to the trace class [9] in general so that the standard operator determinant cannot be validated. Thus, developing a Nyquist-type criterion based on the 2-regularized determinant technique provides us with a natural and much stronger tool in stability analysis in the FDLCP field. In connection with the usual determinant defined on trace class operators, it is worth mentioning that in sampled-data systems, which are periodic (but not included in FDLCP systems) if the signal behavior both at the sampling instants and intersamples is considered, a Nyquist criterion regarding internal stability has been recovered with the transfer operator defined via lifting technique [13] under the assumption that the transfer operator is a trace class operator.

In this paper, using the 2-regularized determinant on Hilbert–Schmidt operators, we first derive some interesting analytic properties of the 2-regularized determinant of what we call the modified harmonic state operator of an FDLCP system. These results have not been explicitly discussed in the literature to the authors' best knowledge and constitute a significant contribution to this study. Second, based on these properties, a 2-regularized Nyquist criterion is established for asymptotic stability of a class of FDLCP systems. This Nyquist criterion is necessary and sufficient, and makes it possible for us to investigate the closed-loop asymptotic stability via the open-loop FDLCP system and the 2-regularized determinant of the corresponding harmonic return difference operator, similar to what we do in the LTI (linear time-invariant) continuous-time case. In spite of the success that the criterion applies to a big class of practical FDLCP systems, however, it brings another problem. Namely, it is nontrivial to implement the criterion numerically because of the 2-regularized determinant on the infinite-dimensional harmonic return difference operator. To resolve the problem, the staircase truncation [30] is applied to the harmonic transfer operator. It is shown that under mild assumptions the truncation convergence can be ensured, and the 2-regularized Nyquist criterion can be implemented via only finite-dimensional computations to any degree of accuracy, and the truncation size can be estimated readily through simple computations.

The following is the outline of this paper. Section 2 gives preliminaries to FDLCP systems, their harmonic state operators and transfer operators, the Toeplitz transformation of periodic functions, and operator determinants. Properties about the 2-regularized determinants of the modified harmonic transfer operators are also derived. In section 3, the 2-regularized Nyquist criterion is established, while its implementation is considered via truncation in section 4. The lossy Mathieu equation is studied to illustrate the results in section 5. Proofs of lemmas, if any, are given in appendices.

In this paper, $\|\cdot\|$ denotes the Euclidean norm of a vector and the norm of a matrix induced by this norm. l_2 is the set of all infinite-dimensional vectors $\underline{x}$ such that $\|\underline{x}\|_{l_2}^2 := \sum_{-\infty}^{+\infty} \|[\underline{x}]_m\|^2 < \infty$, where $[\underline{x}]_m$ is the mth (vector) entry of $\underline{x}$. $\|\cdot\|_{l_2/l_2}$ is the l_2-induced norm. $L_2[0,h]$ is the linear space of all vector measurable functions x defined on $[0,h]$ such that $\|x(\cdot)\|_{L_2[0,h]} := [\int_0^h \|x(t)\|^2\,dt]^{1/2} < \infty$. $F(\cdot) \in L_2[0,h]$ means that F is an h-periodic matrix function, each element of which belongs to $L_2[0,h]$ when its domain is restricted to $[0,h]$. This expression is also used for other function sets defined over $[0,h]$. $\mathbf{C}$ is the field of all complex numbers, and $\mathbf{Z}$ is the ring of all integers.

2. Preliminaries to FDLCP systems and operator determinants. In this section we first review facts about FDLCP systems [29], and then the 2-regularized determinant [5] for Hilbert–Schmidt operators. In particular, we derive properties about the harmonic state operator of FDLCP systems in the 2-regularized determinant sense.

Consider the FDLCP system given by

$$G : \begin{cases} \dot{x} = A(t)x + B(t)u, \\ y = C(t)x, \end{cases} \tag{2.1}$$

where $A(t)$, $B(t)$, and $C(t)$ are h-periodically time-varying matrices. The transition matrix of (2.1) with the initial time t_0 is denoted by $\Phi(t, t_0)$. By the Floquet theorem [14], [18], if $A(t) \in L_2[0, h]$, then $\Phi(t, t_0)$ is continuous with respect to t and has a Floquet factorization $\Phi(t, t_0) = P(t, t_0)e^{Q(t-t_0)}$, where $P(t, t_0)$ is absolutely continuous in t, nonsingular and h-periodic in t and t_0, and Q is a constant matrix. Moreover, the system is asymptotically stable if and only if the eigenvalues of Q lie in the open left-half plane. Without loss of generality, we assume $t_0 = 0$.

Now we review the Toeplitz transformation of periodic functions. Expand $X(t) \in L_2[0, h]$ to its Fourier series $\sum_{m=-\infty}^{+\infty} X_m e^{jm\omega_h t}$ with $\omega_h := 2\pi/h$. The Toeplitz transformation on $X(t)$, denoted by $\mathcal{T}\{X(t)\}$, maps $X(t)$ onto a doubly infinite-dimensional block Toeplitz operator [26] (or block Laurent operator [10, p. 564]) of the form

$$\mathcal{T}\{X(t)\} := \begin{bmatrix} \ddots & \vdots & \vdots & \vdots & \iddots \\ \cdots & X_0 & X_{-1} & X_{-2} & \cdots \\ \cdots & X_1 & X_0 & X_{-1} & \cdots \\ \cdots & X_2 & X_1 & X_0 & \cdots \\ \iddots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} =: \underline{X}. \tag{2.2}$$

We further define $\underline{A} := \mathcal{T}\{A(t)\}$, $\underline{B} := \mathcal{T}\{B(t)\}$, $\underline{C} := \mathcal{T}\{C(t)\}$, $\underline{P} := \mathcal{T}\{P(t, 0)\}$, $\underline{\hat{B}} := \mathcal{T}\{P^{-1}(t, 0)B(t)\}$, $\underline{\hat{C}} := \mathcal{T}\{C(t)P(t, 0)\}$, $\underline{Q} := \operatorname{diag}[\ldots, Q, Q, Q, \ldots]$, and

$$\underline{E}(s) := \operatorname{diag}[\ldots, \varphi_{-2}(s)I, \varphi_{-1}(s)I, \varphi_0(s)I, \varphi_1(s)I, \varphi_2(s)I, \ldots], \tag{2.3}$$

where $\varphi_m(s) := s + jm\omega_h, m \in \mathbf{Z}, s \in \mathbf{C}$. It follows that $\underline{E}(s) = \underline{E}(j0) + s\underline{I}$, where $\underline{I} := \mathcal{T}\{I\}$.

We need the following function sets to validate the Fourier analysis and Toeplitz transformation operations involved (see [29] and [32] for details) and simplify our statements.

$$L_{\mathrm{PCD}}[0, h] := \left\{ f(t) : \begin{matrix} f(t) \text{ is piecewise continuous and} \\ \text{differentiable at a.e. } t \in [0, h] \end{matrix} \right\},$$

$$L_{\mathrm{PCC}}[0, h] := \left\{ f(t) : \begin{matrix} f(t) \text{ is piecewise continuous and its Fourier series} \\ \text{expansion is convergent to } f(t_0) \text{ for a.e. } t_0 \in [0, h] \end{matrix} \right\},$$

$$L_{\mathrm{CAC}}[0, h] := \left\{ f(t) : \begin{matrix} f(t) \text{ is continuous and the Fourier series} \\ \text{expansion of } f(t) \text{ is absolutely convergent} \end{matrix} \right\} \subset L_{\mathrm{PCC}}[0, h].$$

Here PCD stands for piecewise continuous and differentiable, and PCC is short for piecewise continuous and convergent, while CAC is the abbreviation for continuous and absolutely convergent.

Now we state the similarity transformation formulas and eigenvalues of FDLCP systems in terms of the Toeplitz transformation of the system matrices and the Floquet factorization of the transition matrix [29], [31]. To facilitate the statements, let $l_E := \{\underline{x} \in l_2 : \underline{E}(j0)\underline{x} \in l_2\}$. Then, l_E is a proper subset of l_2 and dense in l_2 [29].

LEMMA 2.1. *In the* FDLCP *system* (2.1)*, assume that* $A(t) \in L_{\mathrm{PCD}}[0,h]$ *and* $B(t), C(t) \in L_{\mathrm{PCC}}[0,h]$*. Then,* l_E *is* $\underline{P}$*- and* $\underline{P}^{-1}$*-invariant. Also, the unbounded operators* $\underline{P}(\underline{E}(j0) - \underline{Q})\underline{P}^{-1}$ *and* $\underline{E}(j0) - \underline{A}$ *are densely defined on* l_2 *(or equivalently, well defined on the subset* $l_E \subset l_2$*) and coincide with each other:*

$$\underline{P}(\underline{E}(j0) - \underline{Q})\underline{P}^{-1} = \underline{E}(j0) - \underline{A}. \tag{2.4}$$

Moreover, it holds on the whole Hilbert space l_2 *that* $\hat{\underline{B}} = \underline{P}^{-1}\underline{B}$ *and* $\hat{\underline{C}} = \underline{C}\,\underline{P}$*.*

Furthermore, system (2.1) *is asymptotically stable if and only if the set* Λ *of all eigenvalues of* $\underline{Q} - \underline{E}(j0)$*, i.e.,* $\Lambda = \{\lambda(Q) + jm\omega_h : m \in \mathbf{Z}\}$*, lies in the open left-half plane. It is also true that* $\Lambda = \Lambda_A$ *where* Λ_A *is the set of all eigenvalues of* $\underline{A} - \underline{E}(j0)$*. In the above,* $\lambda(\cdot)$ *denotes the set of all eigenvalues of the matrix* $(\cdot)$*.*

Now we introduce the harmonic transfer operator [26] of system (2.1) given by

$$\underline{G}(s) := \underline{C}(\underline{E}(s) - \underline{A})^{-1}\underline{B} \tag{2.5}$$

in which $\underline{A} - \underline{E}(s)$ is called the harmonic state operator of system (2.1). In view of (2.4) in Lemma 2.1, $\underline{Q} - \underline{E}(s)$ is called the Floquet state operator of (2.1) to distinguish it from $\underline{A} - \underline{E}(s)$. When $\underline{E}^{-1}(s)$ exists, $\underline{I} - \underline{E}^{-1}(s)\underline{A}$ (respectively, $\underline{I} - \underline{E}^{-1}(s)\underline{Q}$) will be called the modified harmonic state operator (respectively, the modified Floquet state operator) of system (2.1).

Now we consider a domain $\Omega \subset \mathbf{C}$ and assume that $s \in \Omega$. Let us further give assumptions A1 and A2 about Ω to facilitate our statements.

A1 The domain Ω is closed and has a simple closed boundary, denoted by $\partial\Omega$, and thus is a simply connected domain on $\mathbf{C}$. Also, it holds that

$$|\mathrm{Im}(s)| < K_\Omega := \omega_h \quad (\forall s \in \Omega). \tag{2.6}$$

Furthermore, $\underline{E}(s) - \underline{Q}$ is an invertible mapping from l_E to l_2 for each $s \in \partial\Omega$.

A2 On the domain Ω, $\underline{E}(s)$ is an invertible mapping from l_E to l_2.

Note that the last assumption of A1 is satisfied if and only if $\partial\Omega$ contains no points in Λ, while A2 is satisfied if and only if Ω does not contain any points in $\Gamma := \{jm\omega_h : m \in \mathbf{Z}\}$. Hence, relation (2.4) tells us that

$$\underline{P}(\underline{E}(s) - \underline{Q})^{-1}\underline{P}^{-1} = (\underline{E}(s) - \underline{A})^{-1} \tag{2.7}$$

for all $s \in \Omega \setminus \Lambda$. That is, $\underline{E}(s) - \underline{A}$ is an invertible mapping from l_E to l_2 for each $s \in \Omega \setminus \Lambda$. (2.7) says that the harmonic transfer operator $\underline{G}(s)$ is well defined on l_2 for all $s \in \Omega\backslash\Lambda$. Lemma 2.2 gives basic facts about $\underline{G}(s)$ that play a key role in developing the Nyquist criterion.

LEMMA 2.2. *In the* FDLCP *system* (2.1)*, let* $A(t) \in L_{\mathrm{PCD}}[0,h]$ *and* $B(t), C(t) \in L_{\mathrm{PCC}}[0,h]$*. Assume that the domain* Ω *satisfies* A1*. Then for each* $s \in \Omega\backslash\Lambda$*,* $(\underline{E}(s) - \underline{Q})^{-1} \in \mathcal{C}_2(l_2)$*, and thus* $\underline{G}(s) \in \mathcal{C}_2(l_2)$*. Furthermore,* $\|\underline{G}(s)\|_2$ *has a uniform upper bound over* $s \in \partial\Omega$*.*

Now we introduce the 2-regularized determinant of Hilbert–Schmidt operators and derive some properties about the 2-regularized determinant of the modified harmonic transfer operators.

Let $\lambda_i(A)$ denote the ith eigenvalue of a linear compact operator $A : l_2 \to l_2$, and $s_i(A) := (\lambda_i(A^*A))^{1/2}$ be its ith singular value. For $p = 1$ and 2, the set of all compact operators $A : l_2 \to l_2$ satisfying $\|A\|_p := (\sum_i s_i(A)^p)^{1/p} < \infty$ is denoted by $\mathcal{C}_1(l_2)$ and $\mathcal{C}_2(l_2)$, respectively. In particular, the operators in $\mathcal{C}_1(l_2)$ are called trace class operators while those in $\mathcal{C}_2(l_2)$ are called Hilbert–Schmidt operators [5]. Clearly, $\mathcal{C}_1(l_2) \subset \mathcal{C}_2(l_2)$. For $A \in \mathcal{C}_1(l_2)$, the operator trace and determinant below are well defined in the sense that the infinite series and product converge

$$\mathrm{tr}(A) := \sum \lambda_i(A), \quad \det(I + A) := \prod (1 + \lambda_i(A)). \tag{2.8}$$

Note that for $A \in \mathcal{C}_2(l_2)$, $R_2(A) := (I + A)\exp\{-A\} - I \in \mathcal{C}_1(l_2)$. Thus, it is justified to define the determinant of $I + R_2(A)$ in the sense of (2.8), denoted by $\det_2(I+A) := \det(I+R_2(A))$, which is called the 2-regularized determinant of $I+A$. For our aim, assume that $B \in \mathcal{C}_2(l_2)$. Then

$$\det_2(I + A) = \prod [(1 + \lambda_i(A)) \exp(-\lambda_i(A))], \tag{2.9}$$

$$\det_2(I + A)\det_2(I + B) = \det_2[(I + A)(I + B)] \exp\{\mathrm{tr}(AB)\}. \tag{2.10}$$

By Proposition 1.3 of [9, p. 98], if $A \in \mathcal{C}_2(l_2)$ and B and C are bounded linear operators on l_2, then BAC belongs to $\mathcal{C}_2(l_2)$ and $\|BAC\|_2 \le \|B\|_{l_2/l_2}\|A\|_2\|C\|_{l_2/l_2}$. Moreover, Theorem 3.1 of [9, p. 43] says that AB and BA have the same nonzero eigenvalues with multiplicity taken into account, i.e., $\det_2(I + AB) = \det_2(I + BA)$.

When establishing a Nyquist-type criterion for FDLCP systems, one might be tempted to talk about some sort of determinant about the harmonic state operator $\underline{A} - \underline{E}(s)$ or the Floquet state operator $\underline{Q} - \underline{E}(s)$ as in the LTI continuous-time case. However, such a determinant notion for these unbounded operators is not readily available in the literature. The following lemma will be a key to get around the difficulty, in which the 2-regularized determinant of these operators premultiplied by $-\underline{E}^{-1}(s)$ (i.e., the modified harmonic and Floquet state operators) are considered.

LEMMA 2.3. *In the* FDLCP *system* (2.1)*, let* $A(t) \in L_{\mathrm{CAC}}[0, h] \cap L_{\mathrm{PCD}}[0, h]$. *Assume that the domain* Ω *satisfies* A1 *and* A2. *Then for each* $s \in \Omega$, $\underline{E}^{-1}(s) \in \mathcal{C}_2(l_2)$, *and thus* $\underline{E}^{-1}(s)\underline{A} \in \mathcal{C}_2(l_2)$, $\underline{E}^{-1}(s)\underline{Q} \in \mathcal{C}_2(l_2)$. *In particular,* $\|\underline{E}^{-1}(s)\|_2$ *has a uniform upper bound over* $s \in \Omega$. *Furthermore,*

$$\det_2[\underline{I} - \underline{E}^{-1}(s)\underline{A}] = g_A(s) \det_2[\underline{I} - \underline{E}^{-1}(s)\underline{Q}], \tag{2.11}$$

where the function $g_A(s)$ *does not vanish for each* $s \in \Omega$ *and is analytic over* Ω. *Also,* $\underline{I} - \underline{E}^{-1}(s)\underline{A}$ *is invertible for each* $s \in \Omega$, *and the inverse of* $\underline{I} - \underline{E}^{-1}(s)\underline{A}$ *is bounded on* l_2.

LEMMA 2.4. *Let* $\lambda_k(Q)$ *denote the* k*th eigenvalue of the* $n \times n$ *matrix* Q. *If the domain* Ω *satisfies* A1 *and* A2, *then the function*

$$\begin{aligned} f_Q(s) &:= \det_2[\underline{I} - \underline{E}^{-1}(s)\underline{Q}] \\ &= \prod_{k=1}^{n} \prod_{m=-\infty}^{\infty} \left(1 - \frac{\lambda_k(Q)}{s + jm\omega_h}\right) \exp\left\{\frac{\lambda_k(Q)}{s + jm\omega_h}\right\} \end{aligned} \tag{2.12}$$

is analytic on Ω, *which has a zero at each point* $\lambda_k(Q) - jm\omega_h$, $k = 1, 2, \ldots, n$, *and* $m \in \mathbf{Z}$ *(i.e., counted up to multiplicity), and has no other zeros on the complex plane.*

Remark 1. Lemma 2.4 says that the set of all zeros of $f_Q(s)$ is equal to the set Λ of all the eigenvalues of $\underline{Q} - \underline{E}(j0)$. This, together with Lemma 2.1, tells us that asymptotic stability of an FDLCP system can be reflected by the function $f_Q(s)$. This is the starting point of establishing a generalized Nyquist criterion via the 2-regularized determinant approach. To relate $f_Q(s)$ to its closed-loop counterpart via the 2-regularized determinant of the harmonic return differential operator will be the task in the following section, by which a Nyquist criterion of FDLCP systems will be derived in a similar fashion to the corresponding result in the LTI continuous-time systems.

In the following, $\partial\Omega$ will be chosen to form a Nyquist contour N_r, which needs to directly pass through the origin or to include the origin in the interior of the region enclosed by N_r (that is, to bypass the origin if there are any eigenvalues of $\underline{Q} - \underline{E}(j0)$ at the origin). However, such an Ω violates the assumption A2, since $0 \in \Gamma$. To surmount this problem, we introduce a shift factor $\rho > 0$ to s and replace assumption A2 by the following assumption.

A2′ On the domain Ω, $\underline{E}(s+\rho)(= \underline{E}(s) + \rho\,\underline{I})$ is an invertible mapping from l_E to l_2.

As we shall see later, assumption A2′ can in fact be essentially simplified to the condition $\rho > 0$ in our context due to the specific choice of the domain Ω given later. The introduction of such $\rho > 0$ is crucial only in the FDLCP setting exactly because $\underline{E}(s+\rho)$ is noninvertible at $s = 0$ if $\rho = 0$; this noninvertibility causes a problem when we try to deal with the 2-regularized determinant of $\underline{I} - \underline{E}^{-1}(s)\underline{A}$. It is easy to see that such a difficulty does not exist in the LTI continuous-time case since in that case $\det(sI - A)$ can be considered directly and s need not be inverted (recall the paragraph just before Lemma 2.3).

Remark 2. Once we introduce $\rho > 0$, we consider $\det_2[\underline{I} - \underline{E}^{-1}(s+\rho)(\underline{A}+\rho\underline{I})]$ and $\det_2[\underline{I} - \underline{E}^{-1}(s+\rho)(\underline{Q}+\rho\underline{I})]$ instead of similar relations in Lemmas 2.3 and 2.4. Hence, what we indeed employ in the subsequent arguments is the accordingly modified versions of Lemmas 2.3 and 2.4. Note that A1 is not affected by this shift factor. Thus, Lemma 2.3 still holds even if A2 is replaced by A2′, provided that $\underline{E}(s)$, $g_A(s)$, $\underline{A}$, and $\underline{Q}$ are also replaced by $\underline{E}(s+\rho)$, $g_{A+\rho I}(s+\rho)$, $\underline{A} + \rho\underline{I}$, and $\underline{Q} + \rho\underline{I}$, respectively. To facilitate the following descriptions, we will refer to this modified result as Lemma 2.3′. Similarly, Lemma 2.4 holds true when A2 is replaced by A2′ if $\underline{E}(s)$, $f_Q(s)$, $\underline{Q}$, $\lambda_k(Q)$, and $jm\omega_h$ are replaced by $\underline{E}(s+\rho)$, $f_{Q+\rho I}(s+\rho)$, $\underline{Q} + \rho\underline{I}$, $\lambda_k(Q+\rho I)$, and $\rho + jm\omega_h$, respectively. This modified result will be referred to as Lemma 2.4′. Moreover, it is obvious that the set of zeros of $f_{Q+\rho I}(s+\rho)$ equals that of the set of zeros of $f_Q(s)$. Clearly, these modified results can be validated even if condition (2.6) is removed from A1, because of the periodicity of $\underline{E}(s)$.

It is worth noticing that the introduction of the shift factor $\rho > 0$ does not cause any approximation effect at all on the stability analysis. This is because the invertibility of $\underline{I} - \underline{E}^{-1}(s+\rho)(\underline{A}+\rho\underline{I})$ is equivalent to that of $\underline{E}^{-1}(s+\rho)\{\underline{I} - \underline{E}^{-1}(s+\rho)(\underline{A}+\rho\underline{I})\} = \underline{E}(s) - \underline{A}$.

3. 2-Regularized Nyquist stability criterion. In this section, we develop a Nyquist criterion by the 2-regularized determinant for stability analysis of the closed-loop FDLCP system when an output feedback is introduced. In system (2.1), let $A(t)$, $B(t)$, $C(t) \in L_{\mathrm{CAC}}[0,h] \cap L_{\mathrm{PCD}}[0,h]$, and an h-periodically time-varying output feedback $u = -K(t)y + v$ is introduced. This leads to the closed-loop FDLCP system

described by

$$G_c : \begin{cases} \dot{x} = A_c(t)x + B(t)v, \\ y = C(t)x. \end{cases} \tag{3.1}$$

Here v is a new reference input and it is assumed that $K(t) \in L_{\mathrm{CAC}}[0,h] \cap L_{\mathrm{PCD}}[0,h]$. Clearly, $A_c(t) := A(t) - B(t)K(t)C(t) \in L_{\mathrm{CAC}}[0,h] \cap L_{\mathrm{PCD}}[0,h]$. These assumptions about $A(t), B(t), C(t)$, and $K(t)$ ensure that the Toeplitz transformation and Lemmas 2.1–2.4 (see also Remark 2) apply to both the open- and closed-loop FDLCP systems.

Now such a question emerges, In what way can one claim asymptotic stability of the closed-loop system G_c by observing the open-loop harmonic transfer operator?

3.1. 2-Regularized determinant relation between open- and closed-loop modified harmonic state operators. As known in the LTI case, one must obtain the relationship between the open- and closed-loop pole polynomials before claiming the Nyquist criterion. The aim of this subsection is to get a similar relationship in the FDLCP case, where Lemma 2.4 and Remark 2 suggest that $f_{Q+\rho I}(s+\rho) = \det_2[\underline{I} - \underline{E}^{-1}(s+\rho)(\underline{Q} + \rho\underline{I})]$, and the corresponding closed-loop counterpart plays the role of pole polynomials. To derive such a relationship, let us define $\underline{K} = \mathcal{T}\{K(t)\}$. Then it follows from Lemma 2.2 that for each $s \in \Omega\backslash\Lambda$, $\underline{K}\,\underline{G}(s)$ is well defined and belongs to $\mathcal{C}_2(l_2)$. Therefore, it makes sense to talk about the $\det_2$ of the return difference operator $\underline{I} + \underline{K}\,\underline{G}(s)$ for each $s \in \Omega\backslash\Lambda$. Noting that $\underline{E}(s+\rho)$ is invertible for all $s \in \Omega$ by A2′, we compute $\det_2$ of $\underline{I} + \underline{K}\,\underline{G}(s)$ as

$$\begin{aligned} \det_2[\underline{I} + \underline{K}\,\underline{G}(s)] &= \det_2[\underline{I} + (\underline{E}(s) - \underline{A})^{-1}\underline{B}\,\underline{K}\,\underline{C}] \\ &= \det_2[\underline{I} + (\underline{I} - \underline{E}^{-1}(s+\rho)(\underline{A} + \rho\underline{I}))^{-1}\underline{E}^{-1}(s+\rho)\underline{B}\,\underline{K}\,\underline{C}] \\ &= \det_2[(\underline{I} - \underline{E}^{-1}(s+\rho)(\underline{A} + \rho\underline{I}))^{-1}(\underline{I} - \underline{E}^{-1}(s+\rho)(\underline{A}_c + \rho\underline{I}))], \end{aligned} \tag{3.2}$$

where $\underline{A}_c := \underline{A} - \underline{B}\,\underline{K}\,\underline{C}$. Note that $\underline{A} - \underline{B}\,\underline{K}\,\underline{C} = \mathcal{T}\{A_c(t)\}$ holds since $B(t), K(t)$, and $C(t)$ belong to $L_{\mathrm{CAC}}[0,h]$ [17]. In (3.2), we used the facts that $(\underline{E}(s) - \underline{A})^{-1} \in \mathcal{C}_2(l_2)$ for each $s \in \Omega\backslash\Lambda$ and that $\underline{K}\,\underline{C}$ is bounded on l_2, where the first fact can be shown by (2.7) and Lemma 2.2.

Before we expand (3.2) via (2.10), we must show that the $\det_2$'s of $(\underline{I} - \underline{E}^{-1}(s+\rho)(\underline{A} + \rho\underline{I}))^{-1}$ and $\underline{I} - \underline{E}^{-1}(s+\rho)(\underline{A}_c + \rho\underline{I})$ are well defined for all $s \in \Omega \setminus \Lambda$. To this end, let us define the infinite-dimensional matrix $\underline{M}(s+\rho)$ such that for each $s \in \Omega\backslash\Lambda$, it holds that

$$(\underline{I} - \underline{E}^{-1}(s+\rho)(\underline{A} + \rho\underline{I}))(\underline{I} + \underline{M}(s+\rho)) = \underline{I}. \tag{3.3}$$

It follows easily that $\underline{M}(s+\rho)$ is indeed given by

$$\underline{M}(s+\rho) = \underline{E}^{-1}(s+\rho)(\underline{A} + \rho\underline{I})(\underline{I} - \underline{E}^{-1}(s+\rho)(\underline{A} + \rho\underline{I}))^{-1}.$$

Then the fact that $\underline{E}^{-1}(s+\rho) \in \mathcal{C}_2(l_2)$ and that $(\underline{A} + \rho\underline{I})(\underline{I} - \underline{E}^{-1}(s+\rho)(\underline{A} + \rho\underline{I}))^{-1}$ is bounded on l_2 (by Lemma 2.3) implies that $\underline{M}(s+\rho) \in \mathcal{C}_2(l_2)$. Similarly, for each $s \in \Omega$, $\underline{E}^{-1}(s+\rho)(\underline{A}_c + \rho\underline{I})$ and $\underline{E}^{-1}(s+\rho)(\underline{A} + \rho\underline{I})$ belong to $\mathcal{C}_2(l_2)$. Thus it makes sense to deal with the $\det_2$'s of $(\underline{I} - \underline{E}^{-1}(s+\rho)(\underline{A} + \rho\underline{I}))^{-1}$, $\underline{I} - \underline{E}^{-1}(s+\rho)(\underline{A}_c + \rho\underline{I})$, and $\underline{I} - \underline{E}^{-1}(s+\rho)(\underline{A} + \rho\underline{I})$ on $\Omega\backslash\Lambda$, separately.

Now taking the $\det_2$ on both sides of (3.3), we obtain

$$\begin{aligned} &\det_2[\underline{I} - \underline{E}^{-1}(s+\rho)(\underline{A} + \rho\underline{I})] \\ &\quad = \frac{\exp\{-\mathrm{tr}(\underline{E}^{-1}(s+\rho)(\underline{A} + \rho\underline{I})\underline{M}(s+\rho))\}}{\det_2[\underline{I} + \underline{M}(s+\rho)]}. \end{aligned} \tag{3.4}$$

Then expanding the $\det_2$ of (3.2) via (2.10) and using (3.4), we have

$$\begin{aligned}\det_2[\underline{I}+\underline{K}\,\underline{G}(s)] &= \det_2[(\underline{I}-\underline{E}^{-1}(s+\rho)(\underline{A}+\rho\underline{I}))^{-1}]\\ &\quad\times\det_2[\underline{I}-\underline{E}^{-1}(s+\rho)(\underline{A}_c+\rho\underline{I})]\\ &\quad\times\exp\{\mathrm{tr}(\underline{M}(s+\rho)\underline{E}^{-1}(s+\rho)(\underline{A}_c+\rho\underline{I}))\}\\ &= \exp\{\Delta(s+\rho)\}\frac{\det_2[\underline{I}-\underline{E}^{-1}(s+\rho)(\underline{A}_c+\rho\underline{I})]}{\det_2[\underline{I}-\underline{E}^{-1}(s+\rho)(\underline{A}+\rho\underline{I})]},\end{aligned}\tag{3.5}$$

where the scalar function $\Delta(s+\rho)$ is given by

$$\begin{aligned}\Delta(s+\rho) &:= \mathrm{tr}(\underline{M}(s+\rho)\underline{E}^{-1}(s+\rho)(\underline{A}_c+\rho\underline{I}))-\mathrm{tr}(\underline{M}(s+\rho)\underline{E}^{-1}(s+\rho)(\underline{A}+\rho\underline{I}))\\ &= -\mathrm{tr}(\underline{E}^{-1}(s+\rho)(\underline{A}+\rho\underline{I})(\underline{E}(s)-\underline{A})^{-1}\underline{B}\,\underline{K}\,\underline{C}).\end{aligned}\tag{3.6}$$

3.2. Nyquist contour and Nyquist locus. Before leading the above arguments to a Nyquist stability criterion, we need to describe in what way an appropriate Nyquist contour, i.e., $\partial\Omega$, should be taken and how the corresponding Nyquist locus should be plotted in the 2-regularized determinant sense.

First let us see how an appropriate Nyquist contour should be taken.

To this purpose, we mention some facts about the eigenvalues of $\underline{Q}-\underline{E}(j0)$ (or equivalently the operator $\underline{A}-\underline{E}(j0)$), which are given in Lemma 2.1. First, all the eigenvalues of $\underline{Q}-\underline{E}(j0)$ are located in a vertical strip region parallel to the imaginary axis; second, the eigenvalues distribution pattern in the horizontal strip

$$\mathbf{C}_F := \{s\in\mathbf{C} : -\omega_h/2 < \mathrm{Im}(s) \le \omega_h/2\},\tag{3.7}$$

which is called the fundamental strip [26], unfolds itself vertically to both $-j\infty$ and $j\infty$ with the period $j\omega_h$. In other words, if we can understand the eigenvalue distribution pattern in $\mathbf{C}_F$, then the whole eigenvalue structure of $\underline{Q}-\underline{E}(j0)$ is clarified. Based on this, a possible Nyquist contour would be the boundary of the right-half fundamental strip of $\mathbf{C}_F$, i.e., $\{s : \mathrm{Re}(s)\ge 0, s\in\mathbf{C}_F\}$. However, since $\underline{G}(s)$ is not well defined for $s\in\Lambda$, the actual Nyquist contour should avoid going through these points in Λ. Hence the Nyquist contour N_r shown in Figure 3.1 is taken. In Figure 3.1 the crosses ($\times$'s) denote possible eigenvalues of the open-loop operator $\underline{Q}-\underline{E}(j0)$ on the boundary of the right-hand half of $\mathbf{C}_F$. It should be stressed that the Nyquist contour N_r bypasses the eigenvalues of $\underline{Q}-\underline{E}(j0)$ on the imaginary axis with $-\omega_h/2 < \mathrm{Im}(\lambda) < \omega_h/2$, if any, from the left-hand side while other eigenvalues on the boundary of the right-half fundamental strip, if any, from the upper-side, via a semicircle with the radius r that is small enough. Thus, if $\underline{Q}-\underline{E}(j0)$ has eigenvalues on the imaginary axis such that $-\omega_h/2 < \mathrm{Im}(\lambda)\le\omega_h/2$, they are to be included in the interior of the region enclosed by N_r. Now let us assume that the right edge of N_r is far enough from the imaginary axis so that there are no eigenvalues of $\underline{Q}-\underline{E}(j0)$ on it, and finally, let us define the domain Ω as the union of the Nyquist contour N_r and the interior of the region enclosed by N_r. Then, it is obvious that Ω satisfies A1 and A2$'$ whenever $\rho > r \ge 0$ and $K_\Omega > \omega_h/2 + r$. Hence the arguments from (3.2) to (3.6) are validated for such a ρ.

Next let us see how the corresponding Nyquist locus can be plotted.

Now segment the Nyquist contour N_r given in Figure 3.1 into four pieces N_{ab}, N_{bc}, N_{cd}, and N_{da} in the obvious fashion and note the following observations. First, since $\det_2[\underline{I}+\underline{K}\underline{G}(s)]$ and $\Delta(s+\rho)$ are $j\omega_h$-periodic in the frequency domain, the plot of $\det_2[\underline{I}+\underline{K}\,\underline{G}(s)]\exp\{-\Delta(s+\rho)\}$ corresponding to N_{ab} will form a closed curve.

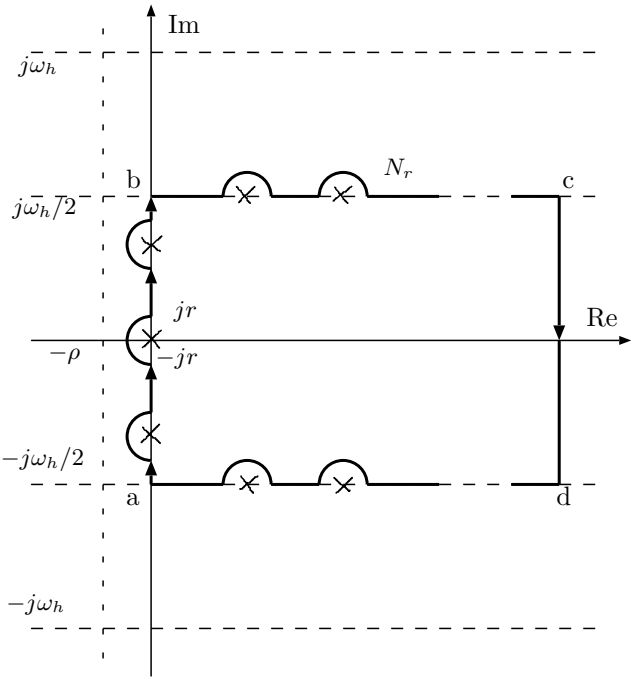


FIG. 3.1. *Nyquist contour* N_r.

Second, for each $s \in N_{bc}$, there is a corresponding complex number $\tilde{s} = s - j\omega_h \in N_{da}$ such that $\det_2[\underline{I} + \underline{K}\,\underline{G}(s)] \exp\{-\Delta(s+\rho)\} = \det_2[\underline{I} + \underline{K}\,\underline{G}(\tilde{s})] \exp\{-\Delta(\tilde{s}+\rho)\}$ due to the same periodicity. The only difference between the plot of $\det_2[\underline{I} + \underline{K}\,\underline{G}(s)] \exp\{-\Delta(s+\rho)\}$ corresponding to N_{bc} and that corresponding to N_{da} is that these two plots are drawn in just opposite directions. To facilitate the discussions, we assume that N_{bc} and N_{da} are taken in such a way that they bypass *closed-loop* eigenvalues, if any, in a similar fashion to what is done for *open-loop* eigenvalues. Clearly, this requirement on N_{bc} and N_{da} can always be satisfied, in principle, and brings no essential difficulty to validate the second assertion we just claimed while keeping A2′. Third, if $\mathrm{Re}(s)$ is made large enough for $s \in N_{cd}$ (i.e., the Nyquist contour N_r is extended to the right far enough so that it encloses all the unstable closed-loop eigenvalues on the fundamental strip), $\det_2[\underline{I} + \underline{K}\,\underline{G}(s)] \exp\{-\Delta(s+\rho)\} \to 1$ for each $s \in N_{cd}$. These facts indicate that the plot segment of $\det_2[\underline{I} + \underline{K}\,\underline{G}(s)] \exp\{-\Delta(s+\rho)\}$ corresponding to N_{bc}, N_{cd}, and N_{da} neither goes through the origin nor contributes to encirclements around the origin. In other words, to investigate the encirclements of $\det_2[\underline{I} + \underline{K}\,\underline{G}(s)] \exp\{-\Delta(s+\rho)\}$ around the origin on N_r, it is enough to see the plot of $\det_2[\underline{I} + \underline{K}\,\underline{G}(s)] \exp\{-\Delta(s+\rho)\}$ corresponding to N_{ab}. In view of this, the plot $\det_2[\underline{I} + \underline{K}\,\underline{G}(s)] \exp\{-\Delta(s+\rho)\} : N_{ab} \to \mathbf{C}$ is called the Nyquist locus of (2.1) when $s \in N_{ab}$ moves in the clockwise direction with respect to N_r.

3.3. 2-Regularized Nyquist stability criterion. On the basis of arguments in the two preceding subsections, we are ready to show the 2-regularized Nyquist criterion, which is the main result of this paper.

THEOREM 3.1. *Assume that* $A(t), B(t), C(t)$ *of the* FDLCP *system* (2.1) *and the feedback matrix* $K(t)$ *belong to* $L_{\mathrm{CAC}}[0,h] \cap L_{\mathrm{PCD}}[0,h]$. *Let* n_{us} *denote the number*

of the unstable eigenvalues of the open-loop Floquet state operator $\underline{Q} - \underline{E}(j0)$ *(or equivalently, the harmonic state operator* $\underline{A} - \underline{E}(j0)$*) in the fundamental strip* $\mathbf{C}_F$ *defined in* (3.7) *counted according to their multiplicity. Take an arbitrary positive number* ρ *and a sufficiently small number* r *according to the open-loop eigenvalue condition of* $\underline{Q} - \underline{E}(j0)$ *on the imaginary axis such that* $\rho > r \geq 0$*, and then consider the harmonic transfer operator* $\underline{G}(s)$ *and the scalar function* $\Delta(s+\rho)$ *defined in* (3.6)*. Then, the closed-loop system* G_c *is asymptotically stable if and only if the Nyquist locus,* $\det_2[\underline{I} + \underline{K}\,\underline{G}(s)] \exp\{-\Delta(s+\rho)\} : N_{ab} \to \mathbf{C}$*, vanishes nowhere on* N_{ab} *and encircles the origin* n_{us} *times in the counterclockwise sense.*

Proof. Under the given assumptions on the system matrices and the feedback gain matrix, Lemma 2.3′ stated in Remark 2 ensures a modified version of (2.11), which reads

$$\det_2[\underline{I} - \underline{E}^{-1}(s+\rho)(\underline{A} + \rho\underline{I})] = g_{A+\rho I}(s+\rho) \det_2[\underline{I} - \underline{E}^{-1}(s+\rho)(\underline{Q} + \rho\underline{I})]$$

for each $s \in \Omega$, where $g_{A+\rho I}(s+\rho)$ is analytic and vanishes nowhere on Ω. By applying Lemma 2.3′ to the closed-loop term $\det_2[\underline{I} - \underline{E}^{-1}(s+\rho)(\underline{A}_c + \rho\underline{I})]$, we readily obtain

$$\det_2[\underline{I} - \underline{E}^{-1}(s+\rho)(\underline{A}_c + \rho\underline{I})] = g_{A_c+\rho I}(s+\rho) \det_2[\underline{I} - \underline{E}^{-1}(s+\rho)(\underline{Q}_c + \rho\underline{I})]$$

for each $s \in \Omega$, where $g_{A_c+\rho I}(s+\rho)$ is analytic and vanishes nowhere on Ω. Here, $\underline{Q}_c := \mathcal{T}\{Q_c\}$ with Q_c being the constant matrix of the Floquet factorization of the transition matrix $\Phi_c(t, t_0)$ of the closed-loop FDLCP system (3.1). Thus, relation (3.5) can be rewritten as

$$\begin{aligned} &\det_2[\underline{I} + \underline{K}\,\underline{G}(s)] \exp\{-\Delta(s+\rho)\} \\ &\quad = \frac{g_{A_c+\rho I}(s+\rho) \det_2[\underline{I} - \underline{E}^{-1}(s+\rho)(\underline{Q}_c + \rho\underline{I})]}{g_{A+\rho I}(s+\rho) \det_2[\underline{I} - \underline{E}^{-1}(s+\rho)(\underline{Q} + \rho\underline{I})]}. \end{aligned} \tag{3.8}$$

Now we concentrate the attention on the right-hand side of (3.8). By Lemma 2.4′ stated in Remark 2, the right-hand side of (3.8) is just $g_{A_c+\rho I}(s+\rho) f_{Q_c+\rho I}(s+\rho) / g_{A+\rho I}(s+\rho) f_{Q+\rho I}(s+\rho) =: d(s)$, which is analytic on Ω except at the zeros of $f_{Q+\rho I}(s+\rho)$ contained in Ω. It is also clear that $d(s)$ is meromorphic by Theorem 15.12 of [24]. Apparently, by the definition of Ω, only finitely many zeros of $f_{Q_c+\rho I}(s+\rho)$ and $f_{Q+\rho I}(s+\rho)$ are contained in Ω, which in particular implies that (3.8) is not identically zero over Ω. These facts imply that the argument principle about complex functions applies to $d(s)$ if the right-hand side of (3.8) never vanishes at each $s \in \partial\Omega (= N_r)$ (this guarantees that there are no zeros of $f_{Q_c+\rho I}(s+\rho)$ located on N_r). Note by Remark 2 that the sets of zeros of $f_{Q+\rho I}(s+\rho)$ and $f_{Q_c+\rho I}(s+\rho)$ are actually independent of ρ and just the sets of the eigenvalues of the open- and closed-loop operators, $\underline{Q} - \underline{E}(j0)$ and $\underline{Q}_c - \underline{E}(j0)$, respectively. Furthermore, it follows readily that the right-hand side of (3.8) never vanishes on the segments N_{bc}, N_{cd}, and N_{da} of N_r because of the assumption that they do not go through the closed-loop eigenvalues. Hence, some straightforward discussions will lead to the desired assertions. □

Remark 3. It should be stressed that n_{us} indicates the number of unstable eigenvalues of $\underline{Q} - \underline{E}(j0)$ in the closed right-half portion of the fundamental strip. However, it is easy to see that n_{us} equals the number of the unstable eigenvalues of Q in the whole closed right-half plane.

3.4. Equivalent interpretation of Theorem 3.1. In this subsection we show that Theorem 3.1 can have a more explicit expression which will provide convenience for implementing Theorem 3.1. To this end, let us return to (3.2) and compute the

$\det_2$ of $\underline{I} + \underline{K}\,\underline{G}(s)$ again but after the harmonic transfer operator $\underline{G}(s)$ has been rewritten as $\underline{G}(s) = \hat{\underline{C}}(\underline{E}(s) - \underline{Q})^{-1}\hat{\underline{B}}$ by means of Lemma 2.1 and (2.7). That is, we compute

$$\det_2[\underline{I} + \underline{K}\,\underline{G}(s)] = \det_2[\underline{I} + \underline{K}\,\hat{\underline{C}}(\underline{E}(s) - \underline{Q})^{-1}\hat{\underline{B}}]. \tag{3.9}$$

It is not hard to see that all the arguments around (3.2)–(3.6) can be repeated on the operator $\underline{K}\,\hat{\underline{C}}(\underline{E}(s) - \underline{Q})^{-1}\hat{\underline{B}} \in \mathcal{C}_2(l_2)$. Then we can conclude that

$$\det_2[\underline{I} + \underline{K}\,\underline{G}(s)]\exp\{-\tilde{\Delta}(s+\rho)\} = \frac{\det_2[\underline{I} - \underline{E}^{-1}(s+\rho)(\underline{\tilde{A}}_c + \rho\underline{I})]}{\det_2[\underline{I} - \underline{E}^{-1}(s+\rho)(\underline{Q} + \rho\underline{I})]} \tag{3.10}$$

with $\underline{\tilde{A}}_c := \underline{Q} - \hat{\underline{B}}\,\underline{K}\,\hat{\underline{C}}$ and

$$\tilde{\Delta}(s+\rho) := -\mathrm{tr}(\underline{E}^{-1}(s+\rho)(\underline{Q} + \rho\underline{I})(\underline{E}(s) - \underline{Q})^{-1}\hat{\underline{B}}\,\underline{K}\,\hat{\underline{C}}). \tag{3.11}$$

Now we observe that $\underline{E}(j0) - \underline{\tilde{A}}_c = \underline{P}^{-1}(\underline{E}(j0) - \underline{A}_c)\underline{P}$, where the similarity transformation formula is used. This equation clearly says that every eigenvalue of $\underline{E}(j0) - \underline{\tilde{A}}_c$ is also an eigenvalue of $\underline{E}(j0) - \underline{A}_c$ and vice versa. This implies that one can test stability of the closed-loop system G_c by that of the closed-loop system $\tilde{G}_c : (\tilde{A}_c(t) := Q - \hat{B}(t)K(t)\hat{C}(t), \hat{B}(t), \hat{C}(t))$. Combining this fact with (3.10), we have the following corollary about Theorem 3.1.

COROLLARY 3.2. *Suppose all the assumptions of Theorem* 3.1 *hold. Take an arbitrary positive number* ρ *and a sufficiently small number* r *according to the open-loop eigenvalue condition of* $\underline{Q} - \underline{E}(j0)$ *on the imaginary axis such that* $\rho > r \geq 0$. *Consider the harmonic transfer operator* $\underline{G}(s)$ *and the scalar function* $\tilde{\Delta}(s+\rho)$ *given in* (3.11). *Then the closed-loop system* G_c *is asymptotically stable if and only if the modified Nyquist locus,* $\det_2[\underline{I} + \underline{K}\,\underline{G}(s)]\exp\{-\tilde{\Delta}(s+\rho)\} : N_{ab} \to \mathbf{C}$ *vanishes nowhere on* N_{ab} *and encircles the origin* n_{us} *times in the counterclockwise sense.*

The Nyquist locus in Theorem 3.1 is defined by $\det_2[\underline{I} + \underline{K}\,\underline{G}(s)]\exp\{-\Delta(s+\rho)\} : N_{ab} \to \mathbf{C}$, while that of Corollary 3.2 is defined by $\det_2[\underline{I} + \underline{K}\,\underline{G}(s)]\exp\{-\tilde{\Delta}(s+\rho)\} : N_{ab} \to \mathbf{C}$. That is, in Corollary 3.2 one only needs to compute $\tilde{\Delta}(s+\rho)$ instead of $\Delta(s+\rho)$. If we further compute $\det_2[\underline{I} + \underline{K}\,\underline{G}(s)]$ by the right-hand relation of (3.9), then the block-diagonal structures of $(\underline{E}(s) - \underline{Q})^{-1}$ and $\underline{E}^{-1}(s+\rho)(\underline{Q} + \rho\underline{I})(\underline{E}(s) - \underline{Q})^{-1}$ will bring us convenience in plotting the modified Nyquist locus $\det_2[\underline{I} + \underline{K}\,\underline{G}(s)]\exp\{-\tilde{\Delta}(s+\rho)\} : N_{ab} \to \mathbf{C}$ as we will see in the next section.

Remark 4. Note that only the "DC-part" of $\hat{B}(t)K(t)\hat{C}(t)$ contributes to $\tilde{\Delta}(s+\rho)$. In other words, if the "DC-part" of $\hat{B}(t)K(t)\hat{C}(t)$ is zero, then the exponential part on the left-hand side of (3.10) can be dropped. As a side note, we point out that $\Delta(s+\rho)$ does not equal $\tilde{\Delta}(s+\rho)$ since $\underline{P}$ and $\underline{E}^{-1}(s)$ do not commute, and $\det_2[\underline{I} - \underline{E}^{-1}(s)\underline{Q}] \neq \det_2[\underline{I} - \underline{E}^{-1}(s)\underline{A}]$ in general.

Remark 5. Since any periodically time-varying state matrix $A(t)$ can be rewritten in the form of $A_{\mathrm{const}} + \tilde{A}(t)$ with A_{const} being a constant matrix, stability of the FDLCP system with the state matrix $A(t)$, no matter this FDLCP system itself is open- or closed-loop, can be easily tested by recasting the stability problem as a closed-loop stability problem with $(A_{\mathrm{const}}, I, I) = (Q, I, I)$ being the open-loop system matrices and $-\tilde{A}(t)$ being (treated as) the feedback gain. This recasting technique means that Corollary 3.2 can be easily applied without computing the transition matrix of any FDLCP models. Having this recasting technique in mind, a finite-dimensional truncated implementation of the modified Nyquist criterion is developed in the following section.

4. Implementation of the 2-regularized Nyquist criterion. In the previous section, generalized Nyquist criteria were developed for asymptotic stability of FDLCP systems. Unfortunately, however, it is hard to implement them directly due to the infinite-dimensionality of the harmonic transfer operator and operators involved in the determinant and trace computations. In this section, we consider the implementation problem of the 2-regularized Nyquist criterion in Corollary 3.2 through the staircase truncations on $\underline{G}(s)$ and other infinite-dimensional operators involved, and then the closed-loop stability analysis is reduced to that of a finite-dimensional LTI continuous-time system in an asymptotic sense.

The staircase truncation is first proposed for the H_∞ norm computation in FDLCP systems [30]. It should be noted that although the same truncation is adopted, the convergence problem in the H_∞ norm computation and that in the Nyquist locus plotting are essentially different. More precisely, in the H_∞ norm case convergence is related to infinite summations, while in the Nyquist locus case of this paper convergence pertains to infinite products. This discrepancy alerts us that just sketching the convergence proof might mislead the reader in understanding the implementation algorithm, and thus we keep the arguments in their complete form. Another benefit to make the convergence arguments in this way is that the inequalities in the convergence arguments can explicitly provide us with estimation formulas of the truncation size, though we will not deal with the size estimation problem in the paper due to the space limitation.

For simplicity, the discussions are given for the case of $\underline{K} = \underline{I}$ throughout this section. This brings no loss of generality if we notice that one can always treat $\underline{K}\,\underline{C}$ as a single operator in the harmonic return difference operator $\underline{I} + \underline{K}\,\underline{G}(s)$.

4.1. Truncation description. In this subsection we describe the staircase truncation. Strictly speaking, the staircase truncation is two-step: first skew truncate $\underline{G}(s)$ to $\underline{G}_{[N]}(s)$, and then truncate $\underline{G}_{[N]}(s)$ in a staircase fashion to $\underline{G}_{[N,M]}(s)$. Namely, we take an integer $N \geq 1$ and approximate $\underline{G}(s) = \underline{\hat{C}}(\underline{E}(s) - \underline{Q})^{-1}\underline{\hat{B}}$ by

$$\underline{G}_{[N]}(s) = \underline{\hat{C}}_{[N]}(\underline{E}(s) - \underline{Q})^{-1}\underline{\hat{B}}_{[N]}, \tag{4.1}$$

where $\underline{\hat{B}}_{[N]} := \mathcal{T}\{\hat{B}_N(t)\}$. Here $\hat{B}_N(t) := \sum_{m=-N}^{N} \hat{B}_m e^{jm\omega_h t}$ with $\{\hat{B}_m\}$ being the Fourier coefficient sequence of $\hat{B}(t)$. Similarly, $\underline{\hat{C}}_{[N]}$ is constructed in terms of $\{\hat{C}_m\}$, which is the Fourier coefficient sequence of $\hat{C}(t)$. Only the skew truncation cannot reduce the $\det_2$ computation to a finite-dimensional one, and thus we introduce the staircase truncation on $\underline{G}(s)$ as follows:

$$\underline{G}_{[N,M]}(j\varphi) = \underline{\hat{C}}_{[N,M]}(\underline{E}_M(s) - \underline{Q}_M)^{-1}\underline{\hat{B}}_{[N,M]}. \tag{4.2}$$

Here, the infinite-dimensional matrix $\underline{\hat{B}}_{[N,M]} := \mathrm{diag}[\ldots, \hat{B}_{NM}, \hat{B}_{NM}, \hat{B}_{NM}, \ldots]$ is defined with

$$\hat{B}_{NM} = \underbrace{\begin{bmatrix} \hat{B}_0 & \cdots & \hat{B}_{-N} & & 0 \\ \vdots & \ddots & & \ddots & \\ \hat{B}_N & & \ddots & & \hat{B}_{-N} \\ & \ddots & & \ddots & \vdots \\ 0 & & \hat{B}_N & \cdots & \hat{B}_0 \end{bmatrix}}_{(2M+1)-\text{blocks}}, \tag{4.3}$$

where we assume $M \geq N+1$. The infinite-dimensional matrix $\underline{\hat{C}}_{[N,M]}$ is defined similarly to $\underline{\hat{B}}_{[N,M]}$ but in terms of the Fourier coefficients of $\hat{C}(t)$. Furthermore, the infinite-dimensional but block-diagonal operators $\underline{E}(s)$ and $\underline{Q}$ are partitioned into diagonal blocks accordingly. That is,

$$\begin{aligned}\underline{Q}_M &:= \operatorname{diag}[\ldots, Q_M, Q_M, Q_M, \ldots](=\underline{Q}),\\ \underline{E}_M(s) &:= \operatorname{diag}[\ldots, E_{M,-1}(s), E_{M0}(s), E_{M1}(s), \ldots](=\underline{E}(s))\end{aligned}$$

with $Q_M = \operatorname{diag}[Q, Q, \ldots, Q]$ being $(2M+1)\times(2M+1)$ and

$$E_{Mm}(s) = \operatorname{diag}[\varphi_{m(2M+1)-M}(s)I, \ldots, \varphi_{m(2M+1)}(s)I, \ldots, \varphi_{m(2M+1)+M}(s)I],$$

where $m \in \mathbf{Z}$.

4.2. Truncation convergence. To state the final result, we need to establish convergence lemmas associated with the staircase truncation on the harmonic transfer operator in the $\det_2$ and trace sense. In this subsection, we show relevant convergence lemmas for the suggested truncation treatment.

LEMMA 4.1. *Assume in the* FDLCP *system* (2.1) *that* $A(t) \in L_{\mathrm{PCD}}[0,h]$ *and* $B(t), C(t) \in L_{\mathrm{CAC}}[0,h]$, *and the domain* Ω *satisfies* A1. *Then for any* $\mu > 0$, *there exists an integer* $N_0(\mu) > 0$ *such that* $|\det_2[\underline{I} + \underline{G}_{[N]}(s)] - \det_2[\underline{I} + \underline{G}(s)]| < \mu$ $(\forall N \geq N_0(\mu), \forall s \in \partial\Omega)$.

On the basis of Lemma 4.1, to show the convergence that $\det_2[\underline{I} + \underline{G}_{[N,M]}(s)] \to \det_2[\underline{I} + \underline{G}(s)]$ as $N, M \to \infty$, it suffices to show that $\det_2[\underline{I} + \underline{G}_{[N,M]}(s)] \to \det_2[\underline{I} + \underline{G}_{[N]}(s)]$ as $M \to \infty$ for each fixed $N > 0$. This is answered by the following lemma.

LEMMA 4.2. *Assume in the* FDLCP *system* (2.1) *that* $A(t) \in L_{\mathrm{PCD}}[0,h]$ *and* $B(t), C(t) \in L_{\mathrm{CAC}}[0,h]$, *and the domain* Ω *satisfies* A1. *Then for any* $\mu > 0$ *and fixed* $N > 0$, *there exists an integer* $M_0(N,\mu) > 0$ *such that* $|\det_2[\underline{I} + \underline{G}_{[N,M]}(s)] - \det_2[\underline{I} + \underline{G}_{[N]}(s)]| < \mu$ $(\forall M \geq M_0(N,\mu), \forall s \in \partial\Omega)$.

Now we apply the staircase truncation on the infinite-dimensional matrix $\underline{\hat{B}}\,\underline{\hat{C}}$ as we do on $\underline{G}(s)$, and get the truncated version $\underline{\widehat{BC}}_{[N,M]}$, which is defined similarly to $\underline{\hat{B}}_{[N,M]}$ but in terms of the Fourier coefficients of $\hat{B}(t)\hat{C}(t)$. Based on this truncation, we further define

$$\begin{aligned}\tilde{\Delta}_{[N,M]}(s+\rho) &:= -\operatorname{tr}\big(\underline{E}^{-1}(s+\rho)(\underline{Q}+\rho\underline{I})(\underline{E}(s)-\underline{Q})^{-1}\underline{\widehat{BC}}_{[N,M]}\big)\\ &= -\sum_m \operatorname{tr}\big(\tilde{\Delta}_{m[N,M]}(s+\rho)\big),\end{aligned} \tag{4.4}$$

where $\underline{I}_M$ is defined similarly to $\underline{Q}_M$ but in terms of the identity matrix I and

$$\tilde{\Delta}_{m[N,M]}(s+\rho) := E_{Mm}^{-1}(s+\rho)(Q_M+\rho I_M)(E_{Mm}(s)-Q_M)^{-1}\widehat{BC}_{NM} \tag{4.5}$$

with $m \in \mathbf{Z}$ and I_M being the $(2M+1)\times(2M+1)$ blockwise identity. The matrix $\widehat{BC}_{NM}$ is defined similarly to $\hat{B}_{NM}$ but in terms of the Fourier coefficients of $\hat{B}(t)\hat{C}(t)$. Then by repeating arguments similar to those in the proofs of Lemmas 4.1 and 4.2, the following lemma can be shown.

LEMMA 4.3. *Assume in the* FDLCP *system* (2.1) *that* $A(t) \in L_{\mathrm{PCD}}[0,h]$ *and* $B(t), C(t) \in L_{\mathrm{CAC}}[0,h]$, *and the domain* Ω *satisfies* A1 *and* A2$'$. *Then for any* $\mu > 0$ *and fixed* $N > 0$, *there exist integers* $N_1(\mu) > 0$ *and* $M_1(N,\mu) > 0$ *such that* $|\tilde{\Delta}_{[N,M]}(s+\rho) - \tilde{\Delta}(s+\rho)| < \mu$ $(\forall N \geq N_1(\mu), M \geq M_1(N,\mu), \forall s \in \partial\Omega)$.

Combining Lemmas 4.1, 4.2, and 4.3, we can get a tight estimation of $\det_2[\underline{I}+\underline{G}(s)]\exp\{-\tilde{\Delta}(s+\rho)\}$ by $\det_2[\underline{I}+\underline{G}_{[N,M]}(s)]\exp\{-\tilde{\Delta}_{[N,M]}(s+\rho)\}$. Now we show that the latter can be reduced to finite-dimensional computations. Indeed, by definition, we have

$$\begin{aligned}&\det_2\big[\underline{I}+\underline{G}_{[N,M]}(s)\big]\exp\big\{-\tilde{\Delta}_{[N,M]}(s+\rho)\big\}\\&\quad=\prod_m\prod_k\big(1+\lambda_k\big(G_{m[N,M]}(s)\big)\big)\exp\big\{-\lambda_k\big(G_{m[N,M]}(s)\big)\big\}\\&\quad\quad\times\exp\Big\{-\sum_m\operatorname{tr}\big(\tilde{\Delta}_{m[N,M]}(s+\rho)\big)\Big\},\end{aligned}\tag{4.6}$$

where the finite-dimensional matrix $G_{m[N,M]}(s)$ is given by

$$G_{m[N,M]}(s):=\hat{C}_{NM}(E_{Mm}(s)-Q_M)^{-1}\hat{B}_{NM}\quad(m\in\mathbf{Z}).\tag{4.7}$$

LEMMA 4.4. *Assume in the* FDLCP *system* (2.1) *that* $A(t)\in L_{\mathrm{PCD}}[0,h]$ *and* $B(t),C(t)\in L_{\mathrm{CAC}}[0,h]$, *and the domain* Ω *satisfies* A1 *and* A2′. *Then for any fixed* N, *it holds uniformly over* $s\in\partial\Omega$ *that*

$$\begin{cases}\lim\limits_{M\to\infty}\prod\limits_{m\neq0}\prod\limits_k\big(1+\lambda_k\big(G_{m[N,M]}(s)\big)\big)\exp\big\{-\lambda_k\big(G_{m[N,M]}(s)\big)\big\}=1,\\ \lim\limits_{M\to\infty}\sum\limits_{m\neq0}\operatorname{tr}\big(\tilde{\Delta}_{m[N,M]}(s+\rho)\big)=0.\end{cases}\tag{4.8}$$

On the basis of Lemma 4.4 and (4.6), it follows readily that for any $\mu>0$ and fixed $N>0$, there exists an integer $M_2(N,\mu)$ such that $\forall M\geq M_2(N,\mu)$ and $\forall s\in\partial\Omega$

$$\begin{aligned}&\big|\det_2\big[\underline{I}+\underline{G}_{[N,M]}(s)\big]\exp\big\{-\tilde{\Delta}_{[N,M]}(s+\rho)\big\}\\&\quad-\det_2\big[I_M+G_{0[N,M]}(s)\big]\exp\big\{-\operatorname{tr}\big(\tilde{\Delta}_{0[N,M]}(s+\rho)\big)\big\}\big|<\mu.\end{aligned}\tag{4.9}$$

A complete proof for (4.9) is given in Appendix B to keep our mainstream discussions clear.

4.3. Finite-dimensional 2-regularized Nyquist criterion. Summarizing the above discussions, we are led immediately to the following theorem, which reduces the Nyquist criterion of Corollary 3.2 to a finite-dimensional one in an asymptotic sense.

THEOREM 4.5. *Suppose in the* FDLCP *system* (2.1) *that* $A(t)$, $B(t)$, $C(t)\in L_{\mathrm{CAC}}[0,h]\cap L_{\mathrm{PCD}}[0,h]$. *Let* n_{us} *denote the number of the unstable eigenvalues of the open-loop Floquet state operator* $\underline{Q}-\underline{E}(j0)$ *in the fundamental strip* $\mathbf{C}_F$ *defined in* (3.7). *Take an arbitrary positive number* ρ *and a sufficiently small number* r *according to the open-loop eigenvalue condition of* $\underline{Q}-\underline{E}(j0)$ *on the imaginary axis such that* $\rho>r\geq0$. *If* N *and* M *are large enough truncation parameters satisfying* $M\geq N+1$, *then the closed-loop system* G_c *is asymptotically stable if and only if the modified Nyquist locus,* $\det_2[I_M+G_{0[N,M]}(s)]\exp\{-\operatorname{tr}(\tilde{\Delta}_{0[N,M]}(s+\rho))\}:N_{ab}\to\mathbf{C}$, *vanishes nowhere on* N_{ab} *and encircles the origin* n_{us} *times in the counterclockwise sense.* $\tilde{\Delta}_{0[N,M]}(s+\rho)$ *and* $G_{0[N,M]}(s)$ *are defined in* (4.5) *and* (4.7), *respectively, with* $m=0$.

Remark 6. Recall the recasting treatment suggested in Remark 5, by which A_{const} can be taken in such a way that the "DC-part" of $\tilde{A}(t)$ is zero. It would be worth noting that $\operatorname{tr}(\tilde{\Delta}_{0[N,M]}(s+\rho))$ will be identically zero in such a recasting treatment, and thus the exponential part can be dropped in the (modified) Nyquist locus (see Remark 4).

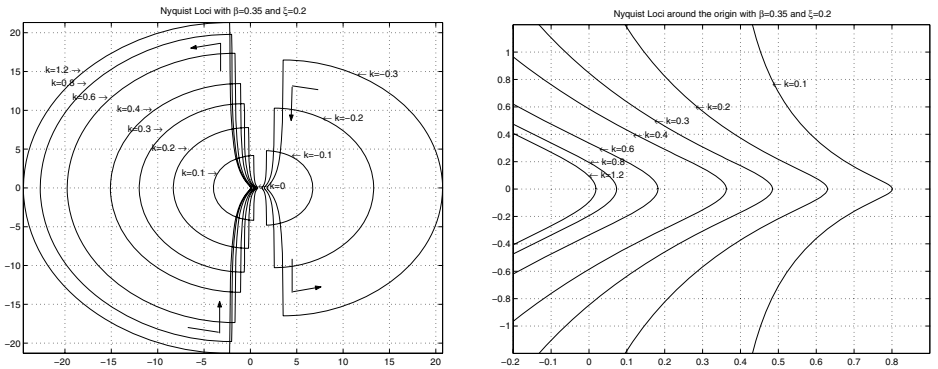
5. Numeric examples. Consider asymptotic stability of the lossy Mathieu equation by means of Theorem 4.5. The lossy Mathieu equation was frequently encountered in such studies as the rolling motion of ships [1], the flapping dynamics of the helicopter rotor blade [26], and the motion of a pendulum with a periodically excited support [12]. A comprehensive study about this differential equation can be found in [23] and [25]. It turns out to be one of the most widely studied FDLCP models in the literature, and hence using the lossy Mathieu equation as our numeric example is reasonable. More precisely, the lossy Mathieu equation is given by

$$\ddot{x}(t) + 2\xi\dot{x}(t) = [1 - 2\beta\cos\omega_h t]u(t), \quad \omega_h = 2 \quad (\text{i.e.}, h = \pi)$$

which leads to the state-space model

$$A(t) = \begin{bmatrix} 0 & 1 \\ 0 & -2\xi \end{bmatrix}, \quad B(t) = \begin{bmatrix} 0 \\ 1 - 2\beta\cos\omega_h t \end{bmatrix}, \quad C(t) = [1 \quad 0],$$

where the open-loop state matrix $A(t)$ is constant but the input matrix $B(t)$ is a π-periodic time-varying matrix, each entry of which is continuous and differentiable on $[0, h]$. In other words, in the open-loop system, $Q = A(t)$ and $P(t, 0) = I$. Now we introduce the output feedback $u(t) = -ky(t)$, where k is a scalar constant. This leads to a closed-loop FDLCP system with a π-periodic time-varying state matrix, and our problem is to test the closed-loop stability by Theorem 4.5.

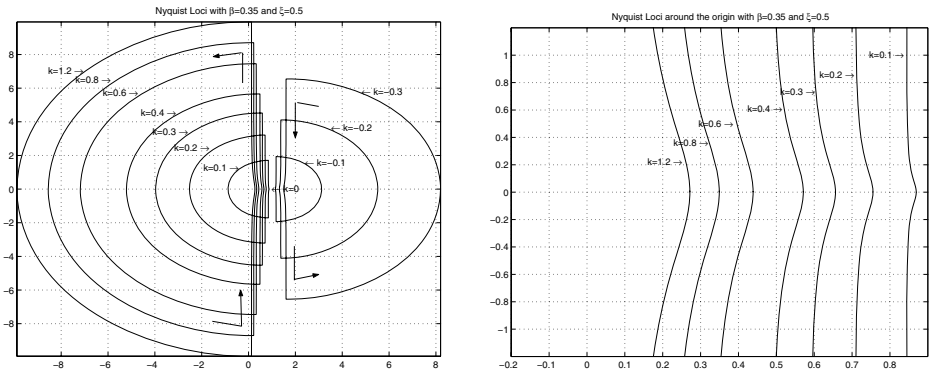


(a) Nyquist loci (b) Nyquist loci around the origin

FIG. 5.1. 2-*Regularized Nyquist loci with the output feedback gain* k *varying from* 1.2 *to* -0.3 *in the case of* $\beta = 0.35$ *and* $\xi = 0.2$ *(*$N_r = N_{0.05}$*, the arrows indicate the Nyquist locus direction).*

It is clear that the open-loop system has a zero eigenvalue. If we take a Nyquist contour as described in Figure 3.1, the corresponding region enclosed by this Nyquist contour has one unstable open-loop eigenvalue of the operator $\underline{Q} - \underline{E}(j0)$, assuming that $\xi > 0$. Note also that the Fourier series expansion of $B(t)$ has nonzero terms only up to the first harmonic wave. Since $P(t, 0) = I$, it follows that the skew truncation can be dispensed with, and only the staircase truncation on the corresponding (open-loop) harmonic transfer operator is enough. Figures 5.1 and 5.2 give the (modified) Nyquist loci under different parameters β, ξ while the output feedback gain k varies from 1.2 to -0.3. In the computations, the staircase truncation parameter $M = 10$, the shift factor $\rho = 0.1$, and the bypassing radius $r = 0.05$ are taken for simplicity. The computation results show that there are no numerically discernible differences among the Nyquist loci when larger M's are taken, for example, when $M = 20$.

From Figure 5.1, it can be asserted that in the case of $\beta = 0.35$ and $\xi = 0.2$, the closed-loop FDLCP system is stable when the feedback gain k is relatively small, while the closed-loop system slides to the stable/unstable boundary when the feedback gain k is overstrong (i.e., $k \geq 1.2$). However, in the case of $\beta = 0.35$ and $\xi = 0.5$, whose Nyquist loci are given in Figure 5.2, stability of the closed-loop FDLCP system has relatively strong robustness to the output feedback gain variation. From Figures 5.1 and 5.2, the Nyquist loci when positive feedback is applied, i.e., $k \leq 0$, tell us that the closed-loop FDLCP systems are unstable in both cases.



(a) Nyquist loci (b) Nyquist loci around the origin

FIG. 5.2. 2-*Regularized Nyquist loci with the output feedback gain* k *varying from* 1.2 *to* -0.3 *in the case of* $\beta = 0.35$ *and* $\xi = 0.5$ *(*$N_r = N_{0.05}$*, the arrows indicate the Nyquist locus direction).*

Finally, we observe that Nyquist loci under different bypassing radii r's can also reveal some structural features of the open-loop FDLCP systems. For example, in the case of $\beta = 0.5$, $\xi = 0.2$, and $k = 0.4$, Figure 5.3 gives the Nyquist loci with the bypassing radius being $r = 0.05$, 0.04, and 0.03, respectively. One can see that as $r \to 0$, the Nyquist locus goes to infinity on the portion corresponding to the bypassing semicircle. This clearly reflects the fact that the open-loop system has an unstable eigenvalue at the origin.

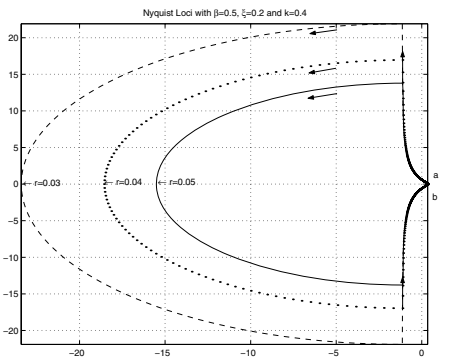


FIG. 5.3. 2-*Regularized Nyquist loci with different bypassing radii* r.

The stability results about the lossy Mathieu equation here coincide with those derived through the approximate modeling approach [31], which also gives necessary and sufficient stability conditions for FDLCP systems via the so-called harmonic

Lyapunov equation. Note that the lossy Mathieu equation may be reexpressed by a periodic differential equation similar to that of Exercise 1.5.6 of [12]. Hence, the stability assertion for Exercise 1.5.6 of [12] confirms that stability analysis through the Nyquist criterion upon the lossy Mathieu equation is effective.

6. Conclusion. In this paper, we established a generalized Nyquist criterion in FDLCP systems by means of the 2-regularized determinant related to the open-loop harmonic transfer operator. This work is inspired by the fact that the harmonic transfer operators of most practical FDLCP systems, which are defined via the Fourier series analysis of the system matrices, are Hilbert–Schmidt operators; that is, the use of the 2-regularized determinants can be validated for a large class of FDLCP systems, while the use of the usual determinant cannot. This criterion makes it possible to test the closed-loop asymptotic stability through open-loop analysis in a much more general FDLCP setting, compared to the Hill-determinant and trace class operator determinant techniques. The Hill-determinant defined on the harmonic transfer operator [26] is hard to validate in general. Moreover, by using the recasting technique suggested in Remark 5, the generalized Nyquist criterion can be applied to both open- and closed-loop FDLCP systems without involving the transition matrix and Floquet factorization computations, and thus can be implemented via finite-dimensional conditions in an asymptotic sense. In addition, it is clear that the 2-regularized Nyquist criterion applies to both SISO (single input/single output) and MIMO (multi input/multi output) cases. Observations (say by Theorem 7.4 of [11]) indicate that the Nyquist locus in the 2-regularized determinant sense is also continuous with regard to the periodically time-varying feedback gain $K(t)$, and thus can be utilized in robustness analysis. This is left as one of our subsequent research topics.

Appendix A. The function $f(n)$ of an integer n is defined by

$$f(n) = \begin{cases} 1, & n = 0, \\ |n|^{-1}, & n \neq 0. \end{cases}$$

Then we have $\sum_{n=N+1}^{\infty} f(n)^2 < \frac{1}{N} \quad (N \geq 1)$ and $\sum_{n=-\infty}^{\infty} f(n)^2 < 5$.

Appendix B.

Proof of Lemma 2.2. Under the assumptions on $A(t), B(t)$, and $C(t)$, the similarity transformation formulas of Lemma 2.1 apply. Thus the harmonic transfer operator can be rewritten as

$$\underline{G}(s) = \hat{\underline{C}}(\underline{E}(s) - \underline{Q})^{-1}\hat{\underline{B}} \tag{B.1}$$

for all $s \in \Omega\backslash\Lambda$. Here, $\hat{\underline{B}}$ and $\hat{\underline{C}}$ are bounded on l_2 by Corollary 2.2 of [9, p. 567]. Also, it is obvious that for each fixed $s \in \Omega\backslash\Lambda$, there exists a number $K(s) > 0$ such that

$$\|(\varphi_m(s)I - Q)^{-1}\| \leq K(s)f(m), \tag{B.2}$$

where f is defined in Appendix A. Noting the block-diagonal structure of $(\underline{E}(s)-\underline{Q})^{-1}$, it follows that $(\underline{E}(s) - \underline{Q})^{-1}$ is compact for $s \in \Omega\backslash\Lambda$. Since $\|(\underline{E}(s) - \underline{Q})^{-1}\|_2^2 = \sum_m \|(\varphi_m(s) - Q)^{-1}\|_2^2 \leq n\sum_m \|(\varphi_m(s) - Q)^{-1}\|^2$, (B.2) tells that $(\underline{E}(s) - \underline{Q})^{-1} \in \mathcal{C}_2(l_2)$ for each fixed $s \in \Omega\backslash\Lambda$. Noting that $\hat{\underline{B}}$ and $\hat{\underline{C}}$ are bounded on l_2, we obtain by (B.1) that $\underline{G}(s) \in \mathcal{C}_2(l_2)$ for each $s \in \Omega\backslash\Lambda$.

It remains to show that $\|\underline{G}(s)\|_2$ is uniformly bounded over $s \in \partial\Omega$. To this end, let us show that there exists a number $K > 0$ independent of s such that

$$\|(\varphi_m(s)I - Q)^{-1}\| \le Kf(m) \tag{B.3}$$

for all $s \in \partial\Omega$. To see this, we note that if $\|Q/z\| < 1$, then

$$\begin{aligned}\|(zI - Q)^{-1}\| &= \frac{1}{|z|}\left\|I - \frac{1}{z}Q + \frac{1}{z^2}Q^2 - \cdots\right\| \\ &\le \frac{1}{|z|}\left(\|I\| + \left\|\frac{1}{z}Q\right\| + \left\|\frac{1}{z^2}Q^2\right\| + \cdots\right) \le \frac{1}{|z| - \|Q\|}.\end{aligned}$$

Since $|\varphi_m(s)| > \|Q\|$ if $|m|\omega_h > |\mathrm{Im}(s)| + \|Q\|$, this inequality says if $|m|\omega_h > |\mathrm{Im}(s)| + \|Q\|$, then

$$\|(\varphi_m(s)I - Q)^{-1}\| \le (|\varphi_m(s)| - \|Q\|)^{-1}. \tag{B.4}$$

Thus it is clear by (2.6) that there exists an integer $m_0 > 0$ such that (B.4) holds for all integers $m \ge m_0$ and $s \in \partial\Omega$. Again from (2.6), there is $K_{m_0} > 0$ independent of s such that $\|(\varphi_m(s)I - Q)^{-1}\| \le K_{m_0} f(m)$ for all $m \ge m_0$ and $s \in \partial\Omega$. Furthermore, for each $s \in \partial\Omega$, $\varphi_m(s)$ is not an eigenvalue of Q by A1. Hence, from (2.6) there is $K'_{m_0} > 0$ such that for all $|m| < m_0$ and $s \in \partial\Omega$, $\|(\varphi_m(s)I - Q)^{-1}\| \le K'_{m_0} f(m)$. Taking $K = \max\{K_{m_0}, K'_{m_0}\}$, (B.3) follows for any $s \in \partial\Omega$. Thus it follows from Appendix A that for any $s \in \partial\Omega$, $\|(\underline{E}(s) - \underline{Q})^{-1}\|_2 \le [\sum_m nK^2 f(m)^2]^{1/2} < K'$ for some $K' > 0$ independent of $s \in \partial\Omega$. Finally, noting that $\underline{\hat{B}}$ and $\underline{\hat{C}}$ are bounded on l_2 and $\|\underline{\hat{C}}(\underline{E}(s) - \underline{Q})^{-1}\underline{\hat{B}}\|_2 \le \|\underline{\hat{C}}\|_{l_2/l_2}\|\underline{\hat{B}}\|_{l_2/l_2}\|(\underline{E}(s) - \underline{Q})^{-1}\|_2$, the uniform boundedness of $\|\underline{\hat{C}}(\underline{E}(s) - \underline{Q})^{-1}\underline{\hat{B}}\|_2$ over $\partial\Omega$ follows readily. □

Proof of Lemma 2.3. By the definition of $\varphi_m(s)$, if we write $s = x + jy$, then we obtain

$$|\varphi_m(s)^{-1}| = \frac{1}{\sqrt{x^2 + (y + m\omega_h)^2}} \le \begin{cases} (|y + m\omega_h|)^{-1} \le K_\varphi f(m) & (m \ne 0), \\ (\sqrt{x^2 + y^2})^{-1} \le K_\varphi f(m) & (m = 0) \end{cases} \tag{B.5}$$

for some $K_\varphi > 0$ independent of $s \in \Omega$, where A1 is used for $m \ne 0$ while A2 is used for $m = 0$. Furthermore, we observe from (B.5) that

$$\|\underline{E}^{-1}(s)\|_2 \le \left[\sum_m nK_\varphi^2 f(m)^2\right]^{1/2} \le K_E < \infty \tag{B.6}$$

for some $K_E > 0$ independent of $s \in \Omega$. The inequality (B.6) says that $\underline{E}^{-1}(s) \in \mathcal{C}_2(l_2)$ for each $s \in \Omega$. Since $\underline{A}$ and $\underline{Q}$ are bounded on l_2, $\underline{E}^{-1}(s)\underline{A}$ and $\underline{E}^{-1}(s)\underline{Q}$ belong to $\mathcal{C}_2(l_2)$ for each $s \in \Omega$.

To see (2.11), we note by Lemma 2.1 that $\underline{E}(s) - \underline{A} = \underline{P}(\underline{E}(s) - \underline{Q})\underline{P}^{-1}$, which implies

$$\underline{I} - \underline{E}^{-1}(s)\underline{A} = \underline{E}^{-1}(s)\underline{P}\,\underline{E}(s)(\underline{I} - \underline{E}^{-1}(s)\underline{Q})\underline{P}^{-1} \tag{B.7}$$

on the subset l_E of l_2. Furthermore, it is already known [29] on l_E that

$$\underline{\tilde{P}} = \underline{E}(s)\underline{P} - \underline{P}\,\underline{E}(s) \tag{B.8}$$

with $\underline{\tilde{P}} := \mathcal{T}\{\frac{d}{dt}P(t,0)\}$. Since l_E is $\underline{P}^{-1}$-invariant by Lemma 2.1, using (B.8) in (B.7) gives that on l_E

$$\begin{aligned}\underline{I} - \underline{E}^{-1}(s)\underline{A} &= (\underline{P} - \underline{E}^{-1}(s)\underline{\tilde{P}})(\underline{I} - \underline{E}^{-1}(s)\underline{Q})\underline{P}^{-1} \\ &= (\underline{I} - \underline{E}^{-1}(s)\underline{\tilde{P}}\,\underline{P}^{-1})(\underline{I} - \underline{P}\,\underline{E}^{-1}(s)\underline{Q}\underline{P}^{-1}). \qquad \text{(B.9)}\end{aligned}$$

Noting that all operators in (B.9) are bounded and l_E is dense in l_2, (B.9) is true on the whole l_2. On the other hand, it is straightforward to see that $\underline{E}^{-1}(s)\underline{\tilde{P}}\,\underline{P}^{-1}$ and $\underline{P}\,\underline{E}^{-1}(s)\underline{Q}\underline{P}^{-1}$ belong to $\mathcal{C}_2(l_2)$ since $\underline{E}^{-1}(s)$ does. These observations validate the following derivations:

$$\begin{aligned}\det_2[\underline{I} - \underline{E}^{-1}(s)\underline{A}] &= \det_2[\underline{I} - \underline{E}^{-1}(s)\underline{\tilde{P}}\,\underline{P}^{-1}]\det_2[\underline{I} - \underline{P}\,\underline{E}^{-1}(s)\underline{Q}\underline{P}^{-1}] \\ &\quad \times \exp\{-\mathrm{tr}(\underline{E}^{-1}(s)\underline{\tilde{P}}\,\underline{P}^{-1}\underline{P}\,\underline{E}^{-1}(s)\underline{Q}\underline{P}^{-1})\} \\ &= \det_2[\underline{I} - \underline{E}^{-1}(s)\underline{\tilde{P}}\,\underline{P}^{-1}]\det_2[\underline{I} - \underline{E}^{-1}(s)\underline{Q}] \\ &\quad \times \exp\{-\mathrm{tr}(\underline{E}^{-1}(s)\underline{\tilde{P}}\,\underline{E}^{-1}(s)\underline{Q}\underline{P}^{-1})\} \\ &=: g_A(s)\det_2[\underline{I} - \underline{E}^{-1}(s)\underline{Q}], \qquad \text{(B.10)}\end{aligned}$$

where $g_A(s) := \det_2[\underline{I} - \underline{E}^{-1}(s)\underline{\tilde{P}}\,\underline{P}^{-1}]\exp\{-\mathrm{tr}(\underline{E}^{-1}(s)\underline{\tilde{P}}\,\underline{E}^{-1}(s)\underline{Q}\underline{P}^{-1})\}$. Hence, the assertions regarding (2.11) will follow if it is shown that $g_A(s)$ does not vanish and is analytic over Ω.

We complete the task in two steps by showing that the two components of $g_A(s)$ vanish nowhere on Ω and are analytic in s over Ω.

Step 1. It is shown that $\det_2[\underline{I} - \underline{E}^{-1}(s)\underline{\tilde{P}}\,\underline{P}^{-1}]$ vanishes nowhere on Ω and is analytic in s over Ω. To see this, we note from (B.8) that on l_E,

$$\underline{I} - \underline{E}^{-1}(s)\underline{\tilde{P}}\,\underline{P}^{-1} = \underline{E}^{-1}(s)\underline{P}\,\underline{E}(s)\underline{P}^{-1}, \qquad \text{(B.11)}$$

which says that $\underline{I} - \underline{E}^{-1}(s)\underline{\tilde{P}}\,\underline{P}^{-1}$ is invertible on l_E for each $s \in \Omega$ since $\underline{E}(s)$ is invertible on l_E, and l_E is $\underline{P}$- and $\underline{P}^{-1}$-invariant (see Lemma 2.1). Since l_E is dense in l_2, this, in particular, implies that $\underline{I} - \underline{E}^{-1}(s)\underline{\tilde{P}}\,\underline{P}^{-1}$ has a dense range in l_2. Furthermore, one can claim that $\underline{I} - \underline{E}^{-1}(s)\underline{\tilde{P}}\,\underline{P}^{-1}$ is one-to-one on l_2 since $(\underline{I} - \underline{E}^{-1}(s)\underline{\tilde{P}}\,\underline{P}^{-1})\underline{x} = 0$ implies that $\underline{x} \in l_E$, and thus $\underline{x} = 0$ by the invertibility of $\underline{I} - \underline{E}^{-1}(s)\underline{\tilde{P}}\,\underline{P}^{-1}$ on l_E. On the basis of these facts, Theorem 2.7.6 of [21, p. 30] tells us that the operator $\underline{I} - \underline{E}^{-1}(s)\underline{\tilde{P}}\,\underline{P}^{-1}$ is actually invertible on the whole l_2. This, together with the invertibility of $\underline{I} - \underline{E}^{-1}(s)\underline{Q}$, implies by (B.9) that $\underline{I} - \underline{E}^{-1}(s)\underline{A}$ is invertible on l_2. On the other hand, Property 1.8(e) of [5, p. 17] ensures that $\det_2[\underline{I} - \underline{E}^{-1}(s)\underline{\tilde{P}}\,\underline{P}^{-1}] \neq 0$ on Ω.

To show that $\det_2[\underline{I} - \underline{E}^{-1}(s)\underline{\tilde{P}}\,\underline{P}^{-1}]$ is analytic in s over Ω, we need some preparations. To this end, let us approximate $\underline{\tilde{P}}$ and $\underline{P}^{-1}$ by $[\underline{\tilde{P}}]_N$ and $[\underline{P}^{-1}]_N$, respectively, as follows:

$$[\underline{\tilde{P}}]_N := \mathcal{T}\left\{\sum_{|m|\le N}\tilde{P}_m e^{jm\omega_h t}\right\}, \quad [\underline{P}^{-1}]_N := \mathcal{T}\left\{\sum_{|m|\le N}\check{P}_m e^{jm\omega_h t}\right\}.$$

Here $\{\tilde{P}_m\}$ and $\{\check{P}_m\}$ are the Fourier coefficients sequences of $\tilde{P}(t,0)$ and $P^{-1}(t,0)$, respectively. Now let us define the operators

$$\underline{K}(s) := -\underline{E}^{-1}(s)\underline{\tilde{P}}\,\underline{P}^{-1}, \quad \underline{K}_N(s) := -\underline{E}^{-1}(s)[\underline{\tilde{P}}]_N[\underline{P}^{-1}]_N.$$

Since $\underline{E}^{-1}(s) \in \mathcal{C}_2(l_2)$, it follows that $\underline{K}_N(s) \in \mathcal{C}_2(l_2)$ for each N and $s \in \Omega$. By the structure of the operators $\underline{\tilde{P}}$ and $\underline{P}^{-1}$ and (B.6), Proposition 1.3 of [9, p. 98] tells us that

$$\|\underline{K}(s)\|_2 \le \|\underline{E}^{-1}(s)\|_2 \|\underline{\tilde{P}}\|_{l_2/l_2} \|\underline{P}^{-1}\|_{l_2/l_2} \le K_1 < \infty, \tag{B.12}$$

$$\begin{aligned}\|\underline{K}_N(s)\|_2 &\le \|\underline{E}^{-1}(s)\|_2 \|[\underline{\tilde{P}}]_N\|_{l_2/l_2} \|[\underline{P}^{-1}]_N\|_{l_2/l_2} \le K_E \sum_{|m|\le N} \|\tilde{P}_m\| \sum_{|m|\le N} \|\check{P}_m\| \\ &\le K_E \sum_{m=-\infty}^{+\infty} \|\tilde{P}_m\| \sum_{m=-\infty}^{+\infty} \|\check{P}_m\| \le K_2 < \infty \end{aligned} \tag{B.13}$$

for some $K_1 > 0$ and $K_2 > 0$ independent of $s \in \Omega$ and N. This is because $\tilde{P}(t,0)$ and $P(t,0)$ belong to $L_{\mathrm{CAC}}[0,h]$ under the assumption about $A(t)$ [29], and thus $\sum_{m=-\infty}^{+\infty} \|\tilde{P}_m\| < \infty$ and $\sum_{m=-\infty}^{+\infty} \|\check{P}_m\| < \infty$. By A2 and the form of $\underline{E}(s)$, $\underline{K}_N(s)$ is analytic in s over Ω in the elementwise sense since each entry of $\underline{K}_N(s)$ is a finite sum due to the skew-strip structure of $[\underline{\tilde{P}}]_N$ and $[\underline{P}^{-1}]_N$. In the following, we say that $\underline{K}_N(s)$ is an analytic $\mathcal{C}_2(l_2)$-valued function in this sense.

Now we show that $\underline{K}(s) - \underline{K}_N(s) \to 0$ in the norm of $\mathcal{C}_2(l_2)$ uniformly over Ω as $N \to \infty$. To see this, we note that

$$\begin{aligned}\|\underline{K}(s) - \underline{K}_N(s)\|_2 &\le \|\underline{E}^{-1}(s)\|_2 \|\underline{\tilde{P}}\,\underline{P}^{-1} - [\underline{\tilde{P}}]_N [\underline{P}^{-1}]_N\|_{l_2/l_2} \\ &\le K_E \big(\|\underline{\tilde{P}} - [\underline{\tilde{P}}]_N\|_{l_2/l_2} \|\underline{P}^{-1}\|_{l_2/l_2} + \|[\underline{\tilde{P}}]_N\|_{l_2/l_2} \|\underline{P}^{-1} - [\underline{P}^{-1}]_N\|_{l_2/l_2} \big). \end{aligned} \tag{B.14}$$

On the other hand, by the structures of $\underline{\tilde{P}} - [\underline{\tilde{P}}]_N$ and $\underline{P}^{-1} - [\underline{P}^{-1}]_N$ and the facts that $\tilde{P}(t,0)$ and $P^{-1}(t,0)$ belong to $L_{\mathrm{CAC}}[0,h]$, it follows readily that $\|\underline{\tilde{P}} - [\underline{\tilde{P}}]_N\|_{l_2/l_2}$ and $\|\underline{P}^{-1} - [\underline{P}^{-1}]_N\|_{l_2/l_2}$ go to zero as $N \to \infty$. These facts, together with the fact that there is an upper bound for $\|[\underline{\tilde{P}}]_N\|_{l_2/l_2}$ independent of N, imply that $\|\underline{K}(s) - \underline{K}_N(s)\|_2 \to 0$ uniformly over Ω as $N \to \infty$. This ensures that $\underline{K}(s)$ is an analytic $\mathcal{C}_2(l_2)$-valued function, which leads us to the desired consequence.

Step 2. It is shown that $\exp\{-\mathrm{tr}(\underline{E}^{-1}(s)\underline{\tilde{P}}\,\underline{E}^{-1}(s)\underline{Q}\underline{P}^{-1})\}$ does not vanish and is analytic over $s \in \Omega$. By Property 1.3(c) of [5, p. 14] and Theorem 2.1 of [9, p. 111], we have

$$\begin{aligned}|\mathrm{tr}(\underline{E}^{-1}(s)\underline{\tilde{P}}\,\underline{E}^{-1}(s)\underline{Q}\underline{P}^{-1})| &\le \|\underline{E}^{-1}(s)\underline{\tilde{P}}\,\underline{E}^{-1}(s)\underline{Q}\underline{P}^{-1}\|_1 \\ &\le \|\underline{E}^{-1}(s)\|_2^2 \|\underline{\tilde{P}}\|_{l_2/l_2} \|\underline{Q}\underline{P}^{-1}\|_{l_2/l_2} \le K_E^2 \|\underline{\tilde{P}}\|_{l_2/l_2} \|\underline{Q}\underline{P}^{-1}\|_{l_2/l_2} \le K_3 < \infty \end{aligned}$$

for some $K_3 > 0$ independent of $s \in \Omega$ since $\underline{\tilde{P}}$, $\underline{Q}$, and $\underline{P}^{-1}$ are bounded on l_2. This inequality says clearly that $\exp\{-\mathrm{tr}(\underline{E}^{-1}(s)\underline{\tilde{P}}\,\underline{E}^{-1}(s)\underline{Q}\underline{P}^{-1})\}$ does not vanish on Ω.

To show that $\exp\{-\mathrm{tr}(\underline{E}^{-1}(s)\underline{\tilde{P}}\,\underline{E}^{-1}(s)\underline{Q}\underline{P}^{-1})\}$ is analytic over $s \in \Omega$, by Remark 10.3 of [24, p. 197], it is enough to show that $\mathrm{tr}(\underline{E}^{-1}(s)\underline{\tilde{P}}\,\underline{E}^{-1}(s)\underline{Q}\underline{P}^{-1}) =: t_A(s)$ is analytic over $s \in \Omega$. To this end, we further define the trace function

$$\mathrm{tr}(\underline{E}^{-1}(s)[\underline{\tilde{P}}]_N \underline{E}^{-1}(s)\underline{Q}[\underline{P}^{-1}]_N) =: t_N(s).$$

It should be pointed out that for each fixed N, $\underline{E}^{-1}(s)[\underline{\tilde{P}}]_N \underline{E}^{-1}(s)\underline{Q}[\underline{P}^{-1}]_N$ belongs to $\mathcal{C}_1(l_2)$ since $\underline{E}^{-1}(s) \in \mathcal{C}_2(l_2)$. Hence, $t_N(s)$ is well defined. Now we observe that

$$|t_A(s) - t_N(s)| \le |\mathrm{tr}(\underline{E}^{-1}(s)(\underline{\tilde{P}} - [\underline{\tilde{P}}]_N)\underline{E}^{-1}(s)\underline{Q}\underline{P}^{-1})|$$

$$
\begin{aligned}
&\quad + |\mathrm{tr}(\underline{E}^{-1}(s)[\underline{\tilde{P}}]_N\underline{E}^{-1}(s)\underline{Q}(\underline{P}^{-1}-[\underline{P}^{-1}]_N))|\\
&\le \|(\underline{E}^{-1}(s)(\underline{\tilde{P}}-[\underline{\tilde{P}}]_N)\underline{E}^{-1}(s)\underline{Q}\underline{P}^{-1})\|_1\\
&\quad + \|\underline{E}^{-1}(s)[\underline{\tilde{P}}]_N\underline{E}^{-1}(s)\underline{Q}(\underline{P}^{-1}-[\underline{P}^{-1}]_N)\|_1\\
&\le \|\underline{E}^{-1}(s)\|_2^2\|\underline{\tilde{P}}-[\underline{\tilde{P}}]_N\|_{l_2/l_2}\|\underline{Q}\underline{P}^{-1}\|_{l_2/l_2}\\
&\quad + \|\underline{E}^{-1}(s)\|_2^2\|[\underline{\tilde{P}}]_N\|_{l_2/l_2}\|\underline{Q}\|_{l_2/l_2}\|\underline{P}^{-1}-[\underline{P}^{-1}]_N\|_{l_2/l_2}.
\end{aligned}
$$

As seen in Step 1 that $\tilde{P}(t,0)$ and $P^{-1}(t,0)$ belong to $L_{\mathrm{CAC}}[0,h]$, it follows that $\|\underline{\tilde{P}}-[\underline{\tilde{P}}]_N\|_{l_2/l_2}$ and $\|\underline{P}^{-1}-[\underline{P}^{-1}]_N\|_{l_2/l_2}$ go to zero as $N\to\infty$. These facts, together with $\|\underline{E}^{-1}(s)\|_2^2\le K_E^2$ and the fact that $\|[\underline{\tilde{P}}]_N\|_{l_2/l_2}$ has an upper bound independent of N, imply that $|t_N(s)-t_A(s)|\to 0$ as $N\to\infty$ uniformly over $s\in\Omega$. This indicates that to complete the proof, it suffices to show that for each fixed N, the function $t_N(s)$ is analytic over $s\in\Omega$.

It is easy to see from the structure of $[\underline{\tilde{P}}]_N$ and $[\underline{P}^{-1}]_N$ that the blockwise (m,m)th entry on the diagonal of the operator $\underline{E}^{-1}(s)[\underline{\tilde{P}}]_N\underline{E}^{-1}(s)\underline{Q}[\underline{P}^{-1}]_N$ is

$$
\varphi_m^{-1}(s)[\tilde{P}_N,\dots,\tilde{P}_{-N}]\,\mathrm{diag}\big[\varphi_{m-N}^{-1}(s),\dots,\varphi_{m+N}^{-1}(s)\big]\begin{bmatrix}Q\check{P}_{-N}\\ \vdots\\ Q\check{P}_N\end{bmatrix}
$$

whose trace is denoted by $t_{Nm}(s)$. Obviously, $t_N(s)=\sum_{m=-\infty}^{\infty}t_{Nm}(s)$. It is straightforward to see that $t_{Nm}(s)$ is analytic over Ω, and

$$
|t_{Nm}(s)|\le nK_N|\varphi_m^{-1}(s)|\max\big\{|\varphi_{m-N}^{-1}(s)|,\dots,|\varphi_{m+N}^{-1}(s)|\big\},
$$

where $K_N=\|[\tilde{P}_N,\dots,\tilde{P}_{-N}]\|\dots\|[\check{P}_{-N}^TQ^T,\dots,\check{P}_N^TQ^T]\|$. Furthermore, by Appendix A and (B.5),

$$
\begin{aligned}
&\left|t_N(s)-\sum_{|m|\le M}t_{Nm}(s)\right|\le\sum_{|m|>M}|t_{Nm}(s)|\\
&\le nK_N\left[\sum_{|m|>M}|\varphi_m^{-1}(s)|^2\right]^{1/2}\left[\sum_{|m|>M}\big(\max\big\{|\varphi_{m-N}^{-1}(s)|,\dots,|\varphi_{m+N}^{-1}(s)|\big\}\big)^2\right]^{1/2}\\
&\le nK_N\left[\sum_{|m|>M}K_\varphi^2f^2(m)\right]^{1/2}\left[\sum_{m=-\infty}^{\infty}\big(\max\big\{|\varphi_{m-N}^{-1}(s)|,\dots,|\varphi_{m+N}^{-1}(s)|\big\}\big)^2\right]^{1/2}\\
&< nK_N\left[\frac{2K_\varphi^2}{M}\right]^{1/2}\left[(2N+1)K_\varphi^2f(0)^2+\sum_{|m|\ge1}K_\varphi^2f^2(m)\right]^{1/2}\\
&< nK_NK_\varphi^2\left[\frac{2}{M}\right]^{1/2}[(2N+1)+5]^{1/2}=:nK_N'\left[\frac{2}{M}\right]^{1/2}\to 0\quad(M\to\infty)
\end{aligned}
$$

since $K_N'<\infty$ for any fixed $N\ge1$. The above arguments say that $\sum_{|m|\le M}t_{Nm}(s)$ converges to $t_N(s)$ uniformly over $s\in\Omega$. Therefore, it follows that $t_N(s)$ is also analytic over $s\in\Omega$.

Finally, let us show the assertion that $(\underline{I}-\underline{E}^{-1}(s)\underline{A})^{-1}$ is bounded on l_2. To this end, we note by (B.8) that $\underline{E}^{-1}(s)\underline{P}^{-1}\underline{E}(s)=\underline{P}^{-1}+\underline{E}^{-1}(s)\underline{P}^{-1}\underline{\tilde{P}}\,\underline{P}^{-1}$ on l_E. Hence, $\underline{E}^{-1}(s)\underline{P}^{-1}\underline{E}(s)$ is bounded on l_E. Obviously, this implies that $\underline{P}\,\underline{E}^{-1}(s)\underline{P}^{-1}\underline{E}(s)$,

which is the inverse of $\underline{I}-\underline{E}^{-1}(s)\underline{\tilde{P}}\,\underline{P}^{-1}$ by (B.11), is bounded on l_E. Since l_E is dense in l_2, it follows that $(\underline{I}-\underline{E}^{-1}(s)\underline{\tilde{P}}\,\underline{P}^{-1})^{-1}$ is also bounded on l_2. On the other hand, by the block-diagonal structure of $(\underline{I}-\underline{E}^{-1}(s)\underline{Q})^{-1}$ and the assumption about Ω, it is straightforward to show that $(\underline{I}-\underline{E}^{-1}(s)\underline{Q})^{-1}$ is also bounded on l_2. Summarizing these facts leads to the desired assertion by (B.9). □

Proof of Lemma 2.4. The second equality of (2.12) follows from the 2-regularized determinant definition and the fact that $\underline{E}^{-1}(s)\underline{Q}$ has an eigenvalue at each point $\lambda_k(Q)/(s+jm\omega_h)$, $m\in\mathbf{Z}$.

To see the assertions about $f_Q(s)$, we consider only the case of $n=1$ without loss of generality. The arguments are given by means of the results in Chapter 15 of [24, pp. 298–303]. More precisely, let us define the function sequence $\{f_m(s)\}$ by

$$f_m(s) := \Big(1-\frac{\lambda(Q)}{s+jm\omega_h}\Big)\exp\Big\{\frac{\lambda(Q)}{s+jm\omega_h}\Big\} \quad (m\in\mathbf{Z}).$$

By A2, the function $f_m(s)$ is analytic on Ω and has a zero at $\lambda(Q)-jm\omega_h$. Note by (B.5) that $|\frac{\lambda(Q)}{s+jm\omega_h}|\le|\lambda(Q)|K_\varphi f(m)$ for any $m\in\mathbf{Z}$ and $s\in\Omega$. Hence, there exists a finite integer $m_0>1$ such that $|\frac{\lambda(Q)}{s+jm\omega_h}|\le 1\ \forall|m|\ge m_0$ and $\forall s\in\Omega$, which implies by Lemma 15.8 of [24, p. 301] that

$$|1-f_m(s)|\le\Big|\frac{\lambda(Q)}{s+jm\omega_h}\Big|^2\le|\lambda(Q)|^2K_\varphi^2f^2(m)\quad(\forall|m|\ge m_0\quad\forall s\in\Omega).$$

The above arguments tell us that

$$\sum_{|m|\ge m_0}|1-f_m(s)|\le\sum_{|m|\ge m_0}|\lambda(Q)|^2K_\varphi^2f^2(m)\le\frac{2}{m_0-1}K_\varphi^2|\lambda(Q)|^2 \tag{B.15}$$

by Appendix A, which implies that $\sum_m|1-f_m(s)|$ is uniformly convergent on Ω. Then it follows by the first conclusion of Theorem 15.6 of [24, p. 300] that $\prod_m f_m(s)$ converges uniformly on Ω, and thus the product $\prod_m f_m(s)$ is analytic on Ω. The zeros property is a direct result of the second conclusion of Theorem 15.6 of [24, p. 300]. □

Proof of Lemma 4.1. First let us show that $\det_2[\underline{I}+\underline{G}_{[N]}(s)]$ is well defined for each N. That is, we must show that $\underline{G}_{[N]}(s)\in\mathcal{C}_2(l_2)$ for each N and $s\in\partial\Omega$. To see this, we note by the assumptions about the system matrices that $\hat{B}(t)$ and $\hat{C}(t)$ belong to $L_{\mathrm{CAC}}[0,h]$ [29]. Furthermore, by the structures of the operators $\underline{\hat{B}}_{[N]}$ and $\underline{\hat{C}}_{[N]}$, it follows that

$$\begin{cases}\|\underline{\hat{B}}_{[N]}\|_{l_2/l_2}\le\displaystyle\sum_{|m|\le N}\|\hat{B}_m\|\le\sum_{m=-\infty}^{+\infty}\|\hat{B}_m\|\le K_B<\infty,\\ \|\underline{\hat{C}}_{[N]}\|_{l_2/l_2}\le K_C<\infty\end{cases} \tag{B.16}$$

for some $K_B>0$ and $K_C>0$ that are independent of N. These facts, together with the definition of $\underline{G}_{[N]}(s)$ and (B.3), imply that for any $s\in\partial\Omega$

$$\begin{aligned}\|\underline{G}_{[N]}(s)\|_2&\le\|\underline{\hat{B}}_{[N]}\|_{l_2/l_2}\|\underline{\hat{C}}_{[N]}\|_{l_2/l_2}\|(\underline{E}(s)-\underline{Q})^{-1}\|_2\\ &\le K_BK_C\Big[n\sum_m K^2f^2(m)\Big]^{1/2}<\sqrt{5n}K_BK_CK<\infty.\end{aligned} \tag{B.17}$$

Equation (B.17) gives that $\underline{G}_{[N]}(s)$ belongs to $\mathcal{C}_2(l_2)$ and $\|\underline{G}_{[N]}(s)\|_2$ has a uniform upper bound for all N and $s \in \partial\Omega$. Similarly we can show that $\underline{G}_{[N,M]}(s) \in \mathcal{C}_2(l_2)$ and $\|\underline{G}_{[N,M]}(s)\|_2$ has a uniform upper bound for all N, M, and $s \in \partial\Omega$.

Now we show the main assertion. It is known from Theorem 7.4 of [11, p. 69] that

$$\begin{aligned} &|\det_2[\underline{I} + \underline{G}(s)] - \det_2[\underline{I} + \underline{G}_{[N]}(s)]| \\ &\quad \le \|\underline{G}(s) - \underline{G}_{[N]}(s)\|_2 \exp\left\{\frac{1}{2}(\|\underline{G}(s)\|_2 + \|\underline{G}_{[N]}(s)\|_2 + 1)^2\right\}. \end{aligned} \tag{B.18}$$

Hence, the uniform boundedness of $\|\underline{G}_{[N]}(s)\|_2$ over N and $s \in \partial\Omega$ together with that of $\|\underline{G}(s)\|_2$ over $s \in \partial\Omega$ says that to show the main convergence, it suffices to show that $\|\underline{G}(s) - \underline{G}_{[N]}(s)\|_2 \to 0$ uniformly for all $s \in \partial\Omega$ as $N \to \infty$. To see this, we observe that

$$\begin{aligned} \|\underline{G}(s) - \underline{G}_{[N]}(s)\|_2 &\le \|\hat{\underline{C}} - \hat{\underline{C}}_{[N]}\|_{l_2/l_2}\|(\underline{E}(s) - \underline{Q})^{-1}\|_2\|\hat{\underline{B}}\|_{l_2/l_2} \\ &\quad + \|\hat{\underline{C}}_{[N]}\|_{l_2/l_2}\|(\underline{E}(s) - \underline{Q})^{-1}\|_2\|\hat{\underline{B}} - \hat{\underline{B}}_{[N]}\|_{l_2/l_2} \\ &< \sqrt{5n}KK_B\|\hat{\underline{C}} - \hat{\underline{C}}_{[N]}\|_{l_2/l_2} + \sqrt{5n}KK_C\|\hat{\underline{B}} - \hat{\underline{B}}_{[N]}\|_{l_2/l_2}, \end{aligned} \tag{B.19}$$

where (B.16) is used. Furthermore, by the skew structure of $\hat{\underline{C}} - \hat{\underline{C}}_{[N]}$, it follows that $\|\hat{\underline{C}} - \hat{\underline{C}}_{[N]}\|_{l_2/l_2} \le \sum_{|m|\ge N}\|\hat{C}_m\| \to 0$ as $(N \to \infty)$ since $\hat{C}(t) \in L_{\mathrm{CAC}}[0,h]$. Similarly, by $\hat{B}(t) \in L_{\mathrm{CAC}}[0,h]$ we have $\|\hat{\underline{B}} - \hat{\underline{B}}_{[N]}\|_{l_2/l_2} \to 0$ as $N \to \infty$. Using these facts in (B.19), the desired convergence follows. ∎

Proof of Lemma 4.2. From the proof of Lemma 4.1, we have $\underline{G}_{[N]}(s), \underline{G}_{[N,M]}(s) \in \mathcal{C}_2(l_2)$ for all N, M, and $s \in \partial\Omega$, and $\|\underline{G}_{[N]}(s)\|_2$ and $\|\underline{G}_{[N,M]}(s)\|_2$ are uniformly bounded from above over N, M, and $s \in \partial\Omega$. Therefore, an inequality similar to (B.18) between $\underline{G}_{[N]}(s)$ and $\underline{G}_{[N,M]}(s)$ can be claimed. This means that to show the result, it suffices to show that

$$\|\underline{G}_{[N,M]}(s) - \underline{G}_{[N]}(s)\|_2 \to 0 \quad \forall s \in \partial\Omega \quad (M \to \infty) \tag{B.20}$$

uniformly for each fixed $N > 0$. To this end, we focus the attention on the inequality

$$\begin{aligned} \|\underline{G}_{[N,M]}(s) - \underline{G}_{[N]}(s)\|_2 &\le \|\check{\underline{C}}_{[N,M]}(\underline{E}(s) - \underline{Q})^{-1}\|_2\|\hat{\underline{B}}_{[N,M]}\|_{l_2/l_2} \\ &\quad + \|\hat{\underline{C}}_{[N]}\|_{l_2/l_2}\|(\underline{E}(s) - \underline{Q})^{-1}\check{\underline{B}}_{[N,M]}\|_2, \end{aligned} \tag{B.21}$$

where $\check{\underline{B}}_{[N,M]} := \hat{\underline{B}}_{[N]} - \hat{\underline{B}}_{[N,M]}$. More explicitly, it is given by the infinite-dimensional matrix

$$\check{\underline{B}}_{[N,M]} := \begin{bmatrix} \ddots & \ddots & \ddots & & & & 0 \\ & \check{B}_{NMl} & 0 & \check{B}_{NMu} & & & \\ & & \check{B}_{NMl} & 0 & \check{B}_{NMu} & & \\ & & & \check{B}_{NMl} & 0 & \check{B}_{NMu} & \\ 0 & & & & \ddots & \ddots & \ddots \end{bmatrix}$$

with the entry matrices $\check{B}_{NMl}$ and $\check{B}_{NMu}$ given by

$$\check{B}_{NMl} = \underbrace{\begin{bmatrix} 0 & \hat{B}_N & \cdots & \hat{B}_1 \\ & \ddots & \ddots & \vdots \\ & & \ddots & \hat{B}_N \\ 0 & & & 0 \end{bmatrix}}_{(2M+1)}, \quad \check{B}_{NMu} = \underbrace{\begin{bmatrix} 0 & & & 0 \\ \hat{B}_{-N} & \ddots & & \\ \vdots & \ddots & \ddots & \\ \hat{B}_{-1} & \cdots & \hat{B}_{-N} & 0 \end{bmatrix}}_{(2M+1)}.$$

The matrix $\underline{\check{C}}_{[N,M]}$ is defined similarly but in terms of $\{\hat{C}_m\}_{m=-N}^{N}$.

Furthermore, by the structure of $\underline{\hat{B}}_{[N,M]}$, it is easy to see that $\|\underline{\hat{B}}_{[N,M]}\|_{l_2/l_2} \le \sum_{|m|\le N} \|\hat{B}_m\| \le \sum_{m=-\infty}^{+\infty} \|\hat{B}_m\| \le K_B < \infty$ as in (B.16). This, together with the fact that $\|\underline{\hat{C}}_{[N]}\|_{l_2/l_2} \le K_C$, implies that to complete the proof of (B.20) via (B.21), it remains to show that as $M \to \infty$

$$\|\underline{\check{C}}_{[N,M]}(\underline{E}(s) - \underline{Q})^{-1}\|_2 \to 0, \quad \|(\underline{E}(s) - \underline{Q})^{-1}\underline{\check{B}}_{[N,M]}\|_2 \to 0 \tag{B.22}$$

uniformly for $s \in \partial\Omega$. Now we show that the convergence of (B.22) is true. To see this, we note that

$$\begin{aligned} &\big\|(\underline{E}(s) - \underline{Q})^{-1}\underline{\check{B}}_{[N,M]}\big\|_2 \\ &\quad \le \big\|(\underline{E}(s) - \underline{Q})^{-1}\underline{\check{B}_l}_{[N,M]}\big\|_2 + \big\|(\underline{E}(s) - \underline{Q})^{-1}\underline{\check{B}_u}_{[N,M]}\big\|_2, \end{aligned} \tag{B.23}$$

where $\underline{\check{B}_l}_{[N,M]}$ and $\underline{\check{B}_u}_{[N,M]}$ are the lower and upper triangle portions of $\underline{\check{B}}_{[N,M]}$, respectively. Hence, by the structures of $\underline{E}_M(j\varphi)$, $\underline{Q}_M$, and $\underline{\check{B}_l}_{[N,M]}$ as well as the fact that the entries of $\check{B}_{NMl}$ are zero except its right-upper blocks, we have

$$\begin{aligned} \big\|(\underline{E}(s) - \underline{Q})^{-1}\underline{\check{B}_l}_{[N,M]}\big\|_2 &= \Big[\sum_m \big\|(E_{Mm}(s) - Q_M)^{-1}\check{B}_{NMl}\big\|_2^2\Big]^{1/2} \\ &\le \Big[\sum_m \big\|\partial_N\big((E_{Mm}(s) - Q_M)^{-1}\big)\big\|_2^2 \cdot \|\check{B}_{NMl}\|^2\Big]^{1/2}, \end{aligned} \tag{B.24}$$

where $\partial_N(\cdot)$ means taking out the first N block columns from $(\cdot)$. Moreover, by similar arguments to the above, it readily follows that $\|\check{B}_{NMl}\| \le K_B$ since $\check{B}_{NMl}$ is a submatrix of $\underline{\hat{B}}_{[N,M]}$. Hence, it is easy to see by (B.3) that under our standing assumption $M \ge N + 1$,

$$\begin{aligned} &\big\|(\underline{E}(s) - \underline{Q})^{-1}\underline{\check{B}_l}_{[N,M]}\big\|_2 \\ &\quad \le K_B\Big[\sum_m \sum_{k=0}^{N-1} n\|((s + j(m(2M+1) - M + k))I - Q)^{-1}\|_2^2\Big]^{1/2} \\ &\quad \le K_B\sqrt{n}\Big[\sum_m \sum_{k=0}^{N-1} K^2 f^2(m(2M+1) - M + k)\Big]^{1/2} \\ &\quad \le K_B\sqrt{n}\Big[NK^2 \sum_m \max_{k\in\{0,1,\ldots,N-1\}}\{f^2(m(2M+1) - M + k)\}\Big]^{1/2} \end{aligned}$$

$$\le KK_B\sqrt{nN}\Big[\sum_m f^2\Big(\min_{k\in\{0,1,\dots,N-1\}}\{|m(2M+1)-M+k|\}\Big)\Big]^{1/2}$$
$$< KK_B\sqrt{nN}\Big[\sum_m f^2(m(M+1))\Big]^{1/2}$$
$$= KK_B\sqrt{nN}\Big[\frac{1}{(M+1)^2}\sum_m f^2(m)\Big]^{1/2} \le \frac{KK_B\sqrt{5nN}}{M+1}. \tag{B.25}$$

Thus, for each fixed N and for any $\mu>0$, there exists an integer $M_0'(N,\mu)>0$ such that

$$\|\hat{\underline{C}}_{[N]}\|_{l_2/l_2}\big\|(\underline{E}(s)-\underline{Q})^{-1}\check{\underline{B}}_{l[N,M]}\big\|_2 < \frac{\mu}{4} \quad (\forall M\ge M_0'(N,\mu) \quad \forall s\in\partial\Omega). \tag{B.26}$$

The above arguments can be repeated on the second term of the right-hand side of (B.23). Hence, for the same $\mu>0$ and $M_0'(N,\mu)$, it is easy to see that

$$\|\hat{\underline{C}}_{[N]}\|_{l_2/l_2}\big\|(\underline{E}(s)-\underline{Q})^{-1}\check{\underline{B}}_{u[N,M]}\big\|_2 < \frac{\mu}{4} \quad (\forall M\ge M_0'(N,\mu) \quad \forall s\in\partial\Omega), \tag{B.27}$$

where we used the fact that $\|\check{B}_{NMl}\|$ and $\|\check{B}_{NMu}\|$ have the same upper bound. From (B.26) and (B.28), it follows that

$$\|\hat{\underline{C}}_{[N]}\|_{l_2/l_2}\big\|(\underline{E}(s)-\underline{Q})^{-1}\check{\underline{B}}_{[N,M]}\big\|_2 < \frac{\mu}{2} \quad (\forall M\ge M_0'(N,\mu) \quad \forall s\in\partial\Omega). \tag{B.28}$$

In a similar way, one can conclude that for each fixed N and any $\mu>0$, there exists an integer $M_0''(N,\mu)>0$ such that

$$\|\hat{\underline{B}}_{[N,M]}\|_{l/2l_2}\big\|\check{\underline{C}}_{[N,M]}(\underline{E}(s)-\underline{Q})^{-1}\big\|_2 < \frac{\mu}{2} \quad (\forall M\ge M_0''(N,\mu) \quad \forall s\in\partial\Omega). \tag{B.29}$$

Then, the desired convergence assertion follows from (B.21), (B.28), and (B.29) by taking $M_0(N,\mu)=\max\{M_0'(N,\mu),M_0''(N,\mu)\}$. □

Proof of Lemma 4.4. To see the first assertion of (4.8), we observe from (B.3) that

$$\|G_{m[N,M]}(s)\| \le \|\hat{C}_{NM}\|\cdot\|(E_{Mm}(s)-Q_M)^{-1}\|\cdot\|\hat{B}_{NM}\|$$
$$\le K\,\|\hat{C}_{NM}\|\cdot\|\hat{B}_{NM}\|\,\max\{f(m(2M+1)-M),\dots,$$
$$f(m(2M+1)),\dots,f(m(2M+1)+M)\}$$
$$< K\,\|\hat{C}_{NM}\|\cdot\|\hat{B}_{NM}\|\,f(|m|(M+1)) \le K\,K_B'K_C'\,f(|m|(M+1)) \tag{B.30}$$

for each $|m|\ge 1$. Here $\|\hat{C}_{NM}\|$ and $\|\hat{B}_{NM}\|$ have upper bounds independent of N and M (see similar arguments around (B.16)), denoted by K_B' and K_C'. Inequality (B.30) says that for each fixed N, there exists a large enough integer M_0 with $M_0\ge N+1$ such that $\|G_{m[N,M]}(s)\|<1$ for all $|m|\ge 1$, $M\ge M_0$, and $s\in\partial\Omega$. Thus, $|\lambda_k(G_{m[N,M]}(s))|<1$, from which Lemma 15.8 of [24] and (B.30) yield that

$$\sum_{m\ne 0}\sum_k |1-(1+\lambda_k(G_{m[N,M]}(s)))\exp\{-\lambda_k(G_{m[N,M]}(s))\}|$$
$$\le \sum_{m\ne 0}\sum_k [KK_B'K_C'f(|m|(M+1))]^2$$
$$= [KK_B'K_C']^2\frac{(2M+1)d_C}{(M+1)^2}\sum_{m\ne 0} f^2(m) < \frac{8d_C[KK_B'K_C']^2}{M+1}, \tag{B.31}$$

where d_C is the row dimension of the output matrix $C(t)$ of the FDLCP system (2.1).

On the other hand, it is straightforward to show by induction that

$$\left|\prod_k (1+a_k) - 1\right| \le \exp\left\{\sum_k |a_k|\right\} - 1, \tag{B.32}$$

where $a_k \in \mathbf{C}$. Since $\sum_{m\neq 0}\sum_k |1-(1+\lambda_k(G_{m[N,M]}(s)))\exp\{-\lambda_k(G_{m[N,M]}(s))\}| \to 0$ as $M\to\infty$ by (B.31), it follows from (B.32) that

$$\prod_{m\neq 0}\prod_k \big(1+\lambda_k\big(G_{m[N,M]}(s)\big)\big)\exp\big\{-\lambda_k\big(G_{m[N,M]}(s)\big)\big\} \to 1$$

uniformly with respect to s. This gives the first assertion.

To see the second assertion of (4.8), we observe

$$\begin{aligned}\|\tilde{\Delta}_{m[N,M]}(s+\rho)\| &\le \big\|E_{Mm}^{-1}(s+\rho)\big\|\,\|Q_M+\rho I_M\|\,\big\|(E_{Mm}(s)-Q_M)^{-1}\big\|\,\|\widehat{BC}_{NM}\| \\ &< KK_\varphi f^2(|m|(M+1))(\|Q\|+\rho)\|\widehat{BC}_{NM}\| \end{aligned}\tag{B.33}$$

for each $|m|\ge 1$. In the derivation of (B.33), we repeated some arguments similar to those in (B.30) and used the fact that $\|E_{Mm}^{-1}(s+\rho)\| \le K_\varphi f(|m|(2M+1)-M) < K_\varphi f(|m|(M+1))$. It is easy to see that under the given assumptions about the system matrices, $\|\widehat{BC}_{NM}\|$ has an upper bound independent of N and M, denoted by K_{BC}. Then it follows that

$$\begin{aligned}\left|\sum_{m\neq 0}\operatorname{tr}(\tilde{\Delta}_{m[N,M]}(s+\rho))\right| &\le \sum_{m\neq 0}(2M+1)n\|\tilde{\Delta}_{m[N,M]}(s+\rho)\| \\ &< \sum_{m\neq 0}(2M+1)nKK_\varphi K_{BC}f^2(|m|(M+1))(\|Q\|+\rho) \\ &< \frac{4(2M+1)nKK_\varphi K_{BC}(\|Q\|+\rho)}{(M+1)^2} \to 0 \quad (M\to\infty)\end{aligned}\tag{B.34}$$

which leads to the desired assertion. This completes the proof. □

Proof of (4.9). By the $\det_2$ definition and the block-diagonal structure of the infinite-dimensional matrix $\underline{G}_{[N,M]}(s)$, it is evident that

$$\begin{aligned}&\big|\det_2\big[\underline{I}+\underline{G}_{[N,M]}(s)\big]\exp\big\{-\tilde{\Delta}_{[N,M]}(s+\rho)\big\} \\ &\quad -\det_2\big[I_M+G_{0[N,M]}(s)\big]\exp\big\{-\operatorname{tr}\big(\tilde{\Delta}_{0[N,M]}(s+\rho)\big)\big\}\big| \\ &\le \big|\det_2\big[I_M+G_{0[N,M]}(s)\big]\exp\big\{-\operatorname{tr}\big(\tilde{\Delta}_{0[N,M]}(s+\rho)\big)\big\}\big| \\ &\quad\times\Bigg[\Bigg|\prod_{m\neq 0}\prod_k\big(1+\lambda_k\big(G_{0[N,M]}(s)\big)\big)\exp\big\{-\lambda_k\big(G_{0[N,M]}(s)\big)\big\}-1\Bigg| \\ &\quad\times\left|\exp\left\{-\sum_{m\neq 0}\operatorname{tr}\big(\tilde{\Delta}_{m[N,M]}(s+\rho)\big)\right\}\right| \\ &\quad+\left|\exp\left\{-\sum_{m\neq 0}\operatorname{tr}\big(\tilde{\Delta}_{m[N,M]}(s+\rho)\big)\right\}-1\right|\Bigg].\end{aligned}$$

This, together with Lemma 4.4, tells us that (4.9) follows if it is shown that $|\det_2[I_M+G_{0[N,M]}(s)]|$ and $|\operatorname{tr}(\tilde{\Delta}_{0[N,M]}(s+\rho))|$ are uniformly bounded over M and $s\in\partial\Omega$.

To see the uniform boundedness of $|\mathrm{tr}(\tilde{\Delta}_{0[N,M]}(s+\rho))|$, we observe by (4.5) that

$$\begin{aligned}
\big|\mathrm{tr}\big(\tilde{\Delta}_{0[N,M]}(s+\rho)\big)\big| &\le \big\|\tilde{\Delta}_{0[N,M]}(s+\rho)\big\|_1 \\
&\le \big\|E_{M0}^{-1}(s+\rho)(Q_M+\rho I_M)\big\|_2 \big\|(E_{M0}(s)-Q_M)^{-1}\widehat{\underline{BC}}_{[N,M]}\big\|_2 \\
&\le \big\|E_{M0}^{-1}(s+\rho)\big\|_2 \|Q_M+\rho I_M\| \cdot \big\|(E_{M0}(s)-Q_M)^{-1}\big\|_2 \big\|\widehat{\underline{BC}}_{[N,M]}\big\| \\
&\le (\|Q\|+\rho)\big\|\widehat{\underline{BC}}_{[N,M]}\big\| \sum_{|i|\le M} nK_\varphi^2 f^2(i) \sum_{|i|\le M} nK^2 f^2(i) \\
&< (5nKK_\varphi)^2(\|Q\|+\rho)\big\|\widehat{\underline{BC}}_{[N,M]}\big\|.
\end{aligned}$$

Hence, the desired uniform boundedness follows from the fact that $\|\widehat{\underline{BC}}_{[N,M]}\|$ has an upper bound independent of M.

To see the uniform boundedness of $|\det_2[I_M + G_{0[N,M]}(s)]|$, we need some preparations. By Theorem 7.4 of [11, p. 69], we have

$$\big|\det_2\big[I_M + G_{0[N,M]}(s)\big]\big| \le \exp\left\{\frac{1}{2}\big\|G_{0[N,M]}(s)\big\|_2\right\}$$

which implies that to see the uniform boundedness of $|\det_2[I_M+G_{0[N,M]}(s)]|$, it suffices to show that $\|G_{0[N,M]}(s)\|_2$ is uniformly bounded over M and $s \in \partial\Omega$. To see this, we note that $\hat{B}_{NM}$ and $\hat{C}_{NM}$ are submatrices of $\underline{\hat{B}}_{[N]}$ and $\underline{\hat{C}}_{[N]}$, respectively. Hence, we immediately have

$$\begin{aligned}
\big\|G_{0[N,M]}(s)\big\|_2 &\le \|\hat{C}_{NM}\| \cdot \big\|(E_{M0}(s)-Q_M)^{-1}\big\|_2 \cdot \|\hat{B}_{NM}\| \\
&\le K_C K_B \sum_{|i|\le M} nK^2 f^2(i) < 5nK^2 K_C K_B.
\end{aligned}$$

This completes the proof. □

REFERENCES

[1] A. Allievi and A. Soudack, *Ship stability via the Mathieu equation,* Internat. J. Control, 51 (1990), pp. 139–167.

[2] B. Bamieh and J. B. Pearson, *A general framework for linear periodic systems with applications to H^∞ sampled-data control,* IEEE Trans. Automat. Control, 37 (1992), pp. 418–435.

[3] S. Bittanti and P. Colaneri, *Stabilization of periodic systems: Overview and advances,* in Proceedings of IFAC Workshop on Periodic Control Systems, Italy, 2001, pp. 157–162.

[4] P. Bolzern and P. Colaneri, *The periodic Lyapunov equation,* SIAM J. Matrix Anal. Appl., 9 (1988), pp. 499–512.

[5] A. Böttcher and B. Silbermann, *Analysis of Toeplitz Operators*, Springer-Verlag, Berlin, 1990.

[6] M. Cantoni and K. Glover, *Existence of right and left representations of the graph for linear periodically time-varying systems,* SIAM J. Control Optim., 38 (2000), pp. 786–802.

[7] J. Dugundji and J. H. Wendell, *Some analysis methods for rotating systems with periodic coefficients,* AIAA J., 21 (1983), pp. 890–897.

[8] M. Farkas, *Periodic Motions,* Springer-Verlag, New York, 1994.

[9] I. Gohberg, S. Goldberg, and M. A. Kaashoek, *Classes of Linear Operators,* Vol. I, Birkhäuser, Basel, 1990.

[10] I. Gohberg, S. Goldberg, and M. A. Kaashoek, *Classes of Linear Operators,* Vol. II, Birkhäuser, Basel, 1993.

[11] I. Gohberg, S. Goldberg, and N. Krupnik, *Traces and Determinants of Linear Operators,* Birkhäuser, Basel, 2000.

[12] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields,* Springer-Verlag, New York, 1983.
[13] T. Hagiwara, *Nyquist stability criterion and positive-realness of sampled-data systems,* Systems Control Lett., 45 (2002), pp. 283–291.
[14] A. Halanay, *Differential Equations: Stability, Oscillations, Time Lags,* Academic Press, New York, 1966.
[15] S. R. Hall and N. M. Wereley, *Generalized Nyquist criterion for linear time systems,* in Proceedings of American Control Conference, San Diego, 1990, pp. 1518–1525.
[16] C. Y. Kao, A. Megretski, and U. T. Jönsson, *A cutting plane algorithm for robustness analysis of periodically time-varying systems,* IEEE Trans. Automat. Control, 46 (2001), pp. 579–592.
[17] G. A. Korn and T. M. Korn, *Mathematical Handbook for Scientist and Engineers,* 2nd ed., McGraw-Hill, New York, 1968.
[18] D. L. Lukes, *Differential Equations: Classical to Controlled,* Academic Press, New York, 1982.
[19] P. Montagnier, R. J. Spiteri, and J. Angeles, *The control of linear time-periodic systems using Floquet–Lyapunov theory,* Internat. J. Control, 77 (2004), pp. 472–490.
[20] A. H. Nayfeh and D. T. Mook, *Nonlinear Oscillations,* John Wiley and Sons, New York, 1979.
[21] A. W. Naylor and G. R. Sell, *Linear Operator Theory in Engineering and Science,* Springer-Verlag, New York, 1982.
[22] M. Pavella and P. G. Murthy, *Transient Stability of Power System—Theory and Practice,* John Wiley and Sons, New York, 1994.
[23] J. A. Richards, *Analysis of Periodically Time-Varying System,* Springer-Verlag, New York, 1983.
[24] W. Rudin, *Real and Complex Analysis,* 3rd ed., McGraw-Hill, New York, 1987.
[25] J. J. Stoker, *Nonlinear Vibrations in Mechanical and Electrical Systems,* Interscience Publishers, New York, 1950.
[26] N. M. Wereley, *Analysis and Control of Linear Periodically Time Varying Systems,* Ph.D. thesis, Department of Aeronautics and Astronautics, M.I.T., 1990.
[27] V. A. Yakubovich and V. M. Starzhinskii, *Linear Differential Equations with Periodic Coefficients,* Vol. I, John Wiley and Sons, New York, 1975.
[28] V. A. Yakubovich, *Dichotomy and absolute stability of nonlinear systems with periodically non-stationary linear part,* Systems Control Lett., 11 (1988), pp. 221–228.
[29] J. Zhou and T. Hagiwara, *Existence conditions and properties of frequency response operators of continuous-time periodic systems,* SIAM J. Control Optim., 40 (2002), pp. 1867–1887.
[30] J. Zhou and T. Hagiwara, *H_2 and H_∞ norm computations of linear continuous-time periodic systems via the skew analysis of frequency response operators,* Automatica, 38 (2002), pp. 1381–1387.
[31] J. Zhou, T. Hagiwara, and M. Araki, *Stability analysis of continuous-time periodic systems via the harmonic analysis,* IEEE Trans. Automat. Control, 47 (2002), pp. 292–298.
[32] J. Zhou, *Harmonic Analysis of Linear Continuous-Time Periodic Systems,* Ph.D. thesis, Department of Electrical Engineering, Kyoto University, Kyoto, Japan, 2002.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 2, pp. 646–672

GENERATING SERIES FOR INTERCONNECTED ANALYTIC NONLINEAR SYSTEMS*

W. STEVEN GRAY† AND YAQIN LI†

Abstract. Given two analytic nonlinear input-output systems represented as Fliess operators, four system interconnections are considered in a unified setting: the parallel connection, product connection, cascade connection, and feedback connection. In each case, the corresponding generating series is produced and conditions for the convergence of the corresponding Fliess operator are given. In the process, an existing notion of a *composition product* for formal power series has its set of known properties significantly expanded. In addition, the notion of a *feedback product* for formal power series is shown to be well defined in a broad context, and its basic properties are characterized.

Key words. Chen–Fliess series, formal power series, nonlinear operators, nonlinear systems

AMS subject classifications. 47H30, 93C10

DOI. 10.1137/S036301290343007X

1. Introduction. Let $X = \{x_0, x_1, \ldots, x_m\}$ denote an alphabet and X^* the set of all words over X (including the empty word $\emptyset$). A formal power series in X is any mapping of the form $X^* \to \mathbb{R}^\ell$, and the set of all such mappings will be denoted by $\mathbb{R}^\ell\langle\langle X\rangle\rangle$. For each $c \in \mathbb{R}^\ell\langle\langle X\rangle\rangle$, one can formally associate an m-input, ℓ-output operator F_c in the following manner. Let $p \geq 1$ and $a < b$ be given. For a measurable function $u : [a, b] \to \mathbb{R}^m$, define $\|u\|_p = \max\{\|u_i\|_p : \ 1 \leq i \leq m\}$, where $\|u_i\|_p$ is the usual L_p-norm for a measurable real-valued function, u_i, defined on $[a, b]$. Let $L_p^m[a, b]$ denote the set of all measurable functions defined on $[a, b]$ having a finite $\|\cdot\|_p$-norm and $B_p^m(R)[a, b] := \{u \in L_p^m[a, b] : \|u\|_p \leq R\}$. With $t_0, T \in \mathbb{R}$ fixed and $T > 0$, define recursively for each $\eta \in X^*$ the mapping $E_\eta : L_1^m[t_0, t_0 + T] \to \mathcal{C}[t_0, t_0 + T]$ by $E_\emptyset = 1$, and

$$E_{x_i\bar{\eta}}[u](t, t_0) = \int_{t_0}^{t} u_i(\tau) E_{\bar{\eta}}[u](\tau, t_0)\, d\tau,$$

where $x_i \in X$, $\bar{\eta} \in X^*$, and $u_0(t) \equiv 1$. The input-output operator corresponding to c is then

$$F_c[u](t) = \sum_{\eta \in X^*} (c, \eta)\, E_\eta[u](t, t_0),$$

which is referred to as a *Fliess operator*. All Volterra operators with analytic kernels, for example, are Fliess operators. In the classical literature, where these operators first appeared [7, 9, 10, 26], it is normally assumed that there exist real numbers $K, M > 0$ such that $|(c, \eta)| \leq KM^{|\eta|}|\eta|!$ for all $\eta \in X^*$, where $|z| = \max\{|z_1|, |z_2|, \ldots, |z_\ell|\}$ when $z \in \mathbb{R}^\ell$, and $|\eta|$ denotes the number of letters in η. This growth condition on the coefficients of c ensures that there exist positive real numbers R and T_0 such that, for all piecewise continuous u with $\|u\|_\infty \leq R$ and $T \leq T_0$, the series defining

*Received by the editors June 18, 2003; accepted for publication (in revised form) December 8, 2004; published electronically September 12, 2005.
http://www.siam.org/journals/sicon/44-2/43007.html
†Department of Electrical and Computer Engineering, Old Dominion University, Norfolk, VA 23529-0246 (gray@ece.odu.edu, yli@odu.edu).

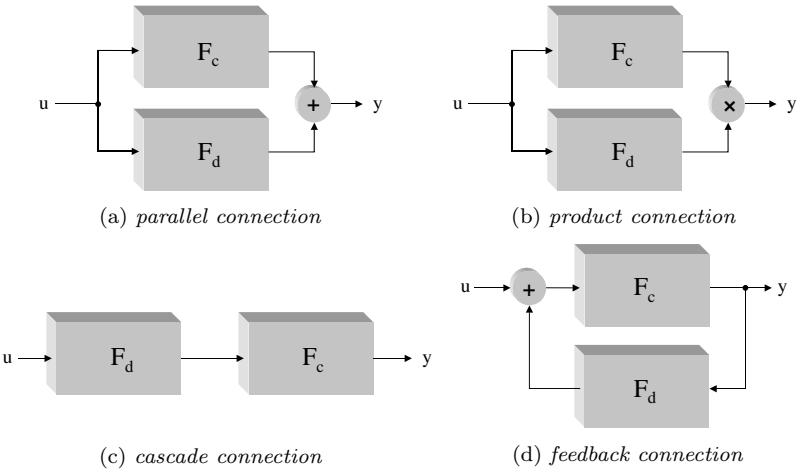


FIG. 1.1. *Elementary system interconnections.*

F_c converges uniformly and absolutely on $[t_0, t_0 + T]$. Therefore, a power series c is said to be *locally convergent* when its coefficients satisfy such a growth condition. The set of all locally convergent series in $\mathbb{R}^\ell\langle\langle X\rangle\rangle$ will be denoted by $\mathbb{R}^\ell_{LC}\langle\langle X\rangle\rangle$. More recently, it was shown in [13] that local convergence also implies that F_c constitutes a well-defined operator from $B^m_p(R)[t_0, t_0 + T]$ into $B^\ell_q(S)[t_0, t_0 + T]$ for sufficiently small $R, S, T > 0$, where the numbers $p, q \in [1, \infty]$ are conjugate exponents, i.e., $1/p + 1/q = 1$ with $(1, \infty)$ being a conjugate pair by convention.

In many applications, input-output systems are interconnected in a variety of ways. Given two Fliess operators F_c and F_d, where $c, d \in \mathbb{R}^\ell_{LC}\langle\langle X\rangle\rangle$, Figure 1.1 shows four elementary interconnections. The product connection is defined componentwise, and in the case of the feedback connection it is assumed that $\ell = m > 0$. The general goal of this paper is to describe in a unified manner the generating series for each elementary interconnection and conditions under which it is locally convergent. The clear antecedent to this work is that of Ferfera, who first described the generating series for such connections (implicitly in the case of feedback) and, in particular, introduced the composition product $c \circ d$ of two formal power series c and d [5, 6]. In each case, however, the local convergence of the new generating series or, equivalently, the convergence of the corresponding Fliess operator, was not explicitly addressed. It is trivial to show that the parallel connection of F_c and F_d always produces a locally convergent generating series when c and d are locally convergent. The same conclusion was later provided in [28] for the product connection via an analysis involving the shuffle product. In this paper, an analogous result is developed for the composition product by producing an explicit expression for one pair of growth constants, $K_{c\circ d}$ and $M_{c\circ d}$. In the process, the set of known properties of the composition product is significantly expanded. (An interesting parallel development has appeared in [3, 11] regarding a composition product for formal power series motivated by the composition of two analytic functions (see, e.g., [18]) rather than two Fliess (integral) operators. Its definition is quite distinct and not clearly related to the composition product described in this paper.)

The feedback connection is a fundamentally more difficult case to analyze. For example, when F_c is a linear operator, the formal solution to the feedback equation

$$y = F_c[u + F_d[y]] \tag{1}$$

is

$$y = F_c[u] + F_c \circ F_d \circ F_c[u] + \cdots.$$

It is not immediately clear that this series converges in any manner and, in particular, converges to another Fliess operator, say, $F_{c@d}$, for some $c@d \in \mathbb{R}^m_{LC}\langle\langle X \rangle\rangle$. When F_c is nonlinear, the problem is further complicated by the fact that operators of the form $I + F_d$, where I denotes the identity map, *never* have a Fliess operator representation. In this paper, the problem is circumvented by introducing a simple variation of the composition product so that an appropriate *feedback product*, $c@d$, is well defined, and $y = F_{c@d}[u]$ satisfies the feedback equation (1) in the sense that every analytic input u produces an analytic output y with (u, y) satisfying (1). In this case, $c@d$ is referred to as being *input-output locally convergent*, and explicit expressions are derived for one set of growth constants, K_{c_y} and M_{c_y}, for the series representation of the output function, c_y.

It should be stated that Ferfera's primary interest in [5, 6] was rational series and their corresponding bilinear realizations. In a state space setting, the issue of local convergence is rather straightforward. If c and d each have finite Lie rank, in addition to being locally convergent, then the mappings F_c and F_d each have a finite-dimensional analytic state space realization, and therefore so does each interconnected system (see [16, 21] for a basic treatment of nonlinear realization theory). The literature then provides that the corresponding generating series can be computed by successive Lie derivatives and, in particular, it must be locally convergent [26, Lemma 4.2]. (Additional analysis of interconnected state space systems using a chronological product together with Hall–Viennot bases appears in [17].) While the state space formalism is clearly dominant in modern control theory, other system descriptions like Volterra series [10, 16, 21] or input-output differential equations [28, 29, 30] are sometimes useful. In such settings, the convergence analysis of interconnected systems is a natural application for the main results of this paper. But even in a pure state space setting, as illustrated by Examples 3.2 and 4.11, knowledge of the growth constants for the generating series of a given interconnection permits one to compute a lower bound on any finite escape time. This is particularly useful in physical problems, like the one described in [12], as it provides computable limitations on the applicability of the underlying mathematical models.

The paper is organized as follows. In section 2 the composition product is introduced and developed independently of the system interconnection problem. First, its various fundamental properties are presented. Then, in preparation for the feedback analysis, it is shown that the composition product produces a contractive mapping on the set of all formal power series using a familiar ultrametric. In section 3, the three *nonrecursive* connections, parallel, product, and cascade, are analyzed primarily by applying results from section 2. In section 4 the feedback connection is considered. The main focus is on showing when the feedback product of two formal power series is well defined and in precisely what sense it is locally convergent.

2. The composition product. The composition product of two formal power series over an alphabet $X = \{x_0, x_1, \ldots, x_m\}$ is defined recursively in terms of the

shuffle product. The shuffle product of two words $\eta, \xi \in X^*$ is defined recursively by

$$\eta \sqcup\!\sqcup \xi = (x_j\eta') \sqcup\!\sqcup (x_k\xi') := x_j[\eta' \sqcup\!\sqcup \xi] + x_k[\eta \sqcup\!\sqcup \xi']$$

with $\emptyset \sqcup\!\sqcup \emptyset = \emptyset$ and $\xi \sqcup\!\sqcup \emptyset = \emptyset \sqcup\!\sqcup \xi = \xi$. It is easily verified that $\eta \sqcup\!\sqcup \xi$ is always a polynomial consisting of words each having length $|\eta|+|\xi|$. The definition is extended to any two series $c, d \in \mathbb{R}\langle\langle X\rangle\rangle$ by

$$c \sqcup\!\sqcup d = \sum_{\eta,\xi\in X^*} [(c,\eta)(d,\xi)]\, \eta \sqcup\!\sqcup \xi. \tag{2}$$

For a fixed $\nu \in X^*$, the coefficient $(\eta \sqcup\!\sqcup \xi, \nu) = 0$ if $|\eta| + |\xi| \neq |\nu|$. Hence, the infinite sum in (2) is well defined since the family of polynomials $\{\eta \sqcup\!\sqcup \xi : \eta, \xi \in X^*\}$ is locally finite [2]. In general, the shuffle product is commutative. It is also associative and distributes over addition. Thus, the vector space $\mathbb{R}\langle\langle X\rangle\rangle$ with the shuffle product forms a commutative $\mathbb{R}$-algebra, the so-called *shuffle algebra*, with multiplicative identity element $\emptyset$. The shuffle product on $\mathbb{R}^\ell\langle\langle X\rangle\rangle$ is defined componentwise, i.e., $(c \sqcup\!\sqcup d, \nu)_i = (c_i \sqcup\!\sqcup d_i, \nu)$ for $i = 1, 2, \ldots, \ell$.

For any $\eta \in X^*$ and $d \in \mathbb{R}^m\langle\langle X\rangle\rangle$, the *composition product* is defined recursively as

$$\eta \circ d = \begin{cases} \eta & : \ |\eta|_{x_i} = 0 \ \ \forall i \neq 0, \\ x_0^{n+1}[d_i \sqcup\!\sqcup (\eta' \circ d)] & : \ \eta = x_0^n x_i \eta', \ \ n \geq 0, \ \ i \neq 0, \end{cases}$$

where $|\eta|_{x_i}$ denotes the number of letters in η equivalent to x_i and $d_i : \xi \mapsto (d, \xi)_i$, the ith component of the coefficient (d, ξ). Consequently, if

$$\eta = x_0^{n_k} x_{i_k} x_0^{n_{k-1}} x_{i_{k-1}} \cdots x_0^{n_1} x_{i_1} x_0^{n_0}, \tag{3}$$

where $i_j \neq 0$ for $j = 1, \ldots, k$, then it follows that

$$\eta \circ d = x_0^{n_k+1}[d_{i_k} \sqcup\!\sqcup x_0^{n_{k-1}+1}[d_{i_{k-1}} \sqcup\!\sqcup \cdots x_0^{n_1+1}[d_{i_1} \sqcup\!\sqcup x_0^{n_0}] \cdots]].$$

Alternatively, for any $\eta \in X^*$, one can uniquely associate a set of right factors $\{\eta_0, \eta_1, \ldots, \eta_k\}$ by the iteration

$$\eta_{j+1} = x_0^{n_{j+1}} x_{i_{j+1}} \eta_j, \ \ \eta_0 = x_0^{n_0}, \ \ i_{j+1} \neq 0, \tag{4}$$

so that $\eta = \eta_k$ with $k = |\eta| - |\eta|_{x_0}$. In which case, $\eta \circ d = \eta_k \circ d$, where

$$\eta_{j+1} \circ d = x_0^{n_{j+1}+1}[d_{i_{j+1}} \sqcup\!\sqcup (\eta_j \circ d)]$$

and $\eta_0 \circ d = x_0^{n_0}$. The theorem below ensures that the composition product of two series described subsequently is well defined.

THEOREM 2.1. *Given a fixed $d \in \mathbb{R}^m\langle\langle X\rangle\rangle$, the family of series $\{\eta \circ d : \eta \in X^*\}$ is locally finite, and therefore summable.*

Proof. Given an arbitrary $\eta \in X^*$ expressed in the form (3), it follows directly that

$$\operatorname{ord}(\eta \circ d) = n_0 + k + \sum_{j=1}^{k} n_j + \operatorname{ord}(d_{i_j}) = |\eta| + \sum_{j=1}^{|\eta|-|\eta|_{x_0}} \operatorname{ord}(d_{i_j}), \tag{5}$$

where the *order* of c is defined as

$$\operatorname{ord}(c) = \begin{cases} \inf\{|\eta| : \eta \in \operatorname{supp}(c)\} & : \quad c \neq 0, \\ \infty & : \quad c = 0, \end{cases}$$

and $\operatorname{supp}(c) := \{\eta \in X^* : (c, \eta) \neq 0\}$ denotes the *support* of c. Hence, for any $\xi \in X^*$,

$$\begin{aligned} I_d(\xi) &:= \{\eta \in X^* : (\eta \circ d, \xi) \neq 0\} \\ &\subset \{\eta \in X^* : \operatorname{ord}(\eta \circ d) \leq |\xi|\} \\ &= \left\{ \eta \in X^* : |\eta| + \sum_{j=1}^{|\eta|-|\eta|_{x_0}} \operatorname{ord}(d_{i_j}) \leq |\xi| \right\}. \end{aligned}$$

Clearly this last set is finite, and thus $I_d(\xi)$ is finite for all $\xi \in X^*$. This fact implies summability. □

For any $c \in \mathbb{R}^\ell\langle\langle X \rangle\rangle$ and $d \in \mathbb{R}^m\langle\langle X \rangle\rangle$, the composition product is defined as

$$c \circ d = \sum_{\eta \in X^*} (c, \eta)\, \eta \circ d.$$

The summation can also be written using the set of all right factors as described by (4). Let X^i be the set of all words in X^* of length i. For each word $\eta \in X^i$, the jth right factor, η_j, has exactly j letters not equal to x_0. Therefore, given any $\nu \in X^*$,

$$(c \circ d, \nu) = \sum_{i=0}^{|\nu|} \sum_{j=0}^{i} \sum_{\eta_j \in X^i} (c, \eta_j)(\eta_j \circ d, \nu). \tag{6}$$

The third summation is understood to be the sum over the set of all possible jth right factors of words of length i. This set has a familiar combinatoric interpretation. A *composition* of a positive integer N is an ordered set of positive integers $\{a_1, a_2, \ldots, a_K\}$ such that $N = a_1 + a_2 + \cdots + a_K$. (For example, the integer 3 has the compositions $1+1+1$, $1+2$, $2+1$, and 3). For a given N and K, it is well known that there are $\mathcal{C}_K(N) = \binom{N-1}{K-1}$ possible compositions. Now each factor $\eta_j \in X^i$, when written in the form

$$\eta_j = x_0^{n_j} x_{i_j} x_0^{n_{j-1}} x_{i_{j-1}} \cdots\ x_0^{n_1} x_{i_1} x_0^{n_0},$$

maps to a unique composition of $i+1$ with $j+1$ elements:

$$i + 1 = (n_0 + 1) + (n_1 + 1) + \cdots + (n_j + 1).$$

Thus, there are exactly $\mathcal{C}_{j+1}(i+1)m^j = \binom{i}{j} m^j$ possible factors η_j in X^i, and the total number of terms in the summations of (6) is $((m+1)^{|\nu|+1} - 1)/m \approx (m+1)^{|\nu|}$. As will be seen shortly, this provides a conservative lower bound on the growth rate of the coefficients of $c \circ d$.

It is easily verified that the composition product is linear in its first argument, but not its second. A special exception are *linear series*. A series $c \in \mathbb{R}^\ell\langle\langle X \rangle\rangle$ is called linear if

$$\operatorname{supp}(c) \subseteq \{\eta \in X^* : \eta = x_0^{n_1} x_i x_0^{n_0},\ \ i \in \{1, 2, \ldots, m\},\ \ n_1, n_0 \geq 0\}.$$

It was shown in [5] that the composition product is associative and distributive from the right over the shuffle product. But in general it is neither commutative nor has an identity element. This lack of an identity element is precisely the reason the identity map I is not realizable as a Fliess operator. Other elementary properties concerning the composition product are summarized below.

LEMMA 2.2. *The following identities hold* ($\mathbb{1}$ *is a column vector with* m *ones*):

1. $0 \circ d = 0$ for all $d \in \mathbb{R}^m\langle\langle X\rangle\rangle$.
2. $c \circ 0 = c_0 := \sum_{n\geq 0}(c, x_0^n)\, x_0^n$. (*Therefore,* $c \circ 0 = 0$ *if and only if* $c_0 = 0$.)
3. $c_0 \circ d = c_0$ for all $d \in \mathbb{R}^m\langle\langle X\rangle\rangle$. (*In particular,* $1 \circ d = 1$.)
4. $c \circ \mathbb{1} = c_{\mathbb{1}} := \sum_{\eta\in X^*}(c, \eta)\, x_0^{|\eta|}$. (*Therefore,* $c \circ \mathbb{1} = c$ *if and only if* $c_0 = c$.)

The set $\mathbb{R}^m\langle\langle X\rangle\rangle$ forms a metric space under the ultrametric

$$\begin{aligned} \text{dist} \quad &: \quad \mathbb{R}^m\langle\langle X\rangle\rangle \times \mathbb{R}^m\langle\langle X\rangle\rangle \to \mathbb{R}^+ \cup \{0\}, \\ &: \quad (c, d) \mapsto \sigma^{\text{ord}(c-d)}, \end{aligned}$$

where $\sigma \in (0, 1)$ is arbitrary [2]. The following theorem states that the composition product on $\mathbb{R}^m\langle\langle X\rangle\rangle \times \mathbb{R}^m\langle\langle X\rangle\rangle$ is continuous in its left argument. (Right argument continuity will be addressed later.)

THEOREM 2.3. *Let* $\{c_i\}_{i\geq 1}$ *be a sequence in* $\mathbb{R}^m\langle\langle X\rangle\rangle$ *with* $\lim_{i\to\infty} c_i = c$. *Then* $\lim_{i\to\infty}(c_i \circ d) = c \circ d$ *for any* $d \in \mathbb{R}^m\langle\langle X\rangle\rangle$.

Proof. Define the sequence of nonnegative integers $k_i = \text{ord}(c_i - c)$ for $i \geq 1$. Since c is the limit of the sequence $\{c_i\}_{i\geq 1}$, the sequence $\{k_i\}_{i\geq 1}$ must have an increasing subsequence $\{k_{i_j}\}$. Now observe that

$$\text{dist}(c_i \circ d, c \circ d) = \sigma^{\text{ord}((c_i - c)\,\circ d)}$$

and

$$\begin{aligned} \text{ord}((c_{i_j} - c) \circ d) &= \text{ord}\left(\sum_{\eta\in\text{supp}(c_{i_j}-c)} (c_{i_j} - c, \eta)\ \eta \circ d\right) \\ &\geq \inf_{\eta\in\text{supp}(c_{i_j}-c)} \text{ord}(\eta \circ d) \\ &\geq \inf_{\eta\in\text{supp}(c_{i_j}-c)} |\eta| + (|\eta| - |\eta|_{x_0})\, \text{ord}(d) \\ &\geq k_{i_j}. \end{aligned}$$

Thus, $\text{dist}(c_{i_j} \circ d, c \circ d) \leq \sigma^{k_{i_j}}$ for all $j \geq 1$, and $\lim_{i\to\infty} c_i \circ d = c \circ d$. ∎

The ultrametric space $(\mathbb{R}^m\langle\langle X\rangle\rangle, \text{dist})$ is known to be complete [2]. Given a fixed $c \in \mathbb{R}^m\langle\langle X\rangle\rangle$, consider the mapping $\mathbb{R}^m\langle\langle X\rangle\rangle \to \mathbb{R}^m\langle\langle X\rangle\rangle : d \mapsto c \circ d$. The goal is to show that this mapping is always a contraction on $\mathbb{R}^m\langle\langle X\rangle\rangle$, i.e., that

$$\text{dist}(c \circ d, c \circ e) \leq \sigma\, \text{dist}(d, e) \quad \forall d, e \in \mathbb{R}^m\langle\langle X\rangle\rangle,$$

so that fixed point theorems can be applied in later analysis [14, 22, 23, 24]. Any $c \in \mathbb{R}^m\langle\langle X\rangle\rangle$ can be written unambiguously in the form

$$c = c_0 + c_1 + \cdots, \tag{7}$$

where $c_k \in \mathbb{R}^m\langle\langle X\rangle\rangle$ has the defining property that $\eta \in \text{supp}(c_k)$ only if $|\eta| - |\eta|_{x_0} = k$. Some of the series c_k may be the zero series. When $c_0 = 0$, c is referred to as being *homogeneous*. When $c_k = 0$ for $k = 0, 1, \ldots, l-1$ and $c_l \neq 0$, then c is called *homogeneous of order* l. In this setting consider the following lemma.

LEMMA 2.4. *For any* c_k *in* (7),

$$\text{dist}(c_k \circ d, c_k \circ e) \leq \sigma^k\, \text{dist}(d, e) \quad \forall d, e \in \mathbb{R}^m\langle\langle X\rangle\rangle.$$

Proof. The proof is by induction for the nontrivial case, where $c_k \neq 0$. First suppose $k = 0$. From the definition of the composition product it follows directly that $\eta \circ d = \eta$ for all $\eta \in \operatorname{supp}(c_0)$. Therefore,

$$c_0 \circ d = \sum_{\eta \in \operatorname{supp}(c_0)} (c_0, \eta)\, \eta \circ d = \sum_{\eta \in \operatorname{supp}(c_0)} (c_0, \eta)\, \eta = c_0,$$

and

$$\operatorname{dist}(c_0 \circ d, c_0 \circ e) = \operatorname{dist}(c_0, c_0) = 0 \leq \sigma^0 \operatorname{dist}(d, e).$$

Now fix any $k \geq 0$ and assume the claim is true for all $c_0, c_1, \ldots, c_k$. In particular, this implies that

$$\operatorname{ord}(c_k \circ d - c_k \circ e) \geq k + \operatorname{ord}(d - e). \tag{8}$$

For any $j \geq 0$, words in $\operatorname{supp}(c_j)$ have the form η_j as defined in (4). Observe then that

$$\begin{aligned}
c_{k+1} \circ d - c_{k+1} \circ e &= \sum_{\eta_{k+1} \in X^*} (c_{k+1}, \eta_{k+1})\, \eta_{k+1} \circ d - (c_{k+1}, \eta_{k+1})\, \eta_{k+1} \circ e \\
&= \sum_{\eta_k, \eta_{k+1} \in X^*} (c_{k+1}, \eta_{k+1}) \; \big[{x_0}^{n_{k+1}+1} [d_{i_{k+1}} \sqcup [\eta_k \circ d]] \\
&\qquad\qquad - {x_0}^{n_{k+1}+1} [e_{i_{k+1}} \sqcup [\eta_k \circ e]] \big] \\
&= \sum_{\eta_k, \eta_{k+1} \in X^*} (c_{k+1}, \eta_{k+1}) \; \big[{x_0}^{n_{k+1}+1} [d_{i_{k+1}} \sqcup [\eta_k \circ d]] \\
&\qquad\qquad - {x_0}^{n_{k+1}+1} [d_{i_{k+1}} \sqcup [\eta_k \circ e]] \\
&\qquad + {x_0}^{n_{k+1}+1} [d_{i_{k+1}} \sqcup [\eta_k \circ e]] - {x_0}^{n_{k+1}+1} [e_{i_{k+1}} \sqcup [\eta_k \circ e]] \big] \\
&= \sum_{\eta_k, \eta_{k+1} \in X^*} (c_{k+1}, \eta_{k+1}) \; \big[{x_0}^{n_{k+1}+1} [d_{i_{k+1}} \sqcup [\eta_k \circ d - \eta_k \circ e]] \\
&\qquad\qquad + {x_0}^{n_{k+1}+1} [(d_{i_{k+1}} - e_{i_{k+1}}) \sqcup [\eta_k \circ e]] \big] ,
\end{aligned}$$

using the fact that the shuffle product distributes over addition. Next, applying the identity (5) and the inequality (8) with $c_k = \eta_k$, it follows that

$$\begin{aligned}
\operatorname{ord}(c_{k+1} \circ d - c_{k+1} \circ e) &\geq \min \Big\{ \inf_{\eta_{k+1} \in \operatorname{supp}(c_{k+1})} n_{k+1} + 1 + \operatorname{ord}(d) + k + \operatorname{ord}(d - e), \\
&\qquad\qquad \inf_{\eta_{k+1} \in \operatorname{supp}(c_{k+1})} n_{k+1} + 1 + \operatorname{ord}(d - e) + |\eta_k| + k \operatorname{ord}(e) \Big\} \\
&\geq k + 1 + \operatorname{ord}(d - e),
\end{aligned}$$

and thus,

$$\operatorname{dist}(c_{k+1} \circ d, c_{k+1} \circ e) \leq \sigma^{k+1} \operatorname{dist}(d, e).$$

Hence, $\operatorname{dist}(c_k \circ d, c_k \circ e) \leq \sigma^k \operatorname{dist}(d, e)$ holds for any $k \geq 0$. ∎

Applying the above lemma leads to the following result.

LEMMA 2.5. *If $c \in \mathbb{R}^m\langle\langle X\rangle\rangle$, then for any series $c_0' \in \mathbb{R}^m\langle\langle X_0\rangle\rangle$,*

$$\operatorname{dist}((c_0' + c) \circ d, (c_0' + c) \circ e) = \operatorname{dist}(c \circ d, c \circ e) \ \ \forall d, e \in \mathbb{R}^m\langle\langle X\rangle\rangle. \tag{9}$$

(*Here X_0 denotes the single letter alphabet $\{x_0\}$.*) *If c is homogeneous of order $l \geq 1$ then*

$$\operatorname{dist}(c \circ d, c \circ e) \leq \sigma^l \operatorname{dist}(d, e) \ \ \forall d, e \in \mathbb{R}^m\langle\langle X\rangle\rangle. \tag{10}$$

Proof. The equality is proved first. Since the ultrametric dist is shift-invariant, observe that

$$\begin{aligned}\operatorname{dist}((c_0' + c) \circ d, (c_0' + c) \circ e) &= \operatorname{dist}(c_0' \circ d + c \circ d, c_0' \circ e + c \circ e) \\ &= \operatorname{dist}(c_0' + c \circ d, c_0' + c \circ e) \\ &= \operatorname{dist}(c \circ d, c \circ e).\end{aligned}$$

The inequality is proved next by first selecting any fixed $l \geq 1$ and showing inductively that it holds for any partial sum $\sum_{i=l}^{l+k} c_i$, where $k \geq 0$. When $k = 0$, Lemma 2.4 implies that

$$\operatorname{dist}(c_l \circ d, c_l \circ e) \leq \sigma^l \operatorname{dist}(d, e).$$

If the result is true for partial sums up to any fixed $k \geq 0$, then using the ultrametric property

$$\operatorname{dist}(d, e) \leq \max\{\operatorname{dist}(d, f), \operatorname{dist}(f, e)\} \ \ \forall d, e, f \in \mathbb{R}^m\langle\langle X\rangle\rangle,$$

it follows that

$$\begin{aligned}&\operatorname{dist}\left(\left(\sum_{i=l}^{l+k+1} c_i\right) \circ d, \left(\sum_{i=l}^{l+k+1} c_i\right) \circ e\right) \\ &= \operatorname{dist}\left(\left(\sum_{i=l}^{l+k} c_i\right) \circ d + c_{l+k+1} \circ d, \left(\sum_{i=l}^{l+k} c_i\right) \circ e + c_{l+k+1} \circ e\right) \\ &\leq \max\left\{\operatorname{dist}\left(\left(\sum_{i=l}^{l+k} c_i\right) \circ d + c_{l+k+1} \circ d, \left(\sum_{i=l}^{l+k} c_i\right) \circ d + c_{l+k+1} \circ e\right),\right. \\ &\qquad \left.\operatorname{dist}\left(\left(\sum_{i=l}^{l+k} c_i\right) \circ d + c_{l+k+1} \circ e, \left(\sum_{i=l}^{l+k} c_i\right) \circ e + c_{l+k+1} \circ e\right)\right\} \\ &= \max\left\{\operatorname{dist}(c_{l+k+1} \circ d, c_{l+k+1} \circ e), \operatorname{dist}\left(\left(\sum_{i=l}^{l+k} c_i\right) \circ d, \left(\sum_{i=l}^{l+k} c_i\right) \circ e\right)\right\} \\ &\leq \max\left\{\sigma^{l+k+1} \operatorname{dist}(d, e), \sigma^l \operatorname{dist}(d, e)\right\} \\ &\leq \sigma^l \operatorname{dist}(d, e).\end{aligned}$$

Hence, the result holds for all $k \geq 0$. Inequality (10) is proved by noting that $c = \lim_{k\to\infty} \sum_{i=l}^{l+k} c_i$ and using the left argument continuity of the composition product, proved in Theorem 2.3, and the continuity of the ultrametric. ∎

The main result regarding contractive mappings is below.

THEOREM 2.6. *For any $c \in \mathbb{R}^m\langle\langle X\rangle\rangle$, the mapping $d \mapsto c \circ d$ is a contraction on $\mathbb{R}^m\langle\langle X\rangle\rangle$.*

Proof. Choose any series $d, e \in \mathbb{R}^m\langle\langle X\rangle\rangle$. If c is homogeneous of order $l \geq 1$, then the result follows directly from (10). Otherwise, observe that, via (9),

$$\operatorname{dist}(c \circ d, c \circ e) = \operatorname{dist}\left(\left(\sum_{l=1}^{\infty} c_i\right) \circ d, \left(\sum_{l=1}^{\infty} c_i\right) \circ e\right) \leq \sigma \operatorname{dist}(d, e). \qquad \square$$

An immediate result of this theorem is the right argument continuity of the composition product.

THEOREM 2.7. *Let $\{d_i\}_{i\geq 1}$ be a sequence in $\mathbb{R}^m\langle\langle X\rangle\rangle$ with $\lim_{i\to\infty} d_i = d$. Then $\lim_{i\to\infty}(c \circ d_i) = c \circ d$ for all $c \in \mathbb{R}^m\langle\langle X\rangle\rangle$.*

Proof. Trivially,

$$\lim_{i\to\infty} \operatorname{dist}(c \circ d_i, c \circ d) \leq \sigma \lim_{i\to\infty} \operatorname{dist}(d_i, d) = 0. \qquad \square$$

The final property considered in this section is local convergence. If all the summands in the defining expression (6) are unity, i.e., c and d have no coefficient growth whatsoever, then earlier combinatoric analysis shows that $(c \circ d, \nu)$ grows at least at the rate $(m+1)^{|\nu|}$. Of course, in general, much faster growth rates are possible when c and d are simply locally convergent. The analysis begins by considering the local convergence of the shuffle product. It provides a point of reference and some important tools. The following theorem was proved in [28].

THEOREM 2.8. *Suppose $c, d \in \mathbb{R}^{\ell}_{LC}\langle\langle X\rangle\rangle$ with growth constants K_c, M_c and K_d, M_d, respectively. Then $c \sqcup\!\sqcup d \in \mathbb{R}^{\ell}_{LC}\langle\langle X\rangle\rangle$ with*

$$|(c \sqcup\!\sqcup d, \nu)| \leq K_c K_d M^{|\nu|}(|\nu|+1)! \ \ \forall \nu \in X^*, \tag{11}$$

where $M = \max\{M_c, M_d\}$.

Noting that $n + 1 \leq 2^n$ for all $n \geq 0$, (11) can be written more conventionally as

$$|(c \sqcup\!\sqcup d, \nu)| \leq K_c K_d (2M)^{|\nu|}|\nu|! \ \ \forall \nu \in X^*.$$

The specific goal here is to show that $c \circ d$ is also locally convergent, when the series c and d are locally convergent, and to produce an inequality analogous to (11). The following properties of the shuffle product are essential.

LEMMA 2.9 (see [28]). *For $c, d \in \mathbb{R}\langle\langle X\rangle\rangle$ and any $\nu \in X^*$,*

1. $$(c \sqcup\!\sqcup d, \nu) = \sum_{\xi,\bar{\xi}\in X^*} (c,\xi)(d,\bar{\xi})(\xi \sqcup\!\sqcup \bar{\xi}, \nu) = \sum_{i=0}^{|\nu|} \sum_{\substack{\xi\in X^i \\ \bar{\xi}\in X^{|\nu|-i}}} (c,\xi)(d,\bar{\xi})(\xi \sqcup\!\sqcup \bar{\xi}, \nu);$$

2. $$\sum_{\substack{\xi\in X^i \\ \bar{\xi}\in X^{|\nu|-i}}} (\xi \sqcup\!\sqcup \bar{\xi}, \nu) = \binom{|\nu|}{i}, \ \ 0 \leq i \leq |\nu|.$$

Now given any $\eta \in X^*$, the set of right factors $\{\eta_0, \eta_1, \ldots, \eta_k\}$ defined by (4) produces a corresponding family of real-valued functions:

$$S_{\eta_0}(n) = \frac{1}{|\eta_0|!}, \ \ n \geq 0,$$

$$S_{\eta_1}(n) = \frac{1}{(n)_{n_1+1}} S_{\eta_0}(n), \ \ 1 \leq |\eta_1| \leq n,$$

$$S_{\eta_j}(n) = \frac{1}{(n)_{n_j+1}} \sum_{i=0}^{n-|\eta_j|} S_{\eta_{j-1}}(n - (n_j+1) - i), \ \ j \leq |\eta_j| \leq n, \ \ 2 \leq j \leq k,$$

where $(n)_i = n!/(n-i)!$ denotes the falling factorial. The next two lemmas form the core of the local convergence proof for the composition product.

LEMMA 2.10. *Suppose $c \in \mathbb{R}^{\ell}_{LC}\langle\langle X\rangle\rangle$ and $d \in \mathbb{R}^{m}_{LC}\langle\langle X\rangle\rangle$ with growth constants K_c, M_c and K_d, M_d, respectively. Then*

$$|(c \circ d, \nu)| \leq K_c\, \psi_{|\nu|}(K_d)\, M^{|\nu|} |\nu|! \;\; \forall \nu \in X^*, \tag{12}$$

where $M = \max\{M_c, M_d\}$, and $\{\psi_n(K_d)\}_{n\geq 0}$ is the set of degree n polynomials in K_d,

$$\psi_n(K_d) = \sum_{i=0}^{n} \sum_{j=0}^{i} \sum_{\eta_j \in X^i} K_d^j\, S_{\eta_j}(n)\, |\eta_j|!, \;\; n \geq 0.$$

Proof. The proof has two main steps. It is first shown that for any integer $l > 0$ and any $\eta \in X^*$ with $|\eta| \leq l$ and right factors $\{\eta_0, \eta_1, \ldots, \eta_k\}$ as defined in (4),

$$|(\eta_j \circ d, \nu)| \leq K_d^j M_d^{-|\eta_j|} M_d^{|\nu|}\, |\nu|!\, S_{\eta_j}(|\nu|) \tag{13}$$

for all $0 \leq j \leq k$ and $|\eta_j| \leq |\nu| \leq l$. (Note that when $|\nu| < |\eta_j|$, the coefficients $(\eta_j \circ d, \nu) = 0$, and $S_{\eta_j}(|\nu|)$ is simply not defined.) This is shown by induction on j. The case $j = 0 < l$ is trivial. When $j = 1 \leq l$, the left-shift operator $x_0^{-(n_1+1)} := (x_0^{n_1+1})^{-1}$ is employed, where, in general, for any $\xi, \nu \in X^*$,

$$\xi^{-1}(\nu) = \begin{cases} \nu' & : \quad \nu = \xi\nu', \\ 0 & : \quad \text{otherwise.} \end{cases}$$

Observe the following for any ν with $|\eta_1| \leq |\nu| \leq l$ and containing the left factor $x_0^{n_1+1}$ (otherwise the claim is trivial):

$$\begin{aligned}
|(\eta_1 \circ d, \nu)| &= \left|(x_0^{n_1+1}(d_{i_1} \sqcup x_0^{n_0}), \nu)\right| \\
&= \left|\Big(d_{i_1} \sqcup x_0^{n_0}, \underbrace{x_0^{-(n_1+1)}(\nu)}_{\nu'}\Big)\right| \\
&= \left|\sum_{\xi \in X^{|\nu'|-n_0}} (d_{i_1}, \xi)(\xi \sqcup x_0^{n_0}, \nu')\right| \\
&\leq \sum_{\xi \in X^{|\nu'|-n_0}} (K_d M_d^{|\xi|} |\xi|!)\, (\xi \sqcup x_0^{n_0}, \nu') \quad (\text{since } 0 \leq |\xi| < l) \\
&\leq K_d M_d^{|\nu'|-n_0} (|\nu'| - n_0)! \binom{|\nu'|}{n_0} \\
&= K_d M_d^{-|\eta_1|} M_d^{|\nu|} |\nu|!\, S_{\eta_1}(|\nu|).
\end{aligned}$$

Now assume that the result holds up to some fixed j, where $1 \leq j \leq k-1$. Then in a similar fashion for $|\eta_{j+1}| \leq |\nu| \leq l$,

$$\begin{aligned}
|(\eta_{j+1} \circ d, \nu)| &= \left|\Big(d_{i_{j+1}} \sqcup (\eta_j \circ d), \underbrace{x_0^{-(n_{j+1}+1)}(\nu)}_{\nu'}\Big)\right| \\
&= \left|\sum_{i=0}^{|\nu'|} \sum_{\substack{\xi \in X^i \\ \bar{\xi} \in X^{|\nu'|-i}}} (d_{i_{j+1}}, \xi)(\eta_j \circ d, \bar{\xi})(\xi \sqcup \bar{\xi}, \nu')\right|.
\end{aligned}$$

Since $(\eta_j \circ d, \bar{\xi}) = 0$ for $|\bar{\xi}| < |\eta_j|$, it follows that, by using the coefficient bounds for d (because $0 \le |\xi| \le l - (j+1)$) and Lemma 2.9 (since $|\eta_j| \le |\bar{\xi}| < l - (n_{j+1}+1)$),

$$\begin{aligned}
|(\eta_{j+1} \circ d, \nu)| &\le \sum_{i=0}^{|\nu'|-|\eta_j|} \sum_{\substack{\xi \in X^i \\ \bar{\xi} \in X^{|\nu'|-i}}} (K_d M_d^{|\xi|} |\xi|!) \cdot \left(K_d^j M_d^{-|\eta_j|} M_d^{|\bar{\xi}|} |\bar{\xi}|!\, S_{\eta_j}(|\bar{\xi}|) \right) (\xi \sqcup\!\sqcup \bar{\xi}, \nu') \\
&= K_d^{j+1} M_d^{-|\eta_{j+1}|} M_d^{|\nu|} \sum_{i=0}^{|\nu'|-|\eta_j|} i!\, (|\nu'|-i)!\, S_{\eta_j}(|\nu'|-i) \binom{|\nu'|}{i} \\
&= K_d^{j+1} M_d^{-|\eta_{j+1}|} M_d^{|\nu|} |\nu|!\, \frac{1}{(|\nu|)_{n_{j+1}+1}} \sum_{i=0}^{|\nu'|-|\eta_j|} S_{\eta_j}(|\nu| - (n_{j+1}+1) - i) \\
&= K_d^{j+1} M_d^{-|\eta_{j+1}|} M_d^{|\nu|} |\nu|!\, S_{\eta_{j+1}}(|\nu|).
\end{aligned}$$

Hence, the claim is true for all $0 \le j \le k$.

In the second step of the proof, the claimed upper bound on $(c \circ d, \nu)$ is produced in terms of the polynomials $\psi_n(K_d)$. Since $\eta \in I_d(\nu)$ only if $|\eta| \le |\nu|$, using the inequality (13), it follows that

$$\begin{aligned}
|(c \circ d, \nu)| &= \left| \sum_{i=0}^{|\nu|} \sum_{j=0}^{i} \sum_{\eta_j \in X^i} (c, \eta_j)(\eta_j \circ d, \nu) \right| \\
&\le \sum_{i=0}^{|\nu|} \sum_{j=0}^{i} \sum_{\eta_j \in X^i} (K_c M^{|\eta_j|} |\eta_j|!) \cdot (K_d^j M^{-|\eta_j|} M^{|\nu|} |\nu|!\, S_{\eta_j}(|\nu|)) \\
&= K_c\, \psi_{|\nu|}(K_d)\, M^{|\nu|} |\nu|!. \qquad \square
\end{aligned}$$

LEMMA 2.11. *For each right factor η_j as defined in* (4) *of a given word $\eta \in X^*$, the following bounds apply:*

$$0 < S_{\eta_j}(n) \le \frac{(\alpha+1)^{n-|\eta_j|+j}}{\alpha^j |\eta_j|!}$$

for any $\alpha > 0$ and all $n \ge |\eta_j|$.

Proof. The proof is again by induction. The $j = 0$ case is trivial. When $j = 1$, observe that

$$\begin{aligned}
S_{\eta_1}(n) &= \frac{1}{(n)_{n_1+1} |\eta_0|!} \\
&\le \frac{1}{(|\eta_1|)_{n_1+1} |\eta_0|!}, \quad n \ge |\eta_1|, \\
&= \frac{1}{|\eta_1|!} \\
&\le \left(\frac{\alpha+1}{\alpha} \right) \frac{(\alpha+1)^{n-|\eta_1|}}{|\eta_1|!}, \quad n \ge |\eta_1|.
\end{aligned}$$

Now suppose the lemma is true up to some fixed $j \ge 1$. Then

$$S_{\eta_{j+1}}(n) = \frac{1}{(n)_{n_{j+1}+1}} \sum_{i=0}^{n-|\eta_{j+1}|} S_{\eta_j}(n - (n_{j+1}+1) - i)$$

$$\leq \frac{1}{(n)_{n_{j+1}+1}} \sum_{i=0}^{n-|\eta_{j+1}|} \frac{(\alpha+1)^{(n-(n_{j+1}+1)-i)-|\eta_j|+j}}{\alpha^j |\eta_j|!}$$

$$\leq \frac{(\alpha+1)^j}{\alpha^j |\eta_{j+1}|!} \sum_{i=0}^{n-|\eta_{j+1}|} (\alpha+1)^{n-|\eta_{j+1}|-i}, \quad n \geq |\eta_{j+1}|,$$

$$\leq \frac{(\alpha+1)^{n-|\eta_{j+1}|+j+1}}{\alpha^{j+1}|\eta_{j+1}|!}.$$

So the result holds for all $j \geq 0$. □

The main local convergence theorem for the composition product follows.

THEOREM 2.12. *Suppose $c \in \mathbb{R}^{\ell}_{LC}\langle\langle X\rangle\rangle$ and $d \in \mathbb{R}^{m}_{LC}\langle\langle X\rangle\rangle$ with growth constants K_c, M_c and K_d, M_d, respectively. Then $c \circ d \in \mathbb{R}^{\ell}_{LC}\langle\langle X\rangle\rangle$ with*

$$|(c \circ d, \nu)| \leq K_c((\phi(mK_d)+1)M)^{|\nu|}(|\nu|+1)! \ \ \forall \nu \in X^*,$$

where $\phi(x) := x/2 + \sqrt{x^2/4 + x}$ and $M = \max\{M_c, M_d\}$.

Proof. In light of Lemma 2.10, the goal is to show that $\psi_n(K_d) \leq (\phi(mK_d)+1)^n(n+1)$ for all $n \geq 0$. Observe that applying Lemma 2.11 gives, for any $\alpha > 0$,

$$\psi_n(K_d) \leq \sum_{i=0}^{n} \sum_{j=0}^{i} \sum_{\substack{\eta_j \in X^i \\ i \geq j}} K_d^j \frac{(\alpha+1)^{n-|\eta_j|+j}}{\alpha^j}$$

$$= (\alpha+1)^n \sum_{i=0}^{n} \sum_{j=0}^{i} \binom{i}{j} \left(\frac{mK_d}{\alpha}\right)^j \left(\frac{1}{\alpha+1}\right)^{i-j}$$

$$= (\alpha+1)^n \sum_{i=0}^{n} \beta^i,$$

where $\beta := mK_d/\alpha + 1/(\alpha+1)$. Setting $\beta = 1$ corresponds to letting $\alpha = \phi(mK_d)$, and the theorem is proved. (Note that $\phi(1) = \phi_g := (1+\sqrt{5})/2$, the golden ratio, and $\phi(mK_d) \approx mK_d$ when $mK_d \gg 1$.) □

Example 2.13. In some cases, the coefficient boundaries given in Theorem 2.12 are conservative; i.e., smaller growth constants might be produced by exploiting particular features of the series under consideration. For example, given linear series $c = \sum_{n\geq 0}(c, x_0^n x_1)x_0^n x_1$ and $d = \sum_{n\geq 0}(d, x_0^n x_1)x_0^n x_1$ in $\mathbb{R}_{LC}\langle\langle X\rangle\rangle$ with $X = \{x_0, x_1\}$, it can be shown directly that, by writing the composition product as a convolution sum and using the fact that $\sum_{k=0}^{n}\binom{n}{k}^{-1} < 3$ for any $n \geq 0$,

$$|(c \circ d, \nu)| < K_c K_d M^{|\nu|}|\nu|! \ \ \forall \nu \in X^*.$$

3. The nonrecursive connections. In this section the generating series are produced for the three nonrecursive interconnections, and their local convergence is characterized.

THEOREM 3.1. *If $c, d \in \mathbb{R}^{\ell}_{LC}\langle\langle X\rangle\rangle$, then each nonrecursive interconnected input-output system shown in* Figure 1.1(a)–(c) *has a Fliess operator representation generated by a locally convergent series as indicated:*

1. $F_c + F_d = F_{c+d}$;
2. $F_c \cdot F_d = F_{c \sqcup\!\sqcup d}$;
3. $F_c \circ F_d = F_{c \circ d}$, *where* $\ell = m$.

Proof.

1. Observe that

$$F_c[u](t) + F_d[u](t) = \sum_{\eta \in X^*} [(c,\eta) + (d,\eta)]\, E_\eta[u](t,t_0) = F_{c+d}[u](t).$$

Since c and d are locally convergent, define $M = \max\{M_c, M_d\}$. Then it follows that

$$|(c+d,\eta)| = |(c,\eta) + (d,\eta)| \leq (K_c + K_d) M^{|\eta|} |\eta|! \ \ \forall \eta \in X^*,$$

or $c + d$ is locally convergent.

2. In light of the componentwise definition of the product interconnection and the shuffle product, it can be assumed without loss of generality that $\ell = 1$. Therefore,

$$\begin{aligned} F_c[u](t) F_d[u](t) &= \sum_{\eta \in X^*} (c,\eta) E_\eta[u](t,t_0) \sum_{\xi \in X^*} (d,\xi) E_\xi[u](t,t_0) \\ &= \sum_{\eta,\xi \in X^*} (c,\eta)(d,\xi)\, E_{\eta \sqcup\!\sqcup \xi}[u](t,t_0) \\ &= F_{c \sqcup\!\sqcup d}[u](t). \end{aligned}$$

Local convergence of $c \sqcup\!\sqcup d$ is provided by Theorem 2.8.

3. It is first shown by induction that $F_\eta \circ F_d = F_{\eta \circ d}$ for any $\eta \in X^*$ and $d \in \mathbb{R}^m_{LC}\langle\langle X \rangle\rangle$. Choose any $\eta \in X^*$, and let $\{\eta_0, \eta_1, \ldots, \eta_k\}$ be the corresponding set of right factors defined in (4). Clearly,

$$(F_{\eta_0} \circ F_d[u])(t) = E_{\eta_0}[u](t,t_0) = F_{\eta_0}[u](t) = F_{\eta_0 \circ d}[u](t).$$

Now assume that

$$(F_{\eta_j} \circ F_d[u])(t) = F_{\eta_j \circ d}[u](t)$$

holds up to some fixed factor η_j. Then

$$\begin{aligned} (F_{\eta_{j+1}} \circ F_d[u])(t) &= E_{x_0^{n_{j+1}} x_{i_{j+1}} \eta_j}[F_d[u]](t,t_0) \\ &= \underbrace{\int_{t_0}^{t} \cdots \int_{t_0}^{\tau_2}}_{n_{j+1}+1 \text{ times}} F_{d_{i_{j+1}}}[u](\tau_1) E_{\eta_j}[F_d[u]](\tau_1,t_0)\, d\tau_1 \cdots d\tau_{n_{j+1}+1} \\ &= \underbrace{\int_{t_0}^{t} \cdots \int_{t_0}^{\tau_2}}_{n_{j+1}+1 \text{ times}} F_{d_{i_{j+1}} \sqcup\!\sqcup (\eta_j \circ d)}[u](\tau_1)\, d\tau_1 \cdots d\tau_{n_{j+1}+1} \\ &= F_{x_0^{n_{j+1}+1}[d_{i_{j+1}} \sqcup\!\sqcup (\eta_j \circ d)]}[u](t) \\ &= F_{\eta_{j+1} \circ d}[u](t). \end{aligned}$$

Thus, the claim holds for $\eta = \eta_{j+1}$ and, by induction, for $\eta = \eta_k$. Finally,

$$\begin{aligned} (F_c \circ F_d[u])(t) &= \sum_{\eta \in X^*} (c,\eta) E_\eta[F_d[u]](t,t_0) = \sum_{\eta \in X^*} (c,\eta) F_{\eta \circ d}[u](t) \\ &= \sum_{\eta \in X^*} (c,\eta) \left[\sum_{\nu \in X^*} (\eta \circ d, \nu) E_\nu[u](t,t_0) \right] \end{aligned}$$

$$
\begin{aligned}
&= \sum_{\nu\in X^*}\left[\sum_{\eta\in X^*}(c,\eta)(\eta\circ d,\nu)\right]E_\nu[u](t,t_0)\\
&= \sum_{\nu\in X^*}(c\circ d,\nu)\,E_\nu[u](t,t_0)\\
&= F_{c\circ d}[u](t).
\end{aligned}
$$

Local convergence of $c\circ d$ was proved in Theorem 2.12. □

Example 3.2. Let $X=\{x_0,x_1\}$, $c=\sum_{k\geq 0}K_cM_c^k k!\,x_1^k$, and $d=\sum_{k\geq 0}K_dM_d^k k!\,x_1^k$, where $K_c,M_c>0$ and $K_d,M_d>0$ are arbitrary growth constants. It is easily verified that the state space systems,

$$
\begin{aligned}
&\dot{z_c}=M_cz_c^2u_c,\quad z_c(0)=1, && \dot{z_d}=M_dz_d^2u_d,\quad z_d(0)=1,\\
&y_c=K_cz_c, && y_d=K_dz_d,
\end{aligned}
$$

realize the operators $F_c:u_c\mapsto y_c$ and $F_d:u_d\mapsto y_d$, respectively, for sufficiently small inputs and intervals of time. Letting $z=[z_c^T\ \ z_d^T]^T$, it follows directly that $F_{c\circ d}$ is realized by

$$\dot{z}=f(z)+g(z)u,\quad z(0)=[1\ \ 1]^T, \tag{14}$$

$$y=h(z), \tag{15}$$

where

$$f(z)=\begin{pmatrix}K_dM_cz_c^2z_d\\0\end{pmatrix},\quad g(z)=\begin{pmatrix}0\\M_dz_d^2\end{pmatrix},\quad h(z)=K_cz_c.$$

The first few coefficients of c, d, and $c\circ d$ are given in Table 3.1 along with the upper bounds on the coefficients of $c\circ d$ predicted by Theorem 2.12. Since these upper bounds hold for *any* series c and d with the given growth constants, they can be

TABLE 3.1
Some coefficients (c,ν), (d,ν), $(c\circ d,\nu)$ *and upper bounds for* $(c\circ d,\nu)$ *in Example* 3.2.

ν	(c,ν)	(d,ν)	$(c\circ d,\nu)$	Upper bounds for $(c\circ d,\nu)$
$\emptyset$	K_c	K_d	K_c	K_c
x_0	0	0	$K_c(K_dM_c)$	$K_c((\phi(K_d)+1)M)\,2!$
x_1	K_cM_c	K_dM_d	0	$K_c((\phi(K_d)+1)M)\,2!$
x_0^2	0	0	$K_c(K_dM_c)^2\,2!$	$K_c((\phi(K_d)+1)M)^2\,3!$
x_0x_1	0	0	$K_c(K_dM_c)M_d$	$K_c((\phi(K_d)+1)M)^2\,3!$
x_1x_0	0	0	0	$K_c((\phi(K_d)+1)M)^2\,3!$
x_1^2	$K_cM_c^2 2!$	$K_dM_d^2 2!$	0	$K_c((\phi(K_d)+1)M)^2\,3!$
x_0^3	0	0	$K_c(K_dM_c)^3\,3!$	$K_c((\phi(K_d)+1)M)^3\,4!$
$x_0^2x_1$	0	0	$K_c(K_dM_c)^2M_d\,2^2$	$K_c((\phi(K_d)+1)M)^3\,4!$
$x_0x_1x_0$	0	0	$K_c(K_dM_c)^2M_d\,2$	$K_c((\phi(K_d)+1)M)^3\,4!$
$x_0x_1^2$	0	0	$K_c(K_dM_c)M_d^2\,2$	$K_c((\phi(K_d)+1)M)^3\,4!$
$x_1x_0^2$	0	0	0	$K_c((\phi(K_d)+1)M)^3\,4!$
$x_1x_0x_1$	0	0	0	$K_c((\phi(K_d)+1)M)^3\,4!$
$x_1^2x_0$	0	0	0	$K_c((\phi(K_d)+1)M)^3\,4!$
x_1^3	$K_cM_c^3 3!$	$K_dM_d^3 3!$	0	$K_c((\phi(K_d)+1)M)^3\,4!$

TABLE 3.2
$T_{\max}$ and t_{esc} for specific examples of $c \circ d$ with $\bar{u} = 1$.

Case	K_c	M_c	K_d	M_d	$M_{c\circ d}$	$T_{\max}$	t_{esc}	$t_{\text{esc}}/T_{\max}$
1	4	2	2	2	7.46	0.03349	0.1967	5.873
2	2	4	2	2	14.93	0.01675	0.1105	6.598
3	2	2	4	2	11.66	0.02145	0.1105	5.152
4	2	2	2	4	14.93	0.01675	0.1580	9.435

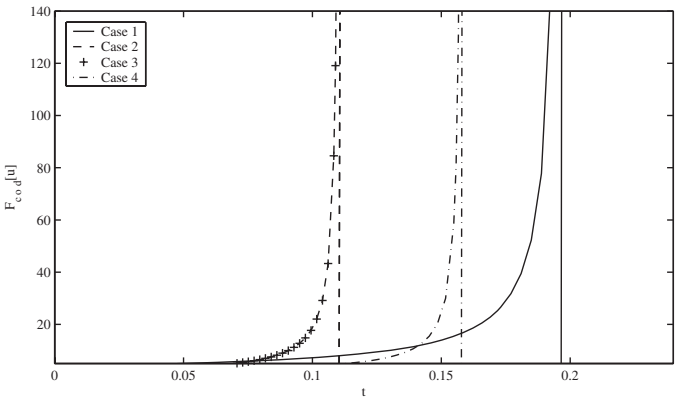


FIG. 3.1. *The output of $F_{c\circ d}[u]$ when $u(t) = \bar{u} = 1$ for Cases 1–4 of Table 3.2.*

conservative in specific cases. In [13] it is shown that given any series $c \in \mathbb{R}^{\ell}_{LC}\langle\langle X\rangle\rangle$, where $X = \{x_0, x_1, \ldots, x_m\}$ and $|(c,\nu)| \le K_c M_c^{|\nu|}\,|\nu|!$ for all $\nu \in X^*$, if

$$\max\{||u||_1, T\} \le \frac{1}{(m+1)^2 M_c},$$

then $F_c[u]$ converges absolutely and uniformly on $[0,T]$. The result still holds if one has the slightly more generous growth condition $|(c,\nu)| \le K_c M_c^{|\nu|}\,(|\nu|+1)!$. For a constant input $u(t) = \bar{u}$, where $|\bar{u}| \ge 1$, define

$$T_{\max} = \frac{1}{(m+1)^2 M_c |\bar{u}|}. \tag{16}$$

Then it follows from Theorem 2.12 that when $m = 1$, $F_{c\circ d}[\bar{u}]$ will always be well defined on at least the interval $[0, T_{\max})$, where

$$T_{\max} = \frac{1}{4M_{c\circ d}|\bar{u}|}$$

and $M_{c\circ d} = (\phi(K_d)+1)\max\{M_c, M_d\}$. Four specific cases are described in Table 3.2. Here each $T_{\max}$ is compared against the finite escape time, t_{esc}, of the state space system (14)–(15) with $u(t) = \bar{u} = 1$, which is determined numerically (see Figure 3.1). In each case, the value of $T_{\max} < t_{\text{esc}}$, but, as expected, $T_{\max}$ is conservative since the coefficient upper bounds for $c \circ d$ are conservative.

Example 3.3. The composition product provides an alternative interpretation of the symbolic calculus of Fliess [8, 10, 19]. Specifically, consider an input-output system represented by F_c with $c \in \mathbb{R}^{\ell}_{LC}\langle\langle X\rangle\rangle$. Any input u, which is analytic at $t = t_0$, can be represented near t_0 by a series $c_u \in \mathbb{R}^{m}_{LC}\langle\langle X_0\rangle\rangle$, i.e., $u = F_{c_u}[v]$ for

some locally convergent series $c_u = \sum_{k\geq 0}(c_u, x_0^k)\, x_0^k$ and all $\nu \in B_p^m(R)[t_0, t_0+T]$. In effect, c_u is the formal Laplace–Borel transform of the input u. (See [20] for more analysis of this example using the formal Laplace–Borel transform.) The analyticity of $y = F_c[u]$ follows from [28, Lemma 2.3.8], and therefore the formal Laplace–Borel transform of y, namely, c_y, can be related to c and c_u via

$$F_{c_y}[v] = y = F_c[F_{c_u}[v]] = F_{c\circ c_u}[v].$$

From [28, Corollary 2.2.4], it follows directly that $c_y = c \circ c_u$.

This last example motivates the following definition.

DEFINITION 3.4. *A series $c \in \mathbb{R}^\ell\langle\langle X\rangle\rangle$ is* input-output locally convergent *if for every $c_u \in \mathbb{R}^m_{LC}\langle\langle X_0\rangle\rangle$ it follows that $c \circ c_u \in \mathbb{R}^\ell_{LC}\langle\langle X_0\rangle\rangle$.*

It is immediate that every locally convergent series is input-output locally convergent, but the converse claim is only known to hold at present in certain special cases.

LEMMA 3.5. *Let $c \in \mathbb{R}^\ell\langle\langle X\rangle\rangle$ be an input-output locally convergent series with nonnegative coefficients. Then c is locally convergent.*

Proof. Set $c_u = \mathbb{1}$ and let K, M be the growth constants for the series $c \circ \mathbb{1}$. Then from Lemma 2.2, property 4,

$$|(c\circ\mathbb{1}, x_0^n)| = \max_i \sum_{\eta\in X^n} (c_i, \eta) \leq KM^n\, n!\ \ \forall n \geq 0.$$

Thus, $|(c,\eta)| = \max_i(c_i,\eta) \leq KM^n\, n!$ for all $n \geq 0$. ∎

LEMMA 3.6. *Let $c \in \mathbb{R}^\ell\langle\langle X\rangle\rangle$ be an input-output locally convergent linear series of the form $c = \sum_{j\geq 0}(c, x_0^j x_{i_j})\, x_0^j x_{i_j}$, where $i_j \in \{1,2,\ldots,m\}$ for all $j \geq 0$. Then c is locally convergent.*

Proof. Again set $c_u = \mathbb{1}$ and let K, M be the growth constants for the series $c \circ \mathbb{1}$. Then

$$|(c\circ\mathbb{1}, x_0^n)| = \max_i |(c_i, x_0^{n-1}x_{i_n})| \leq KM^n\, n!\ \ \forall n \geq 0,$$

and the conclusion follows. ∎

4. The feedback connection. Given any $c, d \in \mathbb{R}^m_{LC}\langle\langle X\rangle\rangle$, the general goal of this section is to determine when there exists a y which satisfies the feedback equation (1) and, in particular, when there exists a generating series e so that $y = F_e[u]$ for all admissible inputs u. In the latter case, the feedback equation becomes equivalent to

$$F_e[u] = F_c[u + F_{d\circ e}[u]], \tag{17}$$

and the *feedback product* of c and d is defined by $c@d = e$. It is assumed throughout that $m > 0$; otherwise the feedback connection is degenerate. An initial obstacle in this analysis is that F_e is required to be the composition of two operators, F_c and $I + F_{d\circ e}$, where the second operator is *never* realizable by a Fliess operator due to the direct feed term I. This does not prevent the composition from being a Fliess operator, but to compensate for the presence of this term a *modified* composition product is needed. Specifically, for any $\eta \in X^*$ and $d \in \mathbb{R}^m\langle\langle X\rangle\rangle$, define the modified composition product as

$$\eta \,\tilde{\circ}\, d = \begin{cases} \eta & : \ |\eta|_{x_i} = 0 \ \ \forall i \neq 0, \\ x_0^n x_i(\eta' \,\tilde{\circ}\, d) + x_0^{n+1}[d_i \sqcup\!\sqcup (\eta' \,\tilde{\circ}\, d)] & : \ \eta = x_0^n x_i \eta', \ \ n \geq 0, \ \ i \neq 0. \end{cases}$$

For $c \in \mathbb{R}^{\ell}\langle\langle X\rangle\rangle$ and $d \in \mathbb{R}^{m}\langle\langle X\rangle\rangle$, the definition is extended as

$$c \,\tilde{\circ}\, d = \sum_{\eta \in X^*} (c, \eta)\ \eta \,\tilde{\circ}\, d.$$

It can be verified in a manner completely analogous to the original composition product that the modified composition product is always well defined (summable), continuous in both arguments, and locally convergent when both c and d are. In particular, the following theorems are central to the analysis in this section.

THEOREM 4.1. *For any $c \in \mathbb{R}^{\ell}_{LC}\langle\langle X\rangle\rangle$ and $d \in \mathbb{R}^{m}_{LC}\langle\langle X\rangle\rangle$, it follows that*

$$F_{c \,\tilde{\circ}\, d}[u] = F_c[u + F_d[u]]$$

for all admissible u.

Proof. The result is verified simply by inserting the direct feed term into the proof of Theorem 3.1, part 3. ∎

THEOREM 4.2. *For any $c \in \mathbb{R}^{m}\langle\langle X\rangle\rangle$, the mapping $d \mapsto c \,\tilde{\circ}\, d$ is a contraction on $\mathbb{R}^{m}\langle\langle X\rangle\rangle$.*

Proof. This is also a minor variation of previous results concerning the composition product, in particular, Lemma 2.4, Lemma 2.5, and Theorem 2.6. The contraction coefficient, σ, is unaffected by the required modifications. ∎

The first main result of this section is given next.

THEOREM 4.3. *Let c, d be fixed series in $\mathbb{R}^{m}\langle\langle X\rangle\rangle$. Then the following propositions hold:*

1. *The mapping*

$$\begin{aligned} S &: \mathbb{R}^{m}\langle\langle X\rangle\rangle \to \mathbb{R}^{m}\langle\langle X\rangle\rangle \\ &: e_i \mapsto e_{i+1} = c \,\tilde{\circ}\, (d \circ e_i) \end{aligned} \tag{18}$$

has a unique fixed point in $\mathbb{R}^{m}\langle\langle X\rangle\rangle$, $c@d = \lim_{i\to\infty} e_i$, which is independent of e_0.

2. *If c, d, and $c@d$ are locally convergent, then $F_{c@d}$ satisfies the feedback equation* (17).

Proof.

1. The mapping S is a contraction since, by Theorems 2.6 and 4.2,

$$\operatorname{dist}(S(e_i), S(e_j)) \le \sigma \operatorname{dist}(d \circ e_i, d \circ e_j) \le \sigma^2 \operatorname{dist}(e_i, e_j).$$

Therefore, the mapping S has a unique fixed point, $c@d$, that is independent of e_0, i.e.,

$$c@d = c \,\tilde{\circ}\, (d \circ (c@d)). \tag{19}$$

2. From the stated assumptions concerning c, d, and $c@d$, it follows that

$$F_{c@d}[u] = F_{c \,\tilde{\circ}\, (d \circ (c@d))}[u] = F_c[u + F_d[F_{c@d}[u]]]$$

for any admissible u. ∎

The obvious question is whether $c@d$ is always locally convergent, or at least input-output locally convergent, when both c and d are locally convergent. The local convergence of c and d guarantees that the feedback system in Figure 1.1(d) is at least *well-posed* in the sense described in [1, 27] since F_c and F_d are well-defined causal analytic operators. That is, there exist sufficiently small $R, S, T > 0$ such that

for any $u \in B_p^m(R)[t_0, t_0+T]$, there exists a $y \in B_q^m(S)[t_0, t_0+T]$ which satisfies the feedback equation (1). But whether $y = F_{c@d}[u]$ on some ball of input functions of nonzero radius over a nonzero interval of time is not immediate. The following example shows that $\mathbb{R}^m_{LC}\langle\langle X\rangle\rangle$ is not a closed subset of $\mathbb{R}^m\langle\langle X\rangle\rangle$ in the ultrametric topology.

Example 4.4. Let $X = \{x_0, x_1\}$ and consider the following sequence of polynomials in $\mathbb{R}^m_{LC}\langle\langle X\rangle\rangle$:

$$e_i = x_1 + 2^2 2!\, x_1^2 + 3^3 3!\, x_1^3 + \cdots + i^i i!\, x_1^i, \quad i \geq 1.$$

Clearly, $e = \lim_{i\to\infty} e_i$ is not locally convergent.

A central issue is whether such an example can be produced by repeated compositions of a locally convergent series. It will be first shown that the answer to this question is *no*. Then the more general case described by (18) is examined. This leads to the main conclusion that the feedback product of two locally convergent series is always input-output locally convergent.

Observe first that if $e = c \circ e$, then it follows that e must have the form $e = \sum_{n\geq 0}(e, x_0^n)\, x_0^n$. Furthermore, since e appears on both sides of the expression $e = c \circ e$, it is possible by repeated substitution to express each coefficient (e, x_0^n) in terms of the coefficients $\{(c, \nu) : |\nu| \leq n\}$. For example, if $X = \{x_0, x_1\}$, the first few coefficients of e are

$$\begin{aligned}
(e,\emptyset) &= (c,\emptyset),\\
(e,x_0) &= (c,x_0) + (c,\emptyset)(c,x_1),\\
(e,x_0^2) &= (c,x_0^2) + (c,x_0)(c,x_1) + (c,\emptyset)(c,x_1)^2 + (c,\emptyset)(c,x_0x_1) + (c,\emptyset)(c,x_1x_0)\\
&\quad +(c,\emptyset)^2(c,x_1^2),\\
(e,x_0^3) &= (c,x_0^2)(c,x_1) + (c,x_0)(c,x_1)^2 + (c,\emptyset)(c,x_1)^3 + (c,\emptyset)(c,x_1)(c,x_0x_1)\\
&\quad +(c,\emptyset)(c,x_1)(c,x_1x_0) + (c,\emptyset)^2(c,x_1)(c,x_1^2) + (c,x_0)(c,x_0x_1)\\
&\quad +(c,\emptyset)(c,x_1)(c,x_0x_1) + 2(c,x_0)(c,x_1x_0) + 2(c,\emptyset)(c,x_1)(c,x_1x_0)\\
&\quad +3(c,\emptyset)(c,x_0)(c,x_1^2) + 3(c,\emptyset)^2(c,x_1)(c,x_1^2) + (c,x_0^3) + (c,\emptyset)(c,x_0^2x_1)\\
&\quad +(c,\emptyset)(c,x_0x_1x_0) + (c,\emptyset)^2(c,x_0x_1^2) + (c,\emptyset)(c,x_1x_0^2) + (c,\emptyset)^2(c,x_1x_0x_1)\\
&\quad +(c,\emptyset)^2(c,x_1^2x_0) + (c,\emptyset)^3(c,x_1^3)\\
&\ \ \vdots
\end{aligned}$$

If c is locally convergent with growth constants K_c, M_c, then

$$\begin{aligned}
|(e,\emptyset)| &\leq K_c,\\
|(e,x_0)| &\leq K_c(K_c+1)M_c,\\
|(e,x_0^2)| &\leq K_c\left(\frac{3}{2}K_c^2 + \frac{5}{2}K_c + 1\right) M_c^2\, 2!,\\
|(e,x_0^3)| &\leq K_c\left(\frac{5}{2}K_c^3 + \frac{35}{6}K_c^2 + \frac{13}{3}K_c + 1\right) M_c^3\, 3!\\
&\ \ \vdots
\end{aligned}$$

This suggests that a variation of inequality (12) is possible, namely, that

$$|(e, x_0^n)| \leq K_c\, \tilde{\psi}_n(K_c)\, M_c^n\, n! \ \ \forall n \geq 0,$$

TABLE 4.1
The first few polynomials $\tilde{S}_{\eta_j}(K_c, n)$ and $\tilde{\psi}_n(K_c)$ when $m = 1$.

n	η_j	$\tilde{S}_{\eta_0}(K_c,n),\dots,\tilde{S}_{\eta_j}(K_c,n)$	$\tilde{\psi}_n(K_c)$
0	$\emptyset$	$\tilde{S}_{\emptyset}(K_c,0)=1$	1
1	x_0	$\tilde{S}_{x_0}(K_c,1)=1$	K_c+2
	x_1	$\tilde{S}_{\emptyset}(K_c,1)=1,\ \tilde{S}_{x_1}(K_c,1)=1$	
2	x_0^2	$\tilde{S}_{x_0^2}(K_c,2)=\frac{1}{2}$	$\frac{3}{2}K_c^2+3K_c+3$
	x_0x_1	$\tilde{S}_{\emptyset}(K_c,2)=1,\ \tilde{S}_{x_0x_1}(K_c,2)=\frac{1}{2}$	
	x_1x_0	$\tilde{S}_{x_0}(K_c,2)=1,\ \tilde{S}_{x_1x_0}(K_c,2)=\frac{1}{2}$	
	x_1^2	$\tilde{S}_{\emptyset}(K_c,2)=1,\ \tilde{S}_{x_1}(K_c,2)=\frac{1}{2}K_c+1,$ $\tilde{S}_{x_1^2}(K_c,2)=\frac{1}{2}$	
3	x_0^3	$\tilde{S}_{x_0^3}(K_c,3)=\frac{1}{6}$	$\frac{5}{2}K_c^3+7K_c^2+6K_c+4$
	$x_0^2x_1$	$\tilde{S}_{\emptyset}(K_c,3)=1,\ \tilde{S}_{x_0^2x_1}(K_c,3)=\frac{1}{6}$	
	$x_0x_1x_0$	$\tilde{S}_{x_0}(K_c,3)=1,\ \tilde{S}_{x_0x_1x_0}(K_c,3)=\frac{1}{6}$	
	$x_0x_1^2$	$\tilde{S}_{\emptyset}(K_c,3)=1,\ \tilde{S}_{x_1}(K_c,3)=\frac{1}{2}K_c^2+K_c+1,$ $\tilde{S}_{x_0x_1^2}(K_c,3)=\frac{1}{6}$	
	$x_1x_0^2$	$\tilde{S}_{x_0^2}(K_c,3)=\frac{1}{2},\ \tilde{S}_{x_1x_0^2}(K_c,3)=\frac{1}{6}$	
	$x_1x_0x_1$	$\tilde{S}_{\emptyset}(K_c,3)=1,\ \tilde{S}_{x_0x_1}(K_c,3)=\frac{1}{6}K_c+\frac{1}{3},$ $\tilde{S}_{x_1x_0x_1}(K_c,3)=\frac{1}{6}$	
	$x_1^2x_0$	$\tilde{S}_{x_0}(K_c,3)=1,\ \tilde{S}_{x_1x_0}(K_c,3)=\frac{1}{3}K_c+\frac{2}{3},$ $\tilde{S}_{x_1^2x_0}(K_c,3)=\frac{1}{6}$	
	x_1^3	$\tilde{S}_{\emptyset}(K_c,3)=1,\ \tilde{S}_{x_1}(K_c,3)=\frac{1}{2}K_c^2+K_c+1,$ $\tilde{S}_{x_1^2}(K_c,3)=\frac{1}{2}K_c+1,\ \tilde{S}_{x_1^3}(K_c,3)=\frac{1}{6}$	

where each $\tilde{\psi}_n(K_c)$ is a polynomial in K_c of degree n. The next lemma establishes the claim using a family of polynomials of the form

$$\tilde{\psi}_n(K_c)=\sum_{i=0}^{n}\sum_{j=0}^{i}\sum_{\eta_j\in X^i}K_c^j\tilde{S}_{\eta_j}(K_c,n)|\eta_j|!,\quad n\geq 0.$$

Given a fixed n, every word η_j in the innermost summation satisfies $j\leq|\eta_j|\leq n$ and has a corresponding set of right factors $\{\eta_0,\eta_1,\dots,\eta_j\}$. When $j>0$, each polynomial $\tilde{S}_{\eta_j}(K_c,n)$ is computed iteratively using its right factors and the previously computed polynomials $\{\tilde{\psi}_0(K_c),\tilde{\psi}_1(K_c),\dots,\tilde{\psi}_{n-1}(K_c)\}$:

$$\begin{aligned}
\tilde{S}_{\eta_0}(K_c,n)&=\frac{1}{|\eta_0|!},\quad 0\leq|\eta_0|\leq n,\\
\tilde{S}_{\eta_1}(K_c,n)&=\frac{1}{(n)_{n_1+1}}\,\tilde{\psi}_{n-|\eta_1|}(K_c)\,\tilde{S}_{\eta_0}(K_c,n),\quad 1\leq|\eta_1|\leq n,\\
\tilde{S}_{\eta_2}(K_c,n)&=\frac{1}{(n)_{n_2+1}}\sum_{i=0}^{n-|\eta_2|}\tilde{\psi}_i(K_c)\,\tilde{S}_{\eta_1}(K_c,n-(n_2+1)-i),\quad 2\leq|\eta_2|\leq n,\\
&\vdots\\
\tilde{S}_{\eta_j}(K_c,n)&=\frac{1}{(n)_{n_j+1}}\sum_{i=0}^{n-|\eta_j|}\tilde{\psi}_i(K_c)\,\tilde{S}_{\eta_{j-1}}(K_c,n-(n_j+1)-i),\quad 2\leq j\leq|\eta_j|\leq n.
\end{aligned}$$

See Table 4.1 for the case where $m=1$.

LEMMA 4.5. *Let $c \in \mathbb{R}^m_{LC}\langle\langle X\rangle\rangle$ with growth constants K_c, M_c, and $e \in \mathbb{R}^m\langle\langle X\rangle\rangle$ such that $e = c \circ e$. Then*

$$|(e, x_0^n)| \leq K_c\, \tilde{\psi}_n(K_c)\, M_c^n\, n!\ \ \forall n \geq 0. \tag{20}$$

Proof. The proof has some elements in common with that of Lemma 2.10, except here it is not assumed a priori that e is locally convergent. The basic approach employs *nested inductions*. The outer induction is on n. It is clear from the discussion above that the claim holds when $n = 0$ and $n = 1$ for $m = 1$. A similar calculation can be done for arbitrary $m \geq 1$. Now suppose (20) holds up to some fixed $n - 1 \geq 1$. Given any η_j, where $j \leq |\eta_j| \leq n$, it will first be shown by induction on j (the inner induction) that

$$|(\eta_j \circ e, x_0^n)| \leq K_c^j M_c^{-|\eta_j|} M_c^n\, n!\, \tilde{S}_{\eta_j}(K_c, n),\ \ 0 \leq j \leq n. \tag{21}$$

The $j = 0$ case is trivial. Suppose $j = 1$. Then $0 \leq n - |\eta_1| \leq n - 1$ and

$$\begin{aligned}
|(\eta_1 \circ e, x_0^n)| &= \left|\left(x_0^{n_1+1}(e_{i_1} \sqcup\!\sqcup x_0^{n_0}), x_0^n\right)\right| \\
&= \left|\left(e_{i_1} \sqcup\!\sqcup x_0^{n_0}, x_0^{n-(n_1+1)}\right)\right| \\
&= \left|\left(e_{i_1}, x_0^{n-|\eta_1|}\right)\left(x_0^{n-|\eta_1|} \sqcup\!\sqcup x_0^{n_0}, x_0^{n-(n_1+1)}\right)\right| \\
&\leq \left(K_c\, \tilde{\psi}_{n-|\eta_1|}(K_c)\, M_c^{n-|\eta_1|}(n - |\eta_1|)!\right)\binom{n-(n_1+1)}{n-|\eta_1|} \\
&= K_c M_c^{-|\eta_1|} M_c^n\, n!\, \tilde{S}_{\eta_1}(K_c, n).
\end{aligned}$$

Now assume that inequality (21) holds up to some fixed j, where $1 \leq j \leq n-1$. Then $0 \leq n - |\eta_{j+1}| \leq n - (j+1)$ and

$$\begin{aligned}
|(\eta_{j+1} \circ e, x_0^n)| &= \left|\left(e_{i_{j+1}} \sqcup\!\sqcup (\eta_j \circ e), x_0^{n-(n_{j+1}+1)}\right)\right| \\
&= \left|\sum_{i=0}^{n-(n_{j+1}+1)} \left(e_{i_{j+1}}, x_0^i\right)\left(\eta_j \circ e, x_0^{n-(n_{j+1}+1)-i}\right)\binom{n-(n_{j+1}+1)}{n-(n_{j+1}+1)-i}\right|.
\end{aligned}$$

Since $(\eta_j \circ e, x_0^{n-(n_{j+1}+1)-i}) = 0$ when $n - (n_{j+1}+1) - i < |\eta_j|$ or, equivalently, $i > n - |\eta_{j+1}|$, it follows that, using the coefficient bound (20) for e (because $0 \leq i \leq n-1$) and the bound (21) for $\eta_j \circ e$,

$$\begin{aligned}
|(\eta_{j+1} \circ e, x_0^n)| \leq & \sum_{i=0}^{n-|\eta_{j+1}|} \left(K_c \tilde{\psi}_i(K_c) M_c^i\, i!\right)\Big(K_c^j M_c^{-|\eta_j|} M_c^{n-(n_{j+1}+1)-i} \\
& \cdot (n-(n_{j+1}+1)-i)!\, \tilde{S}_{\eta_j}(K_c, n-(n_{j+1}+1)-i)\Big) \\
& \cdot \binom{n-(n_{j+1}+1)}{n-(n_{j+1}+1)-i} \\
= & K_c^{j+1} M_c^{-|\eta_{j+1}|} M_c^n\, n!\, \frac{1}{(n)_{n_{j+1}+1}} \\
& \cdot \sum_{i=0}^{n-|\eta_{j+1}|} \tilde{\psi}_i(K_c)\, \tilde{S}_{\eta_j}(K_c, n-(n_{j+1}+1)-i) \\
= & K_c^{j+1} M_c^{-|\eta_{j+1}|} M_c^n\, n!\, \tilde{S}_{\eta_{j+1}}(K_c, n).
\end{aligned}$$

Hence, the claim is true for all $0 \le j \le n$.

To complete the outer induction with respect to n, observe that

$$\begin{aligned}|(e, x_0^n)| = |(c \circ e, x_0^n)| &= \left| \sum_{i=0}^{n} \sum_{j=0}^{i} \sum_{\eta_j \in X^i} (c, \eta_j)(\eta_j \circ e, x_0^n) \right| \\ &\le \sum_{i=0}^{n} \sum_{j=0}^{i} \sum_{\eta_j \in X^i} \left(K_c M_c^{|\eta_j|} |\eta_j|! \right) \left(K_c^j M_c^{-|\eta_j|} M_c^n \, n! \, \tilde{S}_{\eta_j}(K_c, n) \right) \\ &= K_c \, \tilde{\psi}_n(K_c) \, M_c^n \, n!.\end{aligned}$$

Therefore, inequality (20) holds for all $n \ge 0$. ∎

The next lemma provides an upper bound on the growth of the sequence $\tilde{\psi}_n(K_c)$, $n \ge 0$, when K_c is fixed.

LEMMA 4.6. *For any $K_c \ge 1$, it follows that*

$$\tilde{\psi}_n(K_c) \le \phi_g (m K_c (2 + \phi_g) + 1)^n s_n \quad \forall n \ge 0, \tag{22}$$

where $s_0 = 1/\phi_g$, and s_n, $n \ge 1$, is an integer sequence equivalent to the binomial transform of the sequence of Catalan numbers, C_n, $n \ge 1$ (specifically, sequence A007317 *in* [25]).

Proof. The proof has two main parts. First, it is shown by a nested induction that, for any $\epsilon > 0$, there exists a sequence of positive real numbers, $\xi_n(\epsilon)$, such that

$$\tilde{\psi}_n(K_c) \le (m K_c (2 + \epsilon) + 1)^n \xi_n(\epsilon), \quad n \ge 0, \quad K_c \ge 1. \tag{23}$$

Then inequality (22) is produced for $n \ge 1$ by setting $\epsilon = \phi_g$ and showing that $\xi_n(\phi_g) = \phi_g s_n$ when $n \ge 1$. ($n = 0$ is a trivial special case.)

Let $\epsilon > 0$ and define two sequences of positive real numbers, $\xi_n(\epsilon)$ and $\Gamma_n(\epsilon)$, via the recurrence equations

$$\xi_{n+1}(\epsilon) = \xi_n(\epsilon) + \Gamma_{n+1}(\epsilon), \quad n \ge 0, \quad \xi_0 = 1, \quad \Gamma_1 = 1/\epsilon, \tag{24}$$

$$\Gamma_{n+1}(\epsilon) = \frac{1}{\epsilon} \left[\xi_n(\epsilon) + \sum_{i=1}^{n} \xi_i(\epsilon) \Gamma_{n-i+1}(\epsilon) \right], \quad n \ge 1. \tag{25}$$

By definition, $\Gamma_0 = 1$. In light of Table 4.1, inequality (23) clearly holds when $n = 0$ and $n = 1$ for $m = 1$ and $K_c \ge 1$. (It is easily verified to also hold when $m \ge 1$.) Now suppose the inequality holds up to some fixed $n - 1 \ge 1$. Given any word η_j, where $j \le |\eta_j| \le n$, an inner induction with respect to j will now show that

$$\tilde{S}_{\eta_j}(K_c, n) \le \frac{(m K_c (2 + \epsilon) + 1)^{n - |\eta_j|} \, (2 + \epsilon)^j \, \Gamma_{n - |\eta_j|}(\epsilon)}{|\eta_j|!}, \quad 0 \le j \le |\eta_j| \tag{26}$$

(cf. the proof of Lemma 2.11, where some of the computational details are similar). The $j = 0$ case is trivial. Suppose $j = 1$. Since $n - |\eta_1| < n$, it follows that

$$\begin{aligned}\tilde{S}_{\eta_1}(K_c, n) &= \frac{1}{(n)_{n_1+1}} \frac{\tilde{\psi}_{n-|\eta_1|}(K_c)}{|\eta_0|!} \\ &\le \frac{(m K_c (2 + \epsilon) + 1)^{n - |\eta_1|} \, \xi_{n-|\eta_1|}(\epsilon)}{|\eta_1|!} \\ &\le \frac{(m K_c (2 + \epsilon) + 1)^{n - |\eta_1|} \, (2 + \epsilon) \, \Gamma_{n-|\eta_1|}(\epsilon)}{|\eta_1|!}, \quad n \ge |\eta_1|.\end{aligned}$$

This last inequality employs the general properties for any $j \geq 0$ that $\xi_{n-|\eta_j|}(\epsilon) = \Gamma_{n-|\eta_j|}(\epsilon)$ when $n = |\eta_j|$ and

$$\sum_{i=0}^{n-|\eta_j|} \xi_i(\epsilon)\Gamma_{n-|\eta_j|-i}(\epsilon) = (2+\epsilon)\,\Gamma_{n-|\eta_j|}(\epsilon) \tag{27}$$

when $n > |\eta_j|$. Now suppose that inequality (26) holds up to some fixed $j \geq 1$. Then

$$\begin{aligned}
\tilde{S}_{\eta_{j+1}}(K_c, n) &= \frac{1}{(n)_{n_{j+1}+1}} \sum_{i=0}^{n-|\eta_{j+1}|} \tilde{\psi}_i(K_c)\, \tilde{S}_{\eta_j}(K_c, n-(n_{j+1}+1)-i) \\
&\leq \frac{1}{|\eta_{j+1}|!} \sum_{i=0}^{n-|\eta_{j+1}|} (mK_c(2+\epsilon)+1)^i \xi_i(\epsilon) \\
&\qquad \cdot \Big[(mK_c(2+\epsilon)+1)^{n-|\eta_{j+1}|-i}\,(2+\epsilon)^j\,\Gamma_{n-|\eta_{j+1}|-i}(\epsilon)\Big] \\
&= \frac{(mK_c(2+\epsilon)+1)^{n-|\eta_{j+1}|}(2+\epsilon)^j}{|\eta_{j+1}|!} \sum_{i=0}^{n-|\eta_{j+1}|} \xi_i(\epsilon)\Gamma_{n-|\eta_{j+1}|-i}(\epsilon) \\
&= \frac{(mK_c(2+\epsilon)+1)^{n-|\eta_{j+1}|}\,(2+\epsilon)^{j+1}\,\Gamma_{n-|\eta_{j+1}|}(\epsilon)}{|\eta_{j+1}|!}, \quad |\eta_j| < |\eta_{j+1}| \leq n,
\end{aligned}$$

where again identity (27) was used to derive the final equality above. Hence, inequality (26) holds for all $0 \leq j \leq |\eta_j|$. To complete the outer induction with respect to n, observe that

$$\begin{aligned}
\tilde{\psi}_{n+1}(K_c) &= \sum_{i=0}^{n+1} \sum_{j=0}^{i} \sum_{\eta_j \in X^i} K_c^j \tilde{S}_{\eta_j}(K_c, n+1)|\eta_j|! \\
&\leq \sum_{i=0}^{n+1} \sum_{j=0}^{i} \binom{i}{j} \left[\frac{(mK_c(2+\epsilon)+1)^{n+1-i}\,(mK_c(2+\epsilon))^j\,\Gamma_{n+1-i}(\epsilon)}{i!}\right] i! \\
&= (mK_c(2+\epsilon)+1)^{n+1} \sum_{i=0}^{n+1} \Gamma_{n+1-i}(\epsilon) \\
&= (mK_c(2+\epsilon)+1)^{n+1} \xi_{n+1}(\epsilon).
\end{aligned}$$

Thus, inequality (23) must hold for all $n \geq 0$.

Now consider setting $\epsilon = \phi_g$ in the system of equations (24)–(25). Eliminating by substitution the sequence $\Gamma_n(\phi_g)$ gives the recurrence relation

$$\xi_{n+1}(\phi_g) = \phi_g + \frac{1}{\phi_g} \sum_{i=1}^{n} \xi_i(\phi_g)\xi_{n-i+1}(\phi_g), \quad n \geq 1, \quad \xi_1(\phi_g) = \phi_g,$$

or, equivalently,

$$\left(\frac{\xi_{n+1}(\phi_g)}{\phi_g}\right) = 1 + \sum_{i=1}^{n} \left(\frac{\xi_i(\phi_g)}{\phi_g}\right)\left(\frac{\xi_{n-i+1}(\phi_g)}{\phi_g}\right), \quad n \geq 1, \quad \frac{\xi_1(\phi_g)}{\phi_g} = 1.$$

It is known that s_n satisfies the recurrence equation

$$s_{n+1} = 1 + \sum_{i=1}^{n} s_i s_{n-i+1}, \quad n \geq 1, \quad s_1 = 1 \tag{28}$$

(see [25] and the references therein). Hence, the conclusion that $\xi_n(\phi_g) = \phi_g s_n$, $n \geq 1$, is immediate. ∎

The recurrence equation (28) can be derived from the well-known recurrence relation for the Catalan numbers: $C_{n+1} = \sum_{i=0}^{n} C_i C_{n-i}$ with $C_0 = 1$ [4], which in turn is equivalent to Segner's recurrence formula given in the year 1758 as a solution to Euler's polygon division problem [31]. It is also worth noting that the sequence $t_n := \Gamma_n(\phi_g)/\phi_g$, $n \geq 1$, the increments of s_n, is sequence A002212 in [25]. The positive integer sequences C_n, s_n, and t_n each have a variety of combinatoric interpretations in graph theory and the theory of formal languages. Of particular interest to system theorists, for example, is the fact that C_n is equivalent to the number of ways to binary bracket the letters in a word of length $n+1$ [31, 32]. The asymptotic behavior of s_n,

$$s_n \sim \frac{1}{8}\sqrt{\frac{5}{\pi}}\,\frac{5^n}{n^{3/2}}$$

(see [15, sequence 124]), motivates the following central result concerning local convergence.

THEOREM 4.7. *If $c \in \mathbb{R}^m_{LC}\langle\langle X\rangle\rangle$ with growth constants K_c, M_c, and $e = c \circ e$, then $e \in \mathbb{R}^m_{LC}\langle\langle X_0\rangle\rangle$. Specifically, for any $K_c \geq 1$,*

$$|(e, x_0^n)| \leq K_c((mK_c(2+\phi_g)+1)5M_c)^n\ n!\ \ \forall n \geq 0. \tag{29}$$

Proof. The result is trivial when $n = 0$. When $n \geq 1$, it is first necessary to show by induction that $s_{n+1} < 5s_n$. The claim is clearly true when $n = 1$ or $n = 2$. Suppose it is known to hold up to some fixed integer $n+1 \geq 2$. Sequence s_n is known to satisfy another recurrence equation [15, 25]:

$$(n+2)s_{n+2} = (6n+4)s_{n+1} - 5ns_n.$$

Therefore,

$$s_{n+2} < [(6n+4)s_{n+1} - ns_{n+1}]/(n+2) < 5s_{n+1},$$

which proves the claim for all $n \geq 1$. Next, substituting the upper bound $\phi_g s_n \leq 5^n, n \geq 0$, into (22) gives

$$\tilde{\psi}_n(K_c) \leq ((mK_c(2+\phi_g)+1)5)^n\ \ \forall n \geq 0. \tag{30}$$

The theorem is finally proved by simply applying Lemma 4.5. ∎

In most cases the upper bound in (29) is quite conservative because the upper bound (30) is conservative. Figure 4.1 shows $\tilde{\psi}_n(K_c)$ (generated symbolically via MAPLE) and upper bound (30) versus n for various values of K_c.

The final step of the analysis is to use Theorem 4.7 to prove the input-output local convergence of the feedback product.

THEOREM 4.8. *If $c, d \in \mathbb{R}^m_{LC}\langle\langle X\rangle\rangle$, then c@d is input-output locally convergent. Specifically, when $K_c \geq 1$, then*

$$((c@d) \circ b, x_0^n) \leq K_c([mK_c(2+\phi_g)+1][\phi(m(K_b+K_d))+1]10M)^n\ n!$$

for any $b \in \mathbb{R}^m_{LC}\langle\langle X_0\rangle\rangle$ and where $M = \max\{M_b, M_c, M_d\}$.

Proof. Select any series $b \in \mathbb{R}^m_{LC}\langle\langle X_0\rangle\rangle$. It follows from (19) that

$$(c@d) \circ b = (c \,\tilde{\circ}\, (d \circ (c@d))) \circ b = c \circ (b+d) \circ ((c@d) \circ b).$$

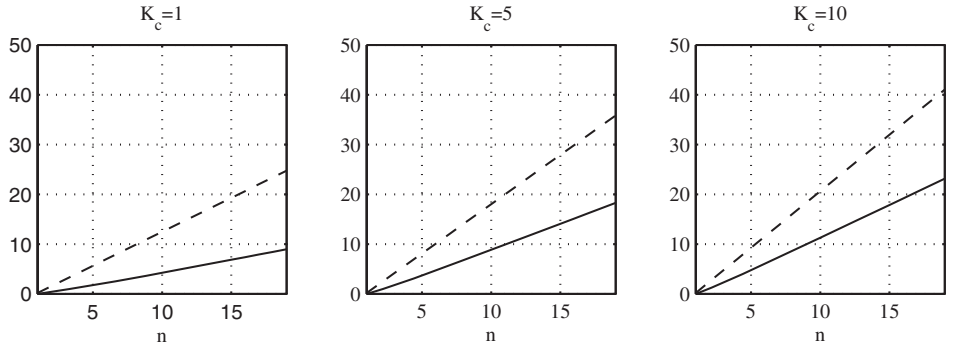


FIG. 4.1. *A plot of* $\log_{10}(\tilde{\psi}_n(K_c))$ (*solid lines*) *and the logarithm* (*base* 10) *of the upper bound in* (30) (*dashed lines*) *versus* n *for various values of* K_c.

Since b, c, and d are all locally convergent, so is the series $c \circ (b + d)$. Now apply Theorem 4.7, replacing c with $c \circ (b + d)$ and e with $(c@d) \circ b$. This implies that $(c@d) \circ b$ is always locally convergent, and therefore $c@d$ must be input-output locally convergent. To produce the given growth condition for the output series, note that

$$K_{c\circ(b+d)} = K_c \qquad M_{c\circ(b+d)} = 2(\phi(m(K_b + K_d)) + 1)M,$$

using Theorem 2.12 and the fact that $n + 1 \leq 2^n$ for all $n \geq 0$. Substituting these growth constants for K_c and M_c, respectively, in Theorem 4.7 produces the desired result. □

Example 4.9. Suppose c and d are linear series in $\mathbb{R}^m_{LC}\langle\langle X\rangle\rangle$. Then $c@d = \lim_{i\to\infty} e_i$, where

$$e_{i+1} = c \,\tilde{\circ}\, (d \circ e_i) = c + (c \circ d) \circ e_i.$$

Setting $e_0 = c$ gives

$$c@d = c + \sum_{k=1}^{\infty} (c \circ d)^{\circ k} \circ c,$$

where $c^{\circ k}$ denotes k copies of c composed $k - 1$ times. It is easily verified since $(c, \emptyset) = 0$ that $((c \circ d)^{\circ k}, \nu) = 0$ for all $k > |\nu|$. Hence,

$$(c@d, \nu) = (c, \nu) + \sum_{k=1}^{|\nu|-1} ((c \circ d)^{\circ k} \circ c, \nu).$$

Example 4.10. For any $c, d \in \mathbb{R}_{LC}\langle\langle X\rangle\rangle$, a self-excited feedback loop can be described by $F_{c@d}[0] = F_{(c@d)\circ 0}[u] = F_{(c@d)_0}[u]$ (cf. Lemma 2.2, property 2). In this case $(c@d)_0 = \lim_{i\to\infty} e_i$, where $e_{i+1} = (c \circ d) \circ e_i$. Using the $m = 0$ version of (16)

TABLE 4.2
Some coefficients (c,ν), (d,ν), *and* $(c@d,\nu)$ *in Example* 4.11.

ν	(c,ν)	(d,ν)	$(c@d,\nu)$
$\emptyset$	K_c	K_d	K_c
x_0	0	0	$K_cK_dM_c$
x_1	K_cM_c	K_dM_d	K_cM_c
x_0^2	0	0	$K_c((K_dM_c)^2\ 2! + K_cK_dM_cM_d)$
x_0x_1	0	0	$K_cK_dM_c^2\ 2!$
x_1x_0	0	0	$K_cK_dM_c^2\ 2!$
x_1^2	$K_cM_c^2 2!$	$K_dM_d^2 2!$	$K_cM_c^2\ 2!$
x_0^3	0	0	$K_c((K_dM_c)^3\ 3! + K_c(K_dM_c)^2M_d\ 7 + K_c^2K_dM_cM_d^2\ 2!)$
$x_0^2x_1$	0	0	$K_c((K_dM_c)^2M_c\ 3! + K_cK_dM_c^2M_d\ 3)$
$x_0x_1x_0$	0	0	$K_c((K_dM_c)^2M_c\ 3! + K_cK_dM_c^2M_d\ 2!)$
$x_0x_1^2$	0	0	$K_cK_dM_c^3\ 3!$
$x_1x_0^2$	0	0	$K_c((K_dM_c)^2M_c\ 3! + K_cK_dM_c^2M_d\ 2!)$
$x_1x_0x_1$	0	0	$K_cK_dM_c^3\ 3!$
$x_1^2x_0$	0	0	$K_cK_dM_c^3\ 3!$
x_1^3	$K_cM_c^3 3!$	$K_dM_d^3 3!$	$K_cM_c^3\ 3!$

(since the closed-loop system has, in effect, no external input) and Theorem 4.7, $F_{(c@d)_0}[u]$ will converge at least on the interval $[0, T_{\max})$, where

$$T_{\max} = \frac{1}{M_{(c@d)_0}} = \frac{1}{(K_{c\circ d}(2+\phi_g)+1)5M_{c\circ d}}.$$

Of course, if the series $(c@d)_0$ can be computed explicitly, a potentially better estimate $T'_{\max} = 1/M'_{(c@d)_0}$ is possible. For example, when $c\circ d = 1+x_1$, it is easily verified that $(c@d)_0 = \sum_{k\geq 0} x_0^k$ so that $F_{c@d}[0](t) = e^t$ for $t \geq 0$. In this case, both $T_{\max} = 0.04331$ and $T'_{\max} = 1$ are very conservative. On the other hand, when $c \circ d = 1 + 2x_1 + 2x_1^2$, it follows that $(c@d)_0 = \sum_{k\geq 0}(k+1)!\, x_0^k$ and $F_{c@d}[0](t) = 1/(1-t)^2$ for $0 \leq t < 1$. Here $T_{\max} = 0.02428$ is less conservative and $T'_{\max} = 1$ is exact.

Example 4.11. Reconsider the state space systems in Example 3.2. The operator $F_{c@d}[u]$ then has the analytic state space realization

$$f(z) = \begin{pmatrix} K_dM_cz_c^2z_d \\ K_cM_dz_cz_d^2 \end{pmatrix}, \quad g(z) = \begin{pmatrix} M_cz_c^2 \\ 0 \end{pmatrix}, \quad h(z) = K_cz_c$$

near $z(0) = [1\ \ 1]^T$. The first few coefficients of $c@d$ are given in Table 4.2. Since $c@d$ is a nonnegative series in this case, local convergence and input-output local convergence are equivalent as a consequence of Lemma 3.5. Setting $u(t) = \bar{u} = 1$ is equivalent to letting $b = 1$ in Theorem 4.8. Therefore, using again the $m = 0$ version of (16) and the growth condition from Theorem 4.8, a lower bound on the finite escape time for this system is

$$T_{\max} = \frac{1}{M_{(c@d)\circ 1}} = \frac{1}{[K_c(2+\phi_g)+1][\phi(1+K_d)+1]10M}.$$

Four specific cases of $T_{\max}$ are given in Table 4.3 and compared against the numerically determined escape times. The conservativeness in these estimates is a consequence of *accumulated* conservativeness in various intermediate upper bounds, for example inequality (30), as compared to the cascade connection in Example 3.2.

TABLE 4.3
$T_{\max}$ and t_{esc} for specific examples of $(c@d)\circ 1$.

Case	K_c	M_c	K_d	M_d	$M_{(c@d)\circ 1}$	$T_{\max}$	t_{esc}	$t_{\text{esc}}/T_{\max}$
1	4	2	2	2	1483	0.6745e−03	0.07556	112.0
2	2	4	2	2	1579	0.6335e−03	0.06606	104.3
3	2	2	4	2	1129	0.8857e−03	0.07387	83.4
4	2	2	2	4	1579	0.6335e−03	0.07556	119.3

REFERENCES

[1] M. ARAKI AND M. SAEKI, *A quantitative condition for the well-posedness of interconnected dynamical systems*, IEEE Trans. Automat. Control, 28 (1983), pp. 569–577.

[2] J. BERSTEL AND C. REUTENAUER, *Les Séries Rationnelles et Leurs Langages*, Springer-Verlag, Paris, 1984.

[3] J. CHAUMAT AND A.-M. CHOLLET, *On composite formal power series*, Trans. AMS, 353 (2001), pp. 1691–1703.

[4] B. CLOITRE, *private communication*, 2004.

[5] A. FERFERA, *Combinatoire du Monoïde Libre Appliquée à la Composition et aux Variations de Certaines Fonctionnelles Issues de la Théorie des Systèmes*, Doctoral dissertation, University of Bordeaux I, 1979.

[6] A. FERFERA, *Combinatoire du monoïde libre et composition de certains systèmes non linéaires*, Astérisque, 75/76 (1980), pp. 87–93.

[7] M. FLIESS, *Fonctionnelles causales non linéaires et indéterminées non commutatives*, Bull. Soc. Math. France, 109 (1981), pp. 3–40.

[8] M. FLIESS, *Développements fonctionnels et calcul symbolique non commutatif*, in Outils et Modèles Mathématiques pour L'Automatique L'Analyse de Systèmes et le Traitement du Signal, vol. 1, I. D. Landau, ed., Centre National de la Recherche Scientifique, Paris, 1981, pp. 359–377.

[9] M. FLIESS, *Réalisation locale des systèmes non linéaires, algèbres de Lie filtrées transitives et séries génératrices non commutatives*, Invent. Math., 71 (1983), pp. 521–537.

[10] M. FLIESS, M. LAMNABHI, AND F. LAMNABHI-LAGARRIGUE, *An algebraic approach to nonlinear functional expansions*, IEEE Trans. Circuits Systems, 30 (1983), pp. 554–570.

[11] X.-X. GAN AND D. KNOX, *On composition of formal power series*, Int. J. Math. Math. Sci., 30 (2002), pp. 761–770.

[12] W. S. GRAY AND B. NABET, *Volterra series analysis and synthesis of a neural network for velocity estimation*, IEEE Trans. Systems Man Cybernet. Part B, 29 (1999), pp. 190–197.

[13] W. S. GRAY AND Y. WANG, *Fliess operators on L_p spaces: Convergence and continuity*, Systems Control Lett., 46 (2002), pp. 67–74.

[14] U. HECKMANNS, *Aspects of Ultrametric Spaces*, Queen's Papers in Pure and Applied Math. 109, Queen's University, Kingston, ON, Canada, 1998.

[15] INRIA ALGORITHMS PROJECT, *Encyclopedia of Combinatorial Structures*, http://algo.inria.fr/encyclopedia/formulaire.html.

[16] A. ISIDORI, *Nonlinear Control Systems*, 3rd ed., Springer-Verlag, London, 1995.

[17] M. KAWSKI, *Calculus of nonlinear interconnections with applications*, in Proceedings of the 39th IEEE Conference on Decision and Control, Sydney, Australia, 2000, pp. 1661–1666.

[18] K. KNOPP, *Infinite Sequences and Series*, Dover Publications, New York, 1956.

[19] M. LAMNABHI, *A new symbolic calculus for the response of nonlinear systems*, Systems Control Lett., 2 (1982), pp. 154–162.

[20] Y. LI AND W. S. GRAY, *The formal Laplace–Borel transform, Fliess operators and the composition product*, in Proceedings of the 36th IEEE Southeastern Symposium on System Theory, Atlanta, GA, 2004, pp. 333–337.

[21] H. NIJMEIJER AND A. J. VAN DER SCHAFT, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, 1990.

[22] S. PRIESS-CRAMPE AND P. RIBENBOIM, *Fixed points, combs, and generalized power series*, Abh. Math. Sem. Univ. Hamburg, 63 (1993), pp. 227–244.

[23] S. PRIESS-CRAMPE AND P. RIBENBOIM, *Fixed point and attractor theorems for ultrametric spaces*, Forum Math., 12 (2000), pp. 53–64.

[24] E. SCHÖRNER, *Ultrametric fixed point theorems and applications*, in Valuation Theory and Its Applications, Vol. II, F.-V. Kuhlmann, S. Kuhlmann, and M. Marshall, eds., AMS, Providence, RI, 2003, pp. 353–359.

[25] N. J. A. Sloane, *The On-Line Encyclopedia of Integer Sequences*, http://www.research.att.com/∼njas/sequences.
[26] H. J. Sussmann, *Lie brackets and local controllability: A sufficient condition for scalar-input systems*, SIAM J. Control Optim., 21 (1983), pp. 686–713.
[27] M. Vidyasagar, *On the well-posedness of large-scale interconnected systems*, IEEE Trans. Automat. Control, 25 (1980), pp. 413–421.
[28] Y. Wang, *Algebraic Differential Equations and Nonlinear Control Systems*, Doctoral dissertation, Rutgers University, New Brunswick, NJ, 1990.
[29] Y. Wang and E. D. Sontag, *Generating series and nonlinear systems: Analytic aspects, local realizability, and I/O representations*, Forum Math., 4 (1992), pp. 299–322.
[30] Y. Wang and E. D. Sontag, *Algebraic differential equations and rational control systems*, SIAM J. Control Optim., 30 (1992), pp. 1126–1149.
[31] E. W. Weisstein et al., *Catalan Number*, MathWorld—A Wolfram Web Resource, http://mathworld.wolfram.com/CatalanNumber.html.
[32] H. S. Wilf, *Generating Functionology*, 2nd ed., Academic Press, San Diego, CA, 1994.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 2, pp. 673–703

A CONNECTION BETWEEN THE MAXIMUM PRINCIPLE AND DYNAMIC PROGRAMMING FOR CONSTRAINED CONTROL PROBLEMS*

AURELIAN CERNEA† AND HÉLÈNE FRANKOWSKA‡

Abstract. We consider the Mayer optimal control problem with dynamics given by a nonconvex differential inclusion, whose trajectories are constrained to a closed set and obtain necessary optimality conditions in the form of the maximum principle together with a relation between the costate and the value function. This additional relation is applied in turn to show that the maximum principle is nondegenerate. We also provide a sufficient condition for the normality of the maximum principle.

To derive these results we use convex linearizations of differential inclusions and convex linearizations of constraints along optimal trajectories. Then duality theory of convex analysis is applied to derive necessary conditions for optimality. In this way we extend the known relations between the maximum principle and dynamic programming from the unconstrained problems to the constrained case.

Key words. differential inclusions, nondegenerate maximum principle, dynamic programming, generalized derivatives, state constraints, variational inclusions

AMS subject classifications. 34A60, 49A24, 49J40, 49K24

DOI. 10.1137/S0363012903430585

1. Introduction.

Consider the following Mayer problem:

$$\text{minimize} \quad g(x(1)) \tag{1.1}$$

over the solutions to the differential inclusion

$$x'(t) \in F(t, x(t)) \quad \text{a.e. in } [0,1], \tag{1.2}$$

satisfying state constraints of the form

$$x(t) \in K \quad \forall\, t \in [0,1] \tag{1.3}$$

and end point constraints of the form

$$x(1) \in K_1, \tag{1.4}$$

$$x(0) = x_0, \tag{1.5}$$

*Received by the editors June 23, 2003; accepted for publication (in revised form) December 14, 2004; published electronically September 12, 2005. This work was supported in part by the European Community's Human Potential Programme under contract HPRN-CT-2002-00281, Evolution Equations.

http://www.siam.org/journals/sicon/44-2/43058.html

†Faculty of Mathematics and Informatics, University of Bucharest, Academiei 14, 010014 Bucharest, Romania (acernea@math.math.unibuc.ro). This author acknowledges the financial support provided through the European Community's Human Potential Programme under contract HPRNCT-2002-00281, Evolution Equations.

‡CNRS, CREA, École Polytechnique, 1, rue Descartes, 75005 Paris, France (franko@shs.polytechnique.fr).

where $K, K_1 \subseteq \mathbf{R}^n$ are closed sets, $g(\cdot) : \mathbf{R}^n \to \mathbf{R}$ is a given function and $F(\cdot,\cdot) : [0,1] \times \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$ is a given set-valued map and $x_0 \in K$. It is well known that the classical Bolza problem in control theory

$$\text{minimize} \quad \left\{ g(x(1)) + \int_0^1 L(t, x(t), x'(t))dt \mid x(\cdot) \text{ solves (1.2)–(1.5)} \right\}$$

is equivalent to the Mayer problem via a simple change of variables. This is the reason why we choose the apparently simpler problem (1.1)–(1.5), instead of a more general Bolza problem.

The value function associated to the problem (1.1)–(1.5) is defined by

(1.6)
$$V(t_0, y_0) = \inf\{g(x(1)) \mid \ x(\cdot) \text{ is a solution to (1.2)–(1.4) on } [t_0, 1],\ x(t_0) = y_0\},$$

with the convention $\inf \emptyset = +\infty$. The value function satisfies the dynamic programming principle. In particular, it is nondecreasing along trajectories to (1.2)–(1.4) and is constant along optimal trajectories. This property can be used to show that the value function is the only solution (in an appropriate sense) of an associated Hamilton–Jacobi–Bellman equation under state constraints (see, for instance, [19, 20]). Such solution is defined using subdifferentials of V. The aim of this paper is to study, using the dynamic programming principle, a sensitivity relation between necessary conditions for optimality and "gradients" of the value function.

Necessary optimality conditions for the problem (1.1)–(1.5) exist in the literature in the form of the maximum principle. When $F(t,x) = f(t,x,U)$ for some smooth enough f, i.e., when the differential inclusion (1.2) is replaced by the control system

$$x'(t) = f(t, x(t), u(t)), \ \ u(t) \in U \quad \text{a.e. in } [0,1],$$

this principle says that for every optimal trajectory/control pair $(\bar{x}, \bar{u})$ there exist $\lambda \in \{0,1\}$, a positive (scalar) Radon measure μ on $[0,1]$ and a μ-integrable function $\nu(\cdot) : [0,1] \to \mathbf{R}^n$, satisfying

$$\nu(t) \in N_K(\bar{x}(t)) \quad \mu - \text{a.e.},$$

(where $N_K(\bar{x}(t))$ denotes the normal cone to K at $\bar{x}(t)$) such that a solution $p(\cdot) : [0,1] \to \mathbf{R}^n$ to the adjoint system

$$-p'(t) = \frac{\partial f}{\partial x}(t, \bar{x}(t), \bar{u}(t))^* \left(p(t) + \int_{[0,t]} \nu(s) d\mu(s) \right),$$

$$-p(1) \in \lambda \nabla g(\bar{x}(1)) + \int_{[0,1]} \nu(s) d\mu(s) + N_{K_1}(\bar{x}(1)) \tag{1.7}$$

satisfies almost everywhere the maximum principle

$$\left\langle p(t) + \int_{[0,t]} \nu(s) d\mu(s), f(t, \bar{x}(t), \bar{u}(t)) \right\rangle = \max_{u \in U} \left\langle p(t) + \int_{[0,t]} \nu(s) d\mu(s), f(t, \bar{x}(t), u) \right\rangle$$

and $(\lambda, p, \mu) \neq 0$. When $\lambda = 0$, the above equalities are more related to constraint qualifications than to optimality and are sometimes called abnormal multiplier rule. When

$\bar{x}(0) \in \partial K$, then there exists a trivial choice of multipliers $0 \neq \zeta \in N_K(\bar{x}(0))$, $\nu = \zeta\delta_0$, $p = -\zeta$, $\lambda = 0$ which forces any feasible trajectory starting at $\bar{x}(0)$ to satisfy the maximum principle (here δ_0 denotes the unit measure concentrated at $\{0\}$.) In [1, 10, 13, 25, 29] the authors investigate the so-called nondegenerate maximum principle, but they still allow $\lambda = 0$. We would like to underline here that calmness of $V(0,\cdot)$ may be used to investigate normality of some maximum principles.

Let $D_x^+V(0,x_0)(\cdot)$ denote the upper directional derivative of $V(0,\cdot)$ at x_0 defined by

$$D_x^+V(0,x_0)(\theta) = \limsup_{s\to 0+,\theta'\to\theta,x_0+s\theta'\in K} \frac{V(0,x_0+s\theta') - V(0,x_0)}{s}.$$

From the results of this paper it follows in particular that if for all $t \in [0,1]$ the Clarke's tangent cones $C_K(\bar{x}(t))$ have nonempty interior, $C_{K_1}(\bar{x}(1)) \cap Int(C_K(\bar{x}(1))) \neq \emptyset$ and for some nonempty open convex subset $\mathcal{F} \subset C_K(\bar{x}(0))$ and $M > 0$

$$D_x^+V(0,\bar{x}(0))(\theta) \geq -M||\theta|| \ \ \forall\, \theta \in \mathcal{F},$$

then the maximum principle is nondegenerate in the sense that for $\psi(t) := \int_{[0,t]} \nu(s)d\mu(s)$ we have

$$\lambda + \sup_{t\in(0,1)} ||p(t) + \psi(t)|| \neq 0.$$

Moreover, if $\bar{x}(1) \in Int(K_1)$, then

$$\lambda + var(\psi,(0,1]) \neq 0,$$

where $var(\psi,(0,1])$ denotes the total variation of ψ on $(0,1]$. Both relations eliminate the above mentioned trivial multipliers for $\bar{x}(0) \in \partial K$. Actually we shall prove this result with a more general choice of tangents (see Corollary 3.7).

A sufficient condition, very similar to the Mangasarian–Fromowitz constraint qualification of mathematical programming, ensuring that $\lambda = 1$ is the existence of a solution w to the linearized along $(\bar{x},\bar{u})$ control system

$$w' = \frac{\partial f}{\partial x}(t,\bar{x}(t),\bar{u}(t))w + v(t), \quad v(t) \in T_{co(f(t,\bar{x}(t),U))}(\bar{x}'(t)), \tag{1.8}$$

satisfying

$$w(t) \in Int(C_K(\bar{x}(t))) \ \ \forall\, t \in [0,1],\ w(1) \in Int(C_{K_1}(\bar{x}(1))), \tag{1.9}$$

where $T_{co(f(t,\bar{x}(t),U))}(\bar{x}'(t))$ denotes the tangent cone of convex analysis to $co(f(t,\bar{x}(t),U))$ at $\bar{x}'(t)$. We shall provide such a condition even in the case of differential inclusions. This condition uses the trajectory/control pair $(\bar{x},\bar{u})$ and is not directly verifiable. Still it can be used to test for optimality a given trajectory/control pair. Furthermore, the cones $\{C_K(\bar{x}(t))\}_{t\in[0,1]}$ may be replaced by larger subsets satisfying Hypothesis 3.3 (see Example 2 in section 3).

In [28], in the context of nonsmooth control systems and smooth state constraints the authors proved that $\lambda = 1$ if an inward pointing condition (involving continuous selections from U) holds true on a neighborhood of the boundary of K and $K_1 = \mathbf{R}^n$. Here we get a similar result by asking a weaker (pointwise) inward pointing condition

just along the optimal trajectory (see Theorem 3.10). Namely we show that $\lambda = 1$ provided $\bar{x}(1) \in Int(K_1)$, for some $\eta > 0$ the signed distance

$$d(x) = \begin{cases} -dist(x, \partial K) & \forall\, x \in K \\ dist(x, \partial K) & \text{otherwise} \end{cases}$$

is of class $C^{1,1}_{loc}$ on $\partial K + \eta B$ and there exists $\rho > 0$ such that for almost all $t \in [0,1]$ with $\bar{x}(t) \in \partial K + \eta B$ we have $\min_{v \in F(t,\bar{x}(t))} \langle \nabla d(\bar{x}(t)), v \rangle \leq -\rho$.

We discuss next the sensitivity relation of the adjoint state $p(0)$ to the value function $V(0,\cdot)$. When there is no end point and state constraints, then $\lambda = 1$ and $\nu = 0$. If moreover $V(t,\cdot)$ is differentiable at $\bar{x}(t)$, then it is well known that

$$p(t) = -\frac{\partial V}{\partial x}(t, \bar{x}(t))$$

and this last relation can be used to get sufficient conditions for optimality (see [5]). Even when $K_1 = K = \mathbf{R}^n$, in general, $V(t,\cdot)$ is not differentiable. Still a relationship was obtained with the gradient of $V(t,\cdot)$ replaced by the superdifferential. Indeed, it was shown in [5, 16, 26, 36] that in addition the adjoint variable p is related to the value function in the following way:

$$-p(t) \in \partial_x^+ V(t, \bar{x}(t)) \;\; \forall\, t \in [0,1], \tag{1.10}$$

where $\partial_x^+ V(t, \bar{x}(t))$ denotes the superdifferential of $V(t,\cdot)$ at $\bar{x}(t)$ (see section 2 for the precise definition). A similar statement was previously obtained in [9] for control systems with $\partial_x^+ V(t, \bar{x}(t))$ replaced by Clarke's generalized gradient $\partial_x^C V(t, \bar{x}(t))$. We underline that, in general, $\partial_x^+ V(t, \bar{x}(t))$ is smaller than $\partial_x^C V(t, \bar{x}(t))$ and $\partial_x^+ V(t,x) = \frac{\partial V}{\partial x}(t,x)$ whenever $V(t,\cdot)$ is differentiable at x.

The aim of this paper is to derive a relation similar to (1.10) for the value function of the constrained problem (1.1)–(1.5) for $t = 0$. This extension is used in turn to investigate the nondegeneracy of the constrained maximum principle. Furthermore, we are dealing with a more general setting of nonconvex differential inclusions.

Necessary optimality conditions for systems given by differential inclusions in terms of adjoint inclusions associated to derivatives of the set-valued map F were obtained in [14, 15, 17, 27] for problems with endpoint constraints and in [6, 7] for problems under state constraints. We underline that such necessary optimality conditions, in general, are not equivalent to necessary conditions expressed in terms of generalized gradients of the Hamiltonian [1, 23, 33], the generalized Jacobians [34] or in terms of the limiting normal cones (sometimes called Euler–Lagrange necessary conditions) [35] (see [22] for several comparison results and examples). In [23] some necessary conditions involving a costate satisfying simultaneously the Euler–Lagrange and Hamiltonian conditions are proved for constrained convex valued differential inclusions. The maximum principle of the present paper is related to the Euler–Lagrange conditions, but we do not take limiting normals. Since the graph of $F(t,\cdot)$ is, in general, nonsmooth and we allow nonconvex values of F, our results are not contained in [23, 29]. In Example 1 provided in section 3 we show that a trajectory to an unconstrained system satisfies necessary conditions of [35], but in the same time does not fulfill our necessary conditions and so is not optimal. So in some situations our necessary conditions lead to a stronger discrimination between nonoptimal and optimal trajectories. Naturally there is a price to pay to have stronger necessary conditions. It consists in the assumption of existence of a "linearization" of F along

$(\bar{x}, \bar{x}')$ by closed convex processes, which are Lipschitz with respect to the state (see Hypothesis 3.2 in section 3). In the penalization approach to nonsmooth constrained problems based on variational principles (see, for instance, [1, 34, 35]) such assumption is not needed. When f is differentiable with respect to the state variable, the Jacobian of f (with respect to x) can be used to get a linearization. However, it is not the only instance when "linearizations" do exist. Example 1 concerns a nonsmooth control systems "linearized" by linear and convex processes; see also remarks following Hypothesis 3.2.

It is impossible in a short paper to provide the full overview, credits and bibliographies to the constrained maximum principle, because this topic was investigated by many authors since the early sixties. We refer to [23, 29, 33, 35] for extended discussions on the constrained maximum principle and further references and to [1, 10, 11, 25] for the Russian bibliography on the subject.

We derive a similar maximum principle for our problem (1.1)–(1.5) and also obtain a relation of $p(0)$ to $V(0, \cdot)$ (see Theorem 3.4). This relation implies in particular that $p(0) = 0$ whenever $\lambda = 0$. When $V(0, \cdot)$ exhibits some additional regularity (for instance, is Lipschitz on a neighborhood of $\bar{x}(0)$ in K), then

$$-p(0) \in \lambda \partial_x V(0, \bar{x}(0)), \tag{1.11}$$

where λ is the same as in (1.7) and $\partial_x V(0, \bar{x}(0))$ denotes a generalized supergradient of $V(0, \cdot)$ at $\bar{x}(0)$ defined by (2.5) in section 2. This supergradient is kind of regularized (in the terminology of [31]) superdifferential. When $V(0, \cdot)$ is locally Lipschitz around $\bar{x}(0) \in Int(K)$, $\partial_x V(0, \bar{x}(0))$ coincides with Clarke's generalized gradient of $V(0, \cdot)$ at $\bar{x}(0)$. In particular, for control systems, if there exists a solution to (1.8), (1.9), then $-p(0) \in \partial_x V(0, \bar{x}(0))$.

Even if (1.11) looks similar to (1.10), it is valid only at the initial point. In general, the relation (1.11) does not hold when $t \neq 0$ and needs a correction term involving measures (see Example 2 of section 3). In Theorem 3.11 (under some additional assumptions) we show how to correct $p(t)$ by an element $r(t)$ in order that

$$-r(t) - p(t) \in \partial_x V(t, \bar{x}(t)). \tag{1.12}$$

We also deduce few corollaries from our main theorem, Theorem 3.4. For instance, a nondegenerate maximum principle for (1.1)–(1.4) in the presence of initial point constraints $x(0) \in K_0$ for some closed set K_0 (Corollary 3.8), the sensitivity relation for the unconstrained case (Corollary 3.5) and a nondegenerate maximum principle which encompass calm problems (Corollary 3.7).

In the literature proofs of the maximum principle for systems under state constraints are based either on a penalization [1, 34, 35] or on an abstract multiplier rule [11, 21]. We proceed by using linearization of differential inclusion (1.2) along the optimal trajectory, as it was done in [15, 17], but then we also "linearize" constraints (along the optimal trajectory as well) and use the variational inclusions from [15] to get a Fermat type inequality. We derive next the maximum principle for "the linear problem" under conical constraints, that is a very simple application of separation theorems and convex duality results. Recently, variational inclusions under state constraints were obtained in [32] under additional assumptions on the boundary of constraints implying that the value function is Lipschitz. In [8] we used them to derive some preliminary results on this topic. In the present paper such Lipschitz regularity requirement is removed, an additional endpoint constraint is added and a new result on normality of the maximum principle is derived.

The paper is organized as follows. In section 2 we recall some notations and preliminary results to be used in what follows. In section 3 we present the main theorems of this paper. Section 4 is devoted to the study of polar of the set of continuous selections from a lower semicontinuous set-valued map and in section 5 we provide proofs of results of section 3.

2. Preliminaries. Denote by B the closed unit ball in $\mathbf{R}^n$. If $Q \subset \mathbf{R}^n$ we denote by $\overline{Q}$ its closure and by $co(Q)$ (resp. $\overline{co}(Q)$) the convex (resp. closed convex) hull of Q. We recall first the following definitions.

DEFINITION 2.1. *Let $K \subset \mathbf{R}^n$ be closed and $x \in K$.*

(i) the contingent cone *to K at x is defined by*

$$T_K(x) = \left\{ v \in \mathbf{R}^n | \quad \liminf_{s \to 0+} \frac{dist(x+sv, K)}{s} = 0 \right\}.$$

(ii) Clarke's tangent cone *to K at x is defined by*

$$C_K(x) = \left\{ v \in \mathbf{R}^n \mid \quad \lim_{s \to 0+, x' \to_K x} \frac{dist(x'+sv, K)}{s} = 0 \right\},$$

where $\to_K$ denotes the convergence in K.

Let Y be a real Banach space. Recall that a set $C \subset Y$ is called a cone if it is nonempty and for all $\lambda \geq 0$ and $v \in C$ we have $\lambda v \in C$. The negative polar cone of a set $Q \subset Y$ is defined by

$$Q^- = \{ y^* \in Y^* \mid \quad \langle y^*, y \rangle \leq 0 \quad \forall\, y \in Q \},$$

where Y^* denotes the dual of Y. The positive polar cone of Q is $Q^+ = -Q^-$. The negative polar of Clarke's tangent cone $N_K(x) := C_K(x)^-$ is also called the normal cone to the set K at $x \in K$.

Let $F : \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$ be a set-valued map. It is called *Lipschitz* around $x_0 \in \mathbf{R}^n$ if there exist $\varepsilon > 0,\ L \geq 0$ such that for any $x,\ y \in x_0 + \varepsilon B$, $F(x) \subset F(y) + L||x-y||B$, where $||\cdot||$ denotes the Euclidean norm on $\mathbf{R}^n$. Define

$$graph(F) := \{ (x, y) \in \mathbf{R}^n \times \mathbf{R}^n \mid y \in F(x) \}.$$

DEFINITION 2.2. *Consider a set-valued map $F : \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$, Lipschitz around x and let $y \in F(x)$. The adjacent derivative of F at (x, y) is the set-valued map $dF(x, y)$ from $\mathbf{R}^n$ into subsets of $\mathbf{R}^n$ defined by*

$$dF(x,y)w = \left\{ v \in \mathbf{R}^n \mid \quad \lim_{s \to 0+} dist\left(v, \frac{F(x+sw) - y}{s} \right) = 0 \right\}.$$

It is well known that $graph(dF(x, y))$ is equal to the adjacent tangent cone to $graph(F)$ at (x, y) (see [3]*).*

In this paper we use closed convex cones $\mathcal{A} \subset graph(dF(x, y))$. Each such convex cone defines a set-valued map $A : \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$ by $v \in A(u)$ if and only if $(u, v) \in \mathcal{A}$.

DEFINITION 2.3. *Let $A : \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$ be a set-valued map. A is called* closed *(resp.,* convex*) process if $graph(A)$ is a closed (resp., convex) cone. Its adjoint process $A^* : \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$ is defined by*

$$A^*(p) = \{ q \in \mathbf{R}^n \mid \langle q, u \rangle \leq \langle p, v \rangle \ \ \forall\ (u, v) \in graph(A) \},$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product in $\mathbf{R}^n$.

Notice that if A is Lipschitz on $\mathbf{R}^n$ with a Lipschitz constant m, then $\sup_{q\in A^*(p)} \|q\| \leq m\|p\|$. For other properties of closed convex processes we refer to [3].

Consider an extended real function $\phi(\cdot) : \mathbf{R}^n \to \mathbf{R} \cup \{\pm\infty\}$. We say that it is *positively homogeneous* if $\phi(0) > -\infty$ and for every $\lambda > 0$, $\theta \in \mathbf{R}^n$ we have $\phi(\lambda\theta) = \lambda\phi(\theta)$. We would like to underline that this definition differs slightly from the one given in [31, pp. 5 and 87], where the authors require $\phi(0) < \infty$. We need to change this notion, since in this paper we deal with functions whose hypographs are cones, while definitions in [31] are adapted to functions whose epigraphs are cones.

Consider a subset $X \subset \mathbf{R}^n$ and an extended real function $h(\cdot) : X \to \mathbf{R} \cup \{\pm\infty\}$. The domain of $h(\cdot)$ is $dom(h) = \{x \in X \mid \ h(x) \in \mathbf{R}\}$. When $h(\cdot)$ is not differentiable, it is possible to define its gradient by taking weaker limits of differential quotients.

DEFINITION 2.4. *Let $x_0 \in dom(h)$. The* superdifferential *of h at x_0 is the closed convex set*

$$\partial^+ h(x_0) := \left\{ p \in \mathbf{R}^n \mid \limsup_{x \to_X x_0} \frac{h(x) - h(x_0) - \langle p, x - x_0\rangle}{||x - x_0||} \leq 0 \right\}.$$

The subdifferential *of h at x_0 is the closed convex set defined by $\partial^- h(x_0) = -\partial^+(-h)(x_0)$.*

When $x_0 \in Int(dom(h))$ and $h(\cdot)$ is Fréchet differentiable at x_0, then $\partial^+ h(x_0) = \{\nabla h(x_0)\}$.

DEFINITION 2.5. *The upper derivative of $h(\cdot)$ at $x_0 \in dom(h)$ in the direction θ is given by*

$$D^+ h(x_0)(\theta) = \limsup_{s\to 0+,\ \theta'\to\theta,\ x_0+s\theta'\in X} \frac{h(x_0 + s\theta') - h(x_0)}{s} \quad \forall\ \theta \in T_X(x_0)$$

and $D^+ h(x_0)(\theta) = -\infty$ for all $\theta \notin T_X(x_0)$.

Notice that $D^+ h(x_0)(\cdot)$ is upper semicontinuous and positively homogeneous. It is known that $\partial^+ h(x_0) = \{p \in \mathbf{R}^n \mid \langle p, \theta\rangle \geq D^+ h(x_0)(\theta)$ for all $\theta \in T_X(x_0)\}$; see [16, Lemma 2.7]. If X is convex and h is concave, then $\partial^+ h(x_0)$ is equal to the supergradient of convex analysis, i.e., $p \in \partial^+ h(x_0)$ if and only if $h(x) \leq h(x_0) + \langle p, x - x_0\rangle$ for all $x \in X$.

In the proof of the results in the next sections we use the following consequence of the separation theorem.

LEMMA 2.6. *Let $h(\cdot) : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ be a positively homogeneous convex function and $C \subseteq \mathbf{R}^n$ be nonempty, convex, and such that for all $\lambda > 0$ and $v \in C$ we have $\lambda v \in C$. Assume that $dom(h) - C = \mathbf{R}^n$ and $h(c) \geq 0$ for all $c \in C$. Then there exists $q_0 \in C^+$ such that $\langle q_0, v\rangle \leq h(v)$ for all $v \in \mathbf{R}^n$.*

Proof. If $dom(h) = \emptyset$, then $h \equiv +\infty$ and the conclusion holds true with $q_0 = 0$. Assume next that $dom(h) \neq \emptyset$. Define $\mathbf{R}^*_- = \{r \in \mathbf{R} \mid r < 0\}$ and $Epi(h) = \{(u, v) \in \mathbf{R}^n \times \mathbf{R} \mid v \geq h(u)\}$. Since $Epi(h) \subset \mathbf{R}^n \times \mathbf{R}$ and $C \times \mathbf{R}^*_- \subset \mathbf{R}^n \times \mathbf{R}$ are convex sets and $(Epi(h)) \cap (C \times \mathbf{R}^*_-) = \emptyset$, there exists $(p, q) \in \mathbf{R}^n \times \mathbf{R}$, $(p, q) \neq 0$ such that

$$\langle p, u\rangle + qv \leq \langle p, c\rangle + qr \tag{2.1}$$

for any $(u, v) \in Epi(h)$ and any $(c, r) \in C \times \mathbf{R}^*_-$. Let $u_0 \in dom(h)$, $c_0 \in C$. If in (2.1) we take $u = \lambda_k u_0, v = \lambda_k h(u_0), r = r_k$ with $\lambda_k \to 0+$, $r_k \to 0-$ we deduce that $0 \leq \langle p, c\rangle$ for all $c \in C$. Hence $p \in C^+$. On the other hand, if in (2.1) we take $u = \lambda_k u_0$, $v = \lambda_k h(u_0) + \rho$, $c = \lambda_k c_0$, $r = r_k$ with $\rho \geq 0$, $\lambda_k \to 0+, r_k \to 0-$ we infer that $q\rho \leq 0$ for all $\rho \geq 0$. Thus $q \leq 0$. Let us assume for a moment that $q = 0$. Then from (2.1) we deduce that

$$p(u - c) \leq 0 \quad \forall\ u \in dom(h),\ c \in C. \tag{2.2}$$

Therefore, from (2.2) and our assumption it follows that $p = 0$, which leads to a contradiction with $(p, q) \neq 0$. Thus $q < 0$. From (2.1) we get

$$\left\langle \frac{p}{|q|}, u \right\rangle - h(u) \leq \left\langle \frac{p}{|q|}, \lambda_k c_0 \right\rangle - r_k \tag{2.3}$$

for any $u \in dom(h)$ and $\lambda_k \to 0+$, $r_k \to 0-$. Passing to the limit with $k \to \infty$ we find that $\langle \frac{p}{|q|}, u \rangle \leq h(u)$. It remains to set $q_0 = \frac{p}{|q|}$ and the lemma is proved. □

Results of the next sections use a lower version of the regular subderivative (see [31, p. 311]).

DEFINITION 2.7. *Let $X \subset \mathbf{R}^n$, $\phi : X \to \mathbf{R} \cup \{\pm\infty\}$ and $x \in dom(\phi)$. The regular superderivative is defined by $\hat{d}\phi(x)(u) = -\infty$ for all $u \notin C_X(x)$ and*

$$\hat{d}\phi(x)(u) = \lim_{\delta \to 0+} \left(\liminf_{x' \to_\phi x,\ s \to 0+} \sup_{u' \in u + \delta B, x' + su' \in X} \left[\frac{\phi(x' + su') - \phi(x')}{s} \right] \right) \ \forall\, u \in C_X(x), \tag{2.4}$$

where $\to_\phi$ denotes the $\phi-$attentive convergence introduced in [31, p. 301]. *This means that the lower limit is taken over all sequences $x_i \to x$, $s_i \to 0+$ such that $\phi(x_i) \to \phi(x)$.*

It follows from [31, pp. 311–313] that $\hat{d}\phi(x)(\cdot)$ is upper semicontinuous and positively homogeneous and if ϕ is locally upper semicontinuous at x, then $\hat{d}\phi(x)(\cdot)$ is concave. Thus, if in addition $dom(d\phi(x)) \neq \emptyset$, then $\hat{d}\phi(x)(\cdot)$ takes values in $[-\infty, +\infty)$ and $\hat{d}\phi(x)(0) = 0$.

When $x \in Int(X)$ and ϕ is locally Lipschitz at x, then Clarke's directional derivative is defined by

$$\phi^0(x)(u) = \limsup_{(x', s) \to (x, 0+)} \frac{\phi(x' + su) - \phi(x')}{s}, \quad u \in \mathbf{R}^n$$

and Clarke's generalized gradient is defined by

$$\partial^C \phi(x) := \{p \in \mathbf{R}^n \mid \quad \langle p, v \rangle \leq \phi^0(x)(v) \quad \forall\, v \in \mathbf{R}^n\}.$$

A generalized supergradient of ϕ at x that we shall use in this paper is defined using $\hat{d}\phi(x)(\cdot)$ by

$$\partial\phi(x) := \{p \in \mathbf{R}^n \mid \quad \langle p, v \rangle \geq \hat{d}\phi(x)(v) \quad \forall\, v \in C_X(x)\}. \tag{2.5}$$

By [31, Theorem 8.24, p. 317] if $\hat{d}\phi(x)(0) = 0$, then $\partial\phi(x) \neq \emptyset$. It is not difficult to show that if ϕ is locally Lipschitz on a neighborhood of $x \in Int(X)$, then $\partial\phi(x)$ coincides with Clarke's generalized gradient of ϕ at x.

Denote by I the interval $[0, 1]$, by $C(I)$ the space of all continuous functions $x(\cdot) : I \to \mathbf{R}^n$ and by $W^{1,p}(I)$ the space of all absolutely continuous functions $x(\cdot) : I \to \mathbf{R}^n$ such that $x'(\cdot) \in L^p(I)$, $p \geq 1$.

The space $NBV(I)$ (normalized bounded variations) is the space of functions f of bounded variation on I, which are continuous from the right on $(0, 1)$ and such that $f(0) = 0$. The norm of $f \in NBV(I)$ is the total variation of f on I denoted by $\|f\|_{TV}$.

Finally, by a solution of the differential inclusion

$$x'(t) \in F(t, x(t)) \tag{2.6}$$

we mean a function $x(\cdot) \in W^{1,1}(I)$ which satisfies (2.6) almost everywhere in I.

Remark. Let $f(\cdot,\cdot,\cdot) : I \times \mathbf{R}^n \times U \to \mathbf{R}^n$ be measurable in the first variable and continuous in the second and third variables. Set $F(t,x) := f(t,x,U)$. It is well known that if U is a complete separable metric space, then the set of solutions to the control system

$$x' = f(t,x,u(t)), \quad u(t) \in U, \quad u(\cdot) \text{ is measurable}$$

coincides with the set of solutions to the differential inclusion (2.6).

3. Main results. In this section we state several results announced in the introduction. Their proofs are provided in section 5.

Consider the Mayer problem (1.1)–(1.5). We assume the following hypotheses.

Hypothesis 3.1. (i) $F(\cdot,\cdot) : I \times \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$ is a set-valued map with nonempty closed values.

(ii) $\forall\, x \in \mathbf{R}^n,\ F(\cdot,x)$ is measurable.

(iii) There exists $c > 0$ such that $\forall\, (t,x) \in I \times \mathbf{R}^n,\ F(t,x) \subset c(1 + ||x||)B$.

(iv) There exists $l(\cdot) \in L^1(I,\mathbf{R})$ such that $F(t,\cdot)$ is $l(t)$-Lipschitz.

(v) $g : \mathbf{R}^n \to \mathbf{R}$ is locally Lipschitz.

Consider a solution $\bar{x}(\cdot)$ to (1.2). We wish to "linearize" F and K along $\bar{x}(\cdot)$. Denote by $\overline{co}F$ the set-valued map, whose value at (t,x) is the closed convex hull of $F(t,x)$, $\overline{co}F(t,x)$. Below we denote by $d_x\, \overline{co}F(t,\bar{x}(t),\bar{x}'(t))v$ the adjacent derivative of $\overline{co}F(t,\cdot)$ at $(\bar{x}(t),\bar{x}'(t))$.

Hypothesis 3.2. There exists a family of closed convex processes $A(t,\cdot) : \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$, $t \in I$, that satisfies

(i) $A(\cdot,v)$ is measurable $\forall\, v \in \mathbf{R}^n$.

(ii) $A(t,v) \subseteq d_x\, \overline{co}F(t,\bar{x}(t),\bar{x}'(t))v \ \ \forall\, v \in \mathbf{R}^n$, for a.e. $t \in I$.

(iii) For some $m \geq 0$, $A(t,\cdot)$ is m-Lipschitz on $\mathbf{R}^n$ for a.e. $t \in I$.

Notice that from (iii) it follows that for almost all $t \in I,\ A(t,\cdot)^*(0) = \{0\}$.

Remark.

(i) Assume that there exists a Carathéodory selection $f(t,x) \in F(t,x)$ such that for a.e. $t \in I$, $\bar{x}'(t) = f(t,\bar{x}(t))$ and $f(t,\cdot)$ is differentiable at $\bar{x}(t)$. If $||f'_x(\cdot,\bar{x}(\cdot))||_\infty < \infty$, then we may take $A(t,v) = f'_x(t,\bar{x}(t))v$ for all $v \in \mathbf{R}^n$.

(ii) If in Hypothesis 3.1 (iv) we have $l \in L^\infty(I)$ and $graph(\overline{co}F(t,\cdot))$ is sleek along $(\bar{x},\bar{x}')$ in the sense that $T_{graph(\overline{co}F(t,\cdot))}(\bar{x}(t),\bar{x}'(t)) = C_{graph(\overline{co}F(t,\cdot))}(\bar{x}(t),\bar{x}'(t))$ for almost all $t \in I$, then, by results of [3, Chapter 5], we may take

$$A(t,v) := d_x\, \overline{co}F(t,\bar{x}(t),\bar{x}'(t))v \quad \forall\, v \in \mathbf{R}^n.$$

Moreover, in this case $graph(A(t,\cdot))$ is equal to $C_{graph(\overline{co}F(t,\cdot))}(\bar{x}(t),\bar{x}'(t))$.

Example 1. Consider the nonsmooth control system

$$\begin{cases} x' = -|x| - 4y + u + v, & x(0) = 0 \\ y' = u, & y(0) = 0 \\ u,\ v \in [0,1] \end{cases}$$

and let $(\bar{x},\bar{y}) \equiv 0$.

As convex processes satisfying Hypothesis 3.2 one may take, for instance, linear process $A_\lambda(t,(w_1,w_2)) = (\lambda w_1 - 4w_2, 0)$ for any $\lambda \in [-1,1]$, or the convex process $A(t,(w_1,w_2)) = (|w_1| - 4w_2 + \mathbf{R}_+, 0)$.

Concerning the constraints K and K_1 we assume the following hypothesis.

Hypothesis 3.3. K and K_1 are closed subsets of $\mathbf{R}^n$, $Int(C_{K_1}(\bar{x}(1))) \neq \emptyset$ and there exists a lower semicontinuous set-valued map $G : I \rightsquigarrow \mathbf{R}^n$ such that for all $t \in I$, $G(t)$ is a closed convex cone with nonempty interior and for every $v \in Int(G(t))$ we can find $\varepsilon > 0$ such that for all $s \in [t-\varepsilon, t+\varepsilon] \cap I$, $\bar{x}(s) + [0,\varepsilon](v + \varepsilon B) \subset K$.

Remark. When for all $t \in I$, $Int(C_K(\bar{x}(t))) \neq \emptyset$, then we may set $G(t) = C_K(\bar{x}(t))$. Indeed by [2, Proposition 7.13], $v \in Int(C_K(\bar{x}(t)))$ if and only if there exists $\varepsilon > 0$ such that for all $x \in K \cap (\bar{x}(t) + \varepsilon B)$, $x + [0,\varepsilon](v + \varepsilon B) \subset K$. However, in general, there may exist sets $G(t)$ larger than Clarke's tangent cone and sets different from it; see Example 2.

Recall that V denotes the value function of problem (1.1)–(1.5) defined by (1.6). Let $D_x^+ V(0, \bar{x}(0))$, $\hat{d}_x V(0, \bar{x}(0))$ and $\partial_x V(0, \bar{x}(0))$ denote, respectively, the directional derivatives from Definitions 2.5 and 2.7, where we have set $X = K$ and the generalized supergradient of $V(0, \cdot)$ at $\bar{x}(0)$, defined by (2.5).

THEOREM 3.4. *Let $\bar{x}(\cdot)$ be an optimal solution to problem (1.1)–(1.5) and assume that Hypotheses 3.1, 3.2, and 3.3 hold true. Further assume that an upper semicontinuous concave positively homogeneous function $\varphi : \mathbf{R}^n \to \mathbf{R} \cup \{-\infty\}$ satisfies $Int(G(0)) \subset dom(\varphi)$ and $\varphi \leq D_x^+ V(0, \bar{x}(0))$.*

Then there exist $\lambda \in \{0, 1\}$, $\psi \in NBV(I)$ and an absolutely continuous function $p(\cdot) : I \to \mathbf{R}^n$ such that $\lambda + ||\psi||_{TV} \neq 0$ and p satisfies the adjoint inclusion

$$p'(t) \in A^*(t, -p(t) - \psi(t)) \quad \text{a.e. in } I, \tag{3.1}$$

the transversality condition

$$p(1) \in -\lambda \partial^C g(\bar{x}(1)) - \psi(1) - N_{K_1}(\bar{x}(1)), \tag{3.2}$$

the maximum principle

$$\langle p(t) + \psi(t), \bar{x}'(t) \rangle = \max_{v \in F(t, \bar{x}(t))} \langle p(t) + \psi(t), v \rangle \quad \text{a.e. in } I \tag{3.3}$$

and the sensitivity relation

$$-p(0) \in \lambda \partial^+ \varphi(0). \tag{3.4}$$

Furthermore,

$$\psi(0+) \in G(0)^-, \quad \psi(t) - \psi(t-) \in G(t)^-, \quad \psi(t) = \int_{[0,t]} \nu(s) d\mu(s) \quad \forall\, t \in (0, 1] \tag{3.5}$$

for a positive (scalar) Radon measure μ on I and a μ-measurable function $\nu(\cdot) : I \to \mathbf{R}^n$ satisfying

$$\nu(s) \in G(s)^- \cap B \quad \mu - a.e.$$

If $C_{K_1}(\bar{x}(1)) \cap Int(G(1)) \neq \emptyset$, then the following nondegeneracy condition holds true

$$\lambda + \sup_{t \in (0,1)} ||p(t) + \psi(t)|| \neq 0 \tag{3.6}$$

and if $\bar{x}(1) \in Int(K_1)$, then

$$\lambda + var(\psi, (0, 1]) \neq 0,$$

where $var(\psi, (0, 1])$ denotes the total variation of ψ on $(0, 1]$.

Moreover, $\lambda = 1$ if there exists a solution to the constrained differential inclusion

$$w' \in \overline{A(t,w) + T_{co(F(t,\bar{x}(t)))}(\bar{x}'(t))}, \quad w(1) \in Int(C_{K_1}(\bar{x}(1))), \quad w(t) \in Int(G(t)) \quad \forall\, t \in I.$$

In particular, if $V(0,\cdot)$ is locally upper semicontinuous at $\bar{x}(0)$ and $Int(G(0)) \subset dom(\hat{d}_x V(0,x_0))$, then (3.4) *may be replaced by*

$$-p(0) \in \lambda \partial_x V(0, \bar{x}(0)). \tag{3.4'}$$

Remark. Several remarks are in order.

(i) If $G(\cdot)$ satisfies Hypothesis 3.3, then for any closed convex cone with nonempty interior $Q \subset G(0)$, the set-valued map $\hat{G}$ defined by $\hat{G}(t) = G(t)$ for $t > 0$ and $\hat{G}(0) = Q$ also satisfies Hypothesis 3.3. This property will be used to prove Corollary 3.7.

(ii) If the subdifferential $\partial_x^- V(0,x_0)$ of $V(0,\cdot)$ at x_0 is nonempty, then any $\zeta \in \partial_x^- V(0,x_0)$ defines a function φ of Theorem 3.4 by $\varphi(\theta) = \langle \zeta, \theta \rangle$ for all $\theta \in G(0)$ and $\varphi(\theta) = -\infty$ otherwise. Then, for $\lambda = 1$, relation (3.4) implies $-p(0) \in \zeta + G(0)^+ \subset \partial_x^- V(0,x_0) + G(0)^+$, which can be seen as another sensitivity relation at points where $\partial_x^- V(0,x_0) \neq \emptyset$.

The subdifferentials of the value function were used in [19, 20] to define solutions to a Hamilton–Jacobi equation associated to the Mayer problem under state constraints.

(iii) The last statement (3.4′) holds true, in particular, when $V(0,\cdot)$ is Lipschitz on a neighborhood of $\bar{x}(0)$ in K and $Int(G(0)) \subset C_K(\bar{x}(0))$. In the case when $K_1 = \mathbf{R}^n$, the local Lipschitz continuity of $V(0,\cdot)$ can be deduced from the Lipschitz dependence of solutions to (1.2), (1.3) on the initial conditions. In [18] this issue was investigated for both smooth and nonsmooth sets of constraints K using the neighboring feasible trajectories theorem; see also [20] for the case of inequality constraints.

(iv) If in the above theorem we assume that $g(\cdot)$ is differentiable at $\bar{x}(1)$, then, by a very slight modification of the proof provided in section 5, inclusion (3.2) can be replaced by

$$p(1) \in -\lambda \nabla g(\bar{x}(1)) - \psi(1) - N_{K_1}(\bar{x}(1)).$$

Notice that, in general, $\{\nabla g(\bar{x}(1))\}$ may be not equal to $\partial^C g(\bar{x}(1))$; so this would lead to a stronger result.

(v) If $\bar{x}(\cdot)$ is optimal, then for all $t_0 \in I$, the restriction $\bar{x}|_{[t_0,1]}$ is optimal for the problem

$$\text{minimize} \quad g(x(1))$$

over the solutions to

$$x' \in F(t,x), \quad x(t_0) = \bar{x}(t_0), \quad x(t) \in K \quad \forall\, t \in [t_0,1], \quad x(1) \in K_1.$$

Thus the above theorem may be used to get the same conclusion on the time interval $[t_0,1]$ with a costate function $q(\cdot)$ and the inclusion

$$-q(t_0) \in \lambda \partial_x V(t_0, \bar{x}(t_0)).$$

Example 1 *(continuation).* Consider the unconstrained minimization problem

$$\text{minimize } x(1)$$

over solutions to the nonsmooth control system defined at the beginning of this section. We check that $(\bar{x}, \bar{y}) \equiv 0$ does not satisfy conclusions of Theorem 3.4 and so is not optimal. Indeed consider the convex process

$$A(t, (w_1, w_2)) = (w_1 - 4w_2, 0).$$

It satisfies Hypothesis 3.2. Then the adjoint system of Theorem 3.4 can be written as

$$\begin{cases} -p_1' = p_1, & p_1(1) = -1 \\ -p_2' = -4p_1, & p_2(1) = 0. \end{cases}$$

Then $p_1(t) = -e^{1-t},\ p_2(t) = 4e^{1-t} - 4$. Notice that

$$\max_{u,v\in[0,1]} \langle (p_1(t), p_2(t)), (u+v, u)\rangle = \max_{u\in[0,1]} (3e^{1-t} - 4)u.$$

When t is sufficiently small, this maximum is strictly positive. Consequently, $(\bar{x}, \bar{y}) \equiv 0$ does not verify the maximum principle of Theorem 3.4.

We show next that the result of [35] does not allow to eliminate the control $u \equiv 0$.

Indeed, set $F(x, y) = \bigcup_{u,v\in[0,1]}(-|x| - 4y + u + v, u)$ and let $N^L_{graph(F)}(0, 0, 0, 0)$ denote the limiting normal cone to $graph(F)$ at $(0, 0, 0, 0)$ (see, for instance, [33] for the definition of the limiting normal cone). It is not difficult to check that $\{(-1, 0, -1, 0), (1, 0, -1, 0)\} \subset N^L_{graph(F)}(0, 0, 0, 0)$. Let $p \equiv (-1, 0)$. Since $(0, 0) \in co\{(-1, 0), (1, 0)\}$, this p satisfies the necessary conditions of [35].

Example 2. Consider a two-dimensional control system

$$x' = u,\ y' = v + u - 1,\ \ (u, v) \in [-1, 1] \times [-1, 1],$$

the closed set of constraints

$$K = (\mathbf{R}_- \times \mathbf{R}_-) \cup (\mathbf{R}_+ \times \mathbf{R}),$$

and the Mayer problem

$$\text{minimize } -y(1)$$

over solutions to the above control system satisfying the state constraints $(x, y)(t) \in K$. Then, by a direct calculation, for all $(x_0, y_0) \in K$,

$$V(t, (x_0, y_0)) = \begin{cases} t - y_0 - 1 & \text{if } x_0 \geq 0 \text{ or } y_0 \leq x_0 \\ t - x_0 - 1 & \text{otherwise.} \end{cases}$$

Notice that $V(t, \cdot)$ is continuously differentiable on $\mathbf{R}^*_+ \times \mathbf{R}$, where $\mathbf{R}^*_+ = \mathbf{R}_+ \backslash \{0\}$. For the initial condition $x(0) = -1/2,\ y(0) = 0$ an optimal trajectory is given by

$$(\bar{x}(t), \bar{y}(t)) = \begin{cases} (t - 1/2, 0) & \text{if } t \leq 1/2 \\ \\ (t - 1/2, t - 1/2) & \text{if } t \geq 1/2. \end{cases}$$

Set

$$G(t) = \begin{cases} \mathbf{R} \times \mathbf{R}_- & \text{if } t < 1/2 \\ \{(x, y) \mid y \leq x,\ y \leq 0\} & \text{if } t = 1/2 \\ \mathbf{R} \times \mathbf{R} & \text{if } t > 1/2. \end{cases}$$

Then, for every $t_0 > 1/2$, $(0,1) = -\frac{\partial V}{\partial(x,y)}(t_0, \bar{x}(t_0))$. On the other hand, the optimal control $\bar{u}(\cdot)$ corresponding to $\bar{x}(\cdot)$ is equal to $(1,0)$ on $[0, \frac{1}{2}]$ and $G(t)^- = \mathbf{R}_+(0,1)$ for all $t \in [0, \frac{1}{2})$, $G(\frac{1}{2})^- = \{(x,y) \mid |x| \le y,\ x \le 0\}$ and $G(t)^- = \{0\}$ for all $t \in (\frac{1}{2}, 1]$. Let $\lambda,\ p,\ \psi$ be as in Theorem 3.4 for $A(t,\cdot) \equiv 0$ and $\varphi(\theta_1, \theta_2) = -\theta_1$ for $(\theta_1, \theta_2) \in \mathbf{R} \times \mathbf{R}_-$.

By (3.1) we have $p(\cdot) \equiv p(1)$ and, by (3.5), for some nonnegative nondecreasing function γ and all $t \in [0, 1/2)$, $\psi(t) = \gamma(t)(0,1)$. We also have $\psi(t) = \psi(\frac{1}{2})$ for all $t \in [\frac{1}{2}, 1]$ and

$$\psi\left(\frac{1}{2}\right) - \gamma\left(\frac{1}{2} -\right)(0,1) \in \{(x,y) \mid |x| \le y,\ x \le 0\}. \tag{3.7}$$

Define $w(t) = (0,-1) + (0,-1)t$. Then $w(\cdot)$ is a solution to the linearized along $(\bar{x}, \bar{u})$ control system and $w(t) \in Int(G(t))$ for all $t \in [0,1]$. Thus, by Theorem 3.4, $\lambda = 1$.

We first claim that $\psi(1) \neq 0$. Indeed otherwise, by (3.7), $\psi \equiv 0$ and $p = (0,1)$. But this contradicts the maximum principle (3.4) on $[0, \frac{1}{2})$.

Consequently, $\psi(1) \neq 0$ and for all $t_0 > 1/2$, $-p(t_0) = (0,1) + \psi(1) \neq \frac{\partial V}{\partial(x,y)}(t_0, \bar{x}(t_0))$. Hence for all $t \in (\frac{1}{2}, 1]$, $-p(t) \notin \partial_x V(t, \bar{x}(t))$.

To complete the study of this example we next show that the multipliers $p(\cdot),\ \psi(\cdot)$ are unique and $p(\cdot) \equiv (1,0)$,

$$\psi(s) = \begin{cases} 0 & s \in [0, \frac{1}{2}) \\ (-1,1) & s \in [\frac{1}{2}, 1], \end{cases}$$

while there are several choices for ν, μ. For instance, μ may be equal to the Lebesgue measure on $\Pi := [0,1] \backslash \{\frac{1}{2}\}$ and $\mu(\{\frac{1}{2}\}) = 1$ and $\nu(\cdot) = 0$ on $\Pi, \nu(\frac{1}{2}) = (-1,1)$. Another possible choice is the Dirac measure $\mu = \delta_{\frac{1}{2}}$ and any Borel measurable $\nu : [0,1] \to \mathbf{R}^2$ satisfying $\nu(\frac{1}{2}) = (-1,1)$.

We first compute ψ using the maximum principle. By (3.1) we have $p(\cdot) \equiv (0,1) - \psi(1)$ and, by (3.5), $\psi(t) = \psi(\frac{1}{2})$ for all $t \in [\frac{1}{2}, 1]$. Notice that (3.5) implies that for all $0 < t < \frac{1}{2}$ we have

$$p(t) + \psi(t) = (0,1) - \int_{(t, \frac{1}{2}]} \nu(s) d\mu(s) = (0,1) - \rho(t)(0,1) - (n_x, n_y)\mu\left(\left\{\frac{1}{2}\right\}\right) \tag{3.8}$$

for $\nu(s) = (\nu_1(s), \nu_2(s)) \in N_K(\bar{x}(s))$ μ-a.e. and $\nu_2(s) \ge 0$ μ-a.e., the nonincreasing function $\rho(t) := \int_{(t, \frac{1}{2})} \nu_2(s) d\mu(s) \ge 0$ and $(n_x, n_y) \in G(\frac{1}{2})^-$. Hence for a.e. $t \in [0, \frac{1}{2}]$,

$$\langle p(t) + \psi(t), (1,0) \rangle = -n_x \mu\left(\left\{\frac{1}{2}\right\}\right) = \max_{(u,v) \in [-1,1]^2} \langle p(t) + \psi(t), (u, u+v-1) \rangle . \tag{3.9}$$

Since ψ is right continuous on $(0, \frac{1}{2})$, (3.9) holds true for all $0 < t < \frac{1}{2}$. Then (3.8) and (3.9) together imply that for every $0 < t < \frac{1}{2}$, $1 - \int_{(t, \frac{1}{2}]} \nu_2(s) d\mu(s) = 1 - \rho(t) - n_y \mu(\{\frac{1}{2}\}) = 0$ and therefore

$$\rho(t) = 0 \ \ \& \ \ n_y \mu\left(\left\{\frac{1}{2}\right\}\right) = 1. \tag{3.10}$$

Set $\beta := n_x\mu(\{\frac{1}{2}\}) \leq 0$. Then $p(t) + \psi(t) = (-\beta, 0)$ and $\psi(t) = \psi(0+)\ \ \forall\, t \in (0, \frac{1}{2})$. By (3.5), $\psi(0+) \in \mathbf{R}_+(0,1)$. This implies that for some $\alpha \geq 0$ we have

$$\psi(s) = \begin{cases} 0 & s = 0 \\ \alpha(0,1) & s \in (0, \frac{1}{2}) \\ \alpha(0,1) + (\beta, 1) & s \in [\frac{1}{2}, 1]. \end{cases}$$

Notice that $p(\cdot) \equiv (0,1) - \psi(1)$ satisfies the maximum principle (3.3) with any choice of $\alpha \geq 0,\ \beta \leq 0$ and ψ defined as above. The next step is to show that (3.4) implies that $\alpha = 0$ and $\beta = -1$. Indeed, $-p(0) = \alpha(0,1) + (\beta, 1) - (0,1) = (\beta, \alpha) \in \partial\varphi(0)$ and $1 = \varphi(-1,0) \leq -\beta,\ -1 = \varphi(1,0) \leq \beta,\ 0 = \varphi(0,-1) \leq -\alpha$. Therefore $\beta = -1,\ \alpha = 0$. Finally, $p(\cdot) \equiv (0,1) - (-1,1) = (1,0)$.

Observe next that

$$-p(t) - \psi(t) = \begin{cases} (-1,0) & t \in [0, \frac{1}{2}) \\ (0,-1) & t \in [\frac{1}{2}, 1]. \end{cases}$$

Consider

$$\varphi_t(\theta) = \begin{cases} \langle (-1,0), \theta \rangle & t \in [0, \frac{1}{2}),\ \ \theta \in G(t) \\ \langle (0,-1), \theta \rangle & t \in [\frac{1}{2}, 1],\ \ \theta \in G(t) \\ -\infty & \text{otherwise.} \end{cases}$$

Then $\varphi_t \leq D_x^+ V(t, (\bar{x}(t), \bar{y}(t)))$ for all $t \in I$ and we obtained the inclusion $-p(t) - \psi(t) \in \partial^+ \varphi_t(0)$ for all $t \in I$.

The natural question arises: Should we expect, in general, such property along optimal trajectories for any choice of upper semicontinuous positively homogeneous concave function $\varphi_t \leq D_x^+ V(t, \bar{x}(t))$? We conjecture that without additional assumptions on constraints, this is false.

In the unconstrained case, however, by Remark (iv) right after Theorem 3.4, we have the inclusion $-p(t) \in \partial_x V(t, \bar{x}(t))$ for all $t \in I$ provided the costate is unique.

COROLLARY 3.5. *Let $\bar{x}(\cdot)$ be an optimal solution to problem* (1.1)–(1.5) *and assume that Hypotheses* 3.1, 3.2, *and* 3.3 *hold true and that $\bar{x}(I) \subset Int(K)$, $\bar{x}(1) \in Int(K_1)$. Then there exists an absolutely continuous function $p(\cdot) : I \to \mathbf{R}^n$ such that*

$$p'(t) \in A^*(t, -p(t)),\ \ \langle p(t), \bar{x}'(t) \rangle = \max_{v \in F(t, \bar{x}(t))} \langle p(t), v \rangle \quad \text{a.e. in } I,$$

$$p(1) \in -\partial^C g(\bar{x}(1)), \quad -p(0) \in \partial_x V(0, \bar{x}(0)).$$

If, in addition, $g(\cdot)$ is differentiable at $\bar{x}(1)$ and the solution to the adjoint inclusion

$$p'(t) \in A^*(t, -p(t)), \quad p(1) = -\nabla g(\bar{x}(1))$$

is unique, then for all $t \in I$, $-p(t) \in \partial_x V(t, \bar{x}(t))$.

When the value function is differentiable, we have a more precise statement than the one of Theorem 3.4.

THEOREM 3.6. *Let $\bar{x}(\cdot)$ be an optimal solution to problem* (1.1)–(1.5) *and assume that Hypotheses* 3.1, 3.2, *and* 3.3 *hold true. If $\bar{x}(0) \in Int(K)$ and $V(0, \cdot)$ is differentiable at $\bar{x}(0)$, then the same conclusions as in Theorem* 3.4 *hold true with* (3.4) *replaced by $p(0) = -\lambda \frac{\partial V}{\partial x}(0, \bar{x}(0))$.*

The next corollary implies that if $V(0,\cdot)$ has some directional derivatives bounded from below and $C_{K_1}(\bar{x}(1)) \cap Int(G(1)) \neq \emptyset$, then the maximum principle is nondegenerate.

COROLLARY 3.7. *Let $\bar{x}(\cdot)$ be an optimal solution to problem* (1.1)–(1.5). *Assume that Hypotheses* 3.1, 3.2, *and* 3.3 *are satisfied and that there exists a nonempty open convex set $\mathcal{F} \subset G(0)$ and $M > 0$ such that for all $\theta \in \mathcal{F}$*

$$D_x^+ V(0, \bar{x}(0))(\theta) \geq -M||\theta||. \tag{3.11}$$

Then we have the same conclusions as in Theorem 3.4 *with $G(0)$ replaced by $\overline{\mathbf{R}_+\mathcal{F}}$ and* (3.4) *replaced by*

$$-p(0) \in \lambda(MB + \mathcal{F}^+).$$

Proof. It is enough to set $\varphi(\theta) = -M||\theta||$ for all $\theta \in \overline{\mathbf{R}_+\mathcal{F}}$ and $\varphi(\theta) = -\infty$ otherwise and apply Theorem 3.4. □

Remark. Assume that our problem (1.1)–(1.5) is calm in the sense that

$$\liminf_{x \to_K \bar{x}(0)} \frac{V(0,x) - V(0,\bar{x}(0))}{||x - \bar{x}(0)||} > -\infty,$$

then (3.11) holds true for some $M > 0$ and all $\theta \in G(0)$.

Consider next the Mayer problem (1.1)–(1.4) with an initial constraint of the form

$$x(0) \in K_0, \tag{3.12}$$

where $K_0 \subset \mathbf{R}^n$ is a given closed set.

Then we have the following maximum principle for problem (1.1)–(1.4), (3.12) with state and both initial and endpoints constraints.

COROLLARY 3.8. *Let $\bar{x}(\cdot)$ be an optimal solution to problem* (1.1)–(1.4), (3.12) *and assume that Hypotheses* 3.1, 3.2, *and* 3.3 *are satisfied and $Int(C_{K_0}(\bar{x}(0))) \cap Int(G(0)) \neq \emptyset$. Then we have the same conclusions as in Theorem* 3.4 *with* (3.4) *replaced by*

$$p(0) \in \lambda(N_{K_0}(\bar{x}(0)) + G(0)^-). \tag{3.13}$$

Proof. Consider the set-valued map $\hat{G}(\cdot)$ defined by $\hat{G}(0) = C_{K_0}(\bar{x}(0)) \cap G(0)$ and $\hat{G}(t) = G(t)$ for all $t \in (0,1]$ and notice that $\hat{G}(\cdot)$ satisfies Hypothesis 3.3. It remains to apply Theorem 3.4 with $G(\cdot)$ replaced by $\hat{G}(\cdot)$ and with the mapping $\varphi(\cdot)$ defined by

$$\varphi(x) = \begin{cases} 0 & \text{if } x \in C_{K_0}(\bar{x}(0)) \cap G(0) \\ -\infty & \text{otherwise.} \end{cases}$$ □

Remark. In the case when we may take $G(t) \equiv \mathbf{R}^n$, if in the statement of Corollary 3.8 we define $q(t) := p(t) + \psi(0+)$, then we obtain the maximum principle with endpoint constraints from [15].

In what follows we show that when the boundary of K is smooth enough and a pointwise inward pointing assumption holds true along the optimal trajectory, then the maximum principle of Theorem 3.4 is normal, i.e., $\lambda = 1$ provided $\bar{x}(1) \in Int(K_1)$. A normal maximum principle (with a different adjoint inclusion and without assuming

Hypothesis 3.2) was derived in [28] in the context of nonsmooth control systems, smooth state constraints and $K_1 = \mathbf{R}^n$ under an inward pointing condition (involving continuous selections from U) imposed on a neighborhood of the boundary of K. We propose here a less restrictive inward pointing assumption.

Hypothesis 3.9. K is closed and

(i) $\exists\ \eta > 0$ such that the signed distance defined by

$$d(x) = \begin{cases} -dist(x, \partial K) & \forall\ x \in K \\ dist(x, \partial K) & \text{otherwise} \end{cases}$$

is of class $C^{1,1}_{loc}$ on $\partial K + \eta B$.

(ii) $\exists\ \rho > 0$ such that for almost all $t \in I$ with $\bar{x}(t) \in \partial K + \eta B$ we have

$$\min_{v \in F(t, \bar{x}(t))} \langle \nabla d(\bar{x}(t)), v \rangle \leq -\rho.$$

For all $x \in \partial K$ we denote by $n(x)$ the outward unit normal to K at x. Then $T_K(x) = (\mathbf{R}_+ n(x))^-$. Under Hypothesis 3.9, $n(x) = \nabla d(x)$ and therefore $n(\cdot)$ is locally Lipschitz on the boundary of K.

Remark. Assume that Hypothesis 3.1 holds true with l independent from time, that Hypothesis 3.9 (i) is satisfied and that there exists $\rho > 0$ such that for all $t \in I$ we have

$$\forall\ x \in \partial K, \quad \min_{v \in F(t,x)} \langle n(x), v \rangle \leq -\rho.$$

Then it is not difficult to show that Hypothesis 3.9 (ii) holds true with ρ replaced by $\rho/2$ and may be a different choice of a constant η. Furthermore, from results of [18] it follows that in this case the value function is locally Lipschitz on K.

THEOREM 3.10. *Let $\bar{x}(\cdot)$ be an optimal solution to problem* (1.1)–(1.5). *Assume that Hypotheses* 3.1, 3.2, *and* 3.9 *hold true and that $\bar{x}(1) \in Int(K_1)$. Then all conclusions of Theorem* 3.4 *are valid with $\lambda = 1$ and $G(t) = T_K(\bar{x}(t))$ for every $t \in I$.*

THEOREM 3.11. *In Theorem* 3.4 *assume that for all $t \in I$, $A(t, \cdot)$ is a linear operator and $A(t, \cdot)$ is $m(t)$-Lipschitz for some $m(\cdot) \in L^1(I)$. Then all the conclusions of Theorem* 3.4 *are valid.*

If there exists a solution to the constrained linear control system

$$w' = A(s, w) + v(s),\ v(s) \in T_{co(F(s, \bar{x}(s)))}(\bar{x}'(s)), \tag{3.14}$$

$$w(s) \in Int(G(s))\ \ \forall\ s \in [0, 1],\ w(1) \in Int(C_{K_1}(\bar{x}(1))) \tag{3.15}$$

and g is differentiable at $\bar{x}(1) \in Int(K_1)$, then $p(1) = -\nabla g(\bar{x}(1)) - \psi(1)$ and for every upper semicontinuous concave function $\varphi_t : \mathbf{R}^n \to \mathbf{R} \cup \{-\infty\}$ satisfying $Int(G(t)) \subset dom(\varphi_t)$ and $\varphi_t \leq D^+_x V(t, \bar{x}(t))$ there exist $\psi_t \in NBV([t, 1])$, a positive Radon measure μ_t, a selection $\nu_t(s) \in G(s)^- \cap B$ μ_t-a.e. such that for every $s \in (t, 1]$

$$\psi_t(s) = \int_{[t,s]} \nu_t(\tau) d\mu_t(\tau),\ \psi_t(s) - \psi_t(s-) \in G(s)^-,\ \psi_t(t+) \in G(t)^-$$

and the solution $r(\cdot) : [t, 1] \to \mathbf{R}^n$ to

$$r'(s) = A^*(s, -r(s) - \psi_t(s) + \psi(s)), \quad r(1) = \psi(1) - \psi_t(1)$$

satisfies

$$-r(t) - p(t) \in \partial^+ \varphi_t(0), \quad \langle r(s) + p(s) + \psi_t(s), \bar{x}'(s) \rangle$$
$$= \max_{v \in co(F(s,\bar{x}(s)))} \langle r(s) + p(s) + \psi_t(s), v \rangle \text{ a.e. in } [t,1].$$

Remark. The mapping r of the above theorem depends on t. That means for each $t > 0$ the costate $p(t)$ is corrected by a solution of the adjoint equation with a different function ψ_t in order $-r(t) - p(t) \in \partial^+ \varphi_t(0)$. Notice that if for all $s \geq t$, $\bar{x}(s) \in Int(K)$ and $\bar{x}(1) \in Int(K_1)$, then $\nu_t = 0$ and the above theorem implies that the solution q to

$$-q'(s) = A^*(s, q(s)), \quad q(1) = -\nabla g(\bar{x}(1))$$

satisfies $-q(t) \in \partial^+ \varphi_t(0)$.

On the other hand, slightly modifying the proof provided in section 5 we get $-q(s) \in \partial_x^+ V(s, \bar{x}(s))$ for all $s \in [t,1]$, which is a known relation of the costate to the value function derived in [5] for unconstrained problems.

4. Polar of the set of continuous selections. Recall that any function $f : [a,b] \to \mathbf{R}^n$ of bounded variation on $[a,b]$ has right and left limits $f(a+)$ and $f(b-)$ (see [12, p. 154]).

The space $NBV([a,b])$ (Normalized Bounded Variations) is the space of functions f of bounded variation on $[a,b]$, which are continuous from the right on (a,b) and such that $f(a) = 0$. The norm of $f \in NBV([a,b])$ is the total variation of f on $[a,b]$ denoted by $\|f\|_{TV}$. If $\beta \in C([a,b])^*$, then there exists a unique $f \in NBV([a,b])$ such that for all $\varphi \in C([a,b]), \beta(\varphi) = \int_a^b \varphi(s) df(s)$ (the Stieltjes integral) and $\|\beta\| = \|f\|_{TV}$ (see, for instance, [24, p. 113]). Conversely, every $f \in NBV([a,b])$ defines an element $\beta_f \in C([a,b])^*$ by setting $\langle \beta_f, \varphi \rangle = \int_a^b \varphi(s) df(s)$ for all $\varphi \in C([a,b])$.

Let $f \in NBV(I)$ and $\varphi \in W^{1,1}(I)$. We have the following integration by parts formula

$$\int_0^1 \varphi(s) df(s) = \langle f(1), \varphi(1) \rangle - \int_0^1 \langle f(s), \varphi'(s) \rangle ds.$$

Recall that any $g \in NBV(I)$ defines a regular finite countably additive measure on (I, Σ), where Σ denotes the σ-field of Borel subsets of I. We briefly recall the corresponding construction that will be used to prove Lemma 4.2 below.

Define $\bar{g}$ by $\bar{g} = g$ on $[0,1)$ and $\bar{g}(1) = 0$. For all $0 < c < d \leq 1$ set $\lambda_g((c,d]) = \bar{g}(d) - \bar{g}(c)$ and $\lambda_g([0,d])) = \bar{g}(d)$. By [12, pp. 141, 142] λ_g can be extended to a regular countably additive measure (again denoted) λ_g on Borel subsets of I such that $var(\lambda_g, (c,d]) = var(\bar{g}, (c,d])$ and $var(\lambda_g, [0,d]) = var(\bar{g}, [0,d])$, where var states for the variation. Define next the measure μ_g on Borel subsets of I by setting $\mu_g(\{0\}) = g(0+)$, $\mu_g(\{1\}) = g(1) - g(1-)$ and for every Borel subset A of $(0,1)$, $\mu_g(A) = \lambda_g(A)$. It is not difficult to verify that μ_g is a finite, countably additive and regular measure on I.

Notice that for every $0 < t < 1$,

$$\mu_g([0,t]) = \mu_g(\{0\}) + \lambda_g((0,t]) = g(0+) + \mu_g(\cup_{j \geq 1}((a_{j+1}, a_j])) = g(0+)$$
$$+ \bar{g}(t) - \bar{g}(0+) = g(t),$$

where $0 < \cdots < a_2 < a_1 = t$ and for all $0 < t < s < 1, \mu_g((t,s]) = g(s) - g(t)$. Furthermore,

$$\mu_g((t,1]) = \mu_g(\{1\}) + \lambda_g((t,1)) = g(1) - g(1-) + \mu_g(\cup_{j\geq 1}((a_j, a_{j+1}])),$$

where $t = a_1 < a_2 < \cdots < 1$. Thus, by countable additivity,

$$\mu_g((t,1]) = g(1) - g(1-) + g(1-) - g(t) = g(1) - g(t).$$

This and the definition of the Stieltjes integral imply that for every $w \in C(I)$,

$$\int_{[0,1]} w(s) d\mu_g(s) = \int_0^1 w(s) dg(s).$$

LEMMA 4.1. *Consider a lower semicontinuous set-valued map $G : I \rightsquigarrow \mathbf{R}^n$ such that for all $t \in I$, $G(t)$ is a closed convex cone. Assume that for all $t \in I$, $Int(G(t)) \neq \emptyset$ and let*

$$\mathcal{C} = \{w(\cdot) \in C(I) \mid \quad w(t) \in G(t) \quad \forall\, t \in I\}. \tag{4.1}$$

Then

$$Int(\mathcal{C}) = \{w(\cdot) \in C(I) \mid \quad w(t) \in Int(G(t)) \quad \forall\, t \in I\} \neq \emptyset.$$

Proof. Define the compact set $\mathcal{D} = \{t \in I \mid G(t) \neq \mathbf{R}^n\}$. If $\mathcal{D} = \emptyset$, then there is nothing to prove. Assume next that $\mathcal{D} \neq \emptyset$ and define $\Gamma(t) := \{p \in G(t)^- \mid ||p|| = 1\}$ for all $t \in \mathcal{D}$.

We claim that $\Gamma(\cdot)$ is upper semicontinuous on $\mathcal{D}$. Indeed, $\Gamma(\cdot)$ has nonempty compact images and is bounded, so it is enough to check that its graph is closed. Consider any $p_i \in \Gamma(t_i) \subset G(t_i)^-$, $p_i \to p$, $t_i \to_{\mathcal{D}} t_0$. Then $||p|| = 1$. Let $v \in G(t_0)$. By the lower semicontinuity of $G(\cdot)$ there exist $v_i \in G(t_i)$, $v_i \to v$. Hence $\langle p_i, v_i\rangle \leq 0$ and, taking the limit, we get $\langle p, v\rangle \leq 0$. Since $v \in G(t_0)$ is arbitrary, $p \in G(t_0)^-$ and our claim is proved.

Fix $\overline{\varepsilon} > 0$ and define for all $t \in \mathcal{D}$, $F_{\overline{\varepsilon}}(t) = \{v \in G(t) \mid \langle p, v\rangle \leq -\overline{\varepsilon} \ \ \forall\, p \in \Gamma(t)\}$. Since $Int(G(t)) \neq \emptyset$, also $F_{\overline{\varepsilon}}(t) \neq \emptyset\ \forall\ t \in \mathcal{D}$. Obviously, $F_{\overline{\varepsilon}}(\cdot)$ has closed convex images and is lower semicontinuous on $\mathcal{D}$. So by Michael's theorem there exists a continuous selection $\bar{f}(t) \in F_{\overline{\varepsilon}}(t) \subset Int(G(t))\ \forall\ t \in \mathcal{D}$. Let f be any continuous extension of $\bar{f}$ on the whole interval I. Then $f(t) \in Int(G(t))$ for all $t \in I$.

Consider any $g \in C(I)$ such that $||g||_\infty \leq \overline{\varepsilon}$. Then for any $t \in \mathcal{D}$ and $p \in \Gamma(t)$ $\langle p, f(t) + g(t)\rangle \leq -\overline{\varepsilon} + \overline{\varepsilon} = 0$, i.e., $f(t) + g(t) \in G(t)\ \ \forall\, t \in I$, implying that $Int(\mathcal{C}) \neq \emptyset$. Notice that if $\varphi \in Int(\mathcal{C})$, then for some $\varepsilon > 0$ and all $t \in I$, $B_\varepsilon(\varphi(t)) \subset G(t)$. Thus $Int(\mathcal{C}) \subset W := \{w(\cdot) \in C(I) \mid w(t) \in Int(G(t))\ \forall\ t \in I\}$. Next, fix any $\overline{\varphi} \in W$. Then, by the lower semicontinuity of $G(\cdot)$ and compactness of I, there exists $\varepsilon > 0$ such that for all $t \in I$, $\overline{\varphi}(t) + \varepsilon B \subset G(t)$. So, $\overline{\varphi} \in Int(\mathcal{C})$. ∎

In the next lemma we consider $NBV(I)$ as the dual of $C(I)$: with every $f \in NBV(I)$ we associate the functional $\beta_f \in C(I)^*$ defined by $\langle \beta_f, x\rangle = \int_0^1 x(s) df(s)$ (the Stieltjes integral).

LEMMA 4.2. *Consider a lower semicontinuous set-valued map $G : I \rightsquigarrow \mathbf{R}^n$ such that for all $t \in I$, $G(t)$ is a closed convex cone with nonempty interior. Let $\mathcal{C}$ be defined by* (4.1) *and $g \in NBV(I)$ be such that $g \in \mathcal{C}^-$. Then there exists a scalar*

positive Radon measure μ on I and a selection $\nu(s) \in G(s)^- \cap B$ $\mu - a.e.$ such that for every $t \in (0,1]$ $g(t) = \int_{[0,t]} \nu(s) d\mu(s)$, $g(t) - g(t-) \in G(t)^-$ and $g(0+) \in G(0)^-$.

Proof. Let μ_g be the regular countably additive finite measure on I associated to g by the construction recalled at the beginning of this section.

Define also for all $t > 0$, $f(t) = var(g, [0,t])$ (total variation of g on $[0,t]$) and set $f(0) = \|g(0+)\|$. Then $f \geq 0$ is increasing, right continuous on $[0,1)$ and has bounded variation. By the construction provided in [12, pp. 141–142], f defines a regular countably additive positive scalar measure μ on Borel subsets of $(0,1)$. Setting $\mu(\{0\}) = f(0)$, $\mu(\{1\}) = f(1)$ we obtain a Radon measure on I. Furthermore, since the total variation is additive, it follows from this construction that μ_g is μ-continuous. By the Radon–Nikodym theorem there exists a unique μ-integrable function ν such that $\int_E \nu(s) d\mu(s) = \mu_g(E)$ for every Borel set $E \subset I$. In particular, this yields $g(t) = \int_{[0,t]} \nu(s) d\mu(s)$ for all $0 < t \leq 1$. For every Borel subset $A \subset I$ denote by $v(\mu_g, A)$ the total variation of μ_g on A. It is not difficult to check that for every open set $\mathcal{O}$ in $[0,1]$ we have $\mu(\mathcal{O}) = v(\mu_g, \mathcal{O})$. Since μ_g and μ are regular also for every Borel subset $A \subset I$, $\mu(A) = v(\mu_g, A)$. □

Claim 1. $\nu(s) \in B$ μ-a.e. Define $\zeta : I \to B$ by $\zeta(s) = 0$ if $\nu(s) = 0$ and $\zeta(s) = \nu(s)/||\nu(s)||$ otherwise. Then for every Borel set $A \subset I$, $\int_A \zeta(s) d\mu_g(s) = \int_A \zeta(s)\nu(s) d\mu(s) = \int_A ||\nu(s)|| d\mu(s)$. Let A be the set of all $s \in I$ such that $||\nu(s)|| > 1$. If $\mu(A) > 0$ then from the last equality we get $\mu(A) < \int_A \zeta(s) d\mu_g(s) \leq v(\mu_g, A)$. The obtained contradiction proves our claim.

We next show that $\nu(s) \in G(s)^-$ *μ-a.e.* Even if it may be deduced from [30, Corollary 6A] we provide the proof of this inclusion for the sake of completeness.

Set $\Delta = \{t \in I \mid g(t) \neq g(t-)\}$. This set is at most countable.

Claim 2. Let $w_0 \in \mathbf{R}^n$ and $0 \leq t_1 < t_2 \leq 1$ be such that $w_0 \in G(s)$ for all $s \in [t_1, t_2]$ and $t_1 \notin \Delta$. We claim that $\langle g(t_2) - g(t_1), w_0 \rangle \leq 0$. Indeed, by Lemma 4.1, there exists a selection $\bar{w}(t) \in Int(G(t))$, $t \in I$. Then for every $\varepsilon > 0$, $\varepsilon \bar{w}(t) \in Int(G(t))$ $\forall\, t \in I$. For all $\delta > 0$ define $a_\delta := \max\{t_1 - \delta, 0\}$, $b_\delta = \min\{t_2 + \delta, 1\}$ and the function $w_\delta \in C(I)$ by

$$w_\delta(s) := \begin{cases} \frac{s - a_\delta}{\delta} w_0 + \frac{t_1 - s}{\delta} \varepsilon \bar{w}(s) & \text{if } s \in [a_\delta, t_1) \\ w_0 & \text{if } s \in [t_1, t_2] \\ \frac{b_\delta - s}{\delta} w_0 + \frac{s - t_2}{\delta} \varepsilon \bar{w}(s) & \text{if } s \in (t_2, b_\delta] \\ \varepsilon \bar{w}(s) & \text{otherwise.} \end{cases}$$

Then $\int_0^1 w_\delta(s) dg(s) \leq 0$ whenever $\delta > 0$ is sufficiently small. Notice that $\int_{t_1}^{t_2} w_0 dg(s) = \langle g(t_2) - g(t_1), w_0 \rangle$. On the other hand if $t_1 - \delta \geq 0$, then

$$\int_{t_1 - \delta}^{t_1} \frac{t_1 - \delta}{\delta} dg(s) = \frac{t_1 - \delta}{\delta} (g(t_1) - g(t_1 - \delta))$$

and, integrating by parts, we get

$$\int_{t_1 - \delta}^{t_1} \frac{s}{\delta} dg(s) = \frac{t_1}{\delta} g(t_1) - \frac{t_1 - \delta}{\delta} g(t_1 - \delta) - \frac{1}{\delta} \int_{t_1 - \delta}^{t_1} g(s) ds.$$

Finally, if $t_2 + \delta \leq 1$, then

$$\int_{(t_2, t_2 + \delta]} \frac{t_2 + \delta}{\delta} dg(s) = \frac{t_2 + \delta}{\delta} (g(t_2 + \delta) - g(t_2+)).$$

Integrating by parts we get

$$\int_{(t_2,t_2+\delta]} \frac{s}{\delta} dg(s) = \frac{t_2+\delta}{\delta} g(t_2+\delta) - \frac{t_2}{\delta} g(t_2+) - \frac{1}{\delta}\int_{t_2}^{t_2+\delta} g(s)ds.$$

Thus from the equality $g(t_2+) = g(t_2)$ we obtain

$$\int_{t_1-\delta}^{t_1} \frac{s+\delta-t_1}{\delta} dg(s) = g(t_1) - \frac{1}{\delta}\int_{t_1-\delta}^{t_1} g(s)ds,$$
$$\int_{(t_2,t_2+\delta]} \frac{\delta+t_2-s}{\delta} dg(s) = -g(t_2) + \frac{1}{\delta}\int_{t_2}^{t_2+\delta} g(s)ds.$$

Since $\lim_{s\to t_2+} g(s) = g(t_2)$, $\lim_{s\to t_1-} g(s) = g(t_1)$, from the inequality $\int_0^1 w_\delta(s)\, dg(s) \le 0$ we deduce that $\langle g(t_2) - g(t_1), w_0\rangle \le \varepsilon\|\bar{w}\|_\infty \|g\|_{TV}$. Passing to the limit when $\varepsilon \to 0+$, we end the proof of our claim.

To show that $\nu(s) \in G(s)^-$ μ-a.e it is enough to consider the case $\mu(I) \neq 0$. Let $\pi(s) \in G(s)$ be the projection of $\nu(s)$ on $G(s)$. Then π is μ-measurable and $||\pi(s)|| \le 1$ μ-a.e. Thus π is μ-integrable. Furthermore, $\pi(s) = 0$ if and only if $\nu(s) \in G(s)^-$. On the other hand, $\int_I \pi(s)d\mu_g(s) = \int_I \pi(s)\nu(s)d\mu(s) = \int_I ||\pi(s)||^2 d\mu(s)$. Hence if $\pi(s) \neq 0$ on a set of positive measure, then for some $\varepsilon > 0$, $\int_I \pi(s)d\mu_g(s) \ge 2\varepsilon$.

Let $\bar{w} \in Int(\mathcal{C})$. We may assume that $\|\bar{w}\|_\infty \le 1$. Let $\delta > 0$ be so that for all $s \in I$, $\bar{w}(s)+\delta B \subset G(s)$. Then $\pi(s)+\frac{\varepsilon}{\mu(I)}(\bar{w}(s)+\delta B) \subset G(s)$. Set $\zeta(s) := \pi(s)+\frac{\varepsilon}{\mu(I)}\bar{w}(s)$. Then $\int_I \zeta(s)d\mu_g(s) \ge \varepsilon$. Since ζ is μ-measurable, there exist μ-measurable functions $\zeta_i : I \to \mathbf{R}^n$ assuming only countable numbers of values and converging uniformly to ζ. Thus for all large i and all $s \in I$, $\zeta_i(s) \in Int(G(s))$. Consequently, for all large i, $\int_I \zeta_i(s)d\mu_g(s) > 0$. Fix i sufficiently large. Then for some $\alpha \in \mathbf{R}^n$ and a Borel set $A \subset I$, we have $\zeta_i(s) = \alpha$ for all $s \in A$ and $\alpha\mu_g(A) > 0$. Notice next that by the lower semicontinuity of $G(\cdot)$ and since $\alpha \in Int(G(s))$ for every $s \in A$ there exists an open set $\mathcal{O}_s \ni s$ such that $\alpha \in Int(G(s'))$ for all $s' \in \mathcal{O}_s$. Set $\mathcal{O}_2 := \bigcup_{s\in A} \mathcal{O}_s$. Since μ_g is regular, there exists an open (in I) set $\mathcal{O}_3 \supset A$ such $v(\mu_g, \mathcal{O}_3\backslash A) < \mu_g(A)/2$. Define the open set $\mathcal{O} := \mathcal{O}_2 \cap \mathcal{O}_3$.

Consequently, $\alpha\mu_g(\mathcal{O}) > \alpha\mu_g(A)/2 > 0$. Since $\mathcal{O}$ is at most a countable union of disjoint intervals $(t_1^i, t_2^i]$ and eventually an interval $[0, t_0]$ with $0 < t_1^i \notin \Delta$, $t_0 > 0$, either $\alpha\mu_g([0,t_0]) = \alpha g(t_0) > 0$ or there exist $0 < t_1 < t_2$ such that $\alpha\mu_g((t_1,t_2]) = \langle\alpha, g(t_2) - g(t_1)\rangle > 0$. In both cases we obtain a contradiction with Claim 2.

To prove jump conditions, observe that $G(t)^- \cap B = [0,1]\Gamma(t)$ for all $t \in \mathcal{D}$, where Γ and $\mathcal{D}$ are defined as in the proof of Lemma 4.1. Thus, by the proof of Lemma 4.1, $t \rightsquigarrow G(t)^- \cap B$ is upper semicontinuous on $\mathcal{D}$. Since $G(t)^- = \{0\}$ whenever $t \notin \mathcal{D}$, we deduce that the set-valued map $t \rightsquigarrow G(t)^- \cap B$ is upper semicontinuous on I. This implies that for all $\varepsilon > 0$ and for all small $t > 0$, $g(t) = \int_{[0,t]} \nu(s)d\mu(s) \in ((G(0)^- \cap B) + \varepsilon B)\mu([0,t])$. Taking the limit when $t \to 0+$ we get $g(0+) \in ((G(0)^- \cap B) + \varepsilon B)\lim_{t\to 0+}\mu([0,t])$. Since this inclusion is valid for all $\varepsilon > 0$ we deduce that $g(0+) \in G(0)^-$. Fix next any $0 < t \le 1$. Again, using the upper semicontinuity of $t \rightsquigarrow G(t)^- \cap B$, we obtain that for every $\varepsilon > 0$ and all $\tau < t$ sufficiently close to t we have $g(t) - g(\tau) = \int_{(\tau,t]} \nu(s)d\mu(s) \in ((G(t)^- \cap B) + \varepsilon B)\mu((\tau,t])$. Taking the limit, first when $\tau \to t-$ and then when $\varepsilon \to 0+$ we deduce that $g(t) - g(t-) \in G(t)^-$. The proof is complete.

5. Proofs of results of section 3. *Proof of Theorem* 3.4. Consider the set-valued map $B(\cdot,\cdot) : I \times \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$ defined by

$$B(t,v) = \overline{A(t,v) + T_{co(F(t,\bar{x}(t)))}(\bar{x}'(t))}.$$

By [15] for almost all $t \in I$, $B(t,\cdot)$ is a closed convex process, Lipschitz with the same Lipschitz constant as $A(t,\cdot)$ and $B(t,v) \subseteq d_x\, \overline{co}F(t,\bar{x}(t),\bar{x}'(t))v$ for all $v \in \mathbf{R}^n$. Thus the family $\{B(t,\cdot)\}_{t\in I}$ satisfies Hypothesis 3.2. It was also proved in [15] that

$$\begin{aligned} (5.1)\qquad & r(t) \in B^*(t,q(t)) \text{ iff } r(t) \in A^*(t,q(t)),\ q(t) \in [T_{co(F(t,\bar{x}(t)))}(\bar{x}'(t))]^+ \\ & = [F(t,\bar{x}(t)) - \bar{x}'(t)]^+. \end{aligned}$$

Set

$$S := \{w(\cdot) \in W^{1,2}(I) \mid \quad w'(t) \in B(t,w(t)) \quad \text{a.e. in } I\},$$

$$(5.2)\qquad \mathcal{C} = \{w(\cdot) \in C(I) \mid w(t) \in G(t) \quad \forall\, t \in I\}, \quad \mathcal{C}_1 = \{w(\cdot) \in C(I) \mid w(1) \in C_{K_1}(\bar{x}(1))\}$$

$$\gamma(x(\cdot)) = (x(0),x(1)), \quad \gamma_1(x(\cdot)) = x(1) \quad \forall\, x(\cdot) \in C(I).$$

Denote by $\bar{S}$ the closure of S in $C(I)$. By Lemma 4.1, $Int(\mathcal{C}) \neq \emptyset$. It is also clear that $Int(\mathcal{C}_1) \neq \emptyset$. Furthermore, since $\mathcal{C}_1 = \gamma_1^{-1}(C_{K_1}(\bar{x}(1)))$, $\mathcal{C}_1^+ = (\gamma_1^{-1}(C_{K_1}(\bar{x}(1))))^+ = \gamma_1^*((C_{K_1}(\bar{x}(1)))^+)$ (see, for instance, [4]), we infer that for every $\beta_1 \in \mathcal{C}_1^-$ there exists $\eta \in N_{K_1}(\bar{x}(1))$ such that all $w \in C(I)$, $\langle \beta_1, w\rangle = \langle \eta, w(1)\rangle$.

If $Int(\mathcal{C}_1) \cap Int(\mathcal{C}) = \emptyset$, then, by the separation theorem, there exists $0 \neq \beta_1 \in \mathcal{C}_1^-$ satisfying $-\beta_1 \in \mathcal{C}^-$. Let $\eta \in N_{K_1}(\bar{x}(1))$ be such that all $w \in C(I)$, $\langle \beta_1, w\rangle = \langle \eta, w(1)\rangle$. By Lemma 4.2 it is enough then to set $\psi(1) = -\eta$, $\psi = 0$ on $[0,1)$, $\lambda = 0$ and $p \equiv 0$ to get the conclusion of Theorem 3.4 in this case.

We assume next that $Int(\mathcal{C}_1) \cap Int(\mathcal{C}) \neq \emptyset$. Then $Int(\mathcal{C}_1 \cap \mathcal{C}) = Int(\mathcal{C}_1) \cap Int(\mathcal{C})$ and therefore $(\mathcal{C} \cap \mathcal{C}_1)^- = \mathcal{C}^- + \mathcal{C}_1^-$ (see, for instance, [4]).

We have two cases.

Case 1. $\bar{S} \cap (Int(\mathcal{C} \cap \mathcal{C}_1)) = \emptyset$. Since $\bar{S}$ and $\mathcal{C} \cap \mathcal{C}_1$ are closed convex cones in $C(I)$ and $Int(\mathcal{C} \cap \mathcal{C}_1) \neq \emptyset$, they can be separated by a closed hyperplane passing through the origin, i.e., there exists $0 \neq \beta \in C(I)^*$ such that

$$(5.3)\qquad \langle \beta, b\rangle \leq 0 \leq \langle \beta, a\rangle \quad \forall\, a \in \bar{S},\ b \in \mathcal{C} \cap \mathcal{C}_1.$$

Thus $\beta \in \mathcal{C}^- + \mathcal{C}_1^-$. Consider $\beta_0 \in \mathcal{C}^-$ and $\beta_1 \in \mathcal{C}_1^-$ such that $\beta = \beta_0 + \beta_1$.

We claim that $\beta_0 \neq 0$. Indeed otherwise $\beta_1 \neq 0$ and $\beta_1 \in \bar{S}^+$. Let $0 \neq \eta \in N_{K_1}(\bar{x}(1))$ be such that for all $w \in C(I)$, $\langle \beta_1, w\rangle = \langle \eta, w(1)\rangle$. Notice that for every $w_1 \in C_{K_1}(\bar{x}(1))$ there exists $w \in S$ such that $w(1) = w_1$. Hence, by (5.3), for all $w_1 \in C_{K_1}(\bar{x}(1))$ we have $\langle \eta, w_1\rangle = 0$. The interior of $C_{K_1}(\bar{x}(1))$ being nonempty, this last equality yields $\eta = 0$. The obtained contradiction proves our claim.

Let $\psi \in NBV(I)$ be such that for all $w \in C(I)$, $\langle \beta_0, w\rangle = \int_0^1 w(s)d\psi(s)$. Then $\|\psi\|_{TV} \neq 0$. By Lemma 4.2 applied to β_0 for a positive Radon measure μ on I and a μ-measurable selection $\nu(t) \in G(t)^- \cap B$ μ-a.e., relations (3.5) hold true.

On the other hand, by (5.3), we have that $\beta \in S^+ \subset W^{1,2}(I)^*$. Set

$$D(x(\cdot)) = x'(\cdot) \quad \forall\, x(\cdot) \in W^{1,2}(I),$$

$$L = \{(x(\cdot), y(\cdot)) \in L^2(I) \times L^2(I) \mid \quad y(s) \in B(s, x(s)) \ \text{ a.e. in } I\}.$$

According to [15] one has

$$S^+ = (1 \times D)^*(L^+), \tag{5.4}$$

$$L^+ = \{(-r(\cdot), q(\cdot)) \in L^2(I) \times L^2(I) \mid \quad r(s) \in B^*(s, q(s)) \text{ a.e. in } I\}. \tag{5.5}$$

Therefore, there exists $(-r, q) \in L^+$ such that for any $x \in W^{1,2}(I)$

$$\langle \beta, x\rangle = \langle (1 \times D)^*(-r, q), x\rangle = \langle (-r, q), (x, x')\rangle.$$

Let $\eta \in N_{K_1}(\bar{x}(1))$ be such that for all $w \in C(I)$, $\langle \beta_1, w\rangle = \langle \eta, w(1)\rangle$. Hence for any $x(\cdot) \in W^{1,2}(I)$

$$\langle \eta, x(1)\rangle = -\int_0^1 r(t)x(t)dt + \int_0^1 q(t)x'(t)dt - \int_0^1 x(t)d\psi(t). \tag{5.6}$$

From (5.6), integrating by parts, we get

$$\int_0^1 \left[q(t) + \int_0^t r(s)ds + \psi(t)\right] x'(t)dt - \left\langle \int_0^1 r(s)ds, x(1)\right\rangle - \langle \psi(1), x(1)\rangle - \langle \eta, x(1)\rangle = 0.$$

The above holds true, in particular, for all $x(\cdot) \in W^{1,2}(I)$ with $x(1) = 0$. Thus, from the Dubois–Raymond lemma it follows that there exists $r_0 \in \mathbf{R}^n$ such that

$$q(t) + \int_0^t r(s)ds + \psi(t) = r_0 \ \text{ a.e. in } I. \tag{5.7}$$

Define $p(t) := \int_0^t r(s)ds - r_0$. We have $q(t) = -p(t) - \psi(t)$ a.e. and from (5.5) and (5.1) we obtain (3.1) and (3.3). Using (5.7) we deduce that for any $x \in W^{1,2}(I)$

$$0 = \left\langle -\int_0^1 r(t)dt, x(1)\right\rangle + \langle r_0, x(1) - x(0)\rangle - \langle \psi(1), x(1)\rangle - \langle \eta, x(1)\rangle.$$

Applying this relation to all $x(\cdot) \in W^{1,2}(I)$ with $x(1) = 0$, we get $r_0 = 0$, $p(0) = 0$ and therefore $-\int_0^1 r(t)dt - \psi(1) - \eta = 0$. Consequently, $-p(1) - \psi(1) = \eta \in N_{K_1}(\bar{x}(1))$ and (3.2), (3.4) are satisfied with $\lambda = 0$.

Case 2. $\bar{S} \cap (Int(\mathcal{C} \cap \mathcal{C}_1)) \neq \emptyset$. Then $Int(\mathcal{C} \cap \mathcal{C}_1) = Int(\mathcal{C}) \cap Int(\mathcal{C}_1)$ and also $S \cap (Int(\mathcal{C} \cap \mathcal{C}_1)) \neq \emptyset$. Since $\mathcal{C} \cap \mathcal{C}_1$ and $\bar{S}$ are closed convex cones in $C(I)$ and $\bar{S} \cap (Int(\mathcal{C} \cap \mathcal{C}_1)) \neq \emptyset$ we infer that

$$(\bar{S} \cap \mathcal{C} \cap \mathcal{C}_1)^+ = \bar{S}^+ + \mathcal{C}^+ + \mathcal{C}_1^+ \tag{5.8}$$

(see, for instance, [4]). Let $\theta \in \mathbf{R}^n$ and consider a solution w to the differential inclusion

$$w' \in B(t, w), \ \ w(0) = \theta. \tag{5.9}$$

From the variational inclusion (Theorem 3.4 in [15]) we know that for all $s_i \to 0+$, $\theta_i \to \theta$ there exist solutions $x_i(\cdot)$ to (1.2) with $x_i(0) = \bar{x}(0) + s_i\theta_i$ such that $\frac{x_i(\cdot) - \bar{x}(\cdot)}{s_i}$ converge uniformly to $w(\cdot)$.

Define $h(\cdot,\cdot) : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$ by

$$h(\theta, v) = g^0(\bar{x}(1))(v) - \varphi(\theta).$$

Set $\mathcal{E} = \gamma(\bar{S} \cap Int(\mathcal{C} \cap \mathcal{C}_1))$. We claim that

$$h(\theta, v) \geq 0 \quad \forall\, (\theta, v) \in \mathcal{E}. \tag{5.10}$$

We first show that

$$h(\theta, v) \geq 0 \quad \forall\, (\theta, v) \in \gamma(S \cap Int(\mathcal{C} \cap \mathcal{C}_1)). \tag{5.11}$$

For this aim fix $w \in S \cap Int(\mathcal{C} \cap \mathcal{C}_1)$. Then, by Lemma 4.1, $w(t) \in Int(G(t))$ for all $t \in I$. From Hypothesis 3.3, continuity of w and compactness of I we deduce that for some $\varepsilon > 0$ and all $t \in I$, $\bar{x}(t) + [0, \varepsilon](w(t) + \varepsilon B) \subset K$.

Let $s_i \to 0+$, $\theta_i \to w(0)$ be such that $\bar{x}(0) + s_i\theta_i \in K$ and

$$D_x^+ V(0, \bar{x}(0))(w(0)) = \limsup_{i \to \infty} \frac{V(0, \bar{x}(0) + s_i\theta_i) - V(0, \bar{x}(0))}{s_i}$$

and x_i be solutions to (1.2) starting at $\bar{x}(0) + s_i\theta_i$ and such that $\frac{x_i(\cdot) - \bar{x}(\cdot)}{s_i}$ converge uniformly to $w(\cdot)$.

Then, by the dynamic programming principle,

$$\varphi(w(0)) \leq D_x^+ V(0, \bar{x}(0))(w(0)) \leq \limsup_{i \to \infty} \frac{g(x_i(1)) - g(\bar{x}(1))}{s_i} \leq g^0(\bar{x}(1))(w(1)).$$

Consequently, $h(w(0), w(1)) \geq 0$.

Consider next $w \in \bar{S} \cap Int(\mathcal{C} \cap \mathcal{C}_1)$ and $w_j \in S$ such that $\lim_{j \to \infty} w_j = w$ in $C(I)$. Then for all large j, $w_j \in S \cap Int(\mathcal{C} \cap \mathcal{C}_1)$ and, by (5.11),

$$h(w_j(0), w_j(1)) \geq 0.$$

Since $w(0) \in Int(G(0)) \subset dom(\varphi)$ and φ is concave, φ is continuous at $w(0)$. Thus passing to the limit in the last inequality when $j \to \infty$, we deduce (5.10).

Let $\overline{w}(\cdot) \in S \cap (Int(\mathcal{C} \cap \mathcal{C}_1))$. By Lemma 4.1 $\overline{w}(0) \in Int(G(0))$ and so $dom(h) - \mathcal{E} = \mathbf{R}^n \times \mathbf{R}^n$. We apply Lemma 2.6 to deduce that there exists $(a, c) \in \mathcal{E}^+$ such that

$$h(\theta, v) \geq \langle (a, c), (\theta, v) \rangle \quad \forall\, (\theta, v) \in \mathbf{R}^n \times \mathbf{R}^n. \tag{5.12}$$

From (5.12) and the definition of $h(\cdot,\cdot)$ we get $g^0(\bar{x}(1))(v) - \langle c, v \rangle - \langle a, \theta \rangle \geq \varphi(\theta)$ for every $(\theta, v) \in \mathbf{R}^n \times \mathbf{R}^n$. If $\theta = 0$ we obtain

$$c \in \partial^C g(\bar{x}(1)). \tag{5.13}$$

Consequently, for any $\theta \in \mathbf{R}^n$ we have $\varphi(\theta) \leq \langle -a, \theta \rangle$ and therefore

$$-a \in \partial^+ \varphi(0). \tag{5.14}$$

Since for every $w \in \bar{S} \cap Int(\mathcal{C} \cap \mathcal{C}_1)$, $\langle (a, c), \gamma(w) \rangle \geq 0$ we get $\gamma^*(a, c) \in (\bar{S} \cap Int(\mathcal{C} \cap \mathcal{C}_1))^+ = (\bar{S} \cap \mathcal{C} \cap \mathcal{C}_1)^+$. Hence, by (5.8), for some $\beta_0 \in \mathcal{C}^-$, $\beta_1 \in \mathcal{C}_1^-$, $\gamma^*(a, c) + \beta_0 + \beta_1 \in \bar{S}^+$. Denote by S^+ the positive polar of S in $W^{1,2}(I)^*$. Then $\gamma^*(a, c) + \beta_0 + \beta_1 \in S^+$. From (5.4), (5.5) it follows that there exist $(-r, q) \in L^+$ such that

$$\langle \gamma^*(a, c), x \rangle + \langle \beta_0 + \beta_1, x \rangle = \langle (1 \times D)^*(-r, q), x \rangle \ \ \forall\, x \in W^{1,2}(I)$$

or, equivalently, for any $x \in W^{1,2}(I)$

(5.15)
$$\langle \eta, x(1)\rangle + \langle a, x(0)\rangle + \langle c, x(1)\rangle = -\int_0^1 r(t)x(t)dt + \int_0^1 q(t)x'(t)dt - \int_0^1 x(t)d\psi(t)$$

for some $\psi \in NBV(I)$, $\eta \in N_{K_1}(\bar{x}(1))$ satisfying $\langle \beta_0, y\rangle = \int_0^1 y(s)d\psi(s)$ for all $y \in C(I)$ and $\langle \beta_1, w\rangle = \langle \eta, w(1)\rangle$ for all $w \in C(I)$.

We take $x(\cdot) \in W^{1,2}(I)$ such that $x(0) = x(1) = 0$ and we show as in Case 1 that there exists $r_0 \in \mathbf{R}^n$ that verifies (5.7). Define $p(t) = \int_0^t r(s)ds - r_0$. Thus $q(t) = -p(t) - \psi(t)$ a.e. in I and from (5.5) and (5.1) we deduce (3.1) and (3.3). Using again (5.15) and integrating by parts we obtain that for any $x \in W^{1,2}(I)$

$$\langle \eta, x(1)\rangle + \langle a, x(0)\rangle + \langle c, x(1)\rangle = -\langle r_0, x(0)\rangle + \langle -p(1) - \psi(1), x(1)\rangle.$$

We take first $x \in W^{1,2}(I)$ with $x(0) = 0$ and we find that $c + \eta = -p(1) - \psi(1)$ and, by (5.13), (3.2) holds true with $\lambda = 1$. From Lemma 4.2 we deduce that ψ satisfies (3.5) for some ν and μ as in the conclusions of our theorem.

If we take $x \in W^{1,2}(I)$ with $x(1) = 0$, then we get $a = -r_0 = p(0)$ and (3.4) follows from (5.14) with $\lambda = 1$.

We next prove (3.6). We already know that $\lambda + \|\psi\|_{TV} \neq 0$. Assume for a moment that $\lambda = 0$ and

$$\sup_{t\in(0,1)} ||p(t) + \psi(t)|| = 0. \tag{5.16}$$

Then $\|\psi\|_{TV} \neq 0$. From (3.1) and Hypothesis 3.2 we get $p'(t) = 0$ a.e. in I and, by (3.4), $p(\cdot) \equiv 0$. Therefore, by (3.2), $-\psi(1) \in N_{K_1}(\bar{x}(1))$ and, by (5.16), $\psi(\cdot) \equiv 0$ on $(0,1)$. Thus $\psi(1-) = 0$ and, via (3.5), we have that $\psi(1) \in G(1)^-$. Hence $\langle \psi(1), v_1\rangle \leq 0$ for all $v_1 \in G(1)$ and $\langle \psi(1), -v_2\rangle \leq 0$ for all $v_2 \in C_{K_1}(\bar{x}(1))$. This implies that $\psi(1) \in (G(1) - C_{K_1}(\bar{x}(1)))^-$. On the other hand, assumption $C_{K_1}(\bar{x}(1)) \cap Int(G(1)) \neq \emptyset$ yields $G(1) - C_{K_1}(\bar{x}(1)) = \mathbf{R}^n$ and we infer that $\psi(1) = 0$. Hence $\sup_{t\in I}||\psi(t)|| = 0$. This yields $\|\psi\|_{TV} = 0$ and so we derive a contradiction and (3.6) follows.

We have to check that when $\bar{x}(1) \in Int(K_1)$,

$$\lambda + var(\psi, (0,1]) \neq 0. \tag{5.17}$$

Indeed, assume for a moment that $\lambda + var(\psi, (0,1]) = 0$. Then $\|\psi\|_{TV} \neq 0$. By (3.2), $p(1) = -\psi(1) = -\psi(0+)$ and for all $t \in (0,1], \psi(t) = \psi(0+)$. Setting $q(t) := p(t) + \psi(0+)$ we deduce from (3.1) that $q'(t) \in A^*(t, -q(t))$ a.e. in I and $q(1) = 0$. But, by Hypothesis 3.2, $\|q'(t)\| \leq m\|q(t)\|$ for a.e. $t \in I$. This and the Gronwall inequality imply that $q(\cdot) \equiv 0$. By (3.4), $p(0) = 0$ and therefore $\psi \equiv 0$ contradicting to $\|\psi\|_{TV} \neq 0$. Inequality (5.17) is proved.

Let us assume next that there exists a solution $w(\cdot) \in W^{1,1}(I)$ to

$$\begin{aligned} w' &\in \overline{A(t,w) + T_{co(F(t,\bar{x}(t)))}(\bar{x}'(t))}, \quad w(1) \in Int(C_{K_1}(\bar{x}(1))), \\ w(t) &\in Int(G(t)) \quad \forall\, t \in I. \end{aligned}$$

We already know that there exist λ, p, ψ as in the statement of our theorem. Assume for a moment that $\lambda = 0$ and set $\eta = -p(1) - \psi(1) \in N_{K_1}(\bar{x}(1))$. Then, by

(3.4), $p(0) = 0$ and $\int_0^1 w(s)d\psi(s) = \int_{[0,1]} w(s)\nu(s)d\mu(s)$. Hence for all $w \in \mathcal{C} \cap \mathcal{C}_1$, $\int_{[0,1]} w(s)d\psi(s) + \langle \eta, w(1)\rangle \leq 0$. On the other hand, by the very definition of the adjoint process, for every $w \in S$ we have $\int_0^1 (p'w + pw' + \psi w')(s)ds \leq 0$. Thus $\langle p(1), w(1)\rangle + \int_0^1 \psi(s)w'(s)ds \leq 0$. Integrating by parts we deduce that $\langle \eta, w(1)\rangle + \int_0^1 w(s)d\psi(s) \geq 0$. The above yields $S \cap (Int(\mathcal{C} \cap \mathcal{C}_1)) = \emptyset$.

To deduce that $\lambda = 1$ it remains to verify that $S \cap (Int(\mathcal{C} \cap \mathcal{C}_1)) \neq \emptyset$. For every $i \geq 1$ define the measurable set $E_i = \{s \in I \mid ||w'(s)|| \leq i\}$. Then the Lebesgue measures $\mu_l(I \backslash E_i)$ converge to zero and $\lim_{i\to\infty} \int_{I\backslash E_i} ||w'(s)||ds = 0$. Consider $v_i(\cdot) \in L^\infty(I)$ defined by $v_i(s) = w'(s)$ for all $s \in E_i$ and $v_i(s) = 0$ otherwise and set $y_i(t) = w(0) + \int_0^t v_i(s)ds$ for all $t \in I$. Then y_i converge uniformly to w. Furthermore, for almost all $s \in E_i$, $dist(y_i'(s), B(s, y_i(s))) \leq m||y_i(s) - w(s)||$ and for almost all $s \in I\backslash E_i$, $dist(y_i'(s), B(s, y_i(s))) \leq m||y_i(s)||$. By the Filippov theorem (see, for instance, [3, p. 401]) there exist $M > 0$ and $w_i \in W^{1,1}(I)$ such that $w_i'(s) \in B(s, w_i(s))$ a.e. in I and

$$||w_i - y_i||_\infty \leq \varepsilon_i := M(||y_i - w||_\infty + \mu_l(I\backslash E_i)), \quad ||w_i'(t) - y_i'(t)|| \leq m\varepsilon_i + m||y_i(t)|| \text{ a.e. in } I.$$

Then $w_i \in W^{1,\infty}(I)$ and the sequence w_i converge uniformly to w. Hence for all large i, $w_i \in S \cap (Int(\mathcal{C} \cap \mathcal{C}_1))$.

Proof of Theorem 3.6. The proof follows by exactly the same arguments as the ones in the proof of Theorem 3.4. The only change is that in Case 2 it can be directly proved that for any $w(\cdot) \in \mathcal{E}$

$$\left\langle \frac{\partial V}{\partial x}(0, \bar{x}(0)), w(0) \right\rangle \leq g^0(\bar{x}(1))(w(1))$$

and then we apply Lemma 2.6 to the convex function $h(\theta, v) = g^0(\bar{x}(1))(v) - \langle \frac{\partial V}{\partial x}(0, \bar{x}(0)), \theta\rangle$.

Proof of Theorem 3.10.

It is not restrictive to assume that $\rho < 1$. Notice that $d(\bar{x}(t)) \leq 0$ for all $t \in I$. Set $G(t) = T_K(\bar{x}(t))$. We claim that $G(\cdot)$ satisfies Hypothesis 3.3. Indeed for all $t \in I$ such that $\bar{x}(t) \in \partial K$,

$$G(t) = \{v \in \mathbf{R}^n \mid \langle n(\bar{x}(t)), v\rangle \leq 0\}$$

and

$$Int\ (G(t)) = \{v \in \mathbf{R}^n \mid \langle n(\bar{x}(t)), v\rangle < 0\}.$$

Furthermore, $G(t) = \mathbf{R}^n$ whenever $\bar{x}(t) \in Int(K)$. Since $n(\cdot)$ is locally Lipschitz on the boundary of K, the set-valued map $I \ni t \rightsquigarrow Int\ (G(t))$ is lower semicontinuous on I. Thus also $G(\cdot)$ is lower semicontinuous on I. If $\bar{x}(t) \in Int\ (K)$, then for all s sufficiently close to t, $\bar{x}(s) \in Int\ (K)$ and therefore for every $v \in \mathbf{R}^n$ there exists $\epsilon > 0$ such that for all $s \in [t - \epsilon, t + \epsilon] \cap I$, $\bar{x}(s) + [0, \epsilon](v + \epsilon B) \subset K$.

Assume next that $t \in I$ is so that $\bar{x}(t) \in \partial K$ and fix $v \in Int\ (G(t))$. Then $\langle n(\bar{x}(t)), v\rangle < 0$. Notice that the oriented distance is Lipschitz with the Lipschitz constant one on η-neighborhood of ∂K.

By the mean value theorem for all $(s, h) \in I \times \mathbf{R}_+$ sufficiently close to $(t, 0)$ there exist $\theta(s, h) \in [0, 1]$ such that

$$d(\bar{x}(s) + hv) = d(\bar{x}(s)) + h\langle \nabla d(\bar{x}(s) + \theta(s, h)hv), v\rangle \leq h\langle \nabla d(\bar{x}(s) + \theta(s, h)hv), v\rangle.$$

By Hypothesis 3.9 (i) there exist $\delta > 0$ such that for all $s \in [t-\delta, t+\delta] \cap I$, $h \in [0,\delta]$, $\langle \nabla d(\bar{x}(s) + \theta(s,h)hv), v\rangle \leq -\delta$. Therefore for all $v' \in v + \delta B$, $d(\bar{x}(s) + hv') \leq d(\bar{x}(s) + hv) + h|v' - v| \leq -h\delta + h\delta = 0$. Consequently, $\bar{x}(s) + [0,\delta](v + \delta B) \subset K$. This proves our claim.

By Theorem 3.4 it is enough to show that there exists a solution to

$$w' \in \overline{A(t,w) + T_{co(F(t,\bar{x}(t)))}(\bar{x}'(t))}, \quad w(t) \in Int(G(t)) \quad \forall\, t \in I.$$

From [3, Theorem 9.5.3] it follows that there exist $L > 0$ and a selection $\gamma(t,w) \in A(t,w)$ such that for all $w \in \mathbf{R}^n$, $\gamma(\cdot, w)$ is measurable, $\gamma(t,\cdot)$ is L-Lipschitz and $\gamma(t,0) = 0$ for all $t \in I$. Thus $|\gamma(t,w)| \leq L|w|$.

Define $\Gamma := \{t \in I \mid \bar{x}(t) \in \partial K + \eta B\}$. By the measurable selection theorem (see, for instance, [3]) there exists a measurable selection $\Gamma \ni s \to v(s) \in F(s, \bar{x}(s))$ such that $\langle \nabla d(\bar{x}(s)), v(s)\rangle \leq -\rho$ for almost all $s \in \Gamma$. We extend v on I by setting $v(s) = \bar{x}'(s)$ for all $s \notin \Gamma$.

Then $\mathbf{R}_+(v(s) - \bar{x}'(s)) \in T_{co(F(s,\bar{x}(s)))}(\bar{x}'(s))$. We prove that there exists a solution w to

$$w' \in \gamma(s,w) + \mathbf{R}_+(v(s) - \bar{x}'(s)) \tag{5.18}$$

such that $w(s) \in Int\ (G(s))$ for all $s \in I$.

If for all $s \in I$, $\bar{x}(s) \in Int(K)$, then any solution to $w' = \gamma(s,w)$ satisfies $w(s) \in Int\ (G(s))$. Assume next that $\bar{x}(I) \cap \partial K \neq \emptyset$ and set $s_0 = \inf\{s \in I \mid \bar{x}(s) \in \partial K\}$. Let $\bar{w}_0 \in Int\ (G(s_0))$ and consider the solution $\bar{w}$ to

$$w'(s) \in \gamma(s,w),\ s \in [0,s_0],\ \bar{w}(s_0) = \bar{w}_0.$$

Then $\bar{w}(s) \in Int\ (G(s))$ for all $s \in [0,s_0]$.

Denote by k a Lipschitz constant of $\nabla d(\bar{x}(\cdot))$ on $\{t \mid \bar{x}(t) \in \partial K + \eta B\}$ and set

$$M = \sup_{s\in I}\ \sup_{v\in F(s,\bar{x}(s))} ||v||, \quad \chi = max\left\{\frac{1}{\eta}, \frac{k+L+1}{\rho}\left(L + 2M\frac{k+L+1}{\rho}\right)\right\}, \quad \delta = \frac{1}{2M\chi}.$$

Then $\eta \geq 1/\chi$. We claim that it is enough to show that for any $s_0 \leq t_0 < 1$ such that $\bar{x}(t_0) \in \partial K$ and any $w_0 \in Int(G(t_0))$ there exist $t_1 \geq t_0 + \min\{1 - t_0, \delta\}$ and a solution to (5.18) defined on $[t_0, t_1]$ with $w(t_0) = w_0$ and $w(s) \in Int\ (G(s))$ for all $s \in [t_0, t_1]$ and either $t_1 = 1$ or $\bar{x}(t_1) \in \partial K$. Indeed if we prove this property, then we can extend $\bar{w}(\cdot)$ on the time interval $[s_0, 1]$ in a finite number of steps.

So fix $t_0 \geq s_0$ such that $\bar{x}(t_0) \in \partial K$ and $w_0 \in Int(G(t_0))$. Then $w_0 \neq 0$. Define

$$t_2 = \max\left\{s \in [t_0,1] \mid \bar{x}([t_0,s]) \subset \partial K + \frac{1}{2\chi}B\right\} > t_0.$$

Then either $t_2 = 1$ or $d(\bar{x}(t_2)) = -\frac{1}{2\chi}$. In this second case

$$-\frac{1}{2\chi} = d(\bar{x}(t_2)) = d(\bar{x}(t_0)) + \int_{t_0}^{t_2} \langle \nabla d(\bar{x}(s)), \bar{x}'(s)\rangle ds.$$

Thus $\frac{1}{2\chi} \leq M(t_2 - t_0)$ and so $t_2 - t_0 \geq 1/2M\chi = \delta$. Consider the solution w to

$$w' = \gamma(s,w) + \frac{k+L+1}{\rho}||w||(v(s) - \bar{x}'(s)), \quad w(t_0) = w_0$$

and set $\xi(s) = \nabla d(\bar{x}(s))$. Notice that $w(t_2) \neq 0$. Fix any $t \in [t_0, t_2]$ with $\bar{x}(t) \in \partial K$. We have to check that $w(t) \in Int\ (G(t))$. By the choice of $v(\cdot)$

$$\langle \xi(t), w(t)\rangle = \langle \xi(t_0), w_0\rangle + \int_{t_0}^{t} \langle \xi, w\rangle'(s)ds \leq \int_{t_0}^{t} ||\xi'(s)|| \cdot ||w(s)||ds + \int_{t_0}^{t} \langle \xi(s), w'(s)\rangle ds$$

$$\leq (k+L)\int_{t_0}^{t} ||w(s)||ds + \frac{k+L+1}{\rho}\int_{t_0}^{t} \langle \nabla d(\bar{x}(s)), v(s)\rangle ||w(s)||ds -$$

$$\frac{k+L+1}{\rho}\int_{t_0}^{t} d(\bar{x})'(s)||w(s)||ds \leq -\int_{t_0}^{t} ||w(s)||ds - \frac{k+L+1}{\rho}\int_{t_0}^{t} d(\bar{x})'(s)||w(s)||ds.$$

On the other hand, integrating by parts, we obtain

$$-\int_{t_0}^{t} d(\bar{x})'(s)||w(s)||ds = -d(\bar{x}(t))||w(t)|| + d(\bar{x}(t_0))||w(t_0)|| + \int_{t_0}^{t} d(\bar{x}(s))||w||'(s)ds$$

$$\leq \int_{t_0}^{t} |d(\bar{x}(s))| \cdot ||w'(s)||ds \leq \int_{t_0}^{t} \left(L + 2M\frac{k+L+1}{\rho}\right) |d(\bar{x}(s))| \cdot ||w(s)||ds.$$

Consequently,

$$-\frac{k+L+1}{\rho}\int_{t_0}^{t} d(\bar{x})'(s)||w(s)||ds \leq \chi \int_{t_0}^{t} |d(\bar{x}(s))| \cdot ||w(s)||ds \leq \frac{1}{2}\int_{t_0}^{t} ||w(s)||ds.$$

This implies that

$$\langle \xi(t), w(t)\rangle \leq -\int_{t_0}^{t} ||w(s)||ds + \frac{1}{2}\int_{t_0}^{t} ||w(s)||ds = -\frac{1}{2}\int_{t_0}^{t} ||w(s)||ds < 0.$$

So we defined $w(\cdot)$ on $[t_0, t_2]$ in such a way that $w(s) \in Int(G(s))$ for all $s \in [t_0, t_2]$. If $t_2 = 1$ or $\bar{x}(t_2) \in \partial K$, then $w(\cdot)$ is as required. Assume next that $t_2 < 1$ and $\bar{x}(t_2) \notin \partial K$. Set $w_2 = w(t_2)$. If for all $s > t_2$, $\bar{x}(s) \in Int(K)$, then w can be extended on $[t_2, 1]$ by the solution to $w' = \gamma(s, w),\ w(t_2) = w_2$. So we obtain a solution w to (5.18) satisfying $w(s) \in Int(G(s))$ for all $s \in [t_0, 1]$. It remains to consider the case when for some $t_2 < s \leq 1$, $\bar{x}(s) \in \partial K$. Define $t_1 = \min\{s > t_2 \mid \bar{x}(s) \in \partial K\}$ and let

$$t_3 = \max\left\{s \in [t_2, t_1] \mid\ |d(\bar{x}(\tau))| = \frac{1}{2\chi}\right\} < t_1.$$

Then for all $s \in [t_3, t_1]$, $|d(\bar{x}(s))| \leq \rho/4M$.

We extend $w(\cdot)$ to the time interval $[t_2, t_3]$ by the solution to $w' = \gamma(s, w),\ w(t_2) = w_2$. Then $w(s) \in Int\ (G(s))$ for all $s \in [t_2, t_3]$. Set $w_3 = w(t_3) \neq 0$ and extend w on the time interval $[t_3, t_1]$ by the solution to

$$w' = \gamma(s, w) + \frac{k+L+1}{\rho}||w||(v(s) - \bar{x}'(s)) + \frac{2||w_3||}{\rho(t_1 - t_3)}(v(s) - \bar{x}'(s)),\ w(t_3) = w_3.$$

Then $w(s) \in Int\ (G(s)) = \mathbf{R}^n$ for all $s \in [t_3, t_1)$. It remains to check that $w(t_1) \in Int(G(t_1))$.

As before, by the choice of $v(\cdot)$,

$$\langle \xi(t_1), w(t_1)\rangle \leq ||w_3|| + \int_{t_3}^{t_1} \langle \xi, w\rangle'(s)ds \leq ||w_3|| + \int_{t_3}^{t_1} k||w(s)||ds + \int_{t_3}^{t_1} \langle \xi(s), w'(s)\rangle ds$$

$$\leq ||w_3|| - \int_{t_3}^{t_1} ||w(s)||ds + \frac{2||w_3||}{\rho(t_1-t_3)} \int_{t_3}^{t_1} \langle \nabla d(\bar{x}(s)), v(s)\rangle ds - \frac{k+L+1}{\rho} \int_{t_3}^{t_1} d(\bar{x})'(s)||w(s)||ds$$

$$-\frac{2||w_3||}{\rho(t_1-t_3)} \int_{t_3}^{t_1} d(\bar{x})'(s)ds \leq -\int_{t_3}^{t_1} ||w(s)||ds + \frac{k+L+1}{\rho} \int_{t_3}^{t_1} d(\bar{x})(s)||w(s)||'ds - ||w_3|| \leq$$

$$-\int_{t_3}^{t_1} ||w(s)||ds + \chi \int_{t_3}^{t_1} |d(\bar{x}(s))| \cdot ||w(s)||ds - ||w_3|| + \frac{4M||w_3||}{\rho(t_1-t_3)} \int_{t_3}^{t_1} |d(\bar{x}(s))|ds \leq$$

$$-\frac{1}{2} \int_{t_3}^{t_1} ||w(s)||ds - ||w_3|| + ||w_3|| < 0.$$

This yields $w(t_1) \in Int(G(t_1))$. The proof is complete.

Proof of Theorem 3.11. In the difference with Theorem 3.4 we have $m(\cdot) \in L^1(I)$ instead of $L^\infty(I)$. For this reason instead of (5.4), (5.5) we use the integration by parts formula. The proof is very similar to the one of Theorem 3.4 so we only sketch it. Let $\mathcal{C}$, $\mathcal{C}_1$, γ have the same meaning as in the proof of Theorem 3.4 and set

$$S := \{w(\cdot) \in W^{1,1}(I) \mid \ w'(s) \in A(s, w(s)) + T_{co(F(s,\bar{x}(s)))}(\bar{x}'(s)) \text{ a.e. in } I\}.$$

Let $\bar{S}$ denote the closure of S in $C(I)$. By Lemma 4.1 $Int(\mathcal{C}) \neq \emptyset$. We only consider the case $Int(\mathcal{C}) \cap Int(\mathcal{C}_1) \cap \bar{S} \neq \emptyset$ since arguments used in the proof of Theorem 3.4 when $Int(\mathcal{C}_1) \cap Int(\mathcal{C}) = \emptyset$ and $Int(\mathcal{C} \cap \mathcal{C}_1) \cap \bar{S} = \emptyset$ correspond to $\lambda = 0$ and are of the same nature as in the proof of Theorem 3.4 via integration by parts arguments given below.

As in Case 2 of the proof of Theorem 3.4 we show that there exist $-a \in \partial^+\varphi(0)$, $c \in \partial^C g(\bar{x}(1))$, $\beta_0 \in \mathcal{C}^-$ and $\beta_1 \in \mathcal{C}_1^-$ such that $\gamma^*(a,c) + \beta_0 + \beta_1 \in S^+$. Let $\psi \in NBV(I)$ and $\eta \in N_{K_1}(\bar{x}(1))$ be such that for all $w \in C(I)$, $\langle \beta_0, w\rangle = \int_0^1 w(s)d\psi(s)$, $\langle \beta_1, w\rangle = \langle \eta, w(1)\rangle$.

By Lemma 4.2 there exist a positive Radon measure μ, a selection $\nu(s) \in N_K(\bar{x}(s)) \cap B$ μ-a.e. such that (3.5) holds true. Furthermore, for all $w \in S$

$$\langle \eta, w(1)\rangle + \langle a, w(0)\rangle + \langle c, w(1)\rangle + \int_0^1 w(t)d\psi(t) \geq 0.$$

Let p solve (3.1) with $p(1) = -c - \psi(1) - \eta$. From the last inequality, integrating by parts, we deduce that

$$\langle a, w(0)\rangle - \langle p(1), w(1)\rangle - \int_0^1 \langle \psi(t), w'(t)\rangle dt \geq 0 \ \ \forall\, w \in S. \tag{5.19}$$

Notice next that for every solution to $w' = A(s,w)$ on $[0,1]$ we have

$$0 = \int_0^1 (\langle p'(t), w(t)\rangle + \langle p(t) + \psi(t), w'(t)\rangle)dt = \langle p(1), w(1)\rangle - \langle p(0), w(0)\rangle + \int_0^1 \langle \psi(t), w'(t)\rangle dt.$$

This and (5.19) imply that for all $w(0) \in \mathbf{R}^n$, $\langle a - p(0), w(0)\rangle \geq 0$. So $a = p(0)$.

To prove (3.3) consider an integrable selection $v(t) \in T_{co(F(t,\bar{x}(t)))}(\bar{x}'(t))$ and let $w \in S$ be such that $w(0) = 0$ and $w'(t) = A(t, w(t)) + v(t)$ a.e. Then

$$\langle p(1), w(1)\rangle = \int_0^1 (\langle p'(t), w(t)\rangle + \langle p(t), w'(t)\rangle)dt.$$

Since p solves (3.1) with $p(1) = -c - \psi(1) - \eta$, from the last equality and (5.19) we deduce

$$\langle p(1), w(1)\rangle = -\int_0^1 \langle \psi(t), w'(t) - v(t)\rangle dt + \int_0^1 \langle p(t), v(t)\rangle dt \geq \langle p(1), w(1)\rangle + \int_0^1 \langle p(t) + \psi(t), v(t)\rangle dt.$$

Consequently, $0 \geq \int_0^1 \langle p(t) + \psi(t), v(t)\rangle dt$. Since $F(t, \bar{x}(t)) - \bar{x}'(t) \subset T_{co(F(t,\bar{x}(t)))}(\bar{x}'(t))$ and $v(t) \in T_{co(F(t,\bar{x}(t)))}(\bar{x}'(t))$ is an arbitrary integrable selection, we deduce (3.3) from the measurable selection theorem. Exactly the same arguments as those used in the proof of Theorem 3.4 imply that if $C_{K_1}(\bar{x}(1)) \cap Int(G(1)) \neq \emptyset$, then $\lambda + \sup_{t\in(0,1)} ||p(t) + \psi(t)|| \neq 0$ and if $\bar{x}(1) \in Int(K_1)$, then $\lambda + var(\psi, (0,1]) \neq 0$, and that $\lambda = 1$ if there exists a solution $\bar{w}(\cdot)$ to (3.14), (3.15).

To prove the last statement, notice that $\bar{w}(\cdot)$ solves the constrained linear control system (3.14), (3.15) also on the time interval $[t,1]$. If $\bar{x}(1) \in Int(K_1)$ and g is differentiable at $\bar{x}(1)$, then by Remark (iv) after Theorem 3.4, $p(1) = -\nabla g(\bar{x}(1)) - \psi(1)$.

From the already proved first statement of Theorem 3.11, replacing the time interval $[0,1]$ by $[t,1]$, we deduce that there exist ψ_t, μ_t, ν_t as in Theorem 3.11 such that a solution $q(\cdot)$ to the adjoint system

$$q' = A^*(s, -q - \psi_t), \quad q(1) = -\nabla g(\bar{x}(1)) - \psi_t(1)$$

satisfies $-q(t) \in \partial^+ \varphi_t(0)$ and $\langle q(s) + \psi_t(s), \bar{x}'(s)\rangle = \max_{v\in co(F(s,\bar{x}(s)))} \langle q(s) + \psi_t(s), v\rangle$ for almost all $s \in [t,1]$. Setting $r(s) = q(s) - p(s)$, we end the proof.

Acknowledgment. The authors are grateful to anonymous referees for constructive suggestions which helped to improve considerably the first version of this manuscript.

REFERENCES

[1] A. V. Arutyunov and S. M. Aseev, *Investigation of the degeneracy phenomenon of the maximum principle for optimal control problems with state constraints*, SIAM J. Control Optim., 35 (1997), pp. 930–952.

[2] J.-P. Aubin and I. Ekeland, Applied Nonlinear Analysis, Wiley-Interscience, New York, 1984.

[3] J.-P. Aubin and H. Frankowska, Set-Valued Analysis, Birkhäuser, Boston, 1990.

[4] J. Borwein, *Weak tangent cones and optimization in a Banach space*, SIAM J. Control Optim., 16 (1978), pp. 512–522.
[5] P. Cannarsa and H. Frankowska, *Some characterizations of optimal trajectories in control theory*, SIAM J. Control Optim., 29 (1991), pp. 1322–1347.
[6] A. Cernea, *Conditions nécessaires d'optimalité pour les solutions d'une inclusion differentielle avec contraintes d'état*, Bull. Polish. Acad. Sci. Math., 43 (1995), pp. 169–173.
[7] A. Cernea, *Necessary optimality conditions for a class of differential inclusions with state constraints*, Rev. Roumaine Math. Pures Appl., 47 (2002), pp. 295–304.
[8] A. Cernea and H. Frankowska, *The connection between the maximum principle and the value function for optimal control problems under state constraints*, Proceedings of the 43rd IEEE Conference on Decision and Control, Paradise Island, Bahamas, 2004, pp. 893–989.
[9] F. H. Clarke and R. B. Vinter, *The relationship between the maximum principle and dynamic programming*, SIAM J. Control Optim., 25 (1987), pp. 1291–1311.
[10] A. Y. Dubovitskii and V. A. Dubovitskii, *Existence criterion for a significant maximum principle for a problem with phase restriction*, Diff. Equations, 31 (1995), pp. 1595–1634.
[11] A. Y. Dubovitskii and A. A. Milyutin, *Extremal problems with constraints*, USSR Comput. Math. and Math. Physics, 5 (1965), pp. 1–80.
[12] N. Dunford and J. Schwartz, Linear Operators, Part I, General Theory, Interscience, New York, 1967.
[13] M. M. A. Ferreira, F. A. C. C. Fontes, and R. B. Vinter, *Nondegenerate necessary conditions for nonconvex optimal control problems with state constraints*, J. Math. Anal. Appl., 233 (1999), pp. 116–129.
[14] H. Frankowska, *Le principe du maximum pour une inclusion différentielle avec des contraintes sur les états initiaux et finaux*, Comptes-Rendus de l'Académie des Sciences, Paris, Série 1, 302 (1986), pp. 599–602.
[15] H. Frankowska, *The maximum principle for an optimal solution to a differential inclusion with end points constraints*, SIAM J. Control Optim., 25 (1987), pp. 145–157.
[16] H. Frankowska, *Optimal trajectories associated with a solution of the contingent Hamlton-Jacobi equation*, Appl. Math. Optim., 19 (1989), pp. 291–311.
[17] H. Frankowska, *Contingent cones to reachable sets of control systems*, SIAM J. Control Optim., 27 (1989), pp. 170–198.
[18] H. Frankowska and F. Rampazzo, *Filippov's and Filippov-Wazewski's theorems on closed domains*, J. Differential Equations, 161 (2000), pp. 449–478.
[19] H. Frankowska and S. Plaskacz, *Semicontinuous solutions of Hamilton-Jacobi-Bellman equations with degenerate state constraints*, J. Math. Anal. Appl., 251 (2000), pp. 818–838.
[20] H. Frankowska and R. B. Vinter, *Existence of neighboring feasible trajectories: Applications to dynamic programming for state-constrained optimal control problems*, J. Optim. Theory Appl., 104 (2000), pp. 21–40.
[21] A. D. Ioffe and V. M. Tichomirov, *Theory of Extremal Problems*, North-Holland, Amsterdam, 1979.
[22] B. Kaskosz and S. Lojasiewicz, *Lagrange-type extremal trajectories in differential inclusions*, Systems Control Lett., 19 (1992), pp. 241–247.
[23] P. Loewen and R. T. Rockafellar, *The adjoint arc in nonsmooth optimization*, Trans. Amer. Math. Soc., 325 (1991), pp. 39–72.
[24] D. B. Luenberger, *Optimization by Vector Space Methods*, John Wiley & Sons, New York-London-Sydney-Toronto, 1969.
[25] A. S. Matveev, *Necessary conditions for an extremum in an optimal-control problem with phase restrictions*, Differ. Equations, 23 (1987), pp. 427–436.
[26] Ş. Mirică, *A proof of Pontryagin's minimum principle using dynamic programming*, J. Math. Anal. Appl., 170 (1992), pp. 501–512.
[27] E. S. Polovinkin and G. V. Smirnov, *An approach to differentiation of many-valued mapping and necessary optimality conditions for optimization of solutions of differential inclusions*, Differ. Equations, 22 (1986), pp. 660–668.
[28] F. Rampazzo and R. B. Vinter, *A theorem on the existence of neighbouring feasible trajectories with application to optimal control*, IMA J. Math. Control and Systems, 16 (1999), pp. 335–351.
[29] F. Rampazzo and R. B. Vinter, *Degenerate optimal control problems with state constraints*, SIAM J. Control Optim., 39 (2000), pp. 989–1007.
[30] R. T. Rockafellar, *Integrals which are convex functionals*, II, Pacific J. Math., 39 (1971), pp. 439–469.

[31] R. T. Rockafellar and R. J.-B. Wets, Variational Analysis, Gründlehren der Mathematischen Wissenschaften 317, Springer-Verlag, Berlin, 1998.
[32] M. Tamzali-Lafond, *Variational inclusions under state constraints*, SIAM J. Control Optim., 42 (2003), pp. 342–362.
[33] R. B. Vinter, *Optimal Control*, Birkhäuser, Boston, 2000.
[34] R. B. Vinter and G. Pappas, *A maximum principle for nonsmooth optimal-control problems with state constraints*, J. Math. Anal. Appl., 89 (1982), pp. 212–232.
[35] R. B. Vinter and H. Zheng, *Necessary conditions for optimal control problems with state constraints*, Trans. Amer. Math. Soc., 350 (1998), pp. 1181–1204.
[36] X. Y. Zhou, *Maximum principle, dynamic programming, and their connection in deterministic control*, J. Optim. Theory Appl., 65 (1990), pp. 363–373.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 2, pp. 704–727

INPUT-TO-STATE STABILITY OF RATE-CONTROLLED BIOCHEMICAL NETWORKS*

MADALENA CHAVES†

Abstract. In this paper, the study of the class of biochemical systems known as zero deficiency networks is extended to the case of time-varying kinetic parameters. We show that the resulting class of nonlinear systems with inputs satisfies a notion of input-to-state stability uniformly over a set of parameters. In particular, the input-to-state stability estimates allow us to characterize the robustness of zero deficiency networks with respect to perturbations in the parameters as well as study their stability when the reaction rates are controlled by an independent process.

Key words. stability, robustness, biochemical networks

AMS subject classifications. 93D25, 92C45

DOI. 10.1137/S0363012903437964

1. Introduction. A biochemical network consists of the interactions among a certain number of species, according to a set of specified reactions that induce a dynamics for the species' concentrations. The time evolution of the species' concentrations is usually modeled by a system of differential equations together with a family of parameters that characterize the reaction rates. These parameters may depend on various external factors and stimuli such as temperature, the concentration of ligands/substrates, or the concentration of an enzyme which may be regulated by an independent dynamics.

Thus biological systems may, in many cases, be viewed as cascades of biochemical networks, where the output of the ith level becomes the input to the $(i+1)$th level of the cascade. This is indeed the structure of many intracellular signal transduction pathways, which are central to biological processes. For example, the binding of a ligand to a cell receptor triggers a sequence of biochemical reactions [12] that ultimately lead to a cell response (such as contraction, motility, or proliferation).

Each level in the cascade may be studied independently as a system with inputs, and in particular we will focus on the input-to-state stability properties of such system. We are also interested in studying the effect of small parameter perturbations (which may be due, for instance, to variations in the room temperature or other experimental setup problems) on the steady-state response of the system. The notion of a parameter robust system should reflect the idea that these *small perturbations* should not greatly affect the qualitative response and guarantee that the output error will also be small. In addition, the input-to-state stability properties of a system provide a framework for the analysis of the stability and convergence of cascades of systems (see, for instance, [15, 2]).

A mathematical model for a certain family of biochemical networks, where the reactions satisfy the mass action kinetics principle, was introduced by Horn and Jackson in 1972 [9] and followed up by the work of Feinberg [6, 7, 8]. These authors

*Received by the editors November 25, 2003; accepted for publication (in revised form) April 5, 2005; published electronically September 12, 2005.
http://www.siam.org/journals/sicon/44-2/43796.html

†Department of Mathematics, Rutgers University, 110 Frelinghuysen Rd., Piscataway, NJ 08854 (madalena@math.rutgers.edu). The author was supported in part by Fundação para a Ciência e a Tecnologia, by Fundação Calouste Gulbenkian, Portugal, and also in part by the BioMaPS Institute at Rutgers University, Aventis (Bridgewater, NJ) and AFOSR (grant F49620-01-1-0063).

developed a rich and beautiful theory on these systems, also known in the literature as *zero deficiency networks*. This model accommodates a wide variety of significant biological systems, including many models for enzymatic mechanisms [14], a model for T-cell receptor signal transduction [13], and receptor–ligand interactions and G-protein coupled receptor activity in cyclic signaling pathways [3, 12, 19]. The models of receptor–ligand interaction are of interest for biomedical applications as well as for drug design: the concentration–response curves associated with some of the basic models [3, 12] may also be analyzed in the context of these zero deficiency networks [5].

The zero deficiency networks may be characterized in terms of a strongly connected graph and the corresponding irreducible matrix, while the mass action kinetics property leads to nonlinear systems with polynomial vector fields. These are the systems we consider in this paper: for these systems, which exhibit multiple steady states, the state space may be viewed as a (disjoint) union of invariant manifolds (which are parallel translates of a given subspace of $\mathbb{R}^n$) so that to each of these invariant manifolds corresponds a distinct (globally) asymptotically stable steady state. The idea of invariant sets of the state space and the existence of steady states, and how these are affected by the external inputs or perturbations will be central to our analysis.

Recently [17], this class of nonlinear systems was studied from a control theory point of view, and the stability and other properties for the system with no inputs were further analyzed. In [17] a formalism is developed for dealing with this type of systems and several results are established which will be frequently referred to in the present work.

We first introduce the class of systems to be studied and recall some basic results (section 2). The definition and characterization of the notion of uniformly semiglobal input-to-state stable systems as well as the statement of the main results are given in section 3. The input-to-state stability estimates are established and the main theorems are proved in sections 5 and 6. In section 4 we show the dependence of steady states of the system on its parameters: the unique steady state in each invariant manifold is an analytic function of the parameters. Section 7 summarizes the main contributions in this paper.

2. Some notation and previous results. Let $n \geq m$ be integers and let $x \in \mathbb{R}^n$. Introduce the *positive orthant*

$$\mathbb{R}^n_{>0} = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_i > 0 \quad \forall\, i\}$$

and the *closed positive orthant*

$$\mathbb{R}^n_{\geq 0} = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_i \geq 0 \quad \forall\, i\}.$$

Assume given two matrices $A = (a_{ij}) \in \mathbb{R}^{m\times m}_{\geq 0}$ and $B \in \mathbb{N}_0^{n\times m}$, where the columns of B are denoted by $b_1, \ldots, b_m$. In our model of biochemical reactions there are n distinct species whose concentrations are given by $x = (x_1, \ldots, x_n)'$ and m *complexes*, each complex denoting a set of reactants or products in a reaction. The complexes are represented by the column vectors $b_1, \ldots, b_m$ with $b_j = (b_{1j}, \ldots, b_{nj})'$ and $b_{lj} \neq 0$ if the species l appears in the complex j. The matrix $A = (a_{ij})$ is the matrix of the kinetic constants, and an entry $a_{ij} \neq 0$ means that complex i is being produced from complex j.

The model for biochemical reaction networks of the Horn–Jackson–Feinberg zero

deficiency type, with mass action kinetics, is as follows:

$$\dot{x} = f_A(x) := \sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij} x_1^{b_{1j}} x_2^{b_{2j}} \cdots x_n^{b_{nj}} (b_i - b_j). \tag{2.1}$$

Since the vector $x \in \mathbb{R}^n$ represents the concentration of each species involved in the reactions, we will be interested only in those trajectories that evolve in the positive orthant. In fact, it is easy to see that the positive orthant $\mathbb{R}^n_{>0}$ is a forward-invariant set for the system (2.1) (see also section 5). We have the following assumptions on A and B.

(a) The matrix $B = (b_1, \ldots, b_m)$ has nonnegative integer entries; it has full column rank and none of its rows vanishes completely.

(b) The matrix $A = (a_{ij})$ has nonnegative entries (without loss of generality, we assume that its diagonal entries are zero, since the corresponding terms would be of the form $a_{ii} x_1^{b_{1i}} \cdots x_n^{b_{ni}} (b_i - b_i) \equiv 0$) and is assumed to be irreducible.

This last property is equivalent to saying that the incidence graph of A is strongly connected, and it describes the following property of the network: there is a chemical pathway connecting every pair of complexes, but the pathway leading from b_i to b_j may be different from the pathway leading from b_j back to b_i (in other words, each individual chemical reaction is not necessarily reversible). Another equivalent definition of irreducibility says that there is an integer k such that all the entries of the matrix $(A + I)^k$ are strictly positive (see [10]).

Example. The simplest model for the interaction of a cell surface receptor with a specific ligand is depicted as $R + L \leftrightharpoons C$. Here we have three species ($n = 3$): receptor (R), ligand (L), and receptor–ligand product (C). There are only two complexes ($m = 2$): $R + L$ and C. Let x_1, x_2, and x_3 denote the concentrations of R, L, and C, respectively. Then the two columns of matrix B are $b_1 = (1, 1, 0)'$ and $b_2 = (0, 0, 1)'$. The matrix A is of size 2 and its nonzero elements are a_{21} (the rate constant for $R + L \to C$) and a_{12} (the rate constant for $C \to R + L$). It is easy to see that $\dot{x}_1 = \dot{x}_2 = -\dot{x}_3 = -a_{21} x_1 x_2 + a_{12} x_3$. A network involving two distinct receptor conformations is shown in Figure 1.

Example. Another simple example is a dimer model, where the ligand binds to two receptors, according to the diagram $2R + L \leftrightharpoons R + C_1 \leftrightharpoons C_2$. In this case $n = 4$ and $m = 3$. Setting $x = (R, L, C_1, C_2)$, the matrix B consists of the vectors $b_1 = (2, 1, 0, 0)'$, $b_2 = (1, 0, 1, 0)'$, and $b_3 = (0, 0, 0, 1)'$. The nonnegative entries of the matrix A are a_{21}, a_{12}, a_{32}, and a_{23}. Again, it is easy to see that properties (a) and (b) are satisfied. At any given time, the concentration of receptors is given by the equation $\dot{R} = -a_{21} R^2 L - (a_{32} - a_{12}) R C_1 + a_{23} C_2$, and the concentration of ligand is given by $\dot{L} = -a_{21} R^2 L + a_{12} R C_1$. The concentrations of C_1 and C_2 are given by $\dot{C}_1 = -(a_{12} + a_{32}) R C_1 + a_{21} R^2 L + a_{23} C_2$ and $\dot{C}_2 = -a_{23} C_2 + a_{32} R C_1$, respectively.

In this paper, we wish to study system (2.1) when the parameters a_{ij} are allowed to be *time variant*. Values of the parameters should be such that, at each time instant, the matrix $A = (a_{ij})$ is irreducible, so we consider the set of irreducible $m \times m$ matrices whose entries are nonnegative:

$$\mathcal{A}_{\geq 0} = \left\{ A \in \mathbb{R}^{m \times m} : A \geq 0 \text{ and } (A + I)^k > 0 \text{ for some power } k \right\}$$

(the inequality $A \geq 0$ (resp., $A > 0$) means that every entry of the matrix on the left-hand side is nonnegative (resp., positive)). Let $|A|_{\text{ecl}}$ denote the matrix norm induced by the vector norm $|\cdot|$ (the Euclidean norm). Throughout this paper, we will

define an *input* $u(\cdot)$ to be a piecewise locally Lipschitz function, with a finite number of discontinuities, that is, there exist $\ell \in \mathbb{N}$ and numbers $0 = T_0 < T_1 < \cdots < T_\ell < T_{\ell+1} = \infty$ such that the function u is locally Lipschitz on each interval (T_{i-1}, T_i): for each $i = 1, \ldots, \ell+1$ and each compact interval $J \subset (T_{i-1}, T_i)$, there exists $\kappa > 0$ such that

$$|u(t) - u(s)| \leq \kappa |t - s| \quad \forall\, s, t \in J. \tag{2.2}$$

In addition, the mass action kinetics model may be generalized as in [17], so we will consider the system with inputs

$$\dot{x} = f(x, u) := \sum_{i=1}^{m} \sum_{j=1}^{m} u_{ij} \theta_1(x_1)^{b_{1j}} \theta_2(x_2)^{b_{2j}} \cdots \theta_n(x_n)^{b_{nj}} (b_i - b_j), \tag{2.3}$$

where the same assumptions on B hold and each map $\theta_i : \mathbb{R} \to [0, +\infty)$ has the following properties:

(c) θ_i is real analytic;
(d) $\theta_i(0) = 0$;
(e) $\int_0^1 |\ln \theta_i(r)|\, dr < \infty$;
(f) its restriction to $\mathbb{R}_{\geq 0}$ is strictly increasing and onto the set $[0, \sigma_i)$, where $0 < \sigma_i \leq +\infty$.

Before stating the last condition that the functions θ_i should satisfy, let us introduce the following vector functions:

$$\rho^{[n]}(x) = (\ln \theta_1(x_1), \ldots, \ln \theta_n(x_n))' \quad \text{and} \quad \exp^{[n]}(v) = (e^{v_1}, \ldots, e^{v_n})'$$

defined on $\mathbb{R}^n_{>0}$ and $\mathbb{R}^n$, respectively. (From now on, we will drop the superscript n of $\rho^{[n]}$ and $\exp^{[n]}$, since its value is usually clear from the context.)

Each θ_i (restricted to $\mathbb{R}_{>0}$) is onto the set $(0, \sigma_i)$, so each function $\rho_i = \ln \theta_i$ (for the restriction of θ_i to $\mathbb{R}_{>0}$) is onto $(-\infty, \bar{\rho}_i)$ with $\bar{\rho}_i = \ln \sigma_i$. Since θ_i (restricted to $\mathbb{R}_{\geq 0}$) is strictly increasing, ρ_i has an inverse function, which is onto $\mathbb{R}_{>0}$: $\rho_i^{-1} : (-\infty, \bar{\rho}_i) \to \mathbb{R}_{>0}$. Each function θ_i should also satisfy

(g) for any given constant p, $\lim_{t \to \ln \sigma_i} \int_a^t \rho_i^{-1}(s)\, ds - pt = +\infty$ for any $a < \ln \sigma_i$.

Note that, for any constant p, there exists $t_0 \in (-\infty, \bar{\rho}_i)$ such that $\rho_i^{-1}(s) > p + 1$ for all $s \geq t_0$. Therefore, when $\sigma_i = +\infty$, condition (g) always holds. The case of mass action kinetics corresponds to $\theta_i(r) = |r|\ \forall\, i$. Another example of interest is the case $\theta_i(r) = \frac{|r|}{k+|r|}$ with $k > 0$. Even though condition (g) will not be used explicitly in this paper, it is necessary to prove some auxiliary results such as Lemma IV.1 in [17], which will be used in section 4.

For the case of a constant matrix A, system (2.3) has been extensively studied. It takes the form

$$\dot{x} = f(x, A) := f_A(x), \tag{2.4}$$

and we next recall some results already established about this system, which are given in [6, 7, 8, 9] (for the particular case of mass action kinetics), and can also be found in [17] (for the general case).

For any two vectors $a, b \in \mathbb{R}^n$, let $\langle a, b \rangle$ denote their dot product. Define the *stoichiometric space*

$$\mathcal{D} = \text{ span } \{b_i - b_j : i, j = 1, \ldots, m\}$$

and also consider its orthogonal space

$$\mathcal{D}^{\perp} := \{v \in \mathbb{R}^n : \ \langle v, d\rangle = 0 \quad \forall\, d \in \mathcal{D}\}.$$

Then, for any $v \in \mathcal{D}^{\perp}$, notice that

$$\langle f_A, v\rangle = \sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij} x_1^{b_{1j}} \cdots x_n^{b_{nj}} \ \langle (b_i - b_j), v\rangle \equiv 0 \tag{2.5}$$

by the definition of v. Hence $\langle x(t), v\rangle$ =constant= $\langle x(0), v\rangle$, and we have

$$\langle x(t) - x(0), v\rangle = 0 \quad \forall\, v \in \mathcal{D}^{\perp} \ \Leftrightarrow \ x(t) - x(0) \in \mathcal{D} \ \Leftrightarrow \ x(t) \in x(0) + \mathcal{D}.$$

Therefore, the parallel translates of the stoichiometric space, $p + \mathcal{D}$ with $p \in \mathbb{R}^n_{>0}$, define invariant manifolds for the system $\dot{x} = f_A(x)$. For each $p \in \mathbb{R}^n_{>0}$, we will define a *positive class* of system (2.4) to be

$$\mathcal{S} := (p + \mathcal{D}) \cap \mathbb{R}^n_{\geq 0} = \{p + d : d \in \mathcal{D}\} \cap \mathbb{R}^n_{\geq 0}.$$

(If $\mathcal{S} \subset \partial\mathbb{R}^n_{\geq 0}$, then we do not consider such $\mathcal{S}$ as a positive class. In this work, we are not concerned with trajectories that evolve on the boundary of the positive orthant.) Note that the positive classes do not depend on the matrix A, only on B.

The equilibria of system (2.4) may be divided into *boundary equilibria*, $E_0 = \{x \in \partial\mathbb{R}^n_{\geq 0} : \ f_A(x) = 0\}$, and *positive equilibria*, $E_{A,+} = \{x \in \mathbb{R}^n_{>0} : \ f_A(x) = 0\}$.

From [17, Proposition VI.3] we know that E_0 depends only on the matrix B and not on A (however, the elements in $E_{A,+}$ may depend on the matrix A). Throughout this paper we will assume that *no boundary equilibria exist in any positive class*, i.e.,

$$\mathcal{S} \cap E_0 = \emptyset \quad \text{for each positive class } \mathcal{S}. \tag{2.6}$$

Under these conditions we have the following result from [17] and also [6].

THEOREM 2.1. *Consider system* (2.4) *and assume that condition* (2.6) *holds. Then, for each positive class* $\mathcal{S}$ *there exists a (unique) state* $\bar{x} = \bar{x}_{\mathcal{S}} \in \mathbb{R}^n_{>0}$ *which is a globally asymptotically stable point relative to* $\mathcal{S}$*, i.e., for each* $x_0 \in \mathcal{S}$*, the solution of* $\dot{x} = f_A(x)$*,* $x(0) = x_0$*, is defined* $\forall\, t \geq 0$*, and* $x(t) \to \bar{x}$ *as* $t \to \infty$*, and* $\forall\, \varepsilon > 0$ *there exists* $\delta > 0$ *such that if* $|\bar{x} - x_0| < \delta$*, then* $|\bar{x} - x(t)| < \varepsilon$ $\forall\, t > 0$.

Throughout this paper, we will assume that the matrix B (which defines the complexes that form the network) is fixed. Each matrix $A \in \mathcal{A}_{\geq 0}$ characterizes a system of the form (2.4), and we will let $x(t, x_0, A)$ denote the solution of the differential equation $\dot{x} = f_A(x)$ at time t, when the initial condition is $x(0) = x_0 \in \mathbb{R}^n_{>0}$. Then, from Theorem 2.1, it follows that each trajectory $x(\cdot, x_0, A)$ converges to the positive equilibrium in the same class as x_0. So we define $\bar{x}(x_0, A)$ to be the unique equilibrium in the same class as x_0, and thus we may also write

$$E_{A,+} = \left\{\bar{x}(x_0, A) : x_0 \in \mathbb{R}^n_{>0}\right\}$$

and introduce the set of all such positive equilibrium points:

$$\mathcal{E} = \bigcup_{A \in \mathcal{A}_{\geq 0}} E_{A,+}.$$

In section 4 we will show that, as a map from $\mathbb{R}^n_{>0} \times \mathcal{A}_{\geq 0}$ to $\mathbb{R}^n$, $\bar{x}(\cdot, \cdot)$ is a real analytic function.

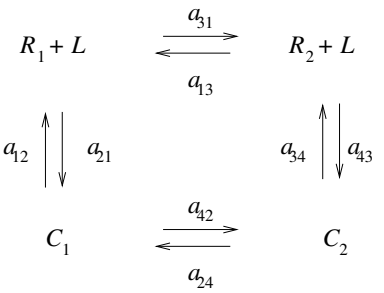


FIG. 1. *A receptor–ligand network.*

2.1. An example. As a motivation for our theoretical results, we will discuss a nominal example. The biochemical network depicted in Figure 1 is a basic model for receptor–ligand interactions at the cell surface [3, 12]. The ligand is denoted by L, two cell receptor conformations are denoted by R_1 and R_2, and the respective receptor–ligand complexes are denoted by C_1 and C_2. These constitute the $n = 5$ individual species, $X = (R_1, R_2, L, C_1, C_2)'$. There are $m = 4$ complexes, $R_1 + L$, C_1, $R_2 + L$, and C_2. This model may be characterized by the matrices

$$A = \begin{pmatrix} 0 & a_{12} & a_{13} & 0 \\ a_{21} & 0 & 0 & a_{24} \\ a_{31} & 0 & 0 & a_{34} \\ 0 & a_{42} & a_{43} & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

where A is clearly irreducible. Under mass action kinetics, the dynamics of the system is

$$\begin{aligned} \dot{R}_1 &= -(a_{21} + a_{31})R_1L + a_{12}C_1 + a_{13}R_2L, \\ \dot{R}_2 &= -(a_{13} + a_{43})R_2L + a_{31}R_1L + a_{34}C_2, \\ \dot{L} &= -a_{21}R_1L - a_{43}R_2L + a_{12}C_1 + a_{34}C_2, \\ \dot{C}_1 &= -(a_{12} + a_{42})C_1 + a_{21}R_1L + a_{24}C_2, \\ \dot{C}_2 &= -(a_{34} + a_{24})C_2 + a_{42}C_1 + a_{43}R_2L. \end{aligned} \tag{2.7}$$

The stoichiometric space is given by

$$\begin{aligned} \mathcal{D} = &\ \operatorname{span}\{b_i - b_j : i, j = 1, \ldots, 4\} \\ = &\ \operatorname{span}\{(1, 0, 1, -1, 0)', (1, -1, 0, 0, 0)', (1, 0, 1, 0, -1)'\}, \end{aligned}$$

and the positive classes are thus characterized by a pair of positive constants (α_1, α_2),

$$L + C_1 + C_2 = \alpha_1, \quad R_1 + R_2 + C_1 + C_2 = \alpha_2,$$

and, incidentally, note that the classes reflect the conservation of the total amount of ligand and the total amount of cell receptors in the system. The boundary equilibria set is given by

$$E_0 = \{(r_1, r_2, 0, 0, 0)', (0, 0, l, 0, 0)' : \ r_1, r_2, l \in [0, +\infty)\},$$

and it is easy to see that each of these equilibrium points implies either $\alpha_1 = 0$ or $\alpha_2 = 0$, which do not define a positive class. Therefore, the "no boundary equilibria" assumption (2.6) holds.

In this receptor–ligand model, the kinetic parameters are assumed to be fixed. A rough numerical estimate of the effect of perturbations on the steady-states shows that, for a sufficiently large (and fixed) T,

$$|x(T, x_0, A) - x(T, x_0, A_0)| \lesssim 0.15\, |A - A_0|_{\text{ecl}}, \tag{2.8}$$

suggesting that the system is indeed parameter-robust and that, moreover, the error is not amplified. Figure 2 shows the effect on the trajectories of the system of random perturbations in the kinetic constants, while Figure 3 justifies estimate (2.8). In this figure, each point $\cdot$ corresponds to the mean square error at steady state for a given error in the kinetic constant ($|A - A_0|_{\text{ecl}}$). To obtain Figure 3, for each $p \in \{10, 20, 30, 40, 50, 60\}$, system (2.7) was simulated 20 times for the same fixed initial condition (x_0), with its kinetic constants randomly perturbed within $p\%$, i.e.,

$$a_{ij} = (1 + \tilde{p}_{ij})a^0_{ij} \quad \text{with} \quad \tilde{p}_{ij} \in [-p/100, p/100].$$

For each simulation, the norms $|A - A_0|_{\text{ecl}}$ and $|x(T, x_0, A) - x(T, x_0, A_0)|$, for a sufficiently large T, were computed, and a point $\cdot$ was plotted. An average of the values $|x(T, x_0, A) - x(T, x_0, A_0)|$ over intervals $|A - A_0|_{\text{ecl}} \in [d_0, d_1]$ of length 0.16 was also computed, resulting in the open squares ($\square$). The solid line represents the best linear fit to these average points with a slope of 0.05. Finally, notice that mostly all points are below the dash-dotted line, that is, they satisfy estimate (2.8).

For both figures, the initial condition was set to $x_0 = (7, 2, 15, 0.5, 0.5)'$ corresponding to a common situation where the amount of ligand is larger than the total amount of receptors, and there are practically no receptor–ligand complexes formed at the beginning of the reaction. The (ideal) values of the parameters were set to

$$A_0 = \begin{pmatrix} 0 & 0.25 & 0.8 & 0 \\ 2.7 & 0 & 0 & 0.45 \\ 0.9 & 0 & 0 & 0.25 \\ 0 & 0.55 & 2.5 & 0 \end{pmatrix}.$$

3. Input-to-state stability and robustness. We wish to study the system with inputs (2.3) and establish general estimates that reflect the stability result obtained numerically for the example in (2.8). To do this, we start by defining appropriate input-to-state stability notions. An important observation on the system is that positive classes are invariant not only under constant inputs but also under any time-variant input map with $u(t) \in \mathcal{A}_{\geq 0}\ \forall t$. This follows from the fact that the matrix B (and hence also the stoichiometric space and its orthogonal space $\mathcal{D}^{\perp}$) is fixed, and also from (2.5). Indeed, let $\mathcal{S}_{\bar{x}}$ be any class (recall that each class may be characterized by a positive equilibrium $\bar{x} \in \mathcal{E}$). Then, for each initial condition $x_0 \in \mathcal{S}_{\bar{x}}$ and input map $u(\cdot)$, the trajectory of system (2.3) evolves in this positive class for all times. Thus, any input-to-state stability estimates only need to hold in that class.

3.1. Definition of input-to-state stability in each invariant subspace. The input-to-state stability notion introduced in Definition 3.2 follows the ideas and the concept of input-to-state stability (ISS) first established in [15] (and see also the notion of input-output-to-state stability introduced in [11]). With the goal of analyzing positive systems, the main difference in our definition of ISS is the introduction of a condition on the completeness of the system with respect to positive states (i.e., those states with all coordinates in the strictly positive half-line). This condition

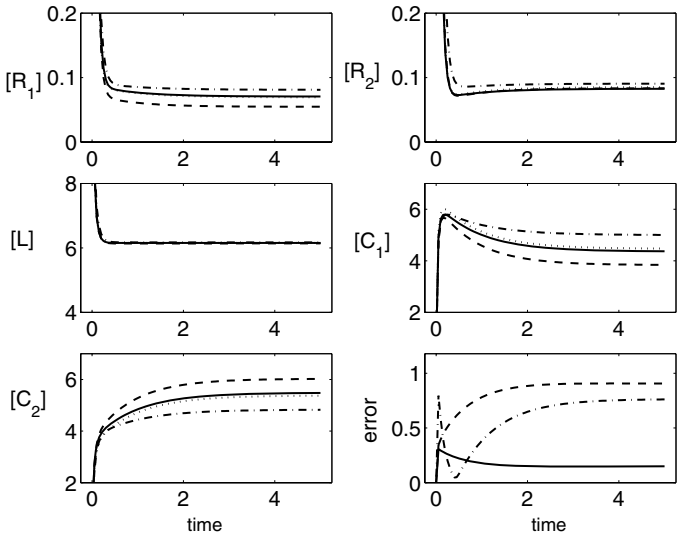


FIG. 2. *The dotted lines represent the trajectory of the system* $\dot{x} = f_{A_0}(x)$ *(*A_0 *as indicated in the text), while the solid, dashed, and dash-dotted lines represent the trajectories of* $\dot{x} = f_A(x)$ *with the entries of* A *randomly chosen within, respectively,* 10%, 20%, *and* 30% *of the (nonzero) entries of* A_0. *The error is computed as the norm* $|x(t, x_0, A_0) - x(t, x_0, A)|$.

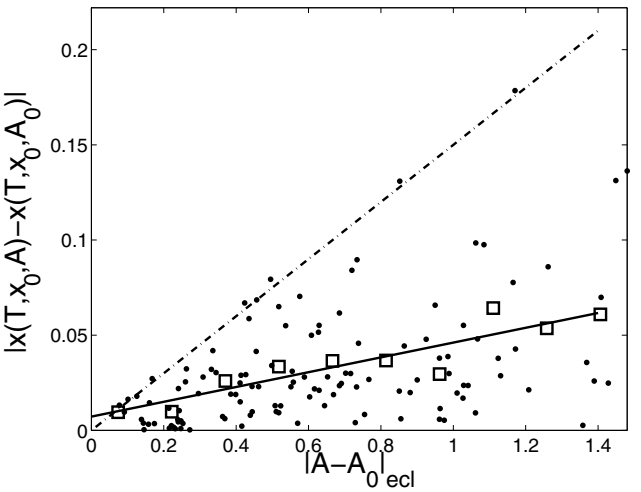


FIG. 3. *The error in the steady states due to random perturbations in the kinetic constants* A_0 *(see explanation in the text).*

plays an important part in the subsequent characterization of the ISS property in terms of an ISS-Lyapunov function: such a function need only be defined on $\mathbb{R}^n_{\geq 0}$ and differentiable on the strictly positive orthant, and it is not required to satisfy a decrease condition except at positive vectors, as stated in Definition 3.3 (in the original characterization, the ISS-Lyapunov function was defined in $\mathbb{R}^n$).

In addition, our notion of ISS is formulated as a *semiglobal* property, in the sense that the input-to-state estimates only hold while the trajectories remain in some preestablished compact set (see also [4]). And it is a *uniform* property, in the sense

that the same functions provide input-to-state estimates for all trajectories evolving in $\cup_{\bar{x}\in P}\mathcal{S}_{\bar{x}}$, where P is a compact subset of $\mathcal{E}$.

We first recall some standard notions that will be used in establishing estimates: a function $\alpha : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is said to be of class $\mathcal{K}$ if it is continuous, strictly increasing, and $\alpha(0) = 0$. The function α is said to be of class $\mathcal{K}_\infty$ if it is of class $\mathcal{K}$ and in addition $\alpha(r) \to +\infty$ as $r \to +\infty$. A function $\beta : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is said to be of class $\mathcal{KL}$ if, for each fixed t, $\beta(\cdot, t)$ is of class $\mathcal{K}$ and for each fixed r, $\beta(r, \cdot)$ is strictly decreasing with $\beta(r, t) \to 0$ as $t \to +\infty$.

For the following definitions, let the system $\dot{x} = f(x, u)$ evolve on a state space $\mathcal{X}$ which is an open subset of $\mathbb{R}^n$ containing $\mathbb{R}^n_{>0}$. Let $\mathbb{U}$ be a subset of $\mathcal{A}_{\geq 0}$, and let $A_0 \in \mathbb{U}$. For any $\bar{x}_0 \in \mathbb{R}^n_{>0}$, let $\mathcal{S}_{\bar{x}_0}$ represent any invariant set for the system $\dot{x} = f(x, u)$ (with $\bar{x}_0 \in \mathcal{S}_{\bar{x}_0}$). Let

$$\|u - A_0\| = \text{ess. sup}\,.\{|u(t) - A_0|_{\text{ecl}} : \ t \in [0, +\infty)\}.$$

DEFINITION 3.1. *A system $\dot{x} = f(x, u)$, with input-value set $\mathbb{U}$, is $\mathbb{R}^n_{>0}$-forward invariant if, for each initial state $x(0) \in \mathbb{R}^n_{>0}$ and each $\mathbb{U}$-valued input $u(\cdot)$, the corresponding maximal solution of $\dot{x} = f(x, u)$ as a differential equation in $\mathcal{X}$, which is defined on an interval $J_{x(0),u} = [0, t_{max})$, has values $x(t) \in \mathbb{R}^n_{>0}$ $\forall\, t \in J_{x(0),u}$. The system is $\mathbb{R}^n_{>0}$-forward complete if it is $\mathbb{R}^n_{>0}$-forward invariant and, for each $x(0) \in \mathbb{R}^n_{>0}$ and $\mathbb{U}$-valued input $u(\cdot)$, $J_{x(0),u} = [0, +\infty)$.*

DEFINITION 3.2. *A system $\dot{x} = f(x, u)$ is uniformly semiglobal input-to-state stable with input-value set $\mathbb{U}$ if*

(i) *the system is $\mathbb{R}^n_{>0}$-forward complete, and*

(ii) *for every compact set $P \subset \mathcal{E}$ and every compact set $F \subset \mathbb{R}^n_{\geq 0}$ containing P, there exist functions $\beta = \beta_{P,F}$ of class $\mathcal{KL}$ and $\varphi = \varphi_{P,F}$ of class $\mathcal{K}_\infty$ such that, for every $\bar{x}_0 \in P \cap E_{A_0,+}$ for some $A_0 \in \mathbb{U}$,*

$$|x(t) - \bar{x}_0| \leq \beta(|x_0 - \bar{x}_0|, t) + \varphi(\|u - A_0\|) \tag{3.1}$$

for each $\mathbb{U}$-valued input $u(\cdot)$ and every initial condition $x_0 \in F \cap \mathcal{S}_{\bar{x}_0}$[1] and $\forall\, t \geq 0$ such that $x(s) \in F$ $\forall\, s \in [0, t]$.

If the functions β, φ given in (ii) *may be chosen independently of the compact F, then the system is uniformly input-to-state stable with input-value set $\mathbb{U}$.*

DEFINITION 3.3. *A continuous function $V : \mathbb{R}^n_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a uniformly semiglobal ISS-Lyapunov function for the system $\dot{x} = f(x, u)$ with input-value set $\mathbb{U}$ if*

(i) *the restriction of V to $\mathbb{R}^n_{>0}$ is continuously differentiable;*

(ii) *for every compact $P \subset \mathcal{E}$, there exist functions $\nu_1 = \nu_{1,P}, \nu_2 = \nu_{2,P} \in \mathcal{K}_\infty$, so that*

$$\nu_1(|x - \bar{x}_0|) \leq V(x) \leq \nu_2(|x - \bar{x}_0|)$$

for each $\bar{x}_0 \in P$ and $\forall\, x \in \mathbb{R}^n_{\geq 0}$;

(iii) *for every compact set $P \subset \mathcal{E}$ and every compact set $F \subset \mathbb{R}^n_{\geq 0}$ containing P, there exist functions $\alpha = \alpha_{P,F}, \gamma = \gamma_{P,F} \in \mathcal{K}_\infty$ such that, for every $\bar{x}_0 \in P \cap E_{A_0,+}$ for some $A_0 \in \mathbb{U}$,*

$$\nabla V(x)\ f(x, u) \leq -\alpha(|x - \bar{x}_0|) + \gamma(|u - A_0|)$$

for every $u \in \mathbb{U}$ and every $x \in F \cap \mathcal{S}_{\bar{x}_0} \cap \mathbb{R}^n_{>0}$.

[1] Since $P \subset F$, the intersection $F \cap \mathcal{S}_{\bar{x}_0}$ is nonempty, containing at least the point $\bar{x}_0$.

If the functions α, γ given in (iii) *may be chosen independently of the compact $F \subset \mathbb{R}^n_{\geq 0}$, then the function V is a uniformly ISS-Lyapunov function for the system.*

We next state without proof that the existence of an ISS-Lyapunov function implies that the system is input-to-state stable (in the sense of the previous definitions). The proof of the lemma is very similar to what is done in the case of the usual definition of an ISS system and follows closely the argument given in [18]. One should keep in mind that the Lyapunov function is differentiable only on the positive orthant, and that the trajectories evolve in invariant classes. (For a similar adaptation of the proof given in [18], see also [4].)

LEMMA 3.4. *Consider an $\mathbb{R}^n_{>0}$-forward complete system $\dot{x} = f(x,u)$ with input-value set $\mathbb{U}$. Suppose that there is a uniformly (semiglobal) ISS-Lyapunov function V for the system. Then, the system is uniformly (semiglobal) input-to-state stable with input-value set $\mathbb{U}$.*

3.2. Main results. As already mentioned, the work of Horn and Jackson, and Feinberg [6, 7, 8, 9] on zero deficiency biochemical networks considers only constant kinetic parameters. This is also the case in the recent work developed in [17, 4]. In other words, so far the focus has been on systems (2.3) with constant inputs, $u_{ij}(t) \equiv a_{ij}$. In this paper, our goal is to study the stability and robustness of zero deficiency networks under time-varying parameters. In order to establish our stability results a "lower bound" on the parameters will be assumed, that is, given any $\varepsilon > 0$, we consider the input-value set to be the following subset of $\mathcal{A}_{\geq 0}$:

$$\mathbb{U}_\varepsilon = \{A \in \mathcal{A}_{\geq 0} : \ a_{ij} \geq \varepsilon \text{ or } a_{ij} = 0\}. \tag{3.2}$$

Note, however, that no upper bound on the values of a_{ij} is required. In addition, recall that the input maps satisfy the regularity condition (2.2). So, we define

$$\mathcal{W} := \{w : [0,+\infty) \to \mathbb{U}_\varepsilon |\ w \text{ is a piecewise locally Lipschitz function}\}. \tag{3.3}$$

The main results state that, first, system (2.3) is uniformly semiglobal ISS, and second, if (2.3) is mass-conservative, then it is also uniformly ISS. The proofs of the theorems are presented in section 6: the ISS properties are established by showing that the system admits a uniformly (semiglobal) ISS-Lyapunov function (section 5.1).

THEOREM 3.5. *System* (2.3) *with the state space $\mathcal{X} = \mathbb{R}^n$, restricted to taking input maps $w \in \mathcal{W}$, is uniformly semiglobal ISS with input-value set $\mathbb{U}_\varepsilon$.*

THEOREM 3.6. *Suppose that system* (2.3) *with state space $\mathcal{X} = \mathbb{R}^n$ satisfies*

$$\exists v \in \mathbb{R}^n_{>0}, \quad v \cdot f(x,u) = 0 \quad \forall\, x \in \mathcal{X} \quad \forall\, u \in \mathcal{A}_{\geq 0}. \tag{3.4}$$

Then the system, when restricted to input maps $w \in \mathcal{W}$, is uniformly ISS with input-value set $\mathbb{U}_\varepsilon$.

We would like to point out that, in the particular case of the constant input $u(t) \equiv A_0$, Theorem 3.6 recovers the global stability result of Theorem 2.1 for mass-conservative systems. In fact, establishing that a given system is uniformly input-to-state stable with input-value set $\mathbb{U}_\varepsilon$ (appropriately chosen) provides an alternative proof of Theorem 2.1. Furthermore, in the case when the input consists of small perturbations around a desired value A_0, for instance, $u(t) = A_0 + \delta(t)$, uniform (global) ISS implies robustness of the system with respect to A_0. In other words, if $\|\delta\| \leq \delta_0$, then we expect the difference between the desired and perturbed steady states of the system to satisfy $|\bar{x} - \bar{x}_0| \lesssim \varphi(\|\delta\|) \leq \varphi(\delta_0)$ (see also the example discussed in section 2.1).

Remark. Condition (3.4) is satisfied by many biochemical systems; in particular, it is satisfied by mass-conservative systems, whose trajectories are a priori constrained to move in a compact subset of $\mathbb{R}^n_{\geq 0}$. A system of the form (2.3) is mass-conservative if and only if

$$v = \sum_{i=1}^{n-m+1} v_i \in \mathbb{R}^n_{>0},$$

where $\{v_1, \ldots, v_{n-m+1}\}$ is a basis of the space $\mathcal{D}^\perp$. Recall that, by definition of the invariant classes, for each v_i there exists a positive constant α_i such that $\langle v_i, x(t)\rangle = \alpha_i$, $\forall\, t \in J$. Then $\langle v, x(t)\rangle = \sum \alpha_i$ for every t, and in those cases where v has all coordinates positive, we immediately have that $x(t)$ evolves in a compact subset of $\mathbb{R}^n_{\geq 0}$ and hence a compact subset of the state space. The example discussed in section 2.1 is mass-conservative with $v = (1, 1, 1, 2, 2)'$.

Remark. One of the main assumptions in the model (2.3) is that, for all times t, the incidence graph of the matrix $u(t)$ is strongly connected or, equivalently, $u(t)$ is irreducible; hence the input u is only allowed to take values in $\mathcal{A}_{\geq 0}$. However, in some ways, the structure of the network may be modified, i.e., new reactions may be added and existing reactions may be removed, provided that *the irreducibility of the matrix $u(t)$ is not violated at any time t.* This is guaranteed by requiring that $u \in \mathbb{U}_\varepsilon$.

As discussed above, Theorems 3.6 and 3.5 hold for input maps that are piecewise locally Lipschitz. These include many of the typical biological inputs such as piecewise constant, periodic, or exponentially decaying signals.

Example. As mentioned in the introduction, changes in the temperature induce changes in the value of the reaction rate constants. These changes are given by the *Arrhenius law* [1]:

$$k = k(T) := F_a\, e^{-\frac{E_a}{RT}},$$

where $F_a > 0$ is the *frequency factor*, E_a is the *activation energy*, T is the temperature (in K), and R is the universal gas constant ($\approx$8.31 $\mathrm{J\,K^{-1}\,mol^{-1}}$). The values F_a and E_a are fixed for each reaction (e.g., for water formation, $\mathrm{OH{+}H_2} \xrightarrow{k} \mathrm{H_2O{+}H}$, $F_a = 8 \times 10^{10}$ $\mathrm{L\,mol^{-1}\,s^{-1}}$, and $E_a = 42 \times 10^3$ $\mathrm{J\,mol^{-1}}$). For most reactions $E_a > 0$, so that k increases with the temperature. Then we have (note that $4/c$ is a Lipschitz constant for the function $e^{-c/x}$)

$$|k(T_1) - k(T_0)| = F_a \left| e^{-\frac{E_a}{RT_1}} - e^{-\frac{E_a}{RT_0}} \right| \leq 4R \frac{F_a}{E_a} |T_1 - T_0|\,.$$

In general, changes in temperature will be reflected in the matrix of kinetic parameters as $\|u^{T_1} - u^{T_0}\| \leq c\,|T_1 - T_0|$ for some $c > 0$. Then, from Theorem 3.6, we expect that a change in temperature from T_0 to T_1 will lead to a deviation in the steady state of order $|\bar{x}_1 - \bar{x}_0| \lesssim \varphi(|T_1 - T_0|)$, where φ is some $\mathcal{K}_\infty$ function.

Example. Consider the model in Figure 1 and assume that the concentration of ligand is regulated from "outside." For instance, $L(t)$ may be experimentally designed to be a piecewise constant function, in order to measure the response of the system to different concentrations of ligand. Or L could be regulated by an independent network. In either case, we would have the following system:

$$\dot{x} = f(x, w), \tag{3.5}$$

where $x = (R_1, R_2, C_1, C_2)'$ and

$$w(t) = \begin{pmatrix} 0 & a_{12} & a_{13}L(t) & 0 \\ a_{21}L(t) & 0 & 0 & a_{24} \\ a_{31}L(t) & 0 & 0 & a_{34} \\ 0 & a_{42} & a_{43}L(t) & 0 \end{pmatrix}.$$

If L is determined by a dynamical system, say $\dot{z} = g(z)$, also of the form (2.1), then we know that $z(t) \to \bar{z}$ for some $\bar{z} \in \mathcal{E}$, and therefore $w(\cdot) \in \mathcal{W}$. An interesting problem for further analysis is whether the convergence of $L(t)$ to some $\bar{L}$ implies that the trajectory $x(t)$ will also converge to some $\bar{x} \in \mathcal{E}$. Another interesting question, which we leave for further research, is whether the cascade system $\dot{x} = f(x, z)$, $\dot{z} = g(z)$ is again input-to-state stable, in the manner developed in [15].

4. Dependence of the steady states on the kinetic parameters. A typical problem concerning cell receptor–ligand interactions, and many other biochemical reactions, is to determine the "dose-response" curves, that is, determine the final concentration of the products, $\bar{C}_1$ or $\bar{C}_2$, as a function of the initial concentration of ligand, L_0 (see [12, 19] and [5]). When translated into mathematical language, this problem involves the characterization of the multiple steady states of system (2.4) and their dependence on the matrix A and classes $\mathcal{S}_{x_0}$.

We recall that, for the case of constant inputs, say $u(t) \equiv A$, the system $\dot{x} = f(x, u)$ with initial condition $x(0) = x_0$ converges to the constant steady state $\bar{x}(x_0, A)$. In contrast, for a general input $u(\cdot)$ one certainly does not expect the system to converge to a constant steady state. However, one may still consider the map $\bar{x} : \mathbb{R}^n_{>0} \times \mathcal{A}_{\geq 0} \to \mathcal{E}$, where $\bar{x}(x_0, A)$ is defined as the unique positive steady state of the system $\dot{x} = f(x, A) = f_A(x)$ in the class $\mathcal{S}_{x_0}$. Then the following is true:

$$\bar{x}(x_0, u(t)) \in \mathcal{E} \quad \forall\, t.$$

We will show that $\bar{x}$ is in fact a real analytic function of x_0 and A. This will help us in the proof of the main results, namely, in section 5, to show that the system is $\mathbb{R}^n_{\geq 0}$-forward complete.

THEOREM 4.1. *Assume that the maps θ_i are real analytic functions. Then the map $\bar{x} : \mathbb{R}^n_{>0} \times \mathcal{A}_{\geq 0} \to \mathcal{E} \subset \mathbb{R}^n_{>0}$ given by $(x_0, A) \mapsto \bar{x}(x_0, A)$ is real analytic.*

To prove this theorem we will use the following alternative expression for f_A (see [17]):

$$f_A(x) = B\tilde{A}\theta_B(x), \tag{4.1}$$

where

$$\theta_B(x) = \begin{pmatrix} \theta_1(x_1)^{b_{11}}\theta_2(x_2)^{b_{21}} \cdots \theta_n(x_n)^{b_{n1}} \\ \theta_1(x_1)^{b_{12}}\theta_2(x_2)^{b_{22}} \cdots \theta_n(x_n)^{b_{n2}} \\ \vdots \\ \theta_1(x_1)^{b_{1m}}\theta_2(x_2)^{b_{2m}} \cdots \theta_n(x_n)^{b_{nm}} \end{pmatrix} = \exp[B'\rho(x)]$$

and

$$\tilde{A} = A + \begin{pmatrix} -\sum_{i=1}^m a_{i1} & 0 & \cdots & 0 \\ 0 & -\sum_{i=1}^m a_{i2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & -\sum_{i=1}^m a_{im} \end{pmatrix}.$$

(Recall that we assumed without loss of generality that all the diagonal entries of A are zero.) Now, given any matrix $G \in \mathbb{R}^{m\times m}$, with entries g_{ij}, define

$$\phi(G) = \left(1 + \sum_{i=1}^{m} g_{ii}^2\right)^{-1} \quad \text{and} \quad M_G = (\phi(G)G + I)^{m-1}.$$

By construction, the diagonal entries of $\phi(G)G+I$ are positive. Introduce the following subset of $\mathbb{R}^{m\times m}$:

$$\mathcal{G} = \{G \in \mathbb{R}^{m\times m} : M_G > 0 \text{ and } \bar{1}\, G = 0\},$$

where the inequality means that every entry of the matrix on the left-hand side is strictly positive, and $\bar{1}$ is the row vector $(1\ 1\cdots 1)$. The set $\mathcal{G}$ may be seen as an open subset of the $(m^2 - m)$-dimensional linear subspace $\{G : \bar{1}\, G = 0\}$ of $\mathbb{R}^{m\times m}$. Define $\mathcal{G}_{\geq 0}$ to be the set of all irreducible matrices which have $\bar{1}\, G = 0$, *nonnegative off-diagonal* entries and *arbitrary diagonal* entries. Note that

$$\mathcal{G}_{\geq 0} = \{G \in \mathcal{G} : G \text{ has nonnegative off-diagonal entries}\}.$$

Then to each matrix $A \in \mathcal{A}_{\geq 0}$, we associate a matrix $\tilde{A} \in \mathcal{G}_{\geq 0}$: clearly, $\bar{1}\tilde{A} = 0$ and so $\tilde{A} \in \mathcal{G}_{\geq 0}$.

For each $G \in \mathcal{G}$ observe that $\bar{1}M_G = \bar{1}(\phi(G)\, G + I)^{m-1} = \bar{1}$ because $\bar{1}\, G = 0$ and $\bar{1}(\phi(G)\, G + I) = \bar{1}$. So, any nonnegative eigenvector, $v \in \mathbb{R}^n_{\geq 0}$, of the matrix M_G must correspond to the eigenvalue $\mu = 1$ since

$$M_G v = \mu v \ \Rightarrow\ \bar{1}(M_G v) = \bar{1}(\mu v) \ \Leftrightarrow\ \bar{1}v = \mu \bar{1}v$$

and $\bar{1}v$ is a positive scalar (since $v \neq (0, \ldots, 0)'$ by the definition of eigenvector).

Since, by definition, M_G is irreducible and has all entries positive, by the Perron–Frobenius theorem we know that the spectral radius of M_G, $\sigma(M_G)$, is an eigenvalue of M_G of algebraic (and hence geometric) multiplicity one. Moreover, an eigenvector associated with $\sigma(M_G)$ can be chosen to have all entries strictly positive (this will be a Perron eigenvector of M_G, and any two such vectors are positive multiples of each other). But, as we have just seen, any positive eigenvector of M_G corresponds to the eigenvalue $\mu = 1$, so we have

$$\sigma(M_G) \equiv 1 \quad \forall\, G \in \mathcal{G}.$$

Define $v_{\text{P}} : \mathcal{G} \to \mathbb{R}^m_{>0}$ to be the map that assigns to each $G \in \mathcal{G}$ the unique Perron eigenvector of M_G, which has its first coordinate equal to 1,

$$v_{\text{P}} = \begin{pmatrix} 1 \\ w_{\text{P}} \end{pmatrix}$$

for some $w_{\text{P}} \in \mathbb{R}^{m-1}_{>0}$. Then the map v_{P} is a rational function on $\mathcal{G}$, as shown in the appendix.

Proof of Theorem 4.1. A function f, defined on an open set $\mathcal{V}$, is real analytic if it admits a power series expansion on a neighborhood of each point of $\mathcal{V}$. If, as in our case, the set $\mathcal{V}$ is not open, then the function f is still called real analytic if it admits an extension to a real analytic function on a neighborhood of $\mathcal{V}$ (see [16]). This is what we will show for the map $\bar{x}(\cdot, \cdot)$.

For each A consider the matrix $\tilde{A} \in \mathcal{G}_{\geq 0}$, constructed from A as indicated above. Then, from (A.1), $\ker \tilde{A} = \text{ span } \{v_{\text{P}}(\tilde{A})\}$. Because B has full column rank, it follows that each equilibrium $\bar{x} \in E_{A,+}$ is characterized by

$$\theta_B(\bar{x}) = c\, v_{\text{P}}(\tilde{A}) \quad \Leftrightarrow \quad B'\rho(\bar{x}) = \rho(c\, v_{\text{P}}(\tilde{A})), \tag{4.2}$$

where c is a positive constant.

Claim. For each A, the element $\bar{z}(A) \in \mathbb{R}^n_{>0}$ given by

$$\bar{z}(A) = \exp\big[B(B'B)^{-1}\rho(v_{\text{P}}(\tilde{A}))\big]$$

is an equilibrium point in $E_{A,+}$.

To prove the claim, note that B has full column rank, so $B'B$ is an invertible matrix and the formula gives $\rho(\bar{z}(A)) = B(B'B)^{-1}\rho(v_{\text{P}}(\tilde{A}))$ or, equivalently,

$$B'\rho(\bar{z}(A)) = B'B(B'B)^{-1}\rho(v_{\text{P}}(\tilde{A})) = \rho(v_{\text{P}}(\tilde{A})).$$

The claim is proved by letting $c = 1$ and $\bar{x} = \bar{z}(A)$ in (4.2).

Now, by Proposition A.1, the map v_{P} is a rational function on $\mathcal{G}$ and, furthermore, $v_{\text{P}}(G) \in \mathbb{R}^m_{>0}$. The functions $\exp(\cdot)$ and $\rho(\cdot)$ are analytic on $\mathbb{R}^n$ and $\mathbb{R}^n_{>0}$, respectively, so it follows that the map $\tilde{A} \mapsto A(\tilde{A}) \mapsto \bar{z}(A)$ from $\mathcal{G}_{\geq 0} \to \mathcal{E}$ is analytic because it admits an analytic extension to $\mathcal{G} \to \mathbb{R}^n_{>0}$. (Denote by $A(\tilde{A})$ the matrix which coincides with A on the off-diagonal entries and has zero in its diagonal.)

Next, from Lemma IV.1 (and proof of Theorem 2) in [17], there is a real analytic map $\varphi(q, w)$, defined on $\mathbb{R}^n_{>0} \times \mathbb{R}^n_{>0}$, such that, for each $q \in \mathbb{R}^n_{>0}$, $x = \varphi(q, \bar{z}(A))$ is the *unique positive equilibrium* of the system $\dot{x} = f_A(x)$ in the same class of q. Let $q = x_0$ and $w = \bar{z}(A)$. We may now conclude that the map $\mathbb{R}^n_{>0} \times \mathcal{G}_{\geq 0} \to \mathcal{E}$ given by

$$(x_0, \tilde{A}) \mapsto \varphi(x_0, \bar{z}(A))$$

is again analytic because it admits an analytic extension to $\mathbb{R}^n_{>0} \times \mathcal{G}$. Therefore,

$$\bar{x}(x_0, A) \equiv \varphi(x_0, \bar{z}(A))$$

is the unique element that belongs to both the class of x_0 and the equilibria set $E_{A,+}$, and we have just shown that the map $\bar{x} : \mathbb{R}^n_{>0} \times \mathcal{A}_{\geq 0} \to \mathcal{E}$ is real analytic. □

5. Existence and completeness of solutions. We now turn our attention to system (2.3) and will show that it is complete in the sense of Definition 3.1.

PROPOSITION 5.1. *Consider system* (2.3)*, with state space* $\mathcal{X} = \mathbb{R}^n$ *and input-value set* $\mathcal{A}_{\geq 0}$*. Then the system is* $\mathbb{R}^n_{>0}$*-forward invariant.*

Proof. Given an initial condition $x(0) = x_0 \in \mathbb{R}^n_{>0}$ and an $\mathcal{A}_{\geq 0}$-valued input $u(t)$, define $F(t, x) := f(x, u(t))$. The existence and uniqueness of a maximal solution to this initial-value problem follows from standard results (such as stated in [16]), by noticing that, for each fixed t, $F(t, x)$ is locally Lipschitz in x and, for each fixed x, it is locally integrable as a function of time. Forward invariance also follows from standard arguments based on the fact that, for $x \in \mathbb{R}^n_{\geq 0}$, if $x_k = 0$ for any k, then $F_k(t, x) \geq 0$. The actual proof is very similar to that of Proposition 3.13 in [4], so we will not reproduce it here. In that proposition, simply take $C = 0$ and replace "a_{ij}" by "u_{ij}" (we only use the fact that $u_{ij} \geq 0$). □

5.1. A Lyapunov function. In order to prove $\mathbb{R}^n_{>0}$-forward completeness of system (2.3), we will need to introduce our candidate ISS-Lyapunov function. Fix any point $\bar{x}$ in $\mathcal{E}$ and recall the notation $\rho_i = \ln \theta_i$. Define

$$V(x,\bar{x}) = \sum_{i=1}^{n} \int_{\bar{x}_i}^{x_i} (\,\rho_i(s) - \rho_i(\bar{x}_i)\,)\, ds. \tag{5.1}$$

This function is introduced and motivated in [17], where it is shown that V is always nonnegative and zero if and only if $x \equiv \bar{x}$. It is also easy to see that $V(x,\bar{x}) \to +\infty$ as $|x-\bar{x}| \to +\infty$. Also, the function V is *proper* in the following sense: for each compact set $P \subset \mathcal{E}$, one can show that there exist two class $\mathcal{K}_\infty$ functions $\nu_1 = \nu_{1,P}$, $\nu_2 = \nu_{2,P}$ such that

$$\nu_1(|x-\bar{x}|) \le V(x,\bar{x}) \le \nu_2(|x-\bar{x}|) \tag{5.2}$$

$\forall\, x \in \mathbb{R}^n_{\ge 0}$ and $\forall\, \bar{x} \in P$. For instance, we may take

$$\nu_1(r) = \inf\{V(x,\bar{x}) : |x-\bar{x}| \ge r,\ x \in \mathbb{R}^n_{\ge 0},\ \bar{x} \in P\}$$

and

$$\nu_2(r) = r + \max\{V(x,\bar{x}) : |x-\bar{x}| \le r,\ x \in \mathbb{R}^n_{\ge 0},\ \bar{x} \in P\}.$$

So, it is easy to see that V satisfies both properties (i) and (ii) of Definition 3.3. In the case of maps $\theta_i(r) = |r|$, the function has the form

$$V(x,\bar{x}) = \sum_{i=1}^{n} x_i\,(\ln x_i - \ln \bar{x}_i) + (\bar{x}_i - x_i). \tag{5.3}$$

Some more notation will be useful. For any $\bar{x} = (\bar{x}_1,\dots,\bar{x}_n)' \in \mathcal{E}$ and $\forall\, x \in \mathbb{R}^n_{>0}$ define

$$q_j(x,\bar{x}) = q_j := \langle b_j, \rho(x) - \rho(\bar{x})\rangle. \tag{5.4}$$

Introduce also the scalar function $\omega : \mathbb{R} \to \mathbb{R}_{\ge 0}$ given by

$$\omega(r) = e^r - 1 - r\,. \tag{5.5}$$

Furthermore, note that

$$\nabla_x V(x,\bar{x}) = \rho(x) - \rho(\bar{x}) = (\ln\theta_1(x_1) - \ln\theta_1(\bar{x}_1),\dots,\ln\theta_n(x_n) - \ln\theta_n(\bar{x}_n)),$$

$$\nabla_{\bar{x}} V(x,\bar{x}) = \left((\bar{x}_1 - x_1)\,\frac{\theta_1'(\bar{x}_1)}{\theta_1(\bar{x}_1)},\dots,(\bar{x}_n - x_n)\,\frac{\theta_n'(\bar{x}_n)}{\theta_n(\bar{x}_n)}\right).$$

Now, given any $A \in \mathcal{A}_{\ge 0}$ and any $\bar{x} \in E_{A,+}$, consider

$$\begin{aligned} \nabla V(x,\bar{x})\, f_A(x) &= \langle \rho(x) - \rho(\bar{x}), f_A(x)\rangle \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij} e^{\langle b_j,\rho(\bar{x})\rangle} e^{q_j}(q_i - q_j) \\ &= -\sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij} e^{\langle b_j,\rho(\bar{x})\rangle} e^{q_j}\omega(q_i - q_j) \\ &=: -W(x,\bar{x}). \end{aligned} \tag{5.6}$$

The third inequality holds because

$$\begin{aligned} e^{q_j}(q_i - q_j) &= e^{q_j}(q_i - q_j) - e^{q_j}(e^{q_i - q_j} - 1) + e^{q_j}(e^{q_i - q_j} - 1) \\ &= -e^{q_j}\omega(q_i - q_j) + (e^{q_i} - e^{q_j}) \end{aligned}$$

and

$$\begin{aligned} &\sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij} e^{\langle b_j, \rho(\bar{x})\rangle}(e^{q_i} - e^{q_j}) \\ &= (e^{q_1}, \dots, e^{q_m})' A \theta_B(\bar{x}) - (e^{q_1}, \dots, e^{q_m})' \operatorname{diag}\left(\sum_i a_{i1}, \dots, \sum_i a_{im} \right) \theta_B(\bar{x}) \\ &= (e^{q_1}, \dots, e^{q_m})' \tilde{A} \theta_B(\bar{x}) = 0, \end{aligned} \tag{5.7}$$

since at steady state, recalling (4.1) and that B has full rank,

$$f(\bar{x}, A) = f_A(\bar{x}) = B\tilde{A}\theta_B(\bar{x}) = 0.$$

An important point to notice is that $-W(x, \bar{x})$ (hence $\nabla V(x, \bar{x}) f_A(x)$) is *always nonpositive*, because $\omega(r) \geq 0 \ \forall\, r$ (with $\omega(r) = 0$ if and only if $r = 0$).

To prove $\mathbb{R}^n_{>0}$-forward completeness, we consider the function $V(x(t), \bar{x}(x_0, u(t)))$ along a trajectory $x(\cdot, x_0, u(\cdot))$, which is the solution of (2.3), when the input is $u(\cdot)$ and the initial condition $x(0) = x_0$. For the next lemma recall that the maps θ_i are onto an interval of the form $[0, \sigma_i)$, where $0 < \sigma_i \leq +\infty$.

LEMMA 5.2. *Given any compact set $P \subset \mathcal{E}$, let $\ell_1, \ell_2 > 0$ be any numbers so that*

$$e^{\frac{1}{\ell_1}} < \frac{\sigma_i}{\theta_i(\bar{x}_i)} \quad \forall\, \bar{x} \in P, \quad \forall\, i = 1, \dots, n \tag{5.8}$$

and

$$\ell_2 \bar{x}_i > |\bar{x}_i - r_\pm| \quad \forall\, \bar{x} \in P, \quad \forall\, i = 1, \dots, n, \tag{5.9}$$

where the numbers $r_\pm$ are defined by the equations $\ln \theta_i(r_\pm) = \ln \theta_i(\bar{x}_i) \pm 1/\ell_1$.

Then

$$|\bar{x}_i - x_i| \leq \ell_1 V(x, \bar{x}) + \ell_2 \bar{x}_i$$

$\forall\, x \in \mathbb{R}^n_{\geq 0}$ and $\forall\, \bar{x} \in P$.

Remark. If $\theta_i(r) = |r| \ \forall\, i = 1, \dots, n$, then $r_\pm = e^{\pm \frac{1}{\ell_1}} \bar{x}_i$, and we may choose ℓ_1 and ℓ_2 independently of P: indeed, condition (5.8) becomes $e^{\frac{1}{\ell_1}} < \infty$ (satisfied by any $\ell_1 > 0$), and condition (5.9) becomes $\ell_2 > |1 - e^{\pm \frac{1}{\ell_1}}|$. For instance, we may pick $\ell_1 = 1$ and $\ell_2 = 2$.

Proof. Pick any compact set $P \subset \mathcal{E}$ and pick any numbers ℓ_1 and ℓ_2 according to (5.8) and (5.9). First, note that the definition of V implies (see (5.1))

$$\int_{\bar{x}_i}^{x_i} \rho_i(s) - \rho_i(\bar{x}_i)\, ds \leq V(x, \bar{x}), \quad i = 1, \dots, n$$

(recall that $\rho_i(s) = \ln \theta_i(s)$). Now fix any $i \in \{1, \dots, n\}$ and put $a = \bar{x}_i$. For $r \geq 0$, $a > 0$, define

$$h(r) = \ell_1 \int_a^r \rho_i(s) - \rho_i(a)\, ds - |a - r| + \ell_2 a.$$

We will show that $h(r) \geq 0 \; \forall \, r \geq 0$. The first derivative of h is piecewise continuous

$$\frac{dh}{dr} = \begin{cases} \ell_1(\rho_i(r) - \rho_i(a)) + 1 & \text{if } 0 \leq r < a, \\ \ell_1(\rho_i(r) - \rho_i(a)) - 1 & \text{if } r > a \end{cases}$$

and the second derivative is

$$\frac{d^2h}{dr^2} = \ell_1 \, \frac{\theta_i'(r)}{\theta_i(r)} > 0 \quad \text{for} \quad r \neq a,$$

where $\theta_i'(r) = d\theta_i/dr > 0$, because θ_i is strictly increasing. Each continuous piece of the derivative has a zero, at the points $r_\pm$,

$$\frac{dh}{dr} = 0 \quad \Leftrightarrow \quad \begin{cases} \rho_i(r_-) = -\frac{1}{\ell_1} + \rho_i(a) & \text{if } 0 < r < a, \\ \rho_i(r_+) = \frac{1}{\ell_1} + \rho_i(a) & \text{if } r > a. \end{cases}$$

Note that, because ρ_i is an increasing function, indeed $r_- < a$ and $r_+ > a$. In addition, from (5.8) it follows that both r_- and r_+ are well defined, since they belong to the domain of θ_i. Since the second derivative of h is always positive for $r \neq a$, h has local minima at the points $r = r_\pm$. By definition of ℓ_2, it follows that the value of h at $r_\pm$ is positive:

$$h(r_\pm) = \ell_1 \int_a^{r_\pm} \rho_i(s) - \rho_i(a) \, ds - |a - r_\pm| + \ell_2 a,$$

since the first term is positive by construction of V and the two other terms satisfy (5.9)

$$-|a - r_\pm| + \ell_2 a > 0.$$

To summarize,

$$\frac{dh}{dr} = \begin{cases} < 0, & 0 \leq r < r_-, \\ > 0, & r_- < r < a, \\ < 0, & a < r < r_+, \\ > 0, & r_+ < r \end{cases}$$

so that h decreases down to a local *positive* minimum at r_-, then increases up to $h(a) > 0$, and decreases again to another local *positive* minimum at r_+, and increases $\forall r > r_+$. Therefore,

$$h(r) > 0 \quad \forall \, r \in [0, +\infty).$$

This finishes the proof, since for each $a = \bar{x}_i$,

$$\begin{aligned} h(r) > 0 \quad \Leftrightarrow \quad |\bar{x}_i - x_i| &\leq \ell_1 \int_{\bar{x}_i}^{x_i} \rho_i(s) - \rho_i(\bar{x}_i) \, ds + \ell_2 \bar{x}_i \\ &\leq \ell_1 V(x, \bar{x}_i) + \ell_2 \bar{x}_i. \qquad \square \end{aligned}$$

5.2. Completeness.

PROPOSITION 5.3. *Consider system* (2.3) *with state space* $\mathcal{X} = \mathbb{R}^n$. *Then the system is* $\mathbb{R}^n_{>0}$*-forward complete, whenever the input map* $u(\cdot)$ *is in* $\mathcal{W}$.

Proof. Pick any input map $u(\cdot)$ in $\mathcal{W}$. Let $x(t)$ be the issuing solution of (2.3), with the initial condition $x(0) = x_0 \in \mathbb{R}^n_{>0}$, and let it be defined on the maximal interval

$[0,T)$. From Proposition 5.1, we already know that $x(t) := x(t, x_0, u(\cdot)) \in \mathbb{R}^n_{>0}\ \forall\, t$ in the interval $[0,T)$. Assuming that $T < +\infty$, we will show that $x(\cdot)$ evolves on a compact subset of $\mathcal{X}\ \forall\, t \in [0,T)$, which is a contradiction. To do this, we consider the function

$$g(t) = V(x(t), \bar{x}(x_0, u(t)))$$

whose derivative is

$$\begin{aligned}\dot{g}(t) &= \nabla_x V(x(t), \bar{x}(x_0,u(t)))\ \frac{d}{dt}[x(t)] + \nabla_{\bar{x}} V(x(t), \bar{x}(x_0,u(t)))\ \frac{d}{dt}[\bar{x}(x_0,u(t))] \\ &= \langle \rho(x) - \rho(\bar{x}), f(x,u)\rangle + \sum_{i=1}^{n} (\bar{x}_i - x_i)\ \dot{\bar{x}}_i\ \frac{\theta_i'(\bar{x}_i)}{\theta_i(\bar{x}_i)},\end{aligned}$$

where, for simplicity, we used $x \equiv x(t)$ and $\bar{x} \equiv \bar{x}(x_0,u(t))$, and $\dot{\bar{x}}_i := \frac{d}{dt}[\bar{x}_i(x_0,u(t))]$.

Now, for almost all $t \in [0,T)$, $u(t)$ takes values in a compact set. So there exist constants $\underline{c}, \overline{c} > 0$ such that

$$\underline{c} \le |\bar{x}_i(x_0,u(t))| \le \overline{c} \quad \text{for almost all}\ \ t \in [0,T).$$

By differentiability of $\bar{x}(\cdot,\cdot)$ (Theorem 4.1), and because u is piecewise locally Lipschitz with finitely many points of discontinuity, there exist positive constants κ and c_1 such that

$$\dot{\bar{x}}_i = \sum_{i,j=1}^{m} \frac{d\bar{x}_i}{du_{ij}} \frac{du_{ij}}{dt} \le \sum_{i,j=1}^{m} \kappa \left| \frac{d\bar{x}_i}{du_{ij}} \right| \le c_1 \quad \text{for almost all}\ \ t \in [0,T).$$

The function θ_i is positive and strictly increasing, so $\theta_i'(r)$ is also positive. Since $\bar{x}_i(\cdot,\cdot)$ takes values in a compact set, there exists $c_2 > 0$ such that

$$\frac{\theta_i'(\bar{x}_i)}{\theta_i(\bar{x}_i)} \le c_2 \quad \text{for almost all}\ \ t \in [0,T).$$

From (5.6), $\langle \rho(x) - \rho(\bar{x}(x_0,u)), f(x,u)\rangle \le 0\ \forall\, x \in \mathbb{R}^n_{>0},\ u \in \mathcal{A}_{\ge 0}$. Then, applying Lemma 5.2, with $P = [\underline{c}, \overline{c}]^n \cap \mathcal{E}$, to the second term on $\dot{g}$, we obtain

$$\dot{g} \le \sum_{i=1}^{n} |\dot{\bar{x}}_i|\ \ \frac{\theta_i'(\bar{x}_i)}{\theta_i(\bar{x}_i)} (\ell_1\, V(x,\bar{x}) + \ell_2 \bar{x}_i),$$

which implies

$$\dot{g}(t) \le \ell_1 c_1 c_2\, g(t) + \ell_2 c_1 c_2 \overline{c} \quad \text{for almost all}\ t \in [0,T).$$

Taking $c_3 = \ell_2 c_1 c_2 \overline{c}$ and $c_4 = \ell_1 c_1 c_2$ and applying Gronwall's lemma, yield

$$g(t) \le c_3 e^{c_4 T} \quad \forall\, t \in [0,T).$$

For the compact set $P = [\underline{c}, \overline{c}]^n \cap \mathcal{E}$, let $\nu_1 = \nu_{1,P}$ be the class $\mathcal{K}_\infty$ function such that $\nu_1(|x - \bar{x}|) \le V(x,\bar{x})\ \forall\, x \in \mathbb{R}^n_{\ge 0}$ and $\bar{x} \in P$. Thus,

$$\nu_1(|x(t) - \bar{x}(x_0,u(t))|) \le g(t) \quad \forall\, t \in [0,T)$$

and, therefore,

$$|x(t)| \le \overline{c} + \nu_1^{-1}\big(c_3 e^{c_4 T}\big)$$

implying that x evolves in a compact subset of the state space, which contradicts $T < +\infty$. ∎

6. ISS estimates. To establish the main results, we now show that the function V is a uniformly semiglobal ISS-Lyapunov function. In section 5.1 it was shown that V satisfies properties (i) and (ii) of Definition 3.3. We next show that it also satisfies property (iii).

For any $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_n)' \in \mathcal{E}$ and any $x \in \mathbb{R}^n_{\geq 0}$ define

$$\pi_j(x, \bar{x}) = \pi_j := \left[\frac{\theta_1(x_1)}{\theta_1(\bar{x}_1)}\right]^{b_{1j}} \left[\frac{\theta_2(x_2)}{\theta_2(\bar{x}_2)}\right]^{b_{2j}} \cdots \left[\frac{\theta_n(x_n)}{\theta_n(\bar{x}_n)}\right]^{b_{nj}} \tag{6.1}$$

and observe that, if $x \in \mathbb{R}^n_{>0}$, from (5.4)

$$\pi_j = e^{q_j}.$$

Using this notation, define the function $\Psi : \mathbb{R}^n_{\geq 0} \times \mathcal{E} \to \mathbb{R}_{\geq 0}$,

$$\Psi(x, \bar{x}) := \sum_{i=1}^{m} \sum_{j=1}^{m} \left(e^{-\pi_i} - e^{-\pi_j}\right)^2,$$

which, from Lemma 3.8 in [4], satisfies

$$\Psi(x, \bar{x}) = 0 \quad \Leftrightarrow \quad x \in E_0 \cup E_{A,+}, \tag{6.2}$$

where E_0 is the set of boundary equilibria and $A \in \mathcal{A}_{\geq 0}$ is such that $\bar{x} \in E_{A,+}$. Recall the function W defined in (5.6): a useful estimate (see Lemma 3.10 in [4]) establishes that, for each fixed $\bar{x}$,

$$W(x, \bar{x}) \geq \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} a_{ij}\, e^{\langle b_j, \rho(\bar{x})\rangle} \left(e^{-\pi_i} - e^{-\pi_j}\right)^2 \quad \forall\, x \in \mathbb{R}^n_{>0}. \tag{6.3}$$

Moreover, since $x \in \mathbb{R}^n_{>0}$, it follows from (6.2) that the expression on the right-hand side is zero if and only if $x \equiv \bar{x}$.

Next, given any $A \in \mathcal{A}_{\geq 0}$, suppose that A_1 is a matrix with entries $a^1_{ij} = 1$ if $a_{ij} > 0$ and $a^1_{ij} = 0$ if $a_{ij} = 0$. Then, for expression (6.3), we can write

$$W(x, \bar{x}) \geq \frac{1}{2} \min_{a_{ij}>0}\{a_{ij}\} \min_j e^{\langle b_j, \rho(\bar{x})\rangle} \sum_{i=1}^{m} \sum_{j=1}^{m} a^1_{ij} \left(e^{-\pi_i} - e^{-\pi_j}\right)^2.$$

Since A_1 is irreducible, we may apply Lemma VIII.1 from [17] to conclude that there exists a positive constant $\kappa(A_1)$ such that

$$\sum_{i=1}^{m} \sum_{j=1}^{m} a^1_{ij} \left(e^{-\pi_i} - e^{-\pi_j}\right)^2 \geq \kappa(A_1) \sum_{i=1}^{m} \sum_{j=1}^{m} \left(e^{-\pi_i} - e^{-\pi_j}\right)^2.$$

Now, define the following subset of $\mathcal{A}_{\geq 0}$,

$$\mathcal{A}^1_{\geq 0} := \{A \in \mathcal{A}_{\geq 0} : \ a_{ij} = 1 \ \text{ or } \ a_{ij} = 0\},$$

and note that its cardinality is finite (in fact, the number of elements in $\mathcal{A}^1_{\geq 0}$ is equal to the number of distinct strongly connected graphs with m vertexes). Then let

$$\kappa_1 := \min\left\{\kappa(A) : \ A \in \mathcal{A}^1_{\geq 0}\right\}.$$

Additionally observe that, given any $\varepsilon > 0$,

$$A \in \mathbb{U}_\varepsilon \quad \text{satisfies} \quad \min_{a_{ij}>0}\{a_{ij}\} = \varepsilon$$

(where $\mathbb{U}_\varepsilon$ is the set defined in (3.2)).

From this discussion, it is easy to establish the following lemmas.

LEMMA 6.1. *For each compact $P \subset \mathcal{E}$ and each $\varepsilon > 0$, there exists a constant $c(P,\varepsilon)$ given by*

$$c(P,\varepsilon) = \frac{1}{2}\,\varepsilon\kappa_1 \min_{\bar{x}\in P} \min_j e^{\langle b_j, \rho(\bar{x})\rangle}$$

such that

$$W(x,\bar{x}) \geq c(P,\varepsilon)\,\Psi(x,\bar{x}) \tag{6.4}$$

$\forall\, x \in \mathbb{R}^n_{>0}$ and any element $\bar{x} \in P$.

LEMMA 6.2. *Let $P \subset \mathcal{E}$ be any compact set. Given any compact subset $F \subset \mathbb{R}^n_{\geq 0}$ containing the set P, there exists a class $\mathcal{K}_\infty$ function, $\alpha = \alpha_{P,F}$, such that*

$$\Psi(x,\bar{x}) \geq \alpha(|x-\bar{x}|)$$

$\forall\, \bar{x} \in P$, $x \in F \cap \mathcal{S}_{\bar{x}}$.

Proof. Pick any compact set $P \subset \mathcal{E}$. Let $F \subset \mathbb{R}^n_{\geq 0}$ be any compact set which contains P, and let R_0 be such that the closed ball $|x| \leq R_0$ contains the set F. Define $R = R_0 + \max_{\bar{x}\in P}\ \bar{x}$. Note that, for every $\bar{x} \in P$, the ball $|x-\bar{x}| \leq R$ also contains the set F. Consider the function $\mathbb{R}^n_{\geq 0} \to \mathbb{R}^n_{\geq 0}$ given by

$$\alpha(r) := \begin{cases} \frac{r}{r+1}\min\{\Psi(x,\bar{x}) :\ \bar{x}\in P, x \in \mathcal{S}_{\bar{x}}, r \leq |x-\bar{x}| \leq R\} & \forall\, 0 \leq r \leq R, \\ \alpha(R)\frac{r}{R} & \forall\, r > R. \end{cases}$$

As discussed above, for $x \in \mathcal{S}_{\bar{x}}$, $\Psi(x,\bar{x}) = 0$ if and only if $x = \bar{x}$. Since the minimum is taken over a compact set, the function α satisfies $\alpha(0) = 0$ and $\alpha(r) > 0$ for $r > 0$. Also clearly, for $R \leq r$, α is strictly increasing and satisfies $\alpha(r) \to +\infty$ as $r \to +\infty$. For $0 \leq r \leq R$, $\alpha(r)$ is also strictly increasing, as a product of a strictly increasing function and a nondecreasing function. By construction, $\Psi(x,\bar{x}) \geq \alpha(|x-\bar{x}|)$ for all $\bar{x} \in P$, $x \in F \cap \mathcal{S}_{\bar{x}}$. Finally, without loss of generality we may assume that α is continuous on $\mathbb{R}_{\geq 0}$ (otherwise, it is possible to construct a continuous $\tilde{\alpha}$, $\tilde{\alpha}(0) = 0$, with $\alpha(r) \geq \tilde{\alpha}(r)$, and $\tilde{\alpha}(r) \to +\infty$ as $r \to +\infty$). □

Pick any $\varepsilon > 0$ and consider the sets $\mathbb{U}_\varepsilon$ and $\mathcal{W}$ as defined in (3.2) and (3.3), respectively. For any point $\bar{x}_0 \in \mathcal{E}$, set $V_0(x) \equiv V(x,\bar{x}_0)$, where V is the function defined in (5.1). As in section 2, let $\mathcal{S}_{\bar{x}_0}$ be the class that contains $\bar{x}_0$.

PROPOSITION 6.3. *Given any compact sets $P \subset \mathcal{E}$ and $F \subset \mathbb{R}^n_{\geq 0}$ containing P, there exist class $\mathcal{K}_\infty$ functions $\alpha = \alpha_{P,F}$ and $\gamma = \gamma_{P,F}$ such that, for every $\bar{x}_0 \in P \cap E_{A_0,+}$ for some $A_0 \in \mathbb{U}_\varepsilon$,*

$$\nabla V_0(x)\ f(x,u) \leq -\alpha(|x-\bar{x}_0|) + \gamma(|u-A_0|_{\text{ecl}})$$

for every $u \in \mathbb{U}_\varepsilon$ and every $x \in F \cap \mathcal{S}_{\bar{x}_0} \cap \mathbb{R}^n_{>0}$.

Proof. Pick any compact sets $P \subset \mathcal{E}$ and $F \subset \mathbb{R}^n_{\geq 0}$ containing P. Let $c(P,\varepsilon)$ be the constant given by Lemma 6.1, and let $\tilde{\alpha} = \tilde{\alpha}_{P,F}$ be the $\mathcal{K}_\infty$ function given by

Lemma 6.2. Now, pick any $\bar{x}_0 \in P \cap E_{A_0,+}$ for some $A_0 = (a^0_{ij}) \in \mathbb{U}_\varepsilon$. Using the notation $q_i \equiv q_i(x, \bar{x}_0)$ (defined in (5.4)), we have

$$\begin{aligned}
\nabla V_0(x)\ f(x,u) &= \sum_{i=1}^m \sum_{j=1}^m u_{ij} e^{\langle b_j, \rho(\bar{x}_0)\rangle} e^{q_j}(q_i - q_j) \\
&= \sum_{i=1}^m \sum_{j=1}^m u_{ij} e^{\langle b_j, \rho(\bar{x}_0)\rangle} e^{q_j}(q_i - q_j - (e^{q_i - q_j} - 1)) \\
&\quad + \sum_{i=1}^m \sum_{j=1}^m u_{ij} e^{\langle b_j, \rho(\bar{x}_0)\rangle} e^{q_j}(e^{q_i - q_j} - 1) \\
&= -\sum_{i=1}^m \sum_{j=1}^m u_{ij} e^{\langle b_j, \rho(\bar{x}_0)\rangle} e^{q_j}\, \omega(q_i - q_j) \\
&\quad + \sum_{i=1}^m \sum_{j=1}^m (u_{ij} - a^0_{ij}) e^{\langle b_j, \rho(\bar{x}_0)\rangle}(e^{q_i} - e^{q_j}),
\end{aligned}$$

where $\omega(r)$ is the function defined in (5.5). The last equality is justified because $\omega(r) \geq 0\ \forall\, r \in \mathbb{R}$ and, by (5.7),

$$\sum_{i=1}^m \sum_{j=1}^m a^0_{ij} e^{\langle b_j, \rho(\bar{x}_0)\rangle}(e^{q_i} - e^{q_j}) = (e^{q_1}, \ldots, e^{q_m})' \tilde{A}_0 \theta_B(\bar{x}_0) = 0.$$

Applying Lemmas 6.1 and 6.2, there is a $\mathcal{K}_\infty$ function $\tilde{\alpha}$ such that

$$\begin{aligned}
\nabla V_0(x)\ f(x,u) \leq &-c(P,\varepsilon)\, \tilde{\alpha}(|x - \bar{x}_0|) \\
&+ |u - A_0|_{\text{ecl}} \sum_{i=1}^m \sum_{j=1}^m e^{\langle b_j, \rho(\bar{x}_0)\rangle}\, |e^{q_i} - e^{q_j}|\,.
\end{aligned}$$

Next, let

$$c_2(P,F) = (m^2 - m) \max_j \max_{\bar{x}_0 \in P} e^{\langle b_j, \rho(\bar{x}_0)\rangle} \max_j \max_{x \in F} e^{q_j}$$

and observe that

$$|u - A_0|_{\text{ecl}} \sum_{i=1}^m \sum_{j=1}^m e^{\langle b_j, \rho(\bar{x}_0)\rangle}\, |e^{q_i} - e^{q_j}| \leq 2\, c_2(P,F)\, |u - A_0|_{\text{ecl}}$$

$\forall\, x \in F$.

Finally, choose $\alpha = \alpha_{P,F}$ to be $\alpha(r) = c(P,\varepsilon)\tilde{\alpha}(r)$ and $\gamma = \gamma_{P,F}$ to be $\gamma(r) = 2\, c_2(P,F)\, r$. □

6.1. Proof of Theorem 3.5. Let $\varepsilon > 0$ be any constant and consider the input-value set $\mathbb{U}_\varepsilon$ defined in (3.2) and the set $\mathcal{W}$ defined in (3.3). Proposition 5.3 shows that system (2.3) is $\mathbb{R}^n_{>0}$-forward complete with respect to input maps in $\mathcal{W}$.

Choose any compact sets $P \subset \mathcal{E}$ and $F \subset \mathbb{R}^n_{\geq 0}$ with $P \subset F$. Pick any element $\bar{x}_0 \in P$ and any matrix $A_0 \in \mathbb{U}_\varepsilon$ so that $\bar{x}_0 = \bar{x}(x_0, A_0)$ for some $x_0 \in F \cap \mathbb{R}^n_{>0}$.[2] Using

[2] If no such A_0 exists, that is, if $P \cap E_{A_0,+} = \emptyset\ \forall\, A_0 \in \mathbb{U}_\varepsilon$, then there is nothing to prove, because the statement of Definition 3.2 is vacuous. But if A_0 exists, then x_0 always exists, for instance, $\bar{x}_0$.

this element $\bar{x}_0$, construct the function $V_0(x) := V(x, \bar{x}_0)$. This function V_0 satisfies properties (i) and (ii) of Definition 3.3 and Proposition 6.3 establishes property (iii). Hence V_0 is a uniformly semiglobal ISS-Lyapunov function for system (2.3).

By Lemma 3.4, it follows that system (2.3) is uniformly semiglobal ISS with the input-value set $\mathbb{U}_\varepsilon$, as we wanted to show.

6.2. Proof of Theorem 3.6. Let $\varepsilon > 0$ be any constant and consider the input value set $\mathbb{U}_\varepsilon$ defined in (3.2) and the set $\mathcal{W}$ defined in (3.3). Proposition 5.3 shows that system (2.3) is $\mathbb{R}^n_{>0}$-forward complete with respect to input maps in $\mathcal{W}$. Assume that system (2.3) is mass conservative, i.e., there exists $v = (v_1, \ldots, v_n)' \in \mathbb{R}^n_{>0}$ so that $\langle v, f(x,u)\rangle = 0$ for every $x \in \mathbb{R}^n$ and every $u \in \mathbb{U}_\varepsilon$.

Choose any compact subset $P \subset \mathcal{E}$ and put

$$F(P) := \text{ closure } \big\{q \in \mathbb{R}^n_{>0} : \ \bar{x}(q, A) \in P \text{ for some } A \in \mathbb{U}_\varepsilon\big\}.$$

Then $F(P)$ is a compact subset of $\mathbb{R}^n_{\geq 0}$ because it is closed and also bounded, since

$$q \in F(P) \ \Rightarrow \ \exists\, \bar{x} \in P \ \text{ such that } \ v_i q_i \leq \langle v, q\rangle = \langle v, \bar{x}\rangle \ \Rightarrow \ q_i \leq \frac{1}{v_i}|v||\bar{x}| \leq \overline{c}\,\frac{|v|}{v_i}$$

$\forall\, i$, where $\overline{c} = \max\{|\bar{x}| : \ \bar{x} \in P\}$. Moreover, given any $\bar{x} \in P$, $F(P)$ contains the whole class $\mathcal{S}_{\bar{x}}$.

Now, pick any element $\bar{x}_0 \in P$ and any matrix $A_0 \in \mathbb{U}_\varepsilon$ so that $\bar{x}_0 = \bar{x}(x_0, A_0)$ for some $x_0 \in F(P) \cap \mathbb{R}^n_{>0}$.[2] Using this element $\bar{x}_0$, construct the function $V_0(x) := V(x, \bar{x}_0)$. This function V_0 satisfies properties (i) and (ii) of Definition 3.3. Moreover, the two $\mathcal{K}_\infty$ functions provided by Proposition 6.3 depend only on P:

$$\alpha = \alpha_{P,F(P)} \equiv \alpha_P \quad \text{ and } \quad \gamma = \gamma_{P,F(P)} \equiv \gamma_P.$$

So, V_0 is in fact a uniformly ISS-Lyapunov function for system (2.3).

By Lemma 3.4, it follows that system (2.3) (when constrained to take input maps in $\mathcal{W}$) is uniformly ISS with the input-value set $\mathbb{U}_\varepsilon$.

7. Conclusions. We have extended the analysis of zero deficiency biochemical networks to the case where the kinetic parameters associated with each reaction rate are assumed to be time-varying inputs. We have shown that these rate controlled biochemical systems are input-to-state stable with respect to an appropriate input set. Thus one may analyze the stability of the biochemical network when the reaction rates are controlled by some independent process; for instance, some reactions may be inhibited or activated through enzymatic activity. Also as a consequence of the ISS property, we conclude that such systems of biochemical networks are robust with respect to small perturbations in the kinetic parameters such as temperature fluctuations.

By definition, the zero deficiency biochemical networks are assumed to be closed systems in the sense that there are no inflows or outflows (such as additive inputs). While the systems we have studied also do not allow any inflows or outflows, we have nevertheless incorporated outside effects, in the form of "multiplicative" inputs, by allowing an independent system to control the reaction rates.

Appendix. The Perron eigenvector $\mathbf{v_P}$. By a *rational function everywhere defined on* $\mathcal{G}$ we mean a function $\psi : \mathcal{G} \to \mathbb{R}^m$ for which every coordinate is a quotient $\psi_i = p_{\text{num}} p_{\text{den}}^{-1}$ of two polynomial functions (on the entries of G) $p_{\text{num}}, p_{\text{den}} : \mathbb{R}^{m\times m} \to \mathbb{R}$ such that $p_{\text{den}}(G) \neq 0\ \forall\, G \in \mathcal{G}$.

PROPOSITION A.1. *The map v_P is a rational function on $\mathcal{G}$.*

Proof. For each $G \in \mathcal{G}$, by an abuse of notation, we write $v_{\rm P}$ for $v_{\rm P}(G)$. We will also drop the subscript and we let $M = M_G$ for simplicity. We have $Mv_{\rm P} = \sigma(M)v_{\rm P} \Leftrightarrow (M-I)v_{\rm P} = 0$. The matrix $M-I$ has rank $m-1$ because $\sigma(M) = 1$ is a simple root of the characteristic polynomial of M. Put $M - I = (N_1\ N)$, where N_1 is the first column of $M-I$ and N is the remaining $m \times (m-1)$ matrix, and notice that

$$(N_1\ N)\begin{pmatrix} 1 \\ w_{\rm P} \end{pmatrix} = 0 \ \Leftrightarrow\ Nw_{\rm P} = -N_1.$$

Claim. The matrix N has full rank.

Suppose the claim is false. Then there exists an element u in the kernel of N, and one can write $N(w_{\rm P} + u) = -N_1$. But if this is true, then it also holds that

$$(M-I)\begin{pmatrix} 1 \\ w_{\rm P} + u \end{pmatrix} = 0$$

which implies $w_{\rm P} + u = w_{\rm P}$, because $v_{\rm P}$ is in fact the unique vector with first coordinate equal to 1 in the kernel of $M-I$. So $u \equiv 0$, which proves the claim.

It follows that $\det(N'N) \neq 0$ for every G, and applying the Moore–Penrose pseudo-inverse of N yields

$$v_{\rm P} = \begin{pmatrix} 1 \\ w_{\rm P} \end{pmatrix} = \begin{pmatrix} 1 \\ -(N'N)^{-1}N'N_1 \end{pmatrix},$$

where N and N_1 are defined from $M = M_G$, as above. This shows that $v_{\rm P}$ is a rational function on $\mathcal{G}$. ∎

For every $G \in \mathcal{G}$, the Perron eigenvector of M_G, $v_{\rm P}$, is also an eigenvector of the matrix G, corresponding to the 0 eigenvalue, and has multiplicity 1. This fact follows from two observations.

1. $\ker(G) \neq \emptyset$, so $\exists v \in \mathbb{R}^m \setminus \{0\}$ such that $Gv = 0$.

This is because $\bar{1}G = 0$, which means that the rows of G are linearly dependent and thus have rank $G \leq m-1$.

2. Any v such that $Gv = 0$ satisfies $v \in$ span $\{v_{\rm P}\}$.

This follows from

$$(\phi(G)\,G + I)v = v \Rightarrow (\phi(G)\,G + I)^{m-1}v = M_G\,v = v,$$

and hence $v \in$ span $\{v_{\rm P}\}$, since $\sigma(M_G) = 1$ is an eigenvalue of M_G, of multiplicity 1.

Therefore, the kernel of G has dimension 1 and is given by

$$\ker(G) = \text{ span}\,\{v_{\rm P}(G)\}. \tag{A.1}$$

Acknowledgment. The author would like to thank Eduardo Sontag for many helpful discussions and suggestions.

REFERENCES

[1] P.W. ATKINS, *Physical Chemistry,* 5th ed., Oxford University Press, Oxford, 1994.

[2] M. ARCAK, D. ANGELI, AND E.D. SONTAG, *A unifying integral ISS framework for stability of nonlinear cascades,* SIAM J. Control Optim., 40 (2002), pp. 1888–1904.

[3] R.P. Bywater, A. Sørensen, P. Røgen, and P.G. Hjorth, *Construction of the simplest model to explain complex receptor activation kinetics,* J. Theoret. Biol., 218 (2002), pp. 139–147.
[4] M. Chaves and E.D. Sontag, *State-estimators for chemical reaction networks of Feinberg–Horn–Jackson zero-deficiency type,* Eur. J. Control, 8 (2002), pp. 343–359.
[5] M. Chaves, E.D. Sontag, and R. Dinerstein, *Steady-states of receptor–ligand dynamics: A theoretical framework,* J. Theoret. Biol., 227 (2003), pp. 413–428.
[6] M. Feinberg, *Mathematical aspects of mass action kinetics,* in Chemical Reactor Theory: A Review, L. Lapidus and N. Amundson, eds., Prentice-Hall, Englewood Cliffs, NJ, 1977, pp. 1–78.
[7] M. Feinberg, *Chemical reaction network structure and the stabiliy of complex isothermal reactors — I. The deficiency zero and deficiency one theorems*, Chem. Engrg. Sci., 42 (1987), pp. 2229–2268.
[8] M. Feinberg, *The existence and uniqueness of steady-states for a class of chemical reaction networks,* Arch. Ration. Mech. Anal., 132 (1995), pp. 311–370.
[9] F.J.M. Horn and R. Jackson, *General mass action kinetics,* Arch. Ration. Mech. Anal., 49 (1972), pp. 81–116.
[10] R.A. Horn and C.R. Johnson, *Matrix Analysis,* Cambridge University Press, New York, 1999.
[11] M. Krichman, E.D. Sontag, and Y. Wang, *Input-output-to-state stability,* SIAM J. Control Optim., 39 (2001), pp. 1874–1928.
[12] D.A. Lauffenburger and J.J. Linderman, *Receptors: Models for Binding, Trafficking, and Signaling,* Oxford University Press, New York, 1993.
[13] T.W. McKeithan, *Kinetic proofreading in T-cell receptor signal transduction,* Proc. Natl. Acad. Sci. USA, 92 (1995), pp. 5042–5046.
[14] D. Siegel and D. MacLean, *Global stability of complex balanced mechanisms,* J. Math. Chem., 27 (2000), pp. 89–110.
[15] E.D. Sontag, *Smooth stabilization implies coprime factorization,* IEEE Trans. Automat. Control, 34 (1989), pp. 435–443.
[16] E.D. Sontag, *Mathematical Control Theory: Deterministic Finite Dimensional Systems,* 2nd ed., Springer-Verlag, New York, 1998.
[17] E.D. Sontag, *Structure and stability of certain chemical networks and applications to the kinetic proofreading model of T-cell receptor signal transduction,* IEEE Trans. Automat. Control, 46 (2001), pp. 1028–1047. Errata in IEEE Trans. Automat. Control, 47 (2002), p. 705.
[18] E.D. Sontag and Y. Wang, *On characterizations of the input-to-state stability property,* Systems Control Lett., 24 (1995), pp. 351–359.
[19] P.J. Woolf, T.P. Kenakin, and J.J. Linderman, *Uncovering biases in high throughput screens of G-protein coupled receptors,* J. Theoret. Biol., 208 (2001), pp. 403–418.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 2, pp. 728–742

AN OPTIMAL OPTICAL FLOW*

KAZUFUMI ITO†

Abstract. The problem of determining optical flow for the image registration problem is discussed. Feedback solutions are proposed, and it is shown that they are optimal for certain optimal control formulations of the problem. Well-posedness of the proposed feedback solutions is analyzed, and numerical findings are presented.

Key words. optical flow, feedback solution

AMS subject classifications. 60H15, 35R60, 47H17

DOI. 10.1137/S0363012904433444

1. Introduction.

In this paper we discuss the image registration problem [8, 4]. Let $I(t,x) \geq 0$ denote the brightness of an image defined on $[0,T] \times \Omega$, where Ω is a bounded domain in R^d, $d = 2, 3$, with sufficiently smooth boundary. Consider the optical flow equation

$$\frac{\partial}{\partial t} I + V \cdot \nabla I = 0, \quad I(0) = I_0, \tag{1.1}$$

where $\nabla I = \mathrm{grad}_x I$ is the gradient of I with respect to $x \in R^d$. Let I_1 be a target image at T. The problem is to find a vector field $V = V(t,x)$ such that $I(T) = I_1$. From this optical flow V, information about the spatial arrangement of an object and its rate of change ought to be determined. Assuming that objects represented in the image are flat surfaces, that illumination is uniform, and that reflectance varies smoothly and has no spatial discontinuities, the image brightness of an object point remains constant in the images when the object moves [8]. That is, the total derivative of I vanishes, which results in (1.1). Optical flow is often a convenient and useful image motion representation, and there has been increasing literature using this approach (e.g., [8, 2, 1, 15, 4]) during the last decade.

In this paper we propose a method to construct such a vector field V, and we test and analyze the proposed methods. We follow the optimal control formulation in [4]: find V that minimizes

$$J(I,V) = \int_0^\tau \int_\Omega \left[\frac{\alpha(|\nabla I(t)|)}{q} |V(t)|^q + \frac{\beta(|\nabla I(t)|)}{p} |I(t) - I_1|^p \right] dxdt \tag{1.2}$$

subject to (1.1), where $\frac{1}{p} + \frac{1}{q} = 1$ and nonnegative functions α, β are chosen appropriately. In order to obtain a smoother vector field $V(t) = V(t,x)$ (say, in $H^1(\Omega)$), for $\delta > 0$ we also consider the following regularized optimal control formulation:

$$\begin{aligned} \min \frac{1}{2} \int_0^\tau \big[&((I - \delta\,\Delta)^{-1}(|I(t) - I_1|^{p-1}\mathrm{sign}(I(t) - I_1)\nabla I), \\ &|I - I_1|^{p-1}\mathrm{sign}(I(t) - I_1)\nabla I)_{L^2(\Omega)} + (|V|^2_{L^2(\Omega)} + \delta\,|\nabla V|^2_{L^2(\Omega)})\big]\, dt, \end{aligned} \tag{1.3}$$

*Received by the editors April 30, 2004; accepted for publication (in revised form) November 16, 2004; published electronically September 12, 2005.
http://www.siam.org/journals/sicon/44-2/43344.html
†Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC 27695-8205 (kito@math.ncsu.edu).

subject to (1.1), where Δ is the Laplace operator.

In [4] an iterative algorithm based on the Lagrange multiplier method is developed for determining the optimal vector field $V(t)$ that minimizes the cost functional, and it is tested numerically. In this paper (section 1.1) the feedback solution for (1.2) in the following forms is derived:

$$V(t,x) = \frac{1}{T-t}\,(I(t)-I_1)\Phi(|\nabla I(t)|)\,\nabla I(t) \tag{1.4}$$

and

$$V(t,x) = |I(t)-I_1|^{p-1}\mathrm{sign}(I(t)-I_1)\,\Phi(|\nabla I(t)|)\,\nabla I(t). \tag{1.5}$$

Let

$$\Psi(s) = \Phi(s)\,s^2, \quad s \in R^+,$$

and assume $\Psi : R^+ \to R^+$ is Lipschitz with $\Psi(0) = 0$. The resulting closed loop equations for I are thus given by

$$I_t + \frac{1}{T-t}\,(I(t)-I_1)\Psi(|\nabla I(t)|) = 0, \quad I(0) = I_0, \tag{1.6}$$

and

$$I_t + |I(t)-I_1|^{p-1}\mathrm{sign}(I(t)-I_1)\Psi(|\nabla I(t)|) = 0, \quad I(0) = I_0, \tag{1.7}$$

respectively. For example, we use in our testing

$$\Psi(s) = \frac{s}{\max(c,s)}, \quad \Phi(s) = \frac{1}{s\max(c,s)}$$

for some $c > 0$. If we let $\Phi(s) = \frac{1}{s}$, then (1.4)–(1.5) reduces to the geometrical motion [11, 13]

$$I_t + \frac{1}{T-t}\,(I(t)-I_1)|\nabla I(t)| = 0, \quad I(0) = I_0,$$

and

$$I_t + |I(t)-I_1|^{p-1}\mathrm{sign}(I(t)-I_1)\,|\nabla I(t)| = 0, \quad I(0) = I_0.$$

The corresponding optimal feedback law (section 1.2) for (1.3) is given by

$$V(t,x) = (I-\delta\,\Delta)^{-1}(\mathrm{sign}(I(t)-I_1)|I(t)-I_1|^{p-1}\nabla I(t)), \tag{1.8}$$

where Δ is the Laplace operator. If $\delta = 0$, then (1.8) reduces to a specific case of (1.5). Because of the nonlocal operation $(I-\delta\,\Delta)^{-1}$ for determining the vector field V at a given time t, the evaluation of (1.8) costs more in terms of its implementation compared to that for (1.5).

The proposed algorithms are of the feedback form; i.e., the vector field V is determined along with the reformed image I by integrating (1.1) with (1.5) or (1.8) in time. Thus it is an alternative to the iterative methods discussed in [8, 4] and references therein. If we solve for $X(t) = X(t;x)$ at each $x \in \Omega$,

$$\frac{d}{dt}X(t) = V(t,X(t)), \quad X(0) = x, \tag{1.9}$$

then $M(x) = X(T;x)$ defines a mapping such that $I_1(x) = I_0(M(x))$ if $I_1 = I(T)$.

In [1, 15] the (stationary optical flow) optimization of the form

$$E(h) = \int_\Omega [|I_0(x - h(x)) - I_1(x)|^2 + C\,\mathrm{trace}(\nabla h\, D(\nabla I_0)\, \nabla h)]\, dx$$

over maps $h : R^2 \to R^2$ with an appropriate matrix weight D is analyzed, and the iterative schemes for finding the optimal map ($M(x) = x - h(x)$) are developed and tested. We also refer to [2] for the alternative approaches and performance comparisons. Finally, we note that our algorithm does not require us to take the derivatives of the images I_0 and I_1 directly.

The following is the key property of the proposed algorithms. Suppose I is a Lipschitz solution of (1.6). By multiplying (1.6) by $I(t) - I_1$ and integrating in x over Ω, we obtain

$$\frac{d}{dt}\frac{1}{2}\int_\Omega |I(t) - I_1|^2\, dx + \int_\Omega \frac{1}{T-t}\, |I(t) - I_1|^2 \Psi(|\nabla I(t)|)\, dx = 0.$$

Thus, $t \to |I(t) - I_1|_{L^2(\Omega)}$ is decreasing and

$$\liminf_{t \to T} \int_\Omega |I(t) - I_1|^2 \Psi(|\nabla I(t)|)\, dx = 0.$$

Moreover, if $\Psi(|\nabla I(t,x)|) \geq \omega > 0$ a.e. in $(0,T) \times \Omega$, then

$$|I(t) - I_1|_{L^2} \leq \left(\frac{T-t}{T}\right)^\omega |I_0 - I_1|_{L^2}$$

and thus $I(T) = I_1$. Similarly, for (1.7),

$$\frac{d}{dt}\frac{1}{2}\int_\Omega |I(t) - I_1|^2\, dx + \int_\Omega |I(t) - I_1|^p \Psi(|\nabla I(t)|)\, dx = 0.$$

If $\Psi(|\nabla I(t,x)|) \geq \omega > 0$ a.e. in $(0,T) \times \Omega$ and $1 \leq p < 2$, then $|I(t) - I_1|_{L^2} = 0$ in a finite time, say $\tau > 0$. Then, we scale the time by $\frac{T}{\tau}$ to obtain $I(T) = I_1$. Similar results also hold for (1.3) and (1.8) (see section 1.2).

Remark 1. Note that any C^1 solution to (1.1) satisfies

$$\max_{x \in \Omega} I(t,x) = \max_{x \in \Omega} I_0(x)$$

for all $t \geq 0$. Thus, in order to have $I(T) = I_1$, it is necessary to have

$$\max_{x \in \Omega} I_0(x) = \max_{x \in \Omega} I_1(x).$$

Otherwise, $\Psi(|\nabla I(t,x)|) \to 0$ as $t \to \infty$ on a subset of Ω, and thus $I(t)$ does not converge to I_1 in a finite time. However, it is observed numerically that $I(t)$ converges to $\min(I_1, \max(I_0))$.

Remark 2. From the above estimate problem, (1.2) as well as (1.3) is the exit problem with $T = \tau$, and the exit time is given by

$$\tau = \inf\left\{t : \int_\Omega |\nabla I(t)|\, |I(t) - I_1|\, dx = 0\right\}.$$

That is, it is possible that it is terminated with a nontrivial subdomain $\tilde{\Omega}$ such that $|\nabla I(\tau, x)| = 0$ and $I(\tau, x) = \max_{x \in \Omega} I(\tau, x) =$ a constant on $\tilde{\Omega}$.

Remark 3. It should be noted that our problem is closely related to the so-called Monge–Kantorovich mass transport problem in the form described in [3]: find a vector field V that minimizes

$$\int_0^T \int_\Omega \rho(t,x)\,|V(t,x)|^2\,dxdt$$

subject to

$$\rho_t + \nabla\cdot(\rho\,V) = 0,$$

$$\rho(0,x) = \rho_0(x), \quad \text{and} \quad \rho(T,x) = \rho_1(x).$$

In [3] an iterative method based on the augmented Lagrangian method is developed for this optimal control problem. It is of interest here to construct a feedback solution to the problem.

Through our numerical testing we observed the following:

- Algorithms (1.5) in general are very efficient, using the numerical integrations outlined in section 2, and work very well for short-range motion including expansion and redistribution. However, for long-range motion we observed cases with a nontrivial subdomain $\tilde{\Omega}$ such that $|\nabla I(\tau,x)| = 0,\ x \in \tilde{\Omega}$.
- Algorithms (1.8) cost more in general but produce smoother motion representations. Also, they work well for long-range motion including large transport. Throughout our numerical tests we did not observe the constant subdomain cases (see Remark 2). However, the convergence of this algorithm is slower in general compared to that for (1.5).

These observations are not surprising because the level-set equation (1.6) works well for the front propagation but may result in a constant subdomain for algorithm (1.5). Because of the nonlocal operation of (1.8), constant regions are prevented, but τ can be ∞ for algorithm (1.8). In conclusion, algorithms (1.5) and (1.8) should be used in the combined manner such that (1.8) is first employed to capture the long-range motions and (1.5) is then used for faster convergence and an accurate representation for the localized motions. This combination is implemented for our numerical tests and is quite effective.

An outline of the paper is as follows. In section 2 we describe numerical integration methods we used to implement the proposed algorithms (1.5) as well as (1.8) and present our numerical tests and findings. In section 3 we present the well-posedness of the proposed algorithm. We conclude the section with the optimality of (1.5) and analysis for the regularized version (1.8).

1.1. Optimality. The feedback solution (1.5) is optimal in the following sense. Consider the optimal control problem

$$\min \quad \int_0^\tau \int_\Omega \left[\frac{\alpha(|\nabla I(t)|)}{q}\,|V(t)|^q + \frac{\beta(|\nabla I(t)|)}{p}|I(t) - I_1|^p \right] dxdt$$

subject to (1.1), where $\frac{1}{p} + \frac{1}{q} = 1$ and $\tau \geq 0$ is the exit time defined by

$$\tau = \inf\left\{ t : \int_\Omega \beta(|\nabla I|)\,|I(t) - I_1|^p = 0 \right\}.$$

Nonnegative functions $\alpha,\ \beta$ should be chosen so that

$$\frac{\beta(s)}{s} = \left(\frac{\alpha(s)}{s}\right)^{1-p}$$

and $\beta = \Psi$, where Ψ appears in (1.6)–(1.7). For example, we have the specific cases

$$\alpha(s) = 1, \quad \beta(s) = s^p, \quad \text{and} \quad \alpha(s) = \beta(s) = s.$$

In our experiments we used

$$\beta(s) = \Psi(s) = \frac{s}{\max(c, s)} = \begin{cases} \dfrac{s}{c} & \text{if } s \le c, \\ \\ 1 & \text{if } s \ge c \end{cases}$$

for some $c > 0$, and the corresponding α is given by

$$\alpha(s) = \begin{cases} c^{q-1} s & \text{if } s \le c, \\ \\ s^q & \text{if } s \ge c. \end{cases}$$

First, we claim that $\mathcal{V}(I) = \frac{1}{2} \int_\Omega |I(x) - I_1(x)|^2$ satisfies the formal Hamilton–Jacobi equation

$$\min_V \int_\Omega \left[-(\mathcal{V}_I, V \cdot \nabla I) + \frac{\alpha(|\nabla I|)}{q} |V|^q + \frac{\beta(|\nabla I|)}{p} |I - I_1|^p \right] dx = 0, \tag{1.10}$$

where $(\mathcal{V}_I, \phi) = (I - I_1, \phi)$. In fact, without loss of generality we let $V = c \frac{\nabla I}{|\nabla I|}$ for almost every x in Ω. Then

$$J = -(\mathcal{V}_I, V \cdot \nabla I) + \frac{\alpha(|\nabla I|)}{q} |V|^q = -c|\nabla I| \, (I - I_1) + \frac{\alpha(|\nabla I|)}{q} |c|^q \tag{1.11}$$

is minimized when

$$c = (|\nabla I| \alpha^{-1}(|\nabla I|)|I - I_1|)^{\frac{1}{q-1}} \operatorname{sign}(I - I_1)$$

and

$$\min \quad J = -\frac{1}{p} |\nabla I|^p \alpha^{-\frac{1}{q-1}}(|\nabla I|)|I - I_1|^p$$

for almost every x in Ω. Since $\frac{1}{q-1} = p - 1$, if we let

$$\Psi(s) = \beta(s) = s^p \alpha(s)^{1-p} \quad \text{and} \quad \Phi(s) = s^{p-2} \alpha(s)^{1-p},$$

then the feedback form

$$V = V(I) = |I - I_1|^{p-1} \operatorname{sign}(I - I_1) \Phi(|\nabla I|) \, \nabla I$$

as defined by (1.5) attains the minimum in (1.11), and $\mathcal{V}$ satisfies (1.10).

Next, we argue the optimality. Let $W \in W^{1\infty}((0, \tau) \times \Omega)$. Then (1.1) has a unique solution $I \in W^{1,\infty}((0, \tau))$; e.g., see [4]. We assume that τ is an exit time. Since $V = V(I)$ as above minimizes J a.e. in Ω, thus

$$\begin{aligned} \frac{d}{dt} \mathcal{V}(I(t)) &= -(I - I_1, W \cdot \nabla I) \\ &= \int_\Omega - \left[\frac{\alpha(|\nabla I|)}{q} |W|^q + \frac{\beta(|\nabla I|)}{p} |I - I_1|^p - \delta(W, V(I)) \right] dx, \end{aligned}$$

where $\delta(\hat{W}, V(I)) > 0$ if $\hat{W} \in R^d$ and $\delta(\hat{W}, V(I)) = 0$ if $\hat{W} = V(I)$. Hence we have

$$\int_0^\tau \int_\Omega \left[\frac{\alpha(|\nabla I|)}{q} |W|^q + \frac{\beta(|\nabla I|)}{p} |I - I_1|^p \right] dxdt$$

$$= \int_\Omega \frac{1}{2} |I_0 - I_1|^2 \, dx + \int_0^\tau \int_\Omega \delta(W, V(I)) \, dxdt.$$

Therefore (1.5) is optimal.

1.2. Regularization. In this section we discuss the regularized versions of (1.3) and (1.8). First, we can show that $\mathcal{V}(I) = \frac{1}{p} \int_\Omega |I(x) - I_1(x)|^p$ satisfies the (formal) Hamilton–Jacobi equation

(1.12)
$$\min_V \int_\Omega \Big[-(\mathcal{V}_I, V \cdot \nabla I) + \frac{1}{2} (|V|^2 + \delta |\nabla V|^2)$$
$$+ \frac{1}{2} ((I - \delta \Delta)^{-1} (|I - I_1|^{p-1} \text{sign}(I(t) - I_1) \nabla I), |I - I_1|^{p-1} \text{sign}(I(t) - I_1) \nabla I) \Big] \, dx = 0,$$

where $(\mathcal{V}_I, \phi) = (|I - I_1|^{p-1} \text{sign}(I - I_1), \phi)$. In fact, it is easy to see that

$$V = (I - \delta \Delta)^{-1} (\text{sign}(I - I_1) |I - I_1|^{p-1} \nabla I)$$

attains the minimum (0) of the quadratic form in (1.12). Thus, the corresponding optimal feedback law is given by (1.8), and its optimality can be argued using exactly the same arguments as in section 1.1.

In general we consider

$$V(t, x) = G(|I(t) - I_1|^{p-1} \text{sign}(I(t) - I_1) \nabla I(t)), \tag{1.13}$$

where

$$G : L^2(\Omega)^d \to L^2(\Omega)^d \quad \text{is a bounded, positive operator.}$$

For example,

$$GV = k_0 V(x) + \int_\Omega k(|x - y|) V(y) \, dy,$$

with a smoothing kernel $k \geq 0$ and $k_0 \geq 0$. For (1.8) we have

$$G = (I - \epsilon \Delta)^{-1}.$$

Suppose that I is a Lipschitz solution of (1.1) with (1.13). Then

$$\frac{d}{dt} \frac{1}{p} \int_\Omega |I(t) - I_1|^p \, dx + (G(|I - I_1|^{p-1} \text{sign}(I - I_1) \nabla I), |I - I_1|^{p-1} \text{sign}(I - I_1) \nabla I) = 0.$$

Thus, $t \to |I - I_1|_{L^p}$ is decreasing. Moreover, if $(GV, V)_{L^2} \geq \gamma |V|^2_{L^2}$ for some $\gamma > 0$, then

$$\frac{d}{dt} \frac{1}{p} \int_\Omega |I(t) - I_1|^p \, dx + \int_\Omega \gamma |I(t) - I_1|^{2p-2} |\nabla I|^2 \, dx \leq 0.$$

Hence, if $|\nabla I(t)|^2 \geq \omega > 0$ a.e. and $1 \leq p < 2$, then $|I(t) - I_1|_{L^p} = 0$ in a finite time $\tau > 0$.

2. Numerical integration and testing. In this section we discuss the numerical integration of the proposed methods and present testing results. We use the Gudnov-type scheme (see, e.g., [12, 16]) for the Hamilton–Jacobi equation on a fixed Cartesian grid with uniform mesh-size h of the square $\Omega = (0,1) \times (0,1)$, and time step-size $\Delta t > 0$ (satisfying the CFL condition); i.e.,

$$0 = \frac{I^{k+1} - I^k}{\Delta t} + c_k$$
$$\begin{cases} \Psi\left(\sqrt{[\max((D_x^-)_{i,j}I^k, -(D_x^+)_{i,j}I^k, 0)]^2 + [\max((D_y^-)_{i,j}I^k, -(D_y^+)_{i,j}I^k, 0)]^2}\right) \\ \qquad \text{for } \ c_k > 0, \\ \Psi\left(\sqrt{[\min((D_x^-)_{i,j}I^k, -(D_x^+)_{i,j}I^k, 0)]^2 + [\min((D_y^-)_{i,j}I^k, -(D_y^+)_{i,j}I^k, 0)]^2}\right) \\ \qquad \text{for } \ c_k < 0 \end{cases}$$

with

$$c_k = |I^k - I_1|^{p-1}(I^k - I_1).$$

If we use the first order backward and forward difference

$$(D_x^-)_{i,j}I = \frac{I_{i,j} - I_{i-1,j}}{h}, \quad (D_x^+)_{i,j}I = \frac{I_{i+1,j} - I_{i,j}}{h},$$

$$(D_y^-)_{i,j}I = \frac{I_{i,j} - I_{i,j-1}}{h}, \quad (D_y^+)_{i,j}I = \frac{I_{i,j+1} - I_{i,j}}{h},$$

then this is a monotone scheme, and its convergence can be argued, for example, as in [7, 12].

We use the third order WENO (weighted essential nonoscillatory) scheme [10] to evaluate the forward and backward differences $(D_x^-)_{i,j}I^k$, $(D_x^+)_{i,j}I^k$ and $(D_y^-)_{i,j}I^k$, $(D_y^+)_{i,j}I^k$. It is advantageous to use the higher order scheme to obtain accurate spatial discretization and reduce the CFL number requirement for the time step-size Δt. We refer to [11] and references therein for further discussion on approximation methods for the Hamilton–Jacobi equation. For the implementation of (1.1) with (1.8) we use the upwinding method based on WENO differences.

2.1. Test results. In our calculations we use (1.5) and (1.7) with $p = \frac{5}{4}$ and

$$\Psi(|\nabla I|) = \frac{|\nabla I|}{\max(c, |\nabla I|)},$$

where $c > 0$ is appropriately chosen (we choose $c = 0.5$ in our computations). Let $\Omega = (0,1) \times (0,1)$ and set $h = 0.01$.

Example 1. The first example [4] is

$$I(t) = \begin{cases} \frac{2}{3}(x+y) - \frac{1}{3}\,t & \text{if } x + y \geq 1, \\ \frac{1}{3}(x+y) + \frac{1}{3} - \frac{1}{3}\,t & \text{if } x + y \leq 1, \end{cases}$$

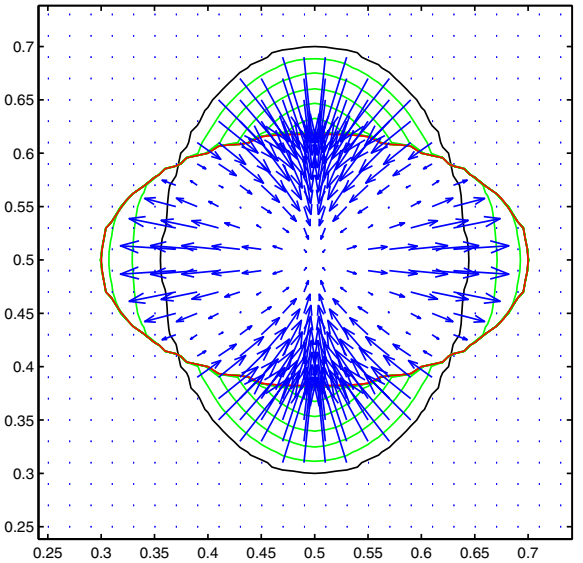


FIG. 1. *Evolution of deforming images from I_0 to I_1, using algorithm* (1.5).

and $I_0 = I(0)$ and $I_1 = I(1)$. In this example the constant field

$$V_0 = \begin{cases} \left(\frac{1}{2}, \frac{1}{2}\right) & \text{if } x + y > 1, \\ (1,1) & \text{if } x + y \le 1 \end{cases}$$

is a solution to problem (1.1). Our algorithm (1.5) with $c \le \frac{\sqrt{2}}{3}$ produces a solution

$$V(t,x) = c(t) \begin{cases} \left(\frac{1}{2}, \frac{1}{2}\right) & \text{if } x + y > 1, \\ (1,1) & \text{if } x + y \le 1, \end{cases}$$

with

$$c(t) = \frac{9}{2}\left(\left(\frac{1}{3}\right)^{\frac{3}{4}} - \frac{3}{4}t\right)^{\frac{1}{3}}.$$

Example 2. This example is a redistribution of image I_0 defined on an ellipse to image I_1 defined on a rotated and resized ellipse:

$$I_0 = \max(0, .041 - 2(x - .5)^2 - (y - .5)^2),$$

$$I_1 = \max(0, .041 - (x - .5)^2 - 3(y - .4)^2).$$

We compare the resulting motions using algorithms (1.5) and (1.8). In Figure 1 we show level curves $\{x \in \Omega : I(t,x) = .001\}$ at uniform time units and the net motion by arrows. The net motion is defined by $M(x) - x \in R^2$ on the support of

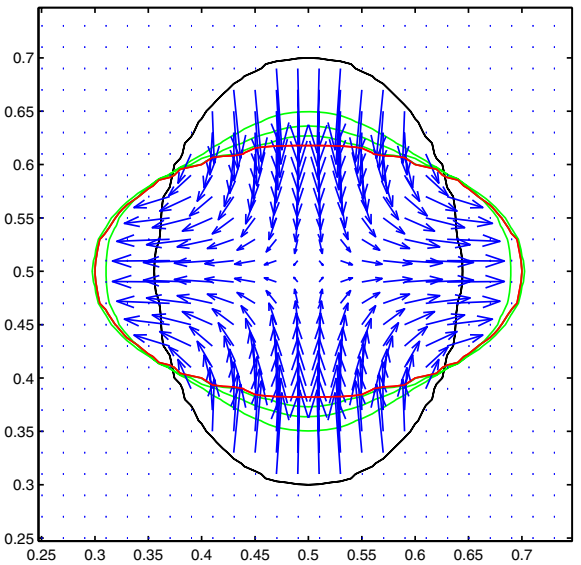


FIG. 2. *Evolution of deforming images from* I_0 *to* I_1*, using algorithm* (1.8).

I_0. The actual trajectory $X(t;x)$ from (1.9) is an oriented curved arc. In Figure 2 the corresponding results for algorithm (1.8) are shown. Algorithm (1.8) is actually terminated by algorithm (1.5) as described in introduction. Algorithm (1.8) results in a smoother motion representation.

For the remaining examples we used the combined algorithms as described in the introduction, i.e., algorithm (1.8) terminated by algorithm (1.5).

Example 3. This example is the same as in Example 2, with noise in I_1. We add i.i.d. (independently and identically distributed) noise uniformly distributed on $.041\,(-.1,.1)$ to I_1 at each pixel. We apply prefilter I_1 by $(I - 1.e^{-4}\,\Delta)^{-1}\,I_1$. It is observed that the final $I(\tau)$ deformed image from I_0 is much smoother that the target prefiltered noisy image I_1 due to the smooth vector field $V(\cdot,x) \in H^2(\Omega)$. Effects of noise can be noticed in Figure 3 compared to Figure 2 (noise-free) but the algorithm performs very well with large noise in the data.

Example 4. This example is for the transport and the redistribution of an image I_0 defined on an ellipse to I_1 defined on a circle:

$$I_0 = \max(0, .041 - 2(x-.5)^2 - (y-.5)^2),$$

$$I_1 = \max(0, .041 - 4(x-.6)^2 - 4(y-.4)^2).$$

The resulting level curves $\{x \in \Omega : I(t,x) = .001\}$ at uniform time units and the net motion by arrows are shown in Figure 4.

Example 5. This example is for the disjointly supported image I_0:

$$I_0 = \max(0, .041 - 4(x-.6)^2 - 4(y-.4)^2, .041 - 4(x-.4)^2 - 4(y-.6)^2)$$

merging into the connected image I_1 as $I_1 = I_0$ in Example 2. In Figure 5 the two disjointed level-sets $\{x \in \Omega : I(t,x) = .005\}$ merge into the center of I_1. The support of $I(t,x)$ remains disconnected under algorithm (1.8) and attached by the algorithm (1.5) phase.

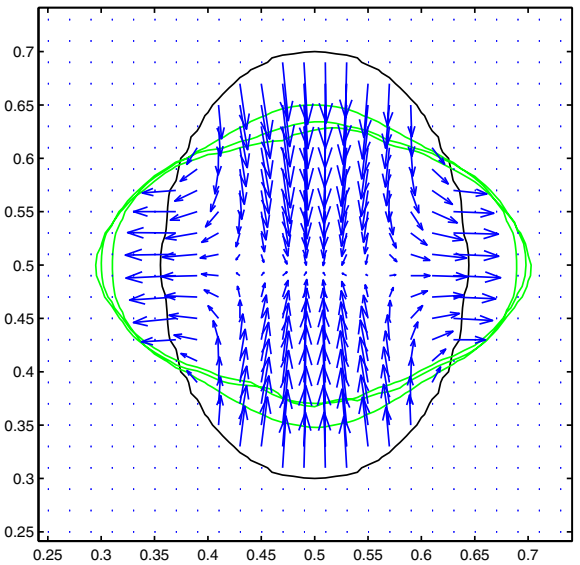


FIG. 3. *Evolution of deforming images from I_0 to I_1, with noise.*

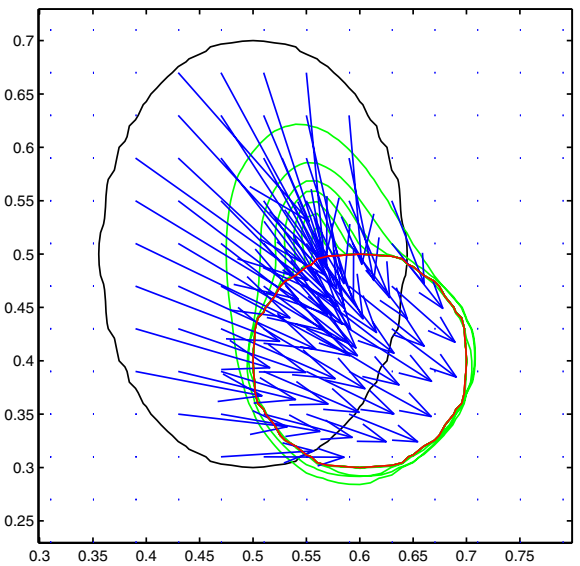


FIG. 4. *Evolution of deforming images from I_0 to I_1, for Example* 4.

Example 6. We tested our algorithm against a consecutive frame of the Yosemite fly-through image sequences by Lynn Quann at SRI. The sequence is generated by taking an aerial image of Yosemite valley and texture mapping it onto a depth map of the valley. It is a benchmark for optimal flow methods. The interest of this test is that one can give a quantitative evaluation of optimal flow methods; for example, the web page http://www.cs.brown.edu/people/black/ contains the files with ground truth flow fields and hints on how the errors should be computed. We have 256×256

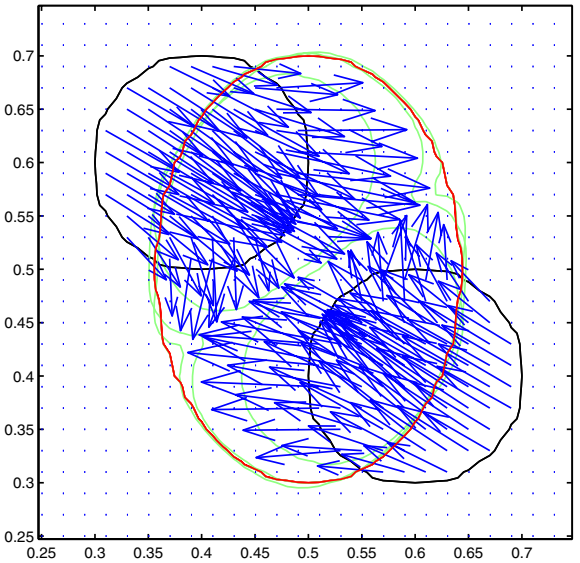


FIG. 5. *Evolution of deforming images from I_0 to I_1 with geometrical change.*

TABLE 1
Comparison between results reported in [2, 1] *with* 100% *density and our result.*

Technique	Average Error	Standard Deviation
Horn and Schunck (original)	31.69°	31.18°
Horn and Schunck (modified)	9.78°	16.19°
Nagel	10.22°	16.51°
Anandan (unthresholded)	13.36°	15.64°
Uras et al. (unthresholded)	8.94°	15.61°
Singh (step 2)	10.03°	13.13°
Alvarez et al.	5.53°	7.40°
Our method	9.24°	10.77°

pixels, and we select $\Delta x = 1$ and normalize the images so that the maximum intensity is one. The sequence of images used for our testings and the resulting net motion are shown in Figure 6. The correct flow field and the estimated flow field are plotted on the top of each other for comparison.

The angular error is calculated in the same way as in [2] using

$$Error = \arccos\left(\frac{u_c u_e + v_c v_e + 1}{\sqrt{(u_c^2 + v_c^2 + 1)(u_e^2 + v_e^2 + 1)}}\right),$$

where (u_c, v_c) denotes the correct flow and (u_e, v_e) is the estimated flow. Our method performs better than the other methods except the one in [1] (see Table 1). As you can see from Figure 6, the majority of errors happen in the cloud region. We have the angular error 4.36° with standard deviation 3.86° if we exclude the cloud region.

3. Well-posedness. In this section we show the well-posedness of the proposed algorithm (1.5). A similar analysis can be done for algorithm (1.8) without much modification. For the clarity of our discussions we assume $\Omega = (0, 1)^d$, and we discuss the periodic boundary condition for I throughout this section. First we consider the

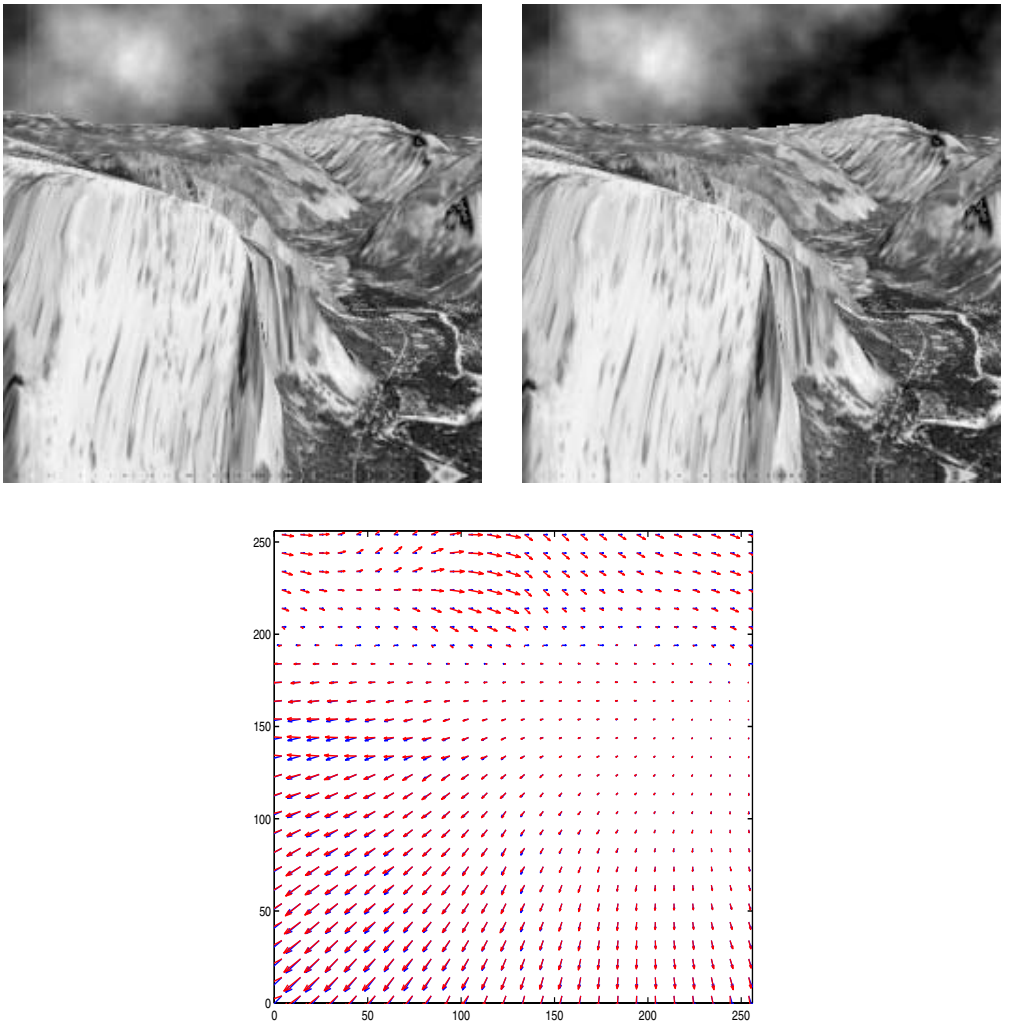


FIG. 6. *Yosemite fly-through images and optical flow.*

regularized problem

$$I_t + \phi(I - I_1)\psi(|\nabla I|) = \epsilon\,\Delta I, \quad I(0) = I_0, \tag{3.1}$$

where $\epsilon > 0$, $\phi : R \to R$ is a monotone increasing bounded Lipschitz function with $\phi(0) = 0$, and $\psi : R^+ \to R^+$ is a Lipschitz continuous function with $\psi(0) = 0$.

Let

$$X = \{I \in H^1(\Omega) : I \text{ is } \Omega \text{ periodic}\},$$

$$V = \{I \in H^2(\Omega) : I,\ \nabla I \text{ are } \Omega \text{ periodic}\}.$$

It can be proved (e.g., see [9]) that A on X defined by

$$AI = -\epsilon\Delta I + \phi(I - I_1)\psi(|\nabla I|)$$

satisfies

$$(AI_1 - AI_2, I_1 - I_2)_X \geq \frac{\epsilon}{2}|I_1 - I_2|_V^2 - \omega\,|I_1 - I_2|_X^2$$

for some $\omega > 0$ and $A : V \to V^* = L^2(\Omega)$ maximum [14], where X is the pivoting space. Note that for $I \in V$

$$\nabla[\phi(I - I_1)\psi(|\nabla I|)] = \phi'(I - I_1)\,(J - \bar{J})\psi(|J|) + \phi(I - I_1)\psi'(|J|)\frac{J}{|J|}\cdot\nabla J, \tag{3.2}$$

where $J = \nabla I$ and $\bar{J} = \nabla I_1$. Thus,

$$|\phi(I(\cdot) - I_1)\psi(|\nabla I(\cdot)|)|_{L^2(0,T;X)} \leq M\,|I(\cdot)|_{L^2(0,T;V)}$$

for some M. It follows from [14] that there exists a unique solution $I(\cdot)$ to (3.1) in $H^1(0,T;X) \cap L^2(0,T;H^3(\Omega)) \cap C(0,T;H^2(\Omega))$.

Let $q \geq 2$. Multiplying (3.1) by $|I|^{q-2}I$ and integrating in x over Ω, we obtain

$$\frac{d}{dt}\int_\Omega |I|^q\,dx + \int_\Omega [q\,|I|^{q-2}I\phi(I - I_1)\psi(|\nabla I|) + q(q-1)\epsilon\,|I|^{q-2}|\nabla I|^2]\,dx = 0.$$

Since $|\phi(I - I_1)\psi(|\nabla I|)| \leq \gamma\,|\nabla I|$ for some γ in Ω, thus

$$\frac{d}{dt}\int_\Omega |I|^q\,dx \leq \int_\Omega \frac{q\gamma^2}{4(q-1)\epsilon}|I|^q\,dx.$$

Hence

$$|I(t)|_{L^q(\Omega)} \leq e^{\frac{\gamma^2}{4(q-1)\epsilon}t}\,|I(0)|_{L^q(\Omega)}.$$

By letting $q \to \infty$, we obtain

$$|I(t)|_{L^\infty(\Omega)} \leq |I(0)|_{L^\infty(\Omega)}. \tag{3.3}$$

Note that, from (3.2), $J = \nabla I$ satisfies

$$J_t + \phi'(I - I_1)\,(J - \bar{J}) + \phi(I - I_1)\psi'(|J|)\frac{J}{|J|}\cdot\nabla J = \epsilon\,\Delta J. \tag{3.4}$$

Let $q \geq 2$. Multiplying (3.4) by $q\,|J|^{q-2}J$ and integrating over Ω, we obtain

$$\frac{d}{dt}\int_\Omega |J|^q\,dx + \int_\Omega [q\,|J|^{q-2}(E_1 + E_2)\cdot J\,dx - q\epsilon\,|J|^{q-2}(J\cdot\Delta J)]\,dx = 0,$$

where

$$E_1 = \phi'(I - I_1)\,(J - \bar{J})\psi(|J|),$$

$$E_2 = \phi(I - I_1)\psi'(|J|)\frac{J}{|J|}\cdot\nabla J.$$

By Green's theorem,

$$-\int_\Omega |J|^{q-2}(J\cdot\Delta J)\,dx = \int_\Omega \left[|J|^{q-2}|\nabla J|^2 + \frac{q-2}{2}\,|J|^{q-4}|\nabla|J|^2|^2\right]dx = 0.$$

Let $|\psi'| \le c_1$ and $0 \le \phi' \le c_2$. Note that

$$|J|^{q-2}J \cdot E_2 \le \frac{c_1}{2}|J|^{q-2}|\nabla|J|^2| \le \epsilon\,\frac{q-2}{2}\,|J|^{q-4}|\nabla|J|^2|^2 + \frac{c_1^2}{4\epsilon(q-2)}|J|^q$$

and

$$|J|^{q-2}J \cdot E_1 = \phi'(|J|^q - |J|^{q-2}J \cdot \bar{J})\psi(|J|) \le c_1c_2|\bar{J}|_\infty|J|^q.$$

Hence we have

$$\frac{d}{dt}\int_\Omega |J|^q\,dx \le \int_\Omega q\left(c_1c_2|\bar{J}|_\infty + \frac{c_1^2}{4\epsilon(q-2)}\right)|J|^q\,dx$$

and thus

$$|J(t)|_{L^\infty(\Omega)} \le e^{c_1c_2|\bar{J}|_\infty t}|J(0)|_{L^\infty(\Omega)}. \tag{3.5}$$

Similarly, for $\dot{I} = \frac{dI}{dt}$ we have

$$\dot{I}_t + \phi'(I - I_1)\,\dot{I}\psi(|J|) + \phi(I - I_1)\psi'(|J|)\frac{J}{|J|} \cdot \nabla\dot{I} = \epsilon\,\Delta\dot{I},$$

assuming $I_0 \in V$. By using the same arguments as above, it can be shown that

$$|\dot{I}(t)|_{L^\infty(\Omega)} \le |\dot{I}(0)|_{L^\infty(\Omega)}. \tag{3.6}$$

From estimate (3.2) it is not necessary to assume that ϕ is bounded. Moreover, in the above it is implicitly assumed that $\psi'(|J|)\frac{J}{|J|}$ is a.e. defined. For the case $\psi(|J|) = |J|$ it is not necessary that such a derivative exists. Thus, we consider a family of regularizations $\psi_\delta(|J|) = \sqrt{\delta^2 + |J|^2} - \delta$ for $\delta > 0$. It can be shown that the estimates (3.3), (3.5), and (3.6) hold uniformly in $\delta > 0$. Let $\{I_\delta\}$ be the corresponding solution to (3.1) with $\psi = \psi_\delta$ and a fixed $\epsilon > 0$. Then $\{I_\delta\}$ is bounded in $W^{1,\infty}((0,T)\times\Omega)\cap L^2(0,T;V)$. Hence it has a weak star–weak convergent subsequence to the limit $I \in W^{1,\infty}((0,T)\times\Omega)\cap L^2(0,T;V)$ as $\delta \to 0^+$, and I is the unique solution to (3.1).

Now we prove that a family of functions $\{I^\epsilon\}$, which are the solution to (3.1) for $\epsilon > 0$, has a convergent subsequence to the limit $I \in W^{1,\infty}((0,T)\times\Omega)$ as $\epsilon \to 0^+$ and that I is a viscosity solution [6, 5] to

$$I_t + \phi(I - I_1)\psi(|\nabla I|) = 0, \quad I(0) = I_0. \tag{3.7}$$

That is, for all $\zeta \in C^1((0,T)\times\Omega)$ if $V - \zeta$ attains a local maximum at $(t_0, x_0) \in (0,T)\times\Omega$, then

$$\zeta_t + \phi(I - I_1)\psi(|\nabla\zeta|) \le 0 \quad \text{at } (t_0, x_0), \tag{3.8a}$$

and if $V - \zeta$ attains a local minimum at $(t_0, x_0) \in (0,T)\times\Omega$, then

$$\zeta_t + \phi(I - I_1)\psi(|\nabla\zeta|) \ge 0 \quad \text{at } (t_0, x_0). \tag{3.8b}$$

First, we note that from (3.3), (3.5)–(3.6)

$$|I^\epsilon|_{W^{1,\infty}((0,T)\times\Omega)} \quad \text{is bounded uniformly in } \epsilon > 0.$$

Thus there exists a subsequence of I^ϵ (denoted by the same symbol) that converges to $I \in W^{1,\infty}((0,T)\times\Omega)$, where the convergence is uniform in $(0,T)\times\Omega$. Next, it can be shown (see, e.g., [9]) that for all $\zeta \in C^2((0,T)\times\Omega)$ if $I^\epsilon - \zeta$ attains a local maximum (minimum, respectively) at $(t_0,x_0)\in R^n$, then

$$\zeta_t + \phi(I^\epsilon - I_1)\psi(|\nabla\zeta|) - \epsilon\,\Delta\zeta \le 0 \quad (\ge 0, \text{ respectively}) \tag{3.9}$$

at (t_0,x_0). We prove (3.8a) first for $\zeta \in C^2((0,T)\times\Omega)$. Assume that for $\zeta \in C^2((0,T)\times\Omega)$ $V^\epsilon - \zeta$ has a local maximum at $x_0\in\Omega$. We can choose $\xi \in C^\infty((0,T)\times\Omega)$ such that $\nabla\xi(t_0,x_0)=0$ and $I^\epsilon-(\zeta-\xi)$ has a strict local maximum at (t_0,x_0). For $\epsilon>0$ sufficiently small, $I^\epsilon-(\zeta-\xi)$ has a local maximum at some $(t_\epsilon,x_\epsilon)\in(0,T)\times\Omega$ and $(t_\epsilon,x_\epsilon)\to(t_0,x_0)$ as $\epsilon\to 0^+$. From (3.9),

$$\zeta_t + \phi(I^\epsilon - I_1)\psi(|\nabla\zeta|) - \epsilon\,\Delta\zeta \le 0$$

at (t_ϵ,x_ϵ). We conclude (3.8a), since $I^\epsilon(t_\epsilon,x_\epsilon)\to I(t_0,x_0)$, $\nabla\zeta(t_\epsilon,x_\epsilon)-\nabla\xi(t_\epsilon,x_\epsilon)\to\nabla\zeta(t_0,x_0)-\nabla\xi(t_0,x_0)=\nabla\zeta_x(t_0,x_0)$, and $\epsilon\,\Delta(\zeta-\xi)_{xx}(t_\epsilon,x_\epsilon)\to 0$ as $\epsilon\to 0$. For $\zeta\in C^1((0,T)\times\Omega)$ exactly the same argument is applied to the convergent sequence $\zeta_n\in C^2((0,T)\times\Omega)$ to ζ in $C^1((0,T)\times\Omega)$ to prove (3.8a).

The uniqueness of the viscosity solution to (3.7) is established in [6, 5].

THEOREM. *Assume that $\phi: R\to R$ is a monotone increasing Lipschitz function with $\phi(0)=0$, and that $\psi: R^+\to R^+$ is a Lipschitz continuous function with $\psi(0)=0$. Equation* (3.7) *has a unique viscosity solution $I\in W^{1,\infty}((0,T)\times\Omega)$, provided that $I_0, I_1\in W^{1,\infty}(\Omega)$.*

REFERENCES

[1] L. ALVAREZ, J. WEICKERT, AND J. SÁNCHEZ, *Reliable estimation of dense optical flow fields with large displacements*, Int. J. Computer Vision, 39 (2000), pp. 41–56.

[2] J. L. BARRON, D. J. FLEET, AND S. BEAUCHEMIN, *Performance of optical flow techniques*, Int. J. Computer Vision, 12 (1994), pp. 43–77.

[3] J. D. BENAMOU AND Y. BRENIER, *A computational fluid mechanics solution to the Monge–Kantrovich mass transfer problem*, Numer. Math., 84 (2000), pp. 375–393.

[4] A. BOZZÌ, K. ITO, AND K. KUNISCH, *Optimal control formulation for determining optical flow*, SIAM J. Sci. Comput., 24 (2002), pp. 818–847.

[5] M. G. CRANDALL, L. C. EVANS, AND P. L. LIONS, *Some properties of viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.

[6] M. G. CRANDALL AND P. L. LIONS, *Viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[7] M. G. CRANDALL AND P. L. LIONS, *Two approximations of solutions of Hamilton–Jacobi equations*, Math. Comp., 43 (1984), pp. 1–19.

[8] B. K. P. HORN AND B. G. SCHUNCK, *Determining optical flow*, Artificial Intelligence, 17 (1981), pp. 185–204.

[9] K. ITO, *Existence of solutions to Hamilton–Jacobi–Bellman equation under quadratic growth conditions*, J. Differential Equations, 176 (2001), pp. 1–28.

[10] G.-S. JIANG AND D. PENG, *Weighted ENO schemes for Hamilton–Jacobi equations*, SIAM J. Sci. Comput., 21 (2000), pp. 2126–2143.

[11] S. OSHER AND C.-W. SHU, *High-order essentially nonoscillatory schemes for Hamilton–Jacobi equations*, SIAM J. Numer. Anal., 28 (1991), pp. 907–922.

[12] E. ROUY AND A. TOURIN, *A viscosity solutions approach to shape-from-shading*, SIAM Numer. Anal., 29 (1992), pp. 867–884.

[13] J. A. SETHIAN, *Level Set Method, Evolving Interfaces in Geometry, Fluid Mechanics, Computer Vision, and Material Science*, Cambridge University Press, Cambridge, UK, 1996.

[14] H. TANABE, *Equations of Evolution*, Pitman, London, 1979.

[15] J. WEICKERT AND C. SCHNÖRR, *A theoretical framework for convex regularizers in PDE-based computation of image motion*, Int. J. Computer Vision, 45 (2001), pp. 245–264.

[16] H. ZHAO, T. CHAN, B. MERRIMAN, AND S. OSHER, *A variational level set approach to multiphase motion*, J. Comput. Phys., 127 (1996), pp. 179–195.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 2, pp. 743–775

OPTIMIZATION PROBLEMS FOR CURVED MECHANICAL STRUCTURES*

VIOREL ARNĂUTU†, JÜRGEN SPREKELS‡, AND DAN TIBA‡§

Abstract. We study optimal design problems for three-dimensional curved rods and for shells under minimal regularity assumptions for the geometry. The results that we establish concern the existence of optimal shapes and the sensitivity analysis. We also compute some numerical examples for the optimization of curved rods. The models used have been investigated in our previous works [A. Ignat, J. Sprekels, and D. Tiba, *Math. Methods Appl. Sci.*, 25 (2002), pp. 835–854], [J. Sprekels and D. Tiba, *Adv. Math. Sci. Appl.*, 12 (2002), pp. 175–190]. A complete study of Kirchhoff–Love arches and of related minimization questions has been performed in [A. Ignat, J. Sprekels, and D. Tiba, *SIAM J. Control Optim.*, 40 (2001), pp. 1107–1133].

Key words. shells and curved rods, minimal regularity, optimal design

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

DOI. 10.1137/S0363012903426252

1. Introduction.

The scientific literature concerning the modeling of curved mechanical structures currently offers a variety of mathematical models for the study of the deformation of such elastic bodies under the impact of various types of internal or external forces and tractions. We refer just to the monographs of Ciarlet [11], Trabucho and Viaño [19], and Antman [2], where very rich material can be found for investigations in this direction.

It is a natural question now to develop a research program concerning the optimization of such objects, including numerical experiments. It should be mentioned that there exist several works of interest discussing such problems, including Chenais and Rousselet [10], Rousselet [16], Myslinski, Piekarski, and Rousselet [14], Sprekels and Tiba [17], and Ignat, Sprekels, and Tiba [12].

In this article, we attempt an analysis of general optimization problems associated with curved rods and shells. The generality of our setting is related to the consideration of a general performance index, of general constraints on the geometry, the relaxation of the regularity assumptions, and the implementation of some numerical experiments. In particular, we are assuming just C^2-regularity, instead of the usual C^3-hypotheses from literature. For shells, we obtain this by using the generalized Naghdi-type model introduced in Sprekels and Tiba [18]. For rods, this is achieved by replacing the classical Frenet frame with a new general algebraic construction that will be introduced in section 2. Other variants of local coordinates systems associated with three-dimensional (3D) curves under low regularity conditions may be found in Cartan [7] (the Darboux frame) or in Ignat, Sprekels, and Tiba [13], from which we

*Received by the editors April 21, 2003; accepted for publication (in revised form) October 13, 2004; published electronically September 12, 2005. This research was supported by the DFG Research Center "Mathematics for key technologies" (FZT 86) in Berlin.

http://www.siam.org/journals/sicon/44-2/42625.html

†Department of Mathematics, University "Al. I. Cuza," RO–6600 Iaşi, Romania (varnautu@uaic.ro).

‡Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstrasse 39, D–10117 Berlin, Germany (sprekels@wias-berlin.de, tiba@wias-berlin.de).

§Institute of Mathematics, Romanian Academy, P. O. Box 1–764, RO–70700 Bucharest, Romania (dtiba@imar.ro).

take the linear model used here. It consists of a system of nine ODEs with null boundary conditions, written in the weak form, which corresponds to the clamped rods case. Other boundary conditions may also be considered. Comparing with the regularity assumptions from the modeling process, we see that the optimization hypotheses are minimal. Let us also mention that both models use the Timoshenko assumption and allow for shear and Poisson effects (i.e., the deformation of the cross section). They generalize the classical Naghdi model for shells, Ciarlet [11], and the model studied by Arunakirinathar and Reddy [3] and Chapelle [8] for curved rods.

Our approach allows us to minimize within the class of curved rods of a prescribed length, which is a natural condition in applications. The prescribed length is also preserved by the variations employed here, according to section 4. We also show how to avoid certain degenerate cases: rods of zero length or with multiple points (see (2.15) below).

Also note that, besides the fact that we have general constraints on the geometry, in certain important examples the parametrization used here allows us to reexpress them in a convex way. The optimization problems considered in this paper are nonconvex, but the convexity of the constraint set is very helpful in the numerical experiments.

The plan of the paper is as follows. We start with the theoretical discussion of optimization problems for curved rods. In section 2, we indicate the necessary preliminaries and the formulation of the problem. In section 3, we prove the existence of the solution (while uniqueness is not valid, in general), and in section 4 we perform a sensitivity study.

A similar program is carried out in sections 5, 6, and 7, in connection with the study of optimal shell configurations. Our basic assumption is that the geometry of the shell can be described by the graph of some mapping in $C^2(\bar{o})$, where $o \subset \mathbb{R}^2$ is a bounded Lipschitz domain; that is, the use of local charts is avoided. While this setting still allows for many applications, it is also helpful as it reduces the complexity of the problem and of the notation.

We underline that, in order to prove the existence of optimal shapes, coercivity inequalities of Korn type, which are uniform with respect to the geometry, have to be established (in sections 3 and 6). In particular, in the case of shells the extension property in Lipschitz domains plays an essential role (see Adams [1]).

In the last section, we present some numerical experiments for optimization problems for 3D curved rods. We underline that it is rather difficult to construct academic examples for geometric optimization problems in three dimensions that also allow a mechanical interpretation. Their computational complexity is quite large, and it seems that here such 3D examples are solved for the first time in the scientific literature. Moreover, the fact that the computed optimal solution has a clear physical interpretation in some of the examples provides a strong validation of the model and of the optimization and approximation methods used here. The numerical treatment of the optimization of shells (which is not considered here) requires very special numerical approximation methods.

2. Description of the curved rods problem. Let $\bar{\theta} = (\theta_1, \theta_2, \theta_3) \in C^k[0, L]^3$, $k \in \mathbb{N}$, be a 3D Jordan curve of length $L > 0$, and let $\bar{t} = (t_1, t_2, t_3) \in C^{k-1}[0, L]^3$ be its tangent vector. We shall always assume that $\bar{\theta}$ originates in the origin of the coordinates system and that it is parametrized with respect to its arc length; i.e., $|\bar{t}(x_3)|_{\mathbb{R}^3} = 1 \; \forall \, x_3 \in [0, L]$.

Then, alternatively, we may consider $\varphi \in C^{k-1}[0, L]$ and $\psi \in C^{k-1}[0, L]$ to be

some spherical coordinates of a unit vector given by $(\sin\varphi\cos\psi, \sin\varphi\sin\psi, \cos\varphi) \in C^{k-1}[0,L]$, which we denote again by $\bar{t}(x_3) = \bar{t}(\varphi(x_3), \psi(x_3))$. The corresponding 3D curve, depending on φ, ψ, is obtained by

$$\bar{\theta}(x_3) = \int_0^{x_3} \bar{t}(\tau)\, d\tau\,, \quad x_3 \in [0,L]\,. \tag{2.1}$$

The arbitrary mappings φ, ψ will play the role of the minimization parameters (controls) in the optimization problems to be studied in sections 3 and 4. Further conditions, called constraints, may be imposed on them later.

Notice that although the polar coordinates may not be uniquely determined in certain cases, relation (2.1) with arbitrary φ, ψ generates a rich class of 3D regular curves having C^k-regularity, which is enough for optimization applications. Later, we will have $k = 2$.

One advantage of the form (2.1) is that the curve is automatically parametrized with respect to its arc length, and that a local frame may be defined by purely algebraic means,

$$\bar{n} = (\cos\varphi\cos\psi\,,\ \cos\varphi\sin\psi\,,\ -\sin\varphi)\,, \tag{2.2}$$

$$\bar{b} = (-\sin\psi\,,\ \cos\psi\,,\ 0)\,, \tag{2.3}$$

in all points of the curve.

We denote by A the orthogonal matrix having the columns $\bar{t}, \bar{n}, \bar{b}$. The geometric meaning of this construction is that we perform a rotation of the global axis system, corresponding to the angles φ and ψ and indicated by A, i.e., $\bar{t} = A(1,0,0)^T$, $\bar{n} = A(0,1,0)^T$, $\bar{b} = A(0,0,1)^T$.

Remark 2.1. It is possible to apply (2.1)–(2.3) to absolutely continuous regular (i.e., with nonzero tangent) curves, after a reparametrization with respect to the arc length. Although we employ the same notation, the vectors $\bar{n}$, $\bar{b}$ are different, in general, from the normal and binormal vectors of the classical Frenet frame obtained under stronger regularity assumptions; see Bloch [5]. Other useful variants of local frames under low smoothness hypotheses may be found in Cartan [7], and Ignat, Sprekels, and Tiba [13].

We introduce the open set (which may be compared with a horizontal "cylinder" of nonconstant thickness)

$$\Omega = \bigcup_{x_3 \in]0,L[} \Big(\omega(x_3) \times \{x_3\}\Big) \subset \mathbb{R}^3, \tag{2.4}$$

where the cross section $\omega(x_3) \subset \mathbb{R}^2$, $x_3 \in [0,L]$, is a bounded, but not necessarily simply connected, domain such that $\omega(x_3) \supset \omega$, with an open set $\omega \subset \mathbb{R}^2$ satisfying the symmetry relations

$$0 = \int_\omega x_1\, dx_1\, dx_2 = \int_\omega x_2\, dx_1\, dx_2 = \int_\omega x_1\, x_2\, dx_1\, dx_2\,. \tag{2.5}$$

The curved rod $\tilde{\Omega}$ associated with $\bar{\theta}$ is obtained by the one-to-one nonlinear geometrical transformation $F : \Omega \to \tilde{\Omega}$,

$$\begin{aligned} &(x_1, x_2, x_3) = \bar{x} \in \Omega \mapsto F\bar{x} = \tilde{x} = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3) \\ &= \bar{\theta}(x_3) + x_1\, \bar{n}(x_3) + x_2\, \bar{b}(x_3) \in \tilde{\Omega} \quad \forall\, \bar{x} \in \Omega\,, \\ &\text{where } \tilde{\Omega} = \{\tilde{x} = F\,\bar{x}\,;\, \bar{x} \in \Omega\}\,. \end{aligned} \tag{2.6}$$

The fact that F is one-to-one follows from (2.10), (2.11) below. In the scientific literature (Trabucho and Viaño [19]), $\tilde{\Omega}$ is also called a "tube," in view of the possible presence of holes in $\omega(\,\cdot\,)$. In what follows, we will always assume that $\varphi, \psi \in C^1[0, L]$, i.e., $k = 2$. Then we get

$$\left\langle \bar{t}(x_3)\,,\, \bar{t}'(x_3) \right\rangle_{\mathbb{R}^3} = \left\langle \bar{n}(x_3)\,,\, \bar{n}'(x_3) \right\rangle_{\mathbb{R}^3} = \left\langle \bar{b}(x_3)\,,\, \bar{b}'(x_3) \right\rangle_{\mathbb{R}^3} = 0\,, \tag{2.7}$$

where $\langle \cdot\,,\, \cdot \rangle_{\mathbb{R}^3}$ denotes the Euclidean inner product in $\mathbb{R}^3$. This yields the "equations of motion" of the considered local frame:

$$\begin{aligned} \bar{t}'(x_3) &= a(x_3)\,\bar{b}(x_3) \,+\, \beta(x_3)\,\bar{n}(x_3)\,, \\ \bar{b}'(x_3) &= -\,a(x_3)\,\bar{t}(x_3) \,+\, c(x_3)\,\bar{n}(x_3)\,, \\ \bar{n}'(x_3) &= -\,\beta(x_3)\,\bar{t}(x_3) \,+\, c(x_3)\,\bar{b}(x_3)\,, \end{aligned} \tag{2.8}$$

with $a\,,\, \beta\,,\, c \in C[0, L]$ expressing the curvature and torsion properties of the curved rod.

The Jacobian of F at $\bar{x} \in \Omega$, denoted by $J(\bar{x}) = D\,F(\bar{x})$, is given by

$$J(\bar{x}) \,=\, \begin{bmatrix} n_1(x_3) & b_1(x_3) & t_1(x_3) \,+\, x_1\,n_1'(x_3) \,+\, x_2\,b_1'(x_3) \\ n_2(x_3) & b_2(x_3) & t_2(x_3) \,+\, x_1\,n_2'(x_3) \,+\, x_2\,b_2'(x_3) \\ n_3(x_3) & b_3(x_3) & t_3(x_3) \,+\, x_1\,n_3'(x_3) \,+\, x_2\,b_3'(x_3) \end{bmatrix}\,. \tag{2.9}$$

By (2.8), (2.9), we have

$$\det J(\bar{x}) \,=\, 1 \,-\, \beta(x_3)\,x_1 \,-\, a(x_3)\,x_2 \quad \forall\, \bar{x} \,\in\, \bar{\Omega}\,. \tag{2.10}$$

Remark 2.2. Relations (2.7)–(2.10) require the existence of second-order derivatives for $\bar{\theta}$ (or of first-order derivatives for $\varphi\,,\, \psi$). The results proved in the next sections show that these assumptions, together with the continuity property for the derivatives, are also sufficient. The same is true for the case of shells; see sections 5–7. For the modeling process, various ways to relax the geometric regularity assumptions have been proposed by Blouza [6], Ignat, Sprekels, and Tiba [13], and Sprekels and Tiba [18].

Usually, in the scientific literature a stronger regularity is required for the parametrization of rods or shells. In the case of the curved rods, the key point in our approach is the use of special local bases as in (2.1)–(2.3) or as in Ignat, Sprekels, and Tiba [13].

If $\omega(x_3)\,,\, x_3 \in [0, L]$, is contained in a sufficiently small disk in $\mathbb{R}^2$, then we may assume that

$$\det J(\bar{x}) \,\geq\, c \,>\, 0 \quad \forall\, \bar{x} \,\in\, \bar{\Omega}\,, \tag{2.11}$$

which justifies the introduction of the curved rod $\tilde{\Omega}$ via the geometric transformation F in (2.6); see Ciarlet [11, Thm. 3.1-1].

We assume that the rod is clamped at both ends, and that it is subjected to body forces $\tilde{f} \in L^2(\tilde{\Omega})^3$ (weight, electromagnetic field, etc.), as well as to surface tractions $\tilde{g} \in L^2(\tilde{\Sigma})^3$ on the lateral surface $\tilde{\Sigma}$ of the rod. On the "inside" lateral face of $\tilde{\Omega}$ (i.e., corresponding to possible holes), we take $\tilde{g} \equiv 0$.

Denote by $\bar{y} : \tilde{\Omega} \to \mathbb{R}^3$ the corresponding displacement of each point $\tilde{x} \in \tilde{\Omega}$. In Ignat, Sprekels, and Tiba [13], the general geometrical assumption that

$$\bar{y}(\tilde{x}) \,=\, \bar{\tau}(x_3) \,+\, x_1\,\bar{N}(x_3) \,+\, x_2\,\bar{B}(x_3) \quad \forall\, \tilde{x} \,\in\, \tilde{\Omega}\,, \tag{2.12}$$

with $\bar{x} = (x_1\,,\, x_2\,,\, x_3) = F^{-1}(\tilde{x})$ and $\bar{\tau}\,,\, \bar{N}\,,\, \bar{B} \in H_0^1(0,L)$ being unknown functions, has been imposed. This is a special case of the so-called *polynomial approximation* of the displacement; see Trabucho and Viaño [19]. Then, the following boundary value problem is obtained from the elasticity problem, where we also introduce the notation $\mathbf{b}_i, i = 1,3$, for later use:

$$
\begin{aligned}
(2.13)\quad & \mathcal{B}(\bar{y},\bar{v}) \,=\, \tilde{\lambda}\,\mathbf{b}_1(\bar{y},\bar{v}) \,+\, \tilde{\mu}\,\mathbf{b}_2(\bar{y},\bar{v}) \,+\, 2\,\tilde{\mu}\,\mathbf{b}_3(\bar{y},\bar{v}) \\
&= \tilde{\lambda} \int_{\Omega} \sum_{i,j=1}^{3} \Bigg[N_i(x_3)\,h_{1i}(\bar{x}) + B_i(x_3)\,h_{2i}(\bar{x}) \\
&\quad + \Big(\tau_i'(x_3) + x_1\,N_i'(x_3) + x_2\,B_i'(x_3)\Big) h_{3i}(\bar{x}) \Bigg] \Bigg[M_j(x_3)\,h_{1j}(\bar{x}) \\
&\quad + D_j(x_3)\,h_{2j}(\bar{x}) + \Big(\mu_j'(x_3) + x_1\,M_j'(x_3) + x_2\,D_j'(x_3)\Big)\,h_{3j}(\bar{x}) \Bigg] \Big|\det J(\bar{x})\Big|\,d\bar{x} \\
&\quad + \tilde{\mu} \int_{\Omega} \sum_{i<j} \Bigg[N_i(x_3)\,h_{1j}(\bar{x}) + B_i(x_3)\,h_{2j}(\bar{x}) + \Big(\tau_i'(x_3) + x_1\,N_i'(x_3) \\
&\qquad + x_2\,B_i'(x_3)\Big)\,h_{3j}(\bar{x}) + N_j(x_3)\,h_{1i}(\bar{x}) + B_j(x_3)\,h_{2i}(\bar{x}) \\
&\quad + \Big(\tau_j'(x_3) + x_1\,N_j'(x_3) + x_2\,B_j'(x_3)\Big)\,h_{3i}(\bar{x}) \Bigg] \Bigg[M_i(x_3)\,h_{1j}(\bar{x}) + D_i(x_3)\,h_{2j}(\bar{x}) \\
&\quad + \Big(\mu_i'(x_3) + x_1\,M_i'(x_3) + x_2\,D_i'(x_3)\Big)\,h_{3j}(\bar{x}) + M_j(x_3)\,h_{1i}(\bar{x}) + D_j(x_3)\,h_{2i}(\bar{x}) \\
&\quad + \Big(\mu_j'(x_3) + x_1\,M_j'(x_3) + x_2\,D_j'(x_3)\Big)\,h_{3i}(\bar{x}) \Bigg] \Big|\det J(\bar{x})\Big|\,d\bar{x} \\
&\quad + 2\,\tilde{\mu} \int_{\Omega} \sum_{i=1}^{3} \Bigg[N_i(x_3)\,h_{1i}(\bar{x}) + B_i(x_3)\,h_{2i}(\bar{x}) + \Big(\tau_i'(x_3) + x_1\,N_i'(x_3) \\
&\qquad + x_2\,B_i'(x_3)\Big)\,h_{3i}(\bar{x}) \Bigg] \Bigg[M_i(x_3)\,h_{1i}(\bar{x}) + D_i(x_3)\,h_{2i}(\bar{x}) \\
&\quad + \Big(\mu_i'(x_3) + x_1\,M_i'(x_3) + x_2\,D_i'(x_3)\Big)\,h_{3i}(\bar{x}) \Bigg] \Big|\det J(\bar{x})\Big|\,d\bar{x} \\
&= \sum_{l=1}^{3} \int_{\Omega} f_l(\bar{x}) \Big(\mu_l(x_3) + x_1\,M_l(x_3) + x_2\,D_l(x_3)\Big)\,|\det J(\bar{x})|\,d\bar{x} \\
&\quad + \sum_{i,j=1}^{3} \sum_{l=1}^{3} \int_{\partial\Omega} g_l(\bar{x}) \Big(\mu_l(x_3) + x_1\,M_l(x_3) + x_2\,D_l(x_3)\Big) |\det J(\bar{x})| \\
&\qquad \times \sqrt{\nu_i(\bar{x})\,g^{ij}(\bar{x})\,\nu_j(\bar{x})}\,d\tau\,.
\end{aligned}
$$

Above, $\tilde{\lambda} \geq 0\,,\, \tilde{\mu} > 0$ are the Lamé coefficients of the material; we have the matrices $(h_{ij}(\bar{x})) = J(\bar{x})^{-1}$, $(g^{ij}(\bar{x})) = (g_{ij}(\bar{x}))^{-1}$, and $(g_{ij}(\bar{x})) = J(\bar{x})^T\,J(\bar{x})$; and $\bar{\mu}\,,\, \bar{M}\,,\, \bar{D} \in H_0^1(0,L)^3$ are arbitrary test functions with $\bar{v}(\bar{x}) = \bar{\mu}(x_3) + x_1\,\bar{M}(x_3) + x_2\,\bar{D}(x_3)$. Further details, and the proof of the coercivity of the bilinear functional $\mathcal{B}(\,\cdot\,,\,\cdot\,)$:

$H_0^1(0,L)^9 \times H_0^1(0,L)^9 \to \mathbb{R}$ given by (2.13), may be found in Ignat, Sprekels, and Tiba [13], where different local bases are used. This yields the existence and the uniqueness of the solution $\bar{y}$ of (2.13) in $H_0^1(0,L)^9$. Equation (2.13) is derived via (2.12) from the usual displacement approach to curved rods (see Trabucho and Viaño [19, Chap. I]):

$$\sum_{i,j=1}^{3} \int\limits_{\tilde{\Omega}} \sigma_{ij}(\bar{u})\, e_{ij}(\bar{v})\, d\tilde{x} \;=\; \sum_{l=1}^{3} \left[\int\limits_{\tilde{\Omega}} \tilde{f}_l\, \bar{v}_l\, d\tilde{x} \;+\; \int\limits_{\tilde{\Sigma}} \tilde{g}_l\, \bar{v}_l\, d\tilde{\tau} \right]$$

$$\forall\, \bar{v} \in \Big\{ \bar{w} \in H^1(\tilde{\Omega})\,;\, \bar{w} \,=\, 0 \text{ on } F(\omega(0) \,\times\, \{0\}) \,\cup\, F(\omega(L) \,\times\, \{L\}) \Big\}$$

with the constitutive law (the stress/strain relation) of linearized elasticity (also known as Hooke's law),

$$\sigma(\bar{u}) \;=\; (\sigma_{ij}(\bar{u}))_{i,j=\overline{1,3}} \;=\; \big(\tilde{\lambda}\, e_{pp}(\bar{u})\, \delta_{ij} \;+\; 2\,\tilde{\mu}\, e_{ij}(\bar{u})\big)_{i,j=\overline{1,3}}\,.$$

In what follows, we shall suppose that $\omega(x_3) = \omega\, \forall\, x_3 \in [0,L]$, with ω satisfying (2.5).

For given $\tilde{f}$, $\tilde{g}$, a general shape optimization problem associated with (2.13) is

$$\text{(P)} \qquad \min_{\varphi,\psi} \Big\{ \Pi(\varphi,\psi) \;=\; j\Big(\bar{\theta}(\varphi,\psi)\,,\, \bar{y}(\varphi,\psi)\Big) \;=\; j(\bar{\theta},\bar{y}) \Big\}\,,$$

subject to $\bar{\theta} \in \mathcal{K}$, where $\mathcal{K} \subset C^2[0,L]^3$ is a closed bounded subset, and $\bar{y} = (\bar{\tau}, \bar{N}, \bar{B}) \in H_0^1(0,L)^9$ is obtained as the solution of (2.13). The condition $\bar{\theta} \in \mathcal{K}$ represents the natural way to impose restrictions on the geometry of the rod in (P). It gives an implicit constraint on $\varphi, \psi \in C^1[0,L]$. We assume that the mappings $j : C^2[0,L]^3 \times H_0^1(0,L)^9 \to \mathbb{R}$ and $\Pi : C^1[0,L]^2 \to \mathbb{R}$ satisfy some regularity properties, to be imposed later. An important example for a cost functional j is the quadratic case. For instance, if

$$j(\bar{\theta},\bar{y}) \;=\; |\tau_1|^2_{H_0^1(0,L)} \;+\; |\tau_2|^2_{H_0^1(0,L)} \;+\; |\tau_3|^2_{H_0^1(0,L)}\,, \tag{2.14}$$

then (P) aims at finding the shape of the curved rod that minimizes the displacement of the line of centroids under prescribed forces and tractions. This is a natural safety requirement in many applications.

Concerning the constraints to which the curved rod may be submitted, we underline that our formalism automatically ensures a prescribed length $L > 0$. This eliminates possible trivial cases, such as a constant $\bar{\theta}$ in $[0,L]$, and is also important from the optimization point of view, since otherwise the cost may depend on L. A simple sufficient condition under which $\bar{\theta}$ has no multiple points (i.e., there are no values $x_3^1, x_3^2 \in [0,L]$, $x_3^1 \neq x_3^2$, such that $\bar{\theta}(x_3^1) = \bar{\theta}(x_3^2)$) is

$$0 \;\le\; \varphi(x_3) \;\le\; \frac{\pi}{2} \;-\; \varepsilon\,, \quad x_3 \;\in\; [0,L]\,, \tag{2.15}$$

with $\varepsilon > 0$ small. This is due to the fact that (2.1) gives $\theta_3'(x_3) = t_3(x_3) = \cos\varphi(x_3) > 0$ in $[0,L]$; i.e., θ_3 is a strictly increasing function in $[0,L]$. Similar other conditions may easily be formulated in accordance with the desired applications. They may be used, for instance, in problems concerning the optimization of strings, where the periodicity condition (for θ_1, θ_2)

$$\int\limits_0^L t_1\, dx_3 \;=\; \int\limits_0^L t_2\, dx_3 \;=\; 0 \tag{2.16}$$

is also important.

Notice that relations (2.14), (2.15) correspond to convex optimization problems, while relation (2.16) is nonlinear in φ, ψ and, consequently, the corresponding set $\mathcal{K}$ is nonconvex. Relation (2.11) should also be included in the definition of $\mathcal{K}$.

Remark 2.3. A very simple variant of representation of the unit tangent vector is $\bar{t} = (u_1, u_2, \sqrt{1 - u_1^2 - u_2^2})$, but this already assumes a prescribed sign for t_3 and requires the more restrictive hypothesis

$$u_1^2 + u_2^2 \leq 1 - \varepsilon$$

for the differentiability of the local frame. However, under this representation relation (2.16) becomes linear, which may be useful in some applications.

3. Existence of optimal curved rods. We prove the following continuous dependence result.

THEOREM 3.1. *Assume that* $\varphi_n \to \varphi$, $\psi_n \to \psi$, *strongly in* $C^1[0, L]$. *If* $\bar{y}_n$, $\check{y}$ *denote the solutions to* (2.13) *associated with* (φ_n, ψ_n) *and* (φ, ψ), *respectively, then*

$$\bar{y}_n \to \check{y} \quad \text{strongly in } H_0^1(0, L)^9. \tag{3.1}$$

Proof. Clearly,

$$\begin{aligned} \bar{t}_n &= (\cos\psi_n \sin\varphi_n, \sin\psi_n \sin\varphi_n, \cos\varphi_n) \\ &\to \bar{t} = (\cos\psi \sin\varphi, \sin\psi \sin\varphi, \cos\varphi) \end{aligned} \tag{3.2}$$

in $C^1[0, L]^3$. Then, (2.1) shows that $\bar{\theta}_n \to \bar{\theta}$ in $C^2[0, L]^3$. By (2.2), (2.3), and with obvious notation, we get that $\bar{n}_n \to \bar{n}$ and $\bar{b}_n \to \bar{b}$ in $C^1[0, L]^3$.

From (2.8) it is easy to infer that

$$a_n = \langle \bar{t}_n', \bar{b}_n \rangle_{\mathbb{R}^3} \to a = \langle \bar{t}', \bar{b} \rangle_{\mathbb{R}^3} \quad \text{strongly in } C[0, L]. \tag{3.3}$$

We also have $\beta_n \to \beta$ and $c_n \to c$ in $C[0, L]$.

Relation (2.10) shows that

$$\det J_n(\bar{x}) \to \det J(\bar{x}) \quad \text{in } C(\bar{\Omega}), \tag{3.4}$$

and from (2.11) we infer that $\{\det J_n(\bar{x})\}$ is bounded from below by some positive constant.

Moreover, (2.9) implies that $J_n(\bar{x}) \to J(\bar{x})$ in $C(\bar{\Omega})^9$ and, likewise, that $J_n^{-1}(\bar{x}) \to J^{-1}(\bar{x})$, by (3.4) and the above observations. In particular, we have

$$h_{ij}^n(\bar{x}) \to h_{ij}(\bar{x}) \quad \text{in } C(\bar{\Omega}) \quad \forall\, i, j = \overline{1, 3}. \tag{3.5}$$

Let $\mathcal{B}_n$ denote the bilinear functional (2.13) with the coefficients h_{ij}^n, $\det J_n$.

LEMMA 3.2. *There are* $c_1 > 0$, $c_2 > 0$ *such that*

$$\mathcal{B}_n(\bar{y}, \bar{y}) \geq c_1 |\bar{y}|^2_{H_0^1(0,L)^9} - c_2 |\bar{y}|^2_{L^2(0,L)^9} \tag{3.6}$$

for any $\bar{y} \in H_0^1(0, L)^9$ *and any* $n \in \mathbb{N}$.

Proof. By (3.4) and (2.11), we have

$$\mathcal{B}_n(\bar{y},\bar{y}) \geq \tilde{\mu}\, c \int\limits_\Omega \sum_{i<j} \Bigg[N_i\, h^n_{1j} + B_i\, h^n_{2j} + \Big(\tau_i' + x_1\, N_i' + x_2\, B_i'\Big)\, h^n_{3j} + N_j\, h^n_{1i} + B_j\, h^n_{2i} + \Big(\tau_j' + x_1\, N_j' + x_2\, B_j'\Big)\, h^n_{3i} \Bigg]^2 d\bar{x} + 2\,\tilde{\mu}\, c \int\limits_\Omega \sum_{i=1}^{3} \Bigg[N_i\, h^n_{1i} + B_i\, h^n_{2i} + \Big(\tau_i' + x_1\, N_i' + x_2 B_i'\Big)\, h^n_{3i} \Bigg]^2 d\bar{x}\,.$$

From the uniform boundedness of the coefficients due to (3.5), and using standard binomial inequalities, we find that

$$\frac{1}{\tilde{\mu}\, c}\, \mathcal{B}_n(\bar{y},\bar{y}) \geq \frac{1}{2} \int\limits_\Omega \sum_{i<j} \Bigg[\Big(\tau_i' + x_1\, N_i' + x_2\, B_i'\Big)\, h^n_{3j} + \Big(\tau_j' + x_1\, N_j' + x_2\, B_j'\Big)\, h^n_{3i} \Bigg]^2 d\bar{x} + \int\limits_\Omega \sum_{i=1}^{3} \Bigg[\Big(\tau_i' + x_1\, N_i' + x_2\, B_i'\Big)\, h^n_{3i} \Bigg]^2 d\bar{x} \;-\; \tilde{c}|\bar{y}|^2_{L^2(0,L)^9}\,.$$

We use the following algebraic identity:

$$\begin{aligned} &\frac{1}{2}\Bigg[\Big(z_1\, h^n_{32} + z_2\, h^n_{31}\Big)^2 + \Big(z_2\, h^n_{33} + z_3\, h^n_{32}\Big)^2 + \Big(z_1\, h^n_{33} + z_3\, h^n_{31}\Big)^2 \Bigg] \\ &\quad + \frac{3}{2}\Bigg[z_1^2 (h^n_{31})^2 + z_2^2 (h^n_{32})^2 + z_3^2 (h^n_{33})^2 \Bigg] \\ &= \frac{1}{2}\Big(z_1^2 + z_2^2 + z_3^2\Big)\Bigg[(h^n_{31})^2 + (h^n_{32})^2 + (h^n_{33})^2 \Bigg] + \frac{1}{2}\Big(z_1\, h^n_{31} + z_2\, h^n_{32}\Big)^2 \\ &\quad + \frac{1}{2}\Big(z_1\, h^n_{31} + z_3\, h^n_{33}\Big)^2 + \frac{1}{2}\Big(z_2\, h^n_{32} + z_3\, h^n_{33}\Big)^2, \end{aligned}$$

with $z_i := \tau_i' + x_1\, N_i' + x_2\, B_i'$, $i = \overline{1,3}$. It follows that

$$\frac{1}{\tilde{\mu}\, c}\mathcal{B}_n(\bar{y},\bar{y}) \;\geq\; \frac{1}{4} \int\limits_\Omega \sum_{i=1}^{3} \Big(\tau_i' + x_1\, N_i' + x_2\, B_i'\Big)^2 \sum_{i=1}^{3} (h^n_{3i})^2\, d\bar{x} - \tilde{c}\,|\bar{y}|^2_{L^2(0,L)^9}\,. \tag{3.7}$$

A direct calculus allows us to find h^n_{ij} and to check that, for some $k > 0$,

$$\sum_{i=1}^{3} (h^n_{3i})^2 \;=\; \Big[\det J_n\Big]^{-2} \sum_{i=1}^{3} (t^n_i)^2 \;=\; \Big[\det J_n\Big]^{-2} \;\geq\; k \;>\; 0 \tag{3.8}$$

since $|\bar{t}_n|_{\mathbb{R}^3} = 1$.

Relations (3.7), (3.8) give

$$\frac{1}{\tilde{\mu}\, c}\mathcal{B}_n(\bar{y},\bar{y}) \;\geq\; \frac{k}{4} \int\limits_\Omega \sum_{i=1}^{3} \Big(\tau_i' + x_1\, N_i' + x_2\, B_i'\Big)^2 d\bar{x} \;-\; \tilde{c}\,|\bar{y}|^2_{L^2(0,L)^9}\,.$$

Performing the computations in the right-hand side and integrating with respect to x_1, x_2, we obtain the inequality (3.6) by means of (2.5). ∎

Proof of Theorem 3.1 *(continued).* We use a contradiction argument to show that the functionals $\mathcal{B}_n$ are uniformly coercive. We assume that there is a sequence $\varepsilon_n \to 0$ and a sequence $\tilde{y}_n \in H_0^1(0,L)^9$, $|\tilde{y}_n|_{H_0^1(0,L)^9} = 1$, such that

$$0 \le \mathcal{B}_n(\tilde{y}_n, \tilde{y}_n) \le \varepsilon_n |\tilde{y}_n|^2_{H_0^1(0,L)^9}. \tag{3.9}$$

Let $\hat{y}$ be the weak limit of $\tilde{y}_n$ in $H_0^1(0,L)^9$, which may be supposed to exist.

We give a detailed computation for the last integral in the definition of $\mathcal{B}_n(\tilde{y}_n, \tilde{y}_n)$:

$$\begin{aligned}
I_n &= 2\,\tilde{\mu} \int_\Omega \sum_{i=1}^{3} \left[\tilde{N}_i^n\, h_{1i}^n + \tilde{B}_i^n\, h_{2i}^n + \left(\tilde{\tau}_i^{n'} + x_1\, \tilde{N}_i^{n'} + x_2\, \tilde{B}_i^{n'} \right) h_{3i}^n \right]^2 \det J^n \, d\bar{x} \\
&= 2\,\tilde{\mu} \int_\Omega \sum_{i=1}^{3} \left[\tilde{N}_i^n\, h_{1i} + \tilde{B}_i^n\, h_{2i} + \left(\tilde{\tau}_i^{n'} + x_1\, \tilde{N}_i^{n'} + x_2\, \tilde{B}_i^{n'} \right) h_{3i} \right]^2 \det J \, d\bar{x} \\
&\quad + 2\,\tilde{\mu} \int_\Omega \sum_{i=1}^{3} \left[\tilde{N}_i^n \left(h_{1i}^n \sqrt{\det J^n} - h_{1i} \sqrt{\det J} \right) + \tilde{B}_i^n \left(h_{2i}^n \sqrt{\det J^n} - h_{2i} \sqrt{\det J} \right) \right. \\
&\qquad \left. + \left(\tilde{\tau}_i^{n'} + x_1\, \tilde{N}_i^{n'} + x_2\, \tilde{B}_i^{n'} \right) \left(h_{3i}^n \sqrt{\det J^n} - h_{3i} \sqrt{\det J} \right) \right] \\
&\quad \times \left[\tilde{N}_i^n \left(h_{1i}^n \sqrt{\det J^n} + h_{1i} \sqrt{\det J} \right) + \tilde{B}_i^n \left(h_{2i}^n \sqrt{\det J^n} + h_{2i} \sqrt{\det J} \right) \right. \\
&\qquad \left. + \left(\tilde{\tau}_i^{n'} + x_1\, \tilde{N}_i^{n'} + x_2\, \tilde{B}_i^{n'} \right) \left(h_{3i}^n \sqrt{\det J^n} + h_{3i} \sqrt{\det J} \right) \right] d\bar{x}\,.
\end{aligned}$$

Here, $\tilde{y}_n = (\tilde{\tau}_n, \tilde{N}_n, \tilde{B}_n)$ belongs to the unit ball in $H_0^1(0,L)^9$. The uniform convergence of the coefficients (see (3.4), (3.5)) shows that the last integral converges to zero. The weak lower semicontinuity of quadratic forms gives

$$\begin{aligned}
\liminf_{n\to\infty} I_n &= \liminf_{n\to\infty}\ 2\tilde{\mu} \int_\Omega \sum_{i=1}^{3} \left[\tilde{N}_i^n\, h_{1i} + \tilde{B}_i^n\, h_{2i} + \left(\tilde{\tau}_i^{n'} + x_1\, \tilde{N}_i^{n'} + x_2\, \tilde{B}_i^{n'} \right) h_{3i} \right]^2 \\
&\qquad \times \det J d\bar{x} \\
&\ge 2\,\tilde{\mu} \int_\Omega \sum_{i=1}^{3} \left[\hat{N}_i\, h_{1i} + \hat{B}_i\, h_{2i} + \left(\hat{\tau}_i' + x_1\, \hat{N}_i' + x_2\, \hat{B}_i' \right) h_{3i} \right]^2 \det J \, d\bar{x}\,,
\end{aligned}$$

where $(\hat{\tau}\,,\, \hat{N}\,,\, \hat{B}) \in H_0^1(0,L)^9$ is the detailed notation of $\hat{y}$.

Computing the other terms in a similar way, we get

$$\begin{aligned}
\liminf_{n\to\infty} \mathcal{B}_n(\tilde{y}_n, \tilde{y}_n) \ge 2\,\tilde{\mu} \int_\Omega \sum_{i=1}^{3} \Bigg[& \hat{N}_i\, h_{1i} + \hat{B}_i\, h_{2i} \\
& + \left(\hat{\tau}_i' + x_1\, \hat{N}_i' + x_2\, \hat{B}_i' \right) h_{3i} \Bigg]^2 \det J \, d\bar{x}
\end{aligned} \tag{3.10}$$

$$
\begin{aligned}
&+\tilde{\lambda}\int_\Omega \sum_{i,j=1}^{3}\left[\hat{N}_i\,h_{1i}+\hat{B}_i\,h_{2i}+\left(\hat{\tau}_i'+x_1\,\hat{N}_i'+x_2\,\hat{B}_i'\right)h_{3i}\right]\\
&\quad\times\left[\hat{N}_j\,h_{1j}+\hat{B}_j\,h_{2j}+\left(\hat{\tau}_j'+x_1\,\hat{N}_j'+x_2\,\hat{B}_j'\right)h_{3j}\right]\det J\,d\bar{x}\\
&+\tilde{\mu}\int_\Omega \sum_{i<j}\left[\hat{N}_i\,h_{1j}+\hat{B}_i\,h_{2j}+\left(\hat{\tau}_i'+x_1\,\hat{N}_i'+x_2\,\hat{B}_i'\right)h_{3j}\right.\\
&\quad\left.+\hat{N}_j\,h_{1i}+\hat{B}_j\,h_{2i}+\left(\hat{\tau}_j'+x_1\,\hat{N}_j'+x_2\,\hat{B}_j'\right)h_{3i}\right]^2\det J\,d\bar{x}\\
&=\mathcal{B}(\hat{y},\hat{y})\,.
\end{aligned}
$$

By assumption (3.9) and by (3.10), we have $\mathcal{B}(\hat{y},\hat{y})=0$.

It is known that such a relation yields $\hat{y}=0$ (see, for instance, Lemma 2.3 in [13]).

We again use inequality (3.9) with Lemma 3.2:

$$
\varepsilon_n\;\geq\;\mathcal{B}_n(\tilde{y}_n,\tilde{y}_n)\;\geq\;c_1\;-\;c_2|\tilde{y}_n|^2_{L^2(0,L)^9}\,, \tag{3.11}
$$

since $|\tilde{y}_n|_{H_0^1(0,L)^9}=1$.

Notice that $\tilde{y}_n\to\hat{y}=0$ strongly in $L^2(0,L)^9$, by the above argument. Then, combining (3.10) and (3.11), we obtain the contradiction $0\geq c_1$. We conclude that there is some $\delta>0$ such that, $\forall\ n\geq 1$,

$$
\mathcal{B}_n(\bar{y},\bar{y})\;\geq\;\delta|\bar{y}|^2_{H_0^1(0,L)^9}\quad\forall\ \bar{y}\;\in\;H_0^1(0,L)^9\,. \tag{3.12}
$$

Let us fix $\bar{v}=\bar{y}_n$ in the state equations (2.13) corresponding to $\mathcal{B}_n(\,\cdot\,,\,\cdot\,)$. Taking (3.12) into account, we immediately obtain that $\{\bar{y}_n\}$ is bounded in $H_0^1(0,L)^9$. We may take a subsequence such that $\bar{y}_n\to\check{y}$ weakly in $H_0^1(0,L)^9$. Due to the uniform convergence of the coefficients h_{ij}^n, $\det J_n$, g_n^{ij}, one may pass to the limit in (2.13) and see that $\check{y}$ is indeed the solution to (2.13) associated with $(\varphi\,,\,\psi)$.

The last step of the proof is to show that the convergence is valid in the strong topology of $H_0^1(0,L)^9$. We subtract the equations corresponding to (τ^n,N^n,B^n) (resp., $(\check{\tau},\check{N},\check{B})$), we intercalate advantageous terms, and, finally, we take test functions of the form $(\tau^n,N^n,B^n)-(\check{\tau},\check{N},\check{B})\in H_0^1(0,L)^9$. We write in detail just the simplest term:

$$
\begin{aligned}
&2\,\tilde{\mu}\int_\Omega\sum_{i=1}^{3}\left[N_i^n\,h_{1i}^n+B_1^n\,h_{2i}^n+\left(\tau_i^{n'}+x_1\,N_i^{n'}+x_2\,B_i^{n'}\right)h_{3i}^n\right]\\
&\times\left[(N_i^n-\check{N}_i)\,h_{1i}^n+(B_i^n-\check{B}_i)\,h_{2i}^n+\left(\tau_i^{n'}-\check{\tau}_i'+x_1(N_i^{n'}-\check{N}_i')+x_2(B_i^{n'}-\check{B}_i')\right)h_{3i}^n\right]\\
&\qquad\times\det J^n\,d\bar{x}-2\,\tilde{\mu}\int_\Omega\sum_{i=1}^{3}\left[\check{N}_i\,h_{1i}+\check{B}_i\,h_{2i}+\left(\check{\tau}_i'+x_1\,\check{N}_i'+x_2\,\check{B}_i'\right)h_{3i}\right]\\
&\times\left[(N_i^n-\check{N}_i)h_{1i}+(B_i^n-\check{B}_i)\,h_{2i}+\left(\tau_i^{n'}-\check{\tau}_i'+x_1(N_i^{n'}-\check{N}_i')+x_2(B_i^{n'}-\check{B}_i)'\right)h_{3i}\right]
\end{aligned}
$$

$$\times \det J\, d\bar{x} = 2\,\tilde{\mu} \int_\Omega \sum_{i=1}^{3} \Big[(N_i^n - \check{N}_i)\, h_{1i} + (B_i^n - \check{B}_i)\, h_{2i}$$

$$+ \Big(\tau_i^{n'} - \check{\tau}_i' + x_1 (N_i^{n'} - \check{N}_i') + x_2 (B_i^{n'} - \check{B}_i') \Big)\, h_{3i} \Big]^2 \det J\, d\bar{x}$$

$$+ 2\,\tilde{\mu} \int_\Omega \sum_{i=1}^{3} \times \Big[N_i^n\, h_{1i} + B_i^n\, h_{2i} + \Big(\tau_i^{n'} + x_1\, N_i^{n'} + x_2\, B_i^{n'} \Big)\, h_{3i} \Big]$$

$$\times \Big[(N_i^n - \check{N}_i)(h_{1i}^n - h_{1i}) + (B_i^n - \check{B}_i)(h_{2i}^n - h_{2i}) + \Big(\tau_i^{n'} - \check{\tau}_i' + x_1 (N_i^{n'} - \check{N}_i')$$

$$+ x_2 (B_i^{n'} - \check{B}_i') \Big) (h_{3i}^n - h_{3i}) \Big] \det J\, d\bar{x}$$

$$+ 2\,\tilde{\mu} \int_\Omega \sum_{i=1}^{3} \Big[N_i^n\, h_{1i} + B_i^n\, h_{2i} + \Big(\tau_i^{n'} + x_1\, N_i^{n'} + x_2\, B_i^{n'} \Big)\, h_{3i} \Big]$$

$$\times \Big[(N_i^n - \check{N}_i)\, h_{1i}^n + (B_i^n - \check{B}_i)\, h_{2i}^n + \Big(\tau_i^{n'} - \check{\tau}_i^{n'} + x_1 (N_i^{n'} - \check{N}_i') + x_2 (B_i^{n'} - \check{B}_i') \Big)\, h_{3i}^n \Big]$$

$$\times \Big[\det J^n - \det J \Big]\, d\bar{x}$$

$$+ 2\,\tilde{\mu} \int_\Omega \sum_{i=1}^{3} \Big[N_i^n (h_{1i}^n - h_{1i}) + B_i^n (h_{2i}^n - h_{2i}) + \Big(\tau_i^{n'} + x_1\, N_i^{n'} + x_2\, B_i^{n'} \Big) (h_{3i}^n - h_{3i}) \Big]$$

$$\times \Big[(N_i^n - \check{N}_i) h_{1i}^n + (B_i^n - \check{B}_i) h_{2i}^n$$

$$+ \Big(\tau_i^{n'} - \check{\tau}_i' + x_1 (N_i^{n'} - \check{N}_i') + x_2 (B_i^{n'} - \check{B}_i') \Big) h_{3i}^n \Big] \det J^n\, d\bar{\tau}.$$

All the terms above, except the first one after the equality sign (the quadratic one), converge to zero due to the weak convergence of (τ^n, N^n, B^n) and to the uniform convergence of the coefficients. Similar computations may be performed for all the integrals in the variational equations, and we conclude that

$$\lim_{n \to \infty} \mathcal{B}(\bar{y}_n - \check{y}\,,\, \bar{y}_n - \check{y}) \;=\; 0\,. \tag{3.13}$$

By (3.12), (3.13) the proof is finished. □

COROLLARY 3.3. *If $\mathcal{K} \subset C^2[0,L]^3$ is generated by a compact in $C^1[0,L]^3$ subset of $\{\varphi, \psi\}$ and $j : C^2(0,L)^3 \times H_0^1(0,L)^9 \to \mathbb{R}$ is lower semicontinuous, then the shape optimization problem* (P) *admits at least one optimal curved rod solution in $\mathcal{K}$.*

Example 3.4 The functional (2.14) satisfies the above conditions, and the constraint (2.15), supplemented by boundedness conditions on ψ and φ', ψ', provides a simple case when Corollary 3.3 may be applied.

4. Sensitivity analysis of curved rods. We first study some differentiability properties of the mapping $(\varphi, \psi) \in C^1[0,L]^2 \mapsto \bar{y} \in H_0^1(0,L)^9$, with $\bar{y}$ being the solution of (2.13) corresponding to (φ, ψ). We consider $(\varphi_\lambda, \psi_\lambda) = (\varphi + \lambda\,\gamma\,,\, \psi +$

$\lambda\,\xi) \in C^1[0,1]^2$, $\lambda \in \mathbb{R}_+$, to be some variation around (φ,ψ), and we denote by $\bar{y}_\lambda = (\bar{\tau}_\lambda, \bar{N}_\lambda, \bar{B}_\lambda) \in H^1_0(0,L)^9$ the corresponding solution of (2.13). Similarly, we denote by $\bar{t}_\lambda, \bar{\theta}_\lambda, \bar{n}_\lambda, \bar{b}_\lambda, a_\lambda, \beta_\lambda, c_\lambda, J_\lambda, h^\lambda_{ij}, g^{ij}_\lambda$ all the quantities defined in section 2, starting from $(\varphi_\lambda, \psi_\lambda)$. Notice that, by our construction, the perturbed curved rod $\bar{\theta}_\lambda$ still has length L and is parametrized with respect to its arc length, i.e., $|\bar{t}_\lambda|_{\mathbb{R}^3} = 1$.

It is elementary, though tedious, to check that all the below listed limits (computed in the indicated "range" spaces) and operators exist and satisfy the indicated properties:

$$\lim_{\lambda\to 0} \frac{\bar{t}_\lambda - \bar{t}}{\lambda} = \tilde{t}(\gamma,\xi)\,,\ \tilde{t} : C^1[0,L]^2 \to C^1[0,L]^3\,, \tag{4.1}$$

$$\lim_{\lambda\to 0} \frac{\bar{\theta}_\lambda - \bar{\theta}}{\lambda} = \tilde{\theta}(\gamma,\xi)\,,\ \tilde{\theta} : C^1[0,L]^2 \to C^2[0,L]^3\,, \tag{4.2}$$

$$\lim_{\lambda\to 0} \frac{\bar{n}_\lambda - \bar{n}}{\lambda} = \tilde{n}(\gamma,\xi)\,,\ \tilde{n} : C^1[0,L]^2 \to C^1[0,L]^3\,, \tag{4.3}$$

$$\lim_{\lambda\to 0} \frac{\bar{b}_\lambda - \bar{b}}{\lambda} = \tilde{b}(\gamma,\xi)\,,\ \tilde{b} : C^1[0,L]^2 \to C^1[0,L]^3\,, \tag{4.4}$$

$$\lim_{\lambda\to 0} \frac{a_\lambda - a}{\lambda} = \tilde{a}(\gamma,\xi)\,,\ \tilde{a} : C^1[0,L]^2 \to C[0,L]\,, \tag{4.5}$$

$$\lim_{\lambda\to 0} \frac{\beta_\lambda - \beta}{\lambda} = \tilde{\beta}(\gamma,\xi)\,,\ \tilde{\beta} : C^1[0,L]^2 \to C[0,L]\,, \tag{4.6}$$

$$\lim_{\lambda\to 0} \frac{c_\lambda - c}{\lambda} = \tilde{c}(\gamma,\xi)\,,\ \tilde{c} : C^1[0,L]^2 \to C[0,L]\,, \tag{4.7}$$

$$\lim_{\lambda\to 0} \frac{\det J_\lambda - \det J}{\lambda} = \tilde{\mathcal{D}}(\gamma,\xi)\,,\ \tilde{\mathcal{D}} : C^1[0,L]^2 \to C(\bar{\Omega})\,, \tag{4.8}$$

$$\lim_{\lambda\to 0} \frac{J_\lambda - J}{\lambda} = \tilde{J}(\gamma,\xi)\,,\ \tilde{J} : C^1[0,L]^2 \to C(\bar{\Omega})^9\,, \tag{4.9}$$

$$\lim_{\lambda\to 0} \frac{J_\lambda^{-1} - J^{-1}}{\lambda} = \tilde{I}(\gamma,\xi)\,,\ \tilde{I} : C^1[0,L]^2 \to C(\bar{\Omega})^9\,, \tag{4.10}$$

$$\lim_{\lambda\to 0} \frac{h^\lambda_{ij} - h_{ij}}{\lambda} = \tilde{h}_{ij}(\gamma,\xi)\,,\ \tilde{h}_{ij} : C^1[0,L]^2 \to C(\bar{\Omega})\,, \tag{4.11}$$

$$\lim_{\lambda\to 0} \frac{g^{ij}_\lambda - g^{ij}}{\lambda} = \tilde{g}^{ij}(\gamma,\xi)\,,\ \tilde{g}^{ij} : C^1[0,L]^2 \to C(\bar{\Omega})\,. \tag{4.12}$$

All the operators $\tilde{t}, \tilde{\theta}, \tilde{n}, \tilde{b}, \tilde{a}, \tilde{\beta}, \tilde{c}, \tilde{\mathcal{D}}, \tilde{J}, \tilde{I}, \tilde{h}_{ij}, \tilde{g}^{ij}$ are linear and bounded in the indicated spaces. For instance, relation (4.1) reads in full detail as

$$\lim_{\lambda\to 0} \lambda^{-1} \Big[\Big(\sin\varphi_\lambda\,\cos\psi_\lambda,\, \sin\varphi_\lambda\,\sin\psi_\lambda,\, \cos\varphi_\lambda\Big) - \Big(\sin\varphi\,\cos\psi,\, \sin\varphi\,\sin\psi,\, \cos\varphi\Big)\Big]$$
$$= \Big(\gamma\,\cos\varphi\,\cos\psi\,-\,\xi\,\sin\varphi\,\sin\psi\,,\,\gamma\,\cos\varphi\,\sin\psi\,+\,\xi\,\sin\varphi\,\cos\psi\,,\,-\gamma\,\sin\varphi\Big)\,.$$

It is valid in $C^1[0,L]^3$; i.e., a similar relation may be written with respect to the derivatives of the above vector functions in $[0,L]$. The linear operator $\tilde{t} : C^1[0,L]^2 \to C^1[0,L]^3$, associated with $\varphi, \psi \in C^1[0,L]$, is

$$\tilde{t}(\gamma,\xi) = \begin{bmatrix} \cos\varphi\,\cos\psi & -\sin\varphi\,\sin\psi \\ \cos\varphi\,\sin\psi & \sin\varphi\,\cos\psi \\ -\sin\varphi & 0 \end{bmatrix} \begin{bmatrix} \gamma \\ \xi \end{bmatrix} \quad \forall\ (\gamma,\xi) \in C^1[0,L]^2\,.$$

By Theorem 3.1, we also have that

$$\bar{y}_\lambda \to \bar{y} \text{ strongly in } H_0^1(0,L)^9\,. \tag{4.13}$$

In order to prove the differentiability properties of $\bar{y}_\lambda$, we subtract the equations for $\bar{y}_\lambda$, $\bar{y}$, divide by λ, and intercalate advantageous terms. Later, we shall also fix test functions of the form $\lambda^{-1}(\bar{y}_\lambda - \bar{y}) \in H_0^1(0,L)^9$.

In the right-hand side of (2.13), it is possible to pass to the limit,

$$\begin{aligned}
&\lim_{\lambda\to 0}\Bigg\{\sum_{l=1}^{3}\int_\Omega f_l(\bar{x})\Big(\mu_l(x_3)+x_1\,M_l(x_3)+x_2\,D_l(x_3)\Big)\frac{\det J_\lambda-\det J}{\lambda}\,d\bar{x}\\
&\quad+\sum_{i,j=1}^{3}\sum_{l=1}^{3}\int_{\partial\Omega} g_l(\bar{x})\Big(\mu_l(x_3)+x_1\,M_l(x_3)+\,x_2\,D_l(x_3)\Big)\\
&\quad\times\frac{\det J_\lambda\sqrt{\nu_i(\bar{x})\,g_\lambda^{ij}\,\nu_j(\bar{x})}-\det J\sqrt{\nu_i(\bar{x})g^{ij}\nu_j(\bar{x})}}{\lambda}\,d\tau\Bigg\}\\
&=\sum_{l=1}^{3}\int_\Omega f_l(\bar{x})\Big(\mu_l(x_3)+x_1\,M_l(x_3)+x_2\,D_l(x_3)\Big)\,\tilde{\mathcal{D}}(\gamma,\xi)\,d\bar{x}\\
&\quad+\sum_{i,j=1}^{3}\sum_{l=1}^{3}\int_{\partial\Omega} g_l(\bar{x})\Big(\mu_l(x_3)+x_1\,M_l(x_3)+x_2\,D_l(x_3)\Big)\Bigg[\tilde{\mathcal{D}}(\gamma,\xi)\sqrt{\nu_i\,g^{ij}\,\nu_j}\\
&\qquad+\det J\frac{\nu_i\,\tilde{g}^{ij}(\gamma,\xi)\nu_j}{2\sqrt{\nu_i\,g^{ij}\,\nu_j}}\Bigg]\,d\tau\,.
\end{aligned}\tag{4.14}$$

We also write the corresponding transformation of the simplest term (i.e., of $\mathbf{b}_3(\,\cdot\,,\,\cdot\,)$) in $\mathcal{B}_\lambda(\,\cdot\,,\,\cdot\,)$, the bilinear functional (2.13) obtained from $(\varphi_\lambda,\psi_\lambda)$:

$$\begin{aligned}
&\frac{1}{\lambda}\Bigg\{2\,\tilde{\mu}\int_\Omega\sum_{i=1}^{3}\Big[N_i^\lambda\,h_{1i}^\lambda+B_i^\lambda\,h_{2i}^\lambda+\Big(\tau_i^{\lambda'}+x_1\,N_i^{\lambda'}+x_2\,B_i^{\lambda'}\Big)\,h_{3i}^\lambda\Big]\\
&\quad\times\Big[M_i\,h_{1i}^\lambda+D_i\,h_{2i}^\lambda+\Big(\mu_i'+x_1\,M_i'+x_2\,d_i'\Big)\,h_{3i}^\lambda\Big]\,\det J_\lambda\,d\bar{x}\\
&\quad-2\,\tilde{\mu}\int_\Omega\sum_{i=1}^{3}\Big[N_i\,h_{1i}+B_i\,h_{2i}+\Big(\tau_i'+x_1\,N_i'+x_2\,B_i'\Big)\,h_{3i}\Big]\\
&\quad\times\Big[M_i\,h_{1i}+D_i\,h_{2i}+\Big(\mu_i'+x_1\,M_i'+x_2\,D_i'\Big)\,h_{3i}\Big]\,\det J\,d\bar{x}\Bigg\}\\
&=2\,\tilde{\mu}\int_\Omega\sum_{i=1}^{3}\Bigg[\frac{N_i^\lambda-N_i}{\lambda}\,h_{1i}+\frac{B_i^\lambda-B_i}{\lambda}h_{2i}\\
&\qquad+\Bigg(\frac{\tau_i^{\lambda'}-\tau_i'}{\lambda}+x_1\frac{N_i^{\lambda'}-N_i'}{\lambda}+x_2\frac{B_i^{\lambda'}-B_{(i'}}{\lambda}\Bigg)\,h_{3i}\Bigg]\\
&\quad\times\Big[M_i\,h_{1i}+D_i\,h_{2i}+\Big(\mu_i'+x_1\,M_i'+x_2\,D_i'\Big)\,h_{3i}\Big]\,\det J\,d\bar{x}
\end{aligned}\tag{4.15}$$

$$
\begin{aligned}
&+2\,\tilde{\mu}\int\limits_{\Omega}\sum_{i=1}^{3}\left[N_i^{\lambda}\,h_{1i}+B_i^{\lambda}\,h_{2i}+\left(\tau_i^{\lambda'}+x_1\,N_i^{\lambda'}+x_2\,B_i^{\lambda'}\right)h_{3i}\right]\\
&\times\left[M_i\frac{h_{1i}^{\lambda}-h_{1i}}{\lambda}+D_i\frac{h_{2i}^{\lambda}-h_{2i}}{\lambda}+\left(\mu_i'+x_1\,M_i'+x_2\,D_i'\right)\frac{h_{3i}^{\lambda}-h_{3i}}{\lambda}\right]\det J\,d\bar{x}\\
&+2\,\tilde{\mu}\int\limits_{\Omega}\sum_{i=1}^{3}\left[N_i^{\lambda}\frac{h_{1i}^{\lambda}\,\det J_{\lambda}-h_{1i}\,\det J}{\lambda}+B_i^{\lambda}\frac{h_{2i}^{\lambda}\,\det J_{\lambda}-h_{2i}\,\det J}{\lambda}\right.\\
&\qquad\left.+\left(\tau_i^{\lambda'}+x_1\,N_i^{\lambda'}+x_2\,B_i^{\lambda'}\right)\frac{h_{3i}^{\lambda}\,\det J_{\lambda}-h_{3i}\,\det J}{\lambda}\right]\\
&\qquad\times\left[M_i\,h_{1i}^{\lambda}+D_i\,h_{2i}^{\lambda}+\left(\mu_i'+x_1\,M_i'+x_2\,D_i'\right)h_{3i}^{\lambda}\right]d\bar{x}\,.
\end{aligned}
$$

The terms $\mathbf{b}_1(\,\cdot\,,\,\cdot\,)$ and $\mathbf{b}_2(\,\cdot\,,\,\cdot\,)$ appearing in (2.13) (after replacing φ and ψ by $\varphi_\lambda\,,\,\psi_\lambda$) can be handled exactly as in (4.15). By summing up (4.14) and (4.15) (including the terms obtained from $\mathbf{b}_1$ and $\mathbf{b}_2$), we get the relation

$$
\mathcal{B}\left(\frac{\bar{y}_\lambda-\bar{y}}{\lambda}\,,\,\bar{v}\right)\,=\,Z_\lambda(\bar{v}) \tag{4.16}
$$

for any test function $\bar{v}\,=\,(\bar{\mu},\bar{M},\bar{D})\,\in\,H_0^1(0,L)^9$, and with some linear bounded operator $Z_\lambda : H_0^1(0,L)^9\to\mathbb{R}$ for any $\lambda\in\mathbb{R}_+$. More precisely, we have

$$
\begin{aligned}
Z_\lambda(\bar{v})\,=\,&\sum_{l=1}^{3}\int\limits_{\Omega}f_l(\bar{x})\Big(\mu_l(x_3)\,+\,x_1\,M_l(x_3)\,+\,x_2\,D_l(x_3)\Big)\,\tilde{\mathcal{D}}(\gamma,\xi)\,d\bar{x}\\
&+\,\sum_{i,j=1}^{3}\sum_{l=1}^{3}\int\limits_{\partial\Omega}g_l(\bar{x})\Big(\mu_l(x_3)\,+\,x_1\,M_l(x_3)\,+\,x_2\,D_l(x_3)\Big)\,\Bigg[\tilde{\mathcal{D}}(\gamma,\xi)\sqrt{\nu_i\,g^{ij}\,\nu_j}\\
&+\,\det J\,\frac{\nu_i\,\tilde{g}^{ij}(\gamma,\xi)\,\nu_j}{2\sqrt{\nu_i\,g^{ij}\,\nu_j}}\Bigg]\,d\tau\,+\,2\,\tilde{\mu}\int\limits_{\Omega}\sum_{i=1}^{3}\Bigg[N_i^{\lambda}\,h_{1i}\,+\,B_i^{\lambda}\,h_{2i}\,+\,\Big(\tau_i^{\lambda'}\,+\,x_1\,N_i^{\lambda'}\\
&+\,x_2\,B_i^{\lambda'}\Big)\,h_{3i}\Bigg]\,\Bigg[M_i\,\frac{h_{1i}^{\lambda}\,-\,h_{1i}}{\lambda}\,+\,D_i\,\frac{h_{2i}^{\lambda}\,-\,h_{2i}}{\lambda}\,+\,(\mu_i'\,+\,x_1\,M_i'\\
&+\,x_2\,D_i')\frac{h_{3i}^{\lambda}\,-\,h_{3i}}{\lambda}\Bigg]\,\det J\,d\bar{x}\,+\,2\,\tilde{\mu}\int\limits_{\Omega}\sum_{i=1}^{3}\Bigg[N_i^{\lambda}\,\frac{h_{1i}^{\lambda}\,\det J_{\lambda}\,-\,h_{1i}\,\det J}{\lambda}\\
&+\,B_i^{\lambda}\,\frac{h_{2i}^{\lambda}\,\det J_{\lambda}\,-\,h_{2i}\,\det J}{\lambda}\,+\,\Big(\tau_i^{\lambda}\,+\,x_1\,N_i^{\lambda'}\\
&+\,x_2\,B_i^{\lambda'}\Big)\,\frac{h_{3i}^{\lambda}\,\det J_{\lambda}\,-\,h_{3i}\,\det J}{\lambda}\Bigg]\,\Big[M_i\,h_{1i}^{\lambda}\,+\,D_i\,h_{2i}^{\lambda}\,+\,(\mu_i'\,+\,x_1\,M_i'\\
&+\,x_2\,D_i')\,h_{3i}^{\lambda}\Big]\,d\bar{x}\,+\,\hat{Z}_\lambda(\bar{v})\,.
\end{aligned}
$$

The operator $\hat{Z}_\lambda : H_0^1(0,L)^9\to\mathbb{R}$ is obtained from the terms $\mathbf{b}_1(\,\cdot\,,\,\cdot\,)$ and $\mathbf{b}_2(\,\cdot\,,\,\cdot\,)$ as explained after (4.15). We do not write it explicitly to save space. The relations

(4.14), (4.15) show that the following estimate is valid:

$$|Z_\lambda(\bar v)| \;\le\; C|\bar v|_{H_0^1(0,L)^9} \tag{4.17}$$

with some constant independent of $\lambda > 0$. Here, we use the differentiability properties of the coefficients, given in (4.1)–(4.12), and the convergence of $\bar y_\lambda$, according to (4.13).

By fixing $\bar v = \lambda^{-1}(\bar y_\lambda - \bar y)$, relations (4.16) and (4.17) show that $\{\frac{\bar y_\lambda - \bar y}{\lambda}\}$ is bounded in $H_0^1(0,L)^9$ for $\lambda > 0$, by the coercivity of $\mathcal{B}$. We may take a weakly convergent subsequence

$$\frac{\bar y_\lambda - \bar y}{\lambda} \to \hat y\,, \quad \text{weakly in } H_0^1(0,L)^9\,. \tag{4.18}$$

As in the previous section, one may see that the convergence is valid in the strong topology of $H_0^1(0,L)^9$. The equation in variations has the form

$$\mathcal{B}(\hat y, \bar v) \;=\; Z(\bar v) \quad \forall\ \bar v \;\in\; H_0^1(0,L)^9\,, \tag{4.19}$$

with $Z(\bar v) = \lim_{\lambda\to 0} Z_\lambda(\bar v)$, which exists by the above discussion. Z depends linearly and boundedly on $(\gamma,\xi) \in C^1[0,L]^2$.

Note that (4.19) has a unique solution $\hat y \in H_0^1(0,L)^9$. We have proved the following result.

PROPOSITION 4.1. *The mapping $(\varphi,\psi) \in C^1[0,L]^2 \mapsto \bar y \in H_0^1(0,L)^9$ is Gâteaux differentiable, and the derivative $\hat y$ satisfies* (4.19).

We introduce now the so-called adjoint system, with unknowns $\bar T = (\bar R, \bar P, \bar Q) \in H_0^1(0,L)^9$ and defined by

$$\mathcal{B}(\bar T, \bar v) \;=\; \nabla_2\, j(\bar\theta, \bar y)(\bar v) \quad \forall\ \bar v \;\in\; H_0^1(0,L)^9\,. \tag{4.20}$$

In (4.20), we assume that $j : C^2[0,L]^3 \times H_0^1(0,L)^9 \to \mathbb{R}$ is Fréchet differentiable and that $\nabla_2 j$ denotes the second component of ∇j or, equivalently, the partial Fréchet differential with respect to $\bar y$. The existence and uniqueness of a solution $\bar T \in H_0^1(0,L)^9$ to (4.20) is obvious, due to the coercivity and boundedness of $\mathcal{B}(\,\cdot\,,\,\cdot\,)$.

PROPOSITION 4.2. *If j is Fréchet differentiable, then the directional derivative of the cost functional* Π *in problem* (P) *at the point $(\varphi,\psi) \in C^1[0,L]^2$ and in the direction $(\gamma,\xi) \in C^1[0,L]^2$ is given by*

$$\begin{aligned}
&\nabla\Pi(\varphi,\psi)\,(\gamma,\xi)\\
&= \nabla_1 j(\bar\theta,\bar y)\,\tilde\theta(\gamma,\xi) + \sum_{l=1}^{3}\int_\Omega f_l(\bar x)\Big(R_l(x_3) + x_1\,P_l(x_3) + x_2\,Q_l(x_3)\Big)\,\tilde{\mathcal{D}}(\gamma,\xi)\,d\bar x\\
&\quad + \sum_{i,j=1}^{3}\sum_{l=1}^{3}\int_{\partial\Omega} g_l(\bar x)\Big(R_l(x_3) + x_1\,P_l(x_3) + x_2\,Q_l(x_3)\Big)\,\tilde{\mathcal{D}}(\gamma,\xi)\sqrt{\nu_i\,g^{ij}\nu_j}\,d\tau\\
&\quad + \sum_{i,j=1}^{3}\sum_{l=1}^{3}\int_{\partial\Omega} g_l(\bar x)\Big(R_l(x_3) + x_1\,P_l(x_3) + x_2\,Q_l(x_3)\Big)\\
&\qquad \times \det J\,\frac{1}{\sqrt{\nu_i\,g^{ij}\nu_j}}\,\nu_i\tilde g^{ij}(\gamma,\xi)\nu_j\,d\tau\\
&\quad - 2\,\tilde\mu\int_\Omega \sum_{i=1}^{3}\Big[N_i\,h_{1i} + B_i\,h_{2i} + \Big(\tau_i' + x_1\,N_i' + x_2\,B_i'\Big)\,h_{3i}\Big]
\end{aligned}$$

$$\times\left[P_i\,\tilde h_{1i}(\gamma,\xi)+Q_i\,\tilde h_{2i}(\gamma,\xi)+\Big(R_i'+x_1\,P_i'+x_2\,Q_i'\Big)\,\tilde h_{3i}(\gamma,\xi)\right]\,\det J\,d\bar x$$

$$-2\,\tilde\mu\int\limits_\Omega\sum_{i=1}^3\left[N_i\,\tilde h_{1i}(\gamma,\xi)+B_i\,\tilde h_{2i}(\gamma,\xi)+\Big(\tau_i'+x_1\,N_i'+x_2\,B_i'\Big)\,\tilde h_{3i}(\gamma,\xi)\right]$$

$$\times\left[P_i\,h_{1i}+Q_i\,h_{2i}+\Big(R_i'+x_1\,P_i'+x_2\,Q_i'\Big)\,h_{3i}\right]\,\det J\,d\bar x$$

$$-2\,\tilde\mu\int\limits_\Omega\sum_{i=1}^3\left[N_i\,h_{1i}+B_i\,h_{2i}+\Big(\tau_i'+x_1\,N_i'+x_2\,B_i'\Big)\,h_{3i}\right]$$

$$\times\left[P_i\,h_{1i}+Q_i\,h_{2i}+\Big(R_i'+x_1\,P_i'+x_2\,Q_i'\Big)\,h_{3i}\right]\tilde{\mathcal D}(\gamma,\xi)\,d\bar x$$

$$-\tilde\lambda\int\limits_\Omega\sum_{i,j=1}^3\left[N_i\,h_{1i}+B_i\,h_{2i}+\Big(\tau_i'+x_1\,N_i'+x_2\,B_i'\Big)\,h_{3i}\right]$$

$$\times\left[P_j\,\tilde h_{1j}(\gamma,\xi)+Q_j\,\tilde h_{2j}(\gamma,\xi)+\Big(R_j'+x_1\,P_j'+x_2\,Q_j'\Big)\,\tilde h_{3j}(\gamma,\xi)\right]\,\det J\,d\bar x$$

$$-\tilde\lambda\int\limits_\Omega\sum_{i,j=1}^3\left[N_i\,\tilde h_{1i}(\gamma,\xi)+B_i\,\tilde h_{2i}(\gamma,\xi)+\Big(\tau_i'+x_1\,N_i'+x_2\,B_i'\Big)\,\tilde h_{3i}(\gamma,\xi)\right]$$

$$\times\left[P_j\,h_{1j}+Q_j\,h_{2j}+\Big(R_j'+x_1\,P_j'+x_2\,Q_j'\Big)\,h_{3j}\right]\,\det J\,d\bar x$$

$$-\tilde\lambda\int\limits_\Omega\sum_{i,j=1}^3\left[N_i\,h_{1i}+B_i\,h_{2i}+\Big(\tau_i'+x_1\,N_i'+x_2\,B_i'\Big)\,h_{3i}\right]$$

$$\times\left[P_j\,h_{1j}+Q_j\,h_{2j}+\Big(R_j'+x_1\,P_j'+x_2\,Q_j'\Big)\,h_{3j}\right]\tilde{\mathcal D}(\gamma,\xi)\,d\bar x$$

$$-\tilde\mu\int\limits_\Omega\sum_{i<j}\left[N_i\,h_{1j}+B_i\,h_{2j}+\Big(\tau_i'+x_1\,N_i'+x_2\,B_i'\Big)\,h_{3j}+N_j\,h_{1i}+B_j\,h_{2i}\right.$$

$$\left.+\Big(\tau_j'+x_1\,N_j'+x_2\,B_j'\Big)\,h_{3i}\right]$$

$$\times\left[P_i\,\tilde h_{1j}(\gamma,\xi)+\,Q_i\,\tilde h_{2j}(\gamma,\xi)+\Big(R_i'+x_1\,P_i'+x_2\,Q_i'\Big)\,\tilde h_{3j}(\gamma,\xi)+P_j\,\tilde h_{1i}(\gamma,\xi)\right.$$

$$\left.+Q_j\,\tilde h_{2i}(\gamma,\xi)+\Big(R_j'+x_1\,P_j'+x_2\,Q_j'\Big)\,\tilde h_{3i}(\gamma,\xi)\right]\,\det J\,d\bar x$$

$$-\tilde\mu\int\limits_\Omega\sum_{i<j}\left[N_i\,\tilde h_{1j}(\gamma,\xi)+B_i\,\tilde h_{2j}(\gamma,\xi)+\Big(\tau_i'+x_1\,N_i'+x_2\,B_i'\Big)\,\tilde h_{3j}(\gamma,\xi)\right.$$

$$
\begin{aligned}
&+ N_j\,\tilde h_{1i}(\gamma,\xi) + B_j\,\tilde h_{2i}(\gamma,\xi) + \Big(\tau_j' + x_1\,N_j' + x_2\,B_j'\Big)\,\tilde h_{3i}(\gamma,\xi)\Big]\\
&\times\Big[P_i\,h_{1j} + Q_i\,h_{2j} + \Big(R_i' + x_1\,P_i' + x_2\,Q_i'\Big)\,h_{3j} + P_j\,h_{1i} + Q_j\,h_{2i}\\
&\quad + \Big(R_j' + x_1\,P_j' + x_2\,Q_j'\Big)\,h_{3i}\Big]\,\det J\,d\bar x\\
-\tilde\mu\int_\Omega \sum_{i<j}\Big[&N_i\,h_{1j} + B_i\,h_{2j} + \Big(\tau_i' + x_1\,N_i' + x_2\,B_i'\Big)\,h_{3j}\\
&+ N_j\,h_{1i} + B_j\,h_{2i} + \Big(\tau_j' + x_1\,N_j' + x_2\,B_j'\Big)\,h_{3i}\Big]\\
&\times\Big[P_i\,h_{1j} + Q_i\,h_{2j} + \Big(R_i' + x_1\,P_i' + x_2\,Q_i'\Big)\,h_{3j} + P_j\,h_{1i} + Q_j\,h_{2i}\\
&\quad + \Big(R_j' + x_1\,P_j' + x_2\,Q_j'\Big)\,h_{3i}\Big]\,\tilde{\mathcal{D}}(\gamma,\xi)\,d\bar x\,.
\end{aligned}
\tag{4.21}
$$

Remark 4.3. In order to compute (4.21), from (φ,ψ) and $(\gamma,\xi)\in C^1[0,L]^2$, one has to compute $\bar\theta\in C^2[0,L]^3$ by (2.1), $\bar y=(\bar\tau,\bar N,\bar B)\in H_0^1(0,L)^9$ by (2.13), $\bar T=(\bar R,\bar P,\bar Q)\in H_0^1(0,L)^9$ by (2.20), and use (4.1)–(4.12), where γ, ξ enter effectively. See the explicit form of $\tilde t(\gamma,\xi)$ given after (4.12). Writing all the operators in (4.1)–(4.12) explicitly, and replacing them in (4.21), would give the "full" expression for $\nabla\Pi(\varepsilon,\psi)(\gamma,\xi)$, which we do not write down to save space. Since spaces of continuous functions are taken into account, it is not advantageous to rewrite (4.21) by using adjoint operators.

Note also that the above argument holds if φ,ψ,γ,ξ are only piecewise continuously differentiable. This is important for the numerical experiments in section 8.

Remark 4.4. Assuming that the cross section of the rod is not constant, one may study optimization problems with respect to the cross section as well, under appropriate regularity conditions.

Let $\mathcal{C}=\{(\varphi,\psi)\in C^1[0,L]^2\,;\,\bar\theta(\varphi,\psi)\in\mathcal{K}\}$ and $u_0=(\varphi_0,\psi_0)\in\mathcal{C}$ be arbitrarily fixed. We denote by

$$
T(\mathcal{C};u_0)=\Big\{u\in C^1[0,L]^2\,;\,u=\lim_{n\to\infty}\lambda_n(u_n-u_0)\,,\ \lambda_n\ge 0\,,\ u_n\in\mathcal{C}\,,\ \text{and } u_n\to u_0\Big\}
$$

the cone of tangents to $\mathcal{C}$ at u_0 (see Barbu and Precupanu [4]). It is known that if $\mathcal{C}$ is convex (see examples (2.15), (2.16) and Remark 2.3), then $T(\mathcal{C};u_0)=\overline{\bigcup_{\lambda>0}\lambda(\mathcal{C}-u_0)}$.

COROLLARY 4.5. *Assume that $u^*=(\varphi^*,\psi^*)$ is a (local) optimum point for* (P). *Then the following statements are valid:*

(i) *If* Π *is Fréchet differentiable on* $C^1[0,L]^2$, *then*

$$
\nabla\Pi(\varphi^*,\psi^*)(\gamma,\xi)\ge 0\quad\forall\,(\gamma,\xi)\,\in\,T(\mathcal{C};u^*)\,.
$$

(ii) *If* $\mathcal{C}$ *is convex, then the directional derivative of* Π *satisfies*

$$
\nabla\Pi(\varphi^*,\psi^*)(\gamma,\xi)\ge 0\quad\forall\,(\gamma,\xi)\,\in\,\mathcal{C}-u^*\,.
$$

Remark 4.6. Corollary 4.5 gives the standard first-order optimality conditions for problem (P) (see Tröltzsch [20]). Relations (4.21), (4.20), etc., indicate the explicit calculation of the directional derivative of the cost functional and will be used in the last section in the numerical experiments.

5. Formulation of the shell optimization problem. Let $o \subset \mathbb{R}^2$ denote a bounded domain, not necessarily simply connected, with Lipschitz boundary ∂o. Define

$$\Omega \,=\, o \times]-\varepsilon\,,\,\varepsilon[\subset \mathbb{R}^3$$

for some "small" $\varepsilon > 0$. We denote by $(x_1, x_2) \in o$ and $x_3 \in]-\varepsilon, \varepsilon[$, $\bar{x} = (x_1, x_2, x_3) \in \Omega$ the independent variables.

Let $p : o \to \mathbb{R}$ be a $C^2(\bar{o})$ mapping, whose graph represents the middle surface $\mathcal{S}$ of a shell. We introduce the geometrical transformation

$$F \,:\, \Omega \,\to\, \mathbb{R}^3\,,$$

$$F(\bar{x}) \,=\, \bar{\pi}(x_1, x_2) \,+\, x_3\, \bar{n}(x_1, x_2)\,, \tag{5.1}$$

with $\bar{\pi} = (\pi_1, \pi_2, \pi_3) = (x_1, x_2, p(x_1, x_2))$, and with $\bar{n} = (n_1, n_2, n_3)$ denoting the normal vector to $\mathcal{S}$ in the point $\bar{\pi}(x_1, x_2)$. Since the tangent vectors $\frac{\partial \bar{\pi}}{\partial x_1} = (1, 0, p_1)$ and $\frac{\partial \bar{\pi}}{\partial x_2} = (0, 1, p_2)$, with $p_1 = \frac{\partial p}{\partial x_1}$ and $p_2 = \frac{\partial p}{\partial x_2}$, are always linearly independent, we may take $\bar{n}$ as the normalization of $\frac{\partial \bar{\pi}}{\partial x_1} \wedge \frac{\partial \bar{\pi}}{\partial x_2}$, that is,

$$\bar{n} \,=\, \frac{1}{\sqrt{1 + p_1^2 + p_2^2}}(-p_1, -p_2, 1)\,. \tag{5.2}$$

Assume that $\partial o = \bar{\gamma}_0 \cup \bar{\gamma}_1$, with $\gamma_0\,,\, \gamma_1$ being nonoverlapping open parts of ∂o such that $\mathrm{meas}\,(\gamma_0) > 0$, and let $\Gamma_0 := \gamma_0 \times]-\varepsilon, \varepsilon[$, $\Gamma_1 := \partial\Omega \setminus \Gamma_0$. We introduce the notation

$$\hat{\Omega} := F(\Omega)\,, \quad \hat{\Gamma}_0 := F(\Gamma_0)\,, \quad \hat{\Gamma}_1 := F(\Gamma_1)\,.$$

We argue later (see (5.9)) that F is a homeomorphism for small ε, and the open set $\hat{\Omega}$ will represent a shell. We assume that body forces $\hat{f} \in L^2(\hat{\Omega})^3$ and surface tractions $\hat{g} \in L^2(\hat{\Gamma}_1)^3$ act on the shell. Our main mechanical assumption is that the corresponding displacement $\hat{u} \in V(\hat{\Omega}) = \{\hat{v} \in H^1(\hat{\Omega})^3\,;\, \hat{v}|_{\hat{\Gamma}_0} = 0\}$ has the form

$$\hat{u}(\hat{x}) \,=\, \bar{u}(x_1, x_2) \,+\, x_3\, \bar{r}(x_1, x_2)\,, \quad x \,\in\, \hat{\Omega}\,. \tag{5.3}$$

Here, $\bar{x} = (x_1, x_2, x_3) = F^{-1}(\hat{x}) \,\in\, \Omega$ and $\bar{u} = (u_1, u_2, u_3)\,,\, \bar{r} = (r_1, r_2, r_3)$ belong to the Hilbert space

$$V(o) \,=\, \{\bar{v} \,=\, (v_1, v_2, v_3) \,\in\, H^1(o)^3\,;\, \bar{v}|_{\gamma_0} \,=\, 0\}\,, \tag{5.4}$$

equipped with the norm

$$|\bar{v}|_{V(o)} := \int_o \left(|\nabla v_1|^2 + |\nabla v_2|^2 + |\nabla v_3|^2\right) dx_1 dx_2\,.$$

If we denote by $\tilde{V}(\hat{\Omega})$ the subspace of $V(\hat{\Omega})$ defined by (5.3), (5.4), we can see that $\tilde{V}(\hat{\Omega})$ can simply be identified with $V(o) \times V(o)$, and we shall do this repeatedly later in this paper.

Clearly, $\bar{u}$ represents the displacement of the middle surface $\mathcal{S}$ of the shell, while $\bar{r}$ is the modification of the points along the normal $\bar{n}(x_1, x_2)$, assumed to remain on a line. The form (5.3) allows for both dilation and contraction of the elastic material; it is a generalization of the classical Naghdi model (see Ciarlet [11] and Blouza [6]).

The Jacobian $J = D\,F$ of F is given by

$$J(\bar{x}) = \begin{bmatrix} 1 + x_3 \dfrac{\partial n_1}{\partial x_1} & x_3 \dfrac{\partial n_1}{\partial x_2} & n_1 \\ x_3 \dfrac{\partial n_2}{\partial x_1} & 1 + x_3 \dfrac{\partial n_2}{\partial x_2} & n_2 \\ p_1 + x_3 \dfrac{\partial n_3}{\partial x_1} & p_2 + x_3 \dfrac{\partial n_3}{\partial x_2} & n_3 \end{bmatrix}. \tag{5.5}$$

As $|\bar{n}|^2_{\mathbb{R}^3} = 1$, we get $\langle \bar{n}\,,\, \frac{\partial \bar{n}}{\partial x_i} \rangle_{\mathbb{R}^3} = 0\,,\ i = 1, 2$, which shows that $\frac{\partial \bar{n}}{\partial x_i}$ can be generated by $\frac{\partial \bar{\pi}}{\partial x_1}$ and $\frac{\partial \bar{\pi}}{\partial x_2}$. We get the relations

$$\frac{\partial \bar{n}}{\partial x_1}(x_1, x_2) = \frac{\partial n_1}{\partial x_1}\frac{\partial \bar{\pi}}{\partial x_1} + \frac{\partial n_2}{\partial x_1}\frac{\partial \bar{\pi}}{\partial x_2}\,, \tag{5.6}$$

$$\frac{\partial \bar{n}}{\partial x_2}(x_1, x_2) = \frac{\partial n_1}{\partial x_2}\frac{\partial \bar{\pi}}{\partial x_1} + \frac{\partial n_2}{\partial x_1}\frac{\partial \bar{\pi}}{\partial x_2}\,, \tag{5.7}$$

which are special cases of the equations of motion of the local frame on the surface $\mathcal{S}$; see Cartan [7]. The coefficients $\frac{\partial n_i}{\partial x_\alpha}\,,\ i = \overline{1,3}\,,\ \alpha = 1, 2$, are related to the curvatures of $\mathcal{S}$.

Equalities (5.5)–(5.7) yield

$$\det J(\bar{x}) = \left[1 + x_3\left(\frac{\partial n_1}{\partial x_1} + \frac{\partial n_2}{\partial x_2}\right) + x_3^2\left(\frac{\partial n_1}{\partial x_1}\frac{\partial n_2}{\partial x_2} - \frac{\partial n_1}{\partial x_2}\frac{\partial n_2}{\partial x_1}\right)\right] \times \sqrt{1 + p_1^2 + p_2^2}\,. \tag{5.8}$$

Since $p \in C^2(\bar{o})$, for "small" $\varepsilon > 0$ we get that

$$\det J(\bar{x}) \geq c > 0 \quad \forall\, \bar{x} \in \Omega\,. \tag{5.9}$$

Let us notice that (5.9) justifies the definition of the shell $\hat{\Omega}$ via the transformation F; see Ciarlet [11, Thm. 3.1-1].

We denote the elements of $J(\bar{x})^{-1}$ by

$$J(\bar{x}) = (h_{ij}(\bar{x}))_{i,j=\overline{1,3}}\,. \tag{5.10}$$

In Sprekels and Tiba [18], the following generalized Naghdi model is obtained:

$$\begin{aligned} &\mathcal{B}\left((\bar{u}, \bar{r})\,,\, (\bar{\mu}, \bar{\rho})\right) \\ &= \tilde{\lambda} \int_\Omega \left\{ \sum_{i=1}^{3} \left[\left(\frac{\partial u_i}{\partial x_1} + x_3 \frac{\partial r_i}{\partial x_1} \right) h_{1i} + \left(\frac{\partial u_i}{\partial x_2} + x_3 \frac{\partial r_i}{\partial x_2} \right) h_{2i} + r_i\, h_{3i} \right] \right\} \\ &\times \left\{ \sum_{j=1}^{3} \left[\left(\frac{\partial \mu_j}{\partial x_1} + x_3 \frac{\partial \rho_j}{\partial x_1} \right) h_{1j} + \left(\frac{\partial \mu_j}{\partial x_2} + x_3 \frac{\partial \rho_j}{\partial x_2} \right) h_{2j} + \rho_j\, h_{3j} \right] \right\} |\det J(\bar{x})|\, d\bar{x} \end{aligned}$$

$$
\begin{aligned}
&+2\,\tilde{\mu}\int_{\Omega}\sum_{i=1}^{3}\left[\left(\frac{\partial u_i}{\partial x_1}+x_3\frac{\partial r_i}{\partial x_1}\right)h_{1i}+\left(\frac{\partial u_i}{\partial x_2}+x_3\frac{\partial r_i}{\partial x_2}\right)h_{2i}+r_i\,h_{3i}\right]\\
&\times\left[\left(\frac{\partial \mu_i}{\partial x_1}+x_3\frac{\partial \rho_i}{\partial x_1}\right)h_{1i}+\left(\frac{\partial \mu_i}{\partial x_2}+x_3\frac{\partial \rho_i}{\partial x_2}\right)h_{2i}+\rho_i\,h_{3i}\right]|\det J(\bar{x})|\,d\bar{x}\\
&+\tilde{\mu}\int_{\Omega}\sum_{1\le i<j\le 3}\left[\left(\frac{\partial u_i}{\partial x_1}+x_3\frac{\partial r_i}{\partial x_1}\right)h_{1j}+\left(\frac{\partial u_i}{\partial x_2}+x_3\frac{\partial r_i}{\partial x_2}\right)h_{2j}+r_i\,h_{3j}\right.\\
&\left.+\left(\frac{\partial u_j}{\partial x_1}+x_3\frac{\partial r_j}{\partial x_1}\right)h_{1i}+\left(\frac{\partial u_j}{\partial x_2}+x_3\frac{\partial r_j}{\partial x_2}\right)h_{2i}+r_j\,h_{3i}\right]\\
&\times\left[\left(\frac{\partial \mu_i}{\partial x_1}+x_3\frac{\partial \rho_i}{\partial x_1}\right)h_{1j}+\left(\frac{\partial \mu_i}{\partial x_2}+x_3\frac{\partial \rho_i}{\partial x_2}\right)h_{2j}+\rho_i\,h_{3j}\right.\\
&\left.+\left(\frac{\partial \mu_j}{\partial x_1}+x_3\frac{\partial \rho_j}{\partial x_1}\right)h_{1i}+\left(\frac{\partial \mu_j}{\partial x_2}+x_3\frac{\partial \rho_j}{\partial x_2}\right)h_{2i}+r_j\,h_{3i}\right]|\det J(\bar{x}|\,d\bar{x}\\
&=\int_{\Omega}\sum_{l=1}^{3}f_l(\mu_l+x_3\,\rho_l)\,|\det J(\bar{x})|\,d\bar{x}+\int_{\Gamma_1}\sum_{l=1}^{3}\sum_{i,j=1}^{3}g_l(\mu_l+x_3\,\rho_l)\,|\det J(\bar{x})|\\
&\times\sqrt{\nu_i(\bar{x})\,g^{ij}(\bar{x})\,\nu_j(\bar{x})}\,d\tau\quad\forall\ (\bar{\mu},\bar{\rho})\ \in\ V(o)^2\,. \qquad (5.11)
\end{aligned}
$$

Here, $\bar{f}(\bar{x})=\hat{f}(F\bar{x})$, $\bar{g}(\bar{x})=\hat{g}(F\bar{x})$, $\bar{x}\in\Omega$, we use the assumed form (5.3) of the displacement, and $\bar{\mu}\in V(o)$, $\bar{\rho}\in V(o)$ are arbitrary test functions. The coefficients g^{ij} are obtained by

$$
\Big(g^{ij}(\bar{x})\Big)_{i,j=\overline{1,3}}\ =\ J(\bar{x})^{-1}\Big[J(\bar{x})^{T}\Big]^{-1}\,, \qquad (5.12)
$$

and $(\nu_i(\bar{x}))_{i=\overline{1,3}}$ is the unit outside normal to Γ_1 at $\bar{x}\in\Gamma_1$.

The coercivity of $\mathcal{B}$ on $V(o)\times V(o)$ was proved by Sprekels and Tiba [18] for ε small enough. This gives the existence and the uniqueness of the solution $(\bar{u},\bar{r})\in V(o)\times V(o)$ to (5.11).

For given $\bar{f}$ and $\bar{g}$ (defined in a sufficiently large ball in $\mathbb{R}^3$), we consider the following general shape optimization problem associated with (5.11):

$$
\min_{p}\ \left\{\Pi(p)\ =\ j\Big(\bar{y}(x_1,x_2)\,,\,p(x_1,x_2)\Big)\right\} \qquad (\mathrm{P}')
$$

with $\bar{y}(x_1,x_2)=(\bar{u}(x_1,x_2),\bar{r}(x_1,x_2))\in V(o)^2$ given by (2.11), and subject to the "control" constraint $p\in\mathcal{K}\subset C^2(\bar{o})$, closed and bounded. Notice that (5.9) should be included in the definition of $\mathcal{K}$. The mapping $j:V(o)^2\times C^2(o)\to\mathbb{R}$ satisfies certain regularity properties to be described later. One classical example is the quadratic case

$$
2j(\bar{y},p)\ =\ |u_1|^2_{V(o)}\ +\ |u_2|^2_{V(o)}\ +\ |u_3|^2_{V(o)}\,. \qquad (5.13)
$$

Then (P′) aims at finding the shape of the shell (the surface $\mathcal{S}$) that minimizes the displacement of the middle surface under prescribed body forces and tractions.

Concerning the constraints to which the shell itself may be submitted and which are abstractly written as $p \in \mathcal{K}$, there is a large variety of examples. We just list

$$0 \le p(x_1, x_2) \quad \forall\, (x_1, x_2) \in o \tag{5.14}$$

(pointwise constraints),

$$\int_o p(x_1, x_2)\, dx_1\, dx_2 \ge c \tag{5.15}$$

(integral constraints). A special integral constraint is to prescribe limits for the area of $\mathcal{S}$:

$$\int_o \sqrt{1 + p_1^2 + p_2^2} \ge \beta\,. \tag{5.16}$$

Although all the examples (5.13)–(5.16) have a convex nature, the shape optimization problem (P′) is strongly nonconvex, since the dependence $p \mapsto \bar{y}$ is nonlinear. (P′) is a control-into-coefficients problem.

6. Existence of optimal shells. First we prove the following continuous dependence result.

THEOREM 6.1. *Assume that* $p_n : \bar{o} \to \mathbb{R}$ *and* $p_n \to p$ *in* $C^2(\bar{o})$. *If* $\bar{y}_n = (\bar{u}_n, \bar{r}_n)$ *and* $\bar{y} = (\bar{u}, \bar{r})$ *are the solutions of* (5.11) *corresponding to* p_n, p, *then* $\bar{y}_n \to \bar{y}$ *strongly in* $V(o)^2$ *for sufficiently small* $\varepsilon > 0$.

Relations (5.1), (5.2), (5.5), and (5.8) give (with obvious notation)

$$\bar{n}_n \to \bar{n} \quad \text{in } C^1(\bar{o})^3\,, \tag{6.1}$$

$$F_n = \bar{\pi}_n + x_3\, \bar{n}_n \to F = \bar{\pi} + x_3\, \bar{n} \quad \text{in } C^1(\bar{\Omega})^3\,, \tag{6.2}$$

$$J_n \to J \quad \text{in } C(\bar{\Omega})^9\,, \tag{6.3}$$

$$\det J_n \to \det J \quad \text{in } C(\bar{\Omega})\,. \tag{6.4}$$

Notice that

$$J(\bar{x}) = \begin{bmatrix} 1 & 0 & n_1 \\ 0 & 1 & n_2 \\ p_1 & p_2 & n_3 \end{bmatrix} \begin{bmatrix} 1 + x_3 \frac{\partial n_1}{\partial x_1} & x_3 \frac{\partial n_1}{\partial x_2} & 0 \\ x_3 \frac{\partial n_2}{\partial x_1} & 1 + x_3 \frac{\partial n_2}{\partial x_2} & 0 \\ 0 & 0 & 1 \end{bmatrix} = S\,R =: S(I + x_3\, M) \tag{6.5}$$

(new matrix notation).

Similarly, we have

$$J_n = S_n\, R_n = S_n(I + x_3\, M_n)\,. \tag{6.6}$$

A simple calculus gives

$$S_n^{-1} = \frac{1}{\sqrt{1 + (p_1^n)^2 + (p_2^n)^2}} \begin{bmatrix} n_3^n - n_2^n\, p_2^n & n_1^n\, p_2^n & -n_1^n \\ n_2^n\, p_1^n & n_3^n - n_1^n\, p_1^n & -n_2^n \\ -p_1^n & -p_2^n & 1 \end{bmatrix}$$

$$\longrightarrow \frac{1}{\sqrt{1 + p_1^2 + p_2^2}} \begin{bmatrix} n_3 - n_2\, p_2 & n_1\, p_2 & -n_1 \\ n_2\, p_1 & n_3 - n_1\, p_1 & -n_2 \\ -p_1 & -p_2 & 1 \end{bmatrix} = S^{-1}\,,$$

strongly in $C^1(\bar{o})$. Moreover,

$$R_n^{-1} = (I + x_3 M_n)^{-1} = I - x_3 M_n + x_3^2 M_n^2 - x_3^3 M_n^3 + \dots \tag{6.7}$$

for ε small. Clearly, we have

$$M_n = \begin{bmatrix} \dfrac{\partial n_1^n}{\partial x_1} & \dfrac{\partial n_1^n}{\partial x_2} & 0 \\ \dfrac{\partial n_2^n}{\partial x_1} & \dfrac{\partial n_2^n}{\partial x_2} & 0 \\ 0 & 0 & 0 \end{bmatrix} \longrightarrow M = \begin{bmatrix} \dfrac{\partial n_1}{\partial x_1} & \dfrac{\partial n_1}{\partial x_2} & 0 \\ \dfrac{\partial n_2}{\partial x_1} & \dfrac{\partial n_2}{\partial x_2} & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{6.8}$$

in $C(\bar{o})$. Relations (6.7) and (6.8) show (by a passage to the limit in the infinite sum, $n \to \infty$) that $R_n^{-1} \to R^{-1}$ in $C(\bar{\Omega})^9$ for ε small.

Then, (6.6) and the above argument give

$$J_n^{-1} \longrightarrow J^{-1} \quad \text{in } C(\bar{\Omega})^9 \,. \tag{6.9}$$

In particular, we have that

$$h_{ij}^n(\bar{x}) \longrightarrow h_{ij}(\bar{x}) \quad \text{in } C(\bar{\Omega}) \quad \forall\, i,j = \overline{1,3}\,, \tag{6.10}$$

$$g_n^{ij}(\bar{x}) \longrightarrow g^{ij}(\bar{x}) \quad \text{in } C(\bar{\Omega}) \quad \forall\, i,j = \overline{1,3}\,, \tag{6.11}$$

according to (5.10), (5.12), (6.3), and (6.4).

Let $\mathcal{B}_n$ denote the bilinear form $\mathcal{B}$ from (5.11) with coefficients h_{ij}^n, $\det J_n$. We show that it has a coercivity constant independent of $n \in \mathbb{N}$ for $\varepsilon > 0$ small enough (again independently of n).

Proposition 6.2. *Assume that $\mathcal{K}$ is bounded in $C^2(\bar{o})$ and that $\varepsilon < \varepsilon(\mathcal{K})$ and $\delta \ll \varepsilon$ are given positive numbers. There are $c = c(\mathcal{K}) > 0$ and $m = m(\mathcal{K}) > 0$ such that*

$$\mathcal{B}_p(\hat{u},\hat{u}) \geq c\Big[\varepsilon|\bar{u}|^2_{V(o)} + \varepsilon^3|\bar{r}|^2_{V(o)}\Big] - \frac{m}{\delta}\Big[|\bar{r}|^2_{L^2(o)^3} + |\bar{u}|^2_{L^2(o)^3}\Big] \tag{6.12}$$

for any $p \in \mathcal{K}$ and any $\hat{u} \in H^1(\hat{\Omega})^3$ given by (5.3).

The constant $\varepsilon(\mathcal{K}) > 0$ depends on $c_i > 0$, $i = 1,2$, defined below in (6.23) and in Lemma 6.3. It should be small enough such that (5.8) is fulfilled, which is possible due to the boundedness of $\mathcal{K}$ in $C^2(\bar{o})$. The precise significance of $\varepsilon(\mathcal{K})\,,\ c(\mathcal{K})\,,\ m(\mathcal{K})$ is indicated in the proof.

The notation $\mathcal{B}_p(\,\cdot\,,\,\cdot\,)$ signifies the bilinear functional (5.11) associated with some $p \in \mathcal{K}$. We prove Proposition 6.2 only for the case $\bar{u}, \bar{r} \in H_0^1(o)^3 = V(o)$, in order to avoid more technical arguments related to the extension of $\hat{u}$ to $H_0^1(\mathbb{R})^3$.

Proof. We consider the mapping $\bar{w} \in H^1(\Omega)^3$, given by

$$\bar{w}(x_1,x_2,x_3) = \bar{u}(x_1,x_2) + x_3\,\bar{r}(x_1,x_2)\,, \tag{6.13}$$

such that $\hat{u}(\hat{x}) = \bar{w}(F^{-1}\hat{x})\,,\ \hat{x} \in \hat{\Omega}\,,\ \bar{x} = F^{-1}\hat{x} \in \Omega$. We denote

$$S^+ = [\varepsilon\,,\,\varepsilon+\delta] \times \bar{o}\,, \quad S^- = [-\varepsilon-\delta\,,\,-\varepsilon] \times \bar{o}\,. \tag{6.14}$$

We extend $\bar{w}$ to $\Omega \cup S^+ \cup S^-$ by $\tilde{w}|_\Omega = \bar{w}$, and

$$\tilde{w}(\bar{x}) \;=\; \delta^{-1}\{[(\varepsilon+\delta)-x_3]\,\bar{u}(x_1,x_2)+\varepsilon\,(\varepsilon+\delta-x_3)\,\bar{r}(x_1,x_2)\} \tag{6.15}$$

for $\bar{x} \in S^+$,

$$\tilde{w}(\bar{x}) \;=\; \delta^{-1}\,\{(\varepsilon+\delta+x_3)\,\bar{u}(x_1,x_2)-\varepsilon\,(\varepsilon+\delta+x_3)\,\bar{r}(x_1,x_2)\} \tag{6.16}$$

for $\bar{x} \in S^-$.

Then, we may extend $\tilde{w}$ by 0 to $\mathbb{R}^3$ as $\bar{u}, \bar{r} \in H_0^1(o)^3$. In the general case of a partially clamped shell, one has to use an extension procedure around $o \subset \mathbb{R}^2$, too (for instance, the Calderon extension (Adams [1]), since ∂o is assumed Lipschitzian).

We may assume that F_p, i.e., the transformation (5.1) associated with any $p \in \mathcal{K}$, is still one-to-one on $\Omega \cup S^+ \cup S^-$, since $\varepsilon+\delta$ is "small" and $\mathcal{K}$ is bounded (see (5.8)). We denote

$$\Sigma_p^+ \;:=\; F_p(S^+)\,, \qquad \Sigma_p^- \;:=\; F_p(S^-)\,. \tag{6.17}$$

Above, the index $p \in \mathcal{K}$ puts into evidence the dependence on p of the geometrical transformation and of the sets. We introduce the extension of $\hat{u} \in H^1(\hat{\Omega}_p)^3$ by

$$\tilde{u}(\hat{x}) \;=\; \tilde{w}\Big(F_p^{-1}(\hat{x})\Big)\,. \tag{6.18}$$

Clearly, it holds that $\tilde{u} \in H_0^1(\hat{\Omega}_p \cup \Sigma_p^+ \cup \Sigma_p^-)$.

As $\mathcal{K}$ is bounded in $C^2(\bar{o})$, there is a ball O in $\mathbb{R}^3$ such that $O \supset \hat{\Omega}_p \cup \Sigma_p^+ \cup \Sigma_p^-$ for any $p \in \mathcal{K}$. We may extend $\tilde{u}$ by 0 to O so that $\tilde{u} \in H_0^1(O)$. We have

$$\mathcal{B}_p(\hat{u},\hat{u}) + \tilde{\mu} \int\limits_{\Sigma_p^+\cup\Sigma_p^-} \sum_{i,j=1}^{3} \Big|\hat{e}_{ij}(\tilde{u})\Big|^2\,d\hat{x} \;\geq\; \tilde{\mu} \int\limits_{O} \sum_{i,j=1}^{3} \Big|\hat{e}_{ij}(\tilde{u})\Big|^2\,d\hat{x}$$

since $\tilde{\lambda} \geq 0\,,\ \tilde{\mu} \geq 0$. Korn's inequality, applied to the last integral, gives that

$$\begin{aligned} \mathcal{B}_p(\hat{u},\hat{u}) &\geq c|\tilde{u}|^2_{H_0^1(O)} \;-\; \tilde{\mu} \int\limits_{\Sigma_p^+\cup\Sigma_p^-} \sum_{i,j=1}^{3} \Big|\hat{e}_{ij}(\tilde{u})\Big|^2\,d\hat{x} \\ &\geq c|\tilde{u}|^2_{H^1(\hat{\Omega}_p)} \;-\; \tilde{\mu} \int\limits_{\Sigma_p^+\cup\Sigma_p^-} \sum_{i,j=1}^{3} \Big|\hat{e}_{ij}(\tilde{u})\Big|^2\,d\hat{x}\,, \end{aligned} \tag{6.19}$$

with $c > 0$ being independent of $p \in \mathcal{K}$.

We have to estimate the last term in (6.19). To this end, we compute

$$\begin{aligned} \int\limits_{\Sigma_p^+\cup\Sigma_p^-} \left|\frac{\partial \tilde{u}_i}{\partial \hat{x}_j}\right|^2 d\hat{x} \;&=\; \int\limits_{\Sigma_p^+\cup\Sigma_p^-} \Big\langle \Big(\frac{\partial \tilde{w}_i}{\partial x_1}\,(\bar{x}(\hat{x}))\,, \\ &\qquad \frac{\partial \tilde{w}_i}{\partial x_2}\,(\bar{x}(\hat{x}))\,,\, \frac{\partial \tilde{w}_i}{\partial x_3}\,(\bar{x}(\hat{x}))\Big)\,,\, \big(d^p_{1j}(\hat{x})\,,\, d^p_{2j}(\hat{x})\,,\, d^p_{3j}(\hat{x})\big)\Big\rangle^2_{\mathbb{R}^3}\, d\hat{x} \\ &=\; \int\limits_{S^+\cup S^-} \Big\langle \Big(\frac{\partial \tilde{w}_i}{\partial x_1}\,,\, \frac{\partial \tilde{w}_i}{\partial x_2}\,,\, \frac{\partial \tilde{w}_i}{\partial x_3}\Big)\,,\, \big(h^p_{1j}\,,\, h^p_{2j}\,,\, h^p_{3j}\big)\Big\rangle^2_{\mathbb{R}^3} \,|\det J_p| d\bar{x}\,, \end{aligned}$$

where $(d_{ij}^p)_{i,j=\overline{1,3}} := D\,F_p^{-1}(\hat{x})$, $(h_{ij}^p)_{i,j=\overline{1,3}} := J_p^{-1}(\bar{x})$, and where we have performed a standard change of variables in the integral (see Sprekels and Tiba [18] for a detailed calculation). Notice that the extension of h_{ij}^p to $S^+ \cup S^-$ is obvious by (5.5).

As $\{\det J_p\}$, $\{h_{ij}^p\}$ are bounded for $p \in \mathcal{K}$, we have to estimate the gradient of $\tilde{w}$ in $L^2(S^+ \cup S^-)$. We compute it in S^+, for example:

$$\frac{\partial \tilde{w}}{\partial x_\alpha} \,=\, \delta^{-1}\left[(\varepsilon+\delta-x_3)\frac{\partial \bar{u}}{\partial x_\alpha} + \varepsilon\,(\varepsilon+\delta-x_3)\frac{\partial \bar{r}}{\partial x_\alpha}\right], \quad \alpha \,=\, 1,2\,, \tag{6.20}$$

$$\frac{\partial \tilde{w}}{\partial x_3} \,=\, -\,\delta^{-1}(\bar{u}+\varepsilon\,\bar{r})\,. \tag{6.21}$$

Thus, we get

$$\left|\frac{\partial \tilde{w}}{\partial x_3}\right|_{L^2(S^+\cup S^-)^3} \le \sqrt{2}\delta^{-\frac{1}{2}}|\bar{u}|_{L^2(\omega)^3} + \sqrt{2}\varepsilon\,\delta^{-\frac{1}{2}}|\bar{r}|_{L^2(\omega)^3}\,.$$

For $\alpha = 1,2$, we have

$$\left|\frac{\partial \tilde{w}}{\partial x_\alpha}\right|_{L^2(S^+\cup S^-)^3} \le \frac{\sqrt{2}}{\sqrt{3}}\,\delta^{\frac{1}{2}}\left|\frac{\partial \bar{u}}{\partial x_\alpha}\right|_{L^2(o)^3} + \frac{\sqrt{2}}{\sqrt{3}}\,\varepsilon\,\delta^{\frac{1}{2}}\left|\frac{\partial \bar{r}}{\partial x_\alpha}\right|_{L^2(o)^3}\,. \tag{6.22}$$

Consequently, we can find some $c_1 > 0$, independent of $p \in \mathcal{K}$, such that

$$\begin{aligned}\mathcal{B}_p(\hat{u},\hat{u}) &\ge c|\hat{u}|^2_{H^1(\hat{\Omega}_p)}\\ &\quad - c_1\left[\delta|\bar{u}|^2_{V(o)} + \varepsilon^2\,\delta|\bar{r}|^2_{V(o)} + \delta^{-1}|\bar{u}|^2_{L^2(o)^3} + \varepsilon^2\,\delta^{-1}|\bar{r}|^2_{L^2(o)^3}\right].\end{aligned} \tag{6.23}$$

Lemma 6.3. *If $\hat{\Omega}_p = F_p(\Omega)$, there are $c_2 > 0$, $c_3 \in \mathbb{R}$, independent of $p \in \mathcal{K}$, such that*

$$\begin{aligned}|\hat{u}|^2_{H^1(\hat{\Omega}_p)} &\ge c_2\Big[\varepsilon|\bar{u}|^2_{V(o)} + \varepsilon^3|\bar{r}|^2_{V(o)}\Big]\\ &\quad - c_3\,\varepsilon|\bar{r}|^2_{L^2(o)^3}\,, \quad \forall\ \hat{u}(\hat{x}) \,=\, \bar{w}(F_p^{-1}\hat{x}) \,\in\, H^1(\Omega_p)\,,\end{aligned}$$

for $\varepsilon \le \varepsilon_0$ and with $\varepsilon_0 > 0$ independent of $p \in \mathcal{K}$.

Proof. The proof of this lemma is quite technical, and we quote Sprekels and Tiba [18, sect. 3] in this respect. It is possible to check that all the constants appearing there may be chosen independently of $p \in \mathcal{K}$. We indicate here just a precise quantitative argument that replaces the qualitative proof of Lemma 3.3 in Sprekels and Tiba [18], in order to preserve the control of the constants. We have

$$\begin{aligned}|\hat{u}|^2_{H^1(\hat{\Omega}_p)} = \int_\Omega \sum_{i,j=1}^3 \Bigg[&\left(\frac{\partial u_i}{\partial x_1} + x_3\frac{\partial r_i}{\partial x_1}\right) h_{1j}^p(\bar{x}) + \left(\frac{\partial u_i}{\partial x_2} + x_3\frac{\partial r_i}{\partial x_2}\right) h_{2j}^p(\bar{x})\\ &+ r_i(\bar{x})\,h_{3j}^p(\bar{x})\Bigg]^2 |\det J_p(\bar{x})|\,d\bar{x}\,,\end{aligned} \tag{6.24}$$

after the change of variables via $F_p : \Omega \to \hat{\Omega}_p$.

We define the quadratic form

$$Q_p(\bar u,\bar r)=2\,\varepsilon\int\limits_o\sum_{i,j=1}^{3}\left(\frac{\partial u_i}{\partial x_1}\,h_{1j}^{p,0}+\frac{\partial u_i}{\partial x_2}\,h_{2j}^{p,0}+r_1\,h_{3j}^{p,0}\right)^2\sqrt{1+p_1^2+p_2^2}\;dx_1\,dx_2$$

$$+\frac{2\,\varepsilon^3}{3}\int\limits_o\sum_{i,j=1}^{3}\left(\frac{\partial r_i}{\partial x_1}\,h_{1j}^{p,0}+\frac{\partial r_i}{\partial x_2}\,h_{2j}^{p,0}\right)^2\sqrt{1+p_1^2+p_2^2}\;dx_1\,dx_2\,, \tag{6.25}$$

and we estimate it first. Here, $(h_{ij}^{p,0})$ are the elements of the matrix S_p^{-1} (see (6.5), (6.6)); that is, they constitute an approximation of (h_{ij}^{p}). Taking into account the structure of S_p^{-1}, we get

$$\frac{\partial r_i}{\partial x_1}=\frac{p_1}{\sqrt{1+p_1^2+p_2^2}}\left(\frac{\partial r_i}{\partial x_1}\,h_{13}^{p,0}+\frac{\partial r_i}{\partial x_2}\,h_{23}^{p,0}\right)$$

$$+\frac{1}{\sqrt{1+p_1^2+p_2^2}}\left(\frac{\partial r_i}{\partial x_1}\,h_{11}^{p,0}+\frac{\partial r_i}{\partial x_2}\,h_{21}^{p,0}\right)\,, \tag{6.26}$$

and similarly for $\frac{\partial r_i}{\partial x_2}$, $\frac{\partial u_i}{\partial x_\alpha}$, $i=\overline{1,3}$, $\alpha=1,2$.

Then, simple algebraic manipulations in (6.25), (6.26), involving the triangle inequality (and the fact that the coefficients of the parentheses in the right-hand side of (6.26) are less than one), put into evidence a constant, independent of $p\in\mathcal{K}$, such that

$$Q_p(\bar u,\bar r)\;\geq\;c\Big(\varepsilon|\bar u|^2_{V(o)}+\varepsilon^3|\bar r|^2_{V(o)}-\varepsilon|\bar r|^2_{L^2(o)^3}\Big)\,,\quad c\;>\;0\,. \tag{6.27}$$

Taking the difference between (6.24) and (6.25), estimates similar to those of Sprekels and Tiba [18, sect. 3] show that it will be dominated by the right-hand side in (6.27) for ε small. This ends the proof of Lemma 6.3. □

Combining this difference with (6.23), we get (6.12), for $\delta\ll\varepsilon$, and the proof of Proposition 6.2 is finished. □

PROPOSITION 6.4. *Let $\tilde{\mathcal{K}}\subset\mathcal{K}$ be a compact subset. There are $\hat\varepsilon>0$ such that for $\varepsilon<\hat\varepsilon$ there is $c_\varepsilon>0$, independent of $p\in\tilde{\mathcal{K}}$, and*

$$\mathcal{B}_p(\hat u,\hat u)\;\geq\;c_\varepsilon\Big[|\bar u|^2_{V(o)}\;+\;|\bar r|^2_{V(o)}\Big]\,,\quad \hat u(\hat x)\;=\;\bar w(F_p^{-1}(\hat x))\;\in\;H^1(\hat\Omega_p) \tag{6.28}$$

for any $p\in\tilde{\mathcal{K}}$.

Proof. We fix $\hat\varepsilon$ and $\varepsilon<\hat\varepsilon$, $\delta\ll\varepsilon$, such that (6.12) is valid.

Assume that (6.28) is false; i.e., there is no $c_\varepsilon>0$ with the indicated property. Therefore, for any $a>0$, there is $p_a\in\mathcal{K}$ and $\bar u_a$, $\bar r_a$, $\hat u_a(\hat x)=\bar w_a(F_{p_a}^{-1}(\hat x))$ such that

$$0\;\leq\;\mathcal{B}_{p_a}(\hat u_a,\hat u_a)\;\leq\;a\Big[|\bar u_a|^2_{V(o)}\;+\;|\bar r_a|^2_{V(o)}\Big]\,. \tag{6.29}$$

In (6.29), we can assume that $|(\bar u_a,\bar r_a)|_{V(o)^2}=1$, and, consequently, that $\mathcal{B}_{p_a}(\hat u_a,\hat u_a)\to 0$ for $a\to 0$. Moreover, we can suppose that $\bar u_a\to\hat u$, $\bar r_a\to\hat r$, both weakly in $V(o)$, and $p_a\to\hat p\in\tilde{\mathcal{K}}$ strongly in $C^2(\bar o)$, due to the compactness of $\tilde{\mathcal{K}}$. In particular, we get $h_{ij}^a\to\hat h_{ij}$ strongly in $C(\bar\Omega)$, where $(h_{ij}^a)_{i,j=1,3}=J_{p_a}^{-1}$, $(\hat h_{ij})_{i,j=\overline{1,3}}=J_{\hat p}^{-1}$.

It is simple to see, due to the uniform convergence of the coefficients h_{ij}^a, that

$$\mathcal{B}_{p_a}(\hat u_a,\hat u_a)\;-\;\mathcal{B}_{\hat p}(\hat u_a,\hat u_a)\;\to\;0 \tag{6.30}$$

(see (5.11)). The weak lower semicontinuity in $H^1(o)^3 \times H^1(o)^3$ of $\mathcal{B}_{\hat{p}}(\,\cdot\,,\,\cdot\,)$ and (6.29), (6.30) show that

$$0 \geq \liminf_{a\to 0} \mathcal{B}_{p_a}(\hat{u}_a, \hat{u}_a) = \liminf_{a\to 0} \mathcal{B}_{\hat{p}}(\hat{u}_a, \hat{u}_a) \geq \mathcal{B}_{\hat{p}}\Big((\hat{u}\,,\,\hat{r})\,;\,(\hat{u}\,,\,\hat{r})\Big) \geq 0\,. \tag{6.31}$$

Clearly, (6.31) shows that $\mathcal{B}_{\hat{p}}((\hat{u}\,,\,\hat{r})\,;\,(\hat{u}\,,\,\hat{r})) = 0$, and the coercivity of $\mathcal{B}_{\hat{p}}$ gives $\hat{u} = 0\,,\ \hat{r} = 0$, according to Sprekels and Tiba [18]. We conclude that $\bar{u}_a \to 0\,,\ \bar{r}_a \to 0$, both weakly in $V(o)$ and strongly in $L^2(o)^3$.

We combine (6.29) and (6.12) to obtain that

$$\begin{aligned} a &\geq c\Big[\varepsilon|\bar{u}_a|^2_{V(o)} + \varepsilon^2|\bar{r}_a|^2_{V(o)}\Big] - \frac{m}{\delta}\Big[|\bar{r}_a|^2_{L^2(o)^3} + |\bar{u}_a|^2_{L^2(o)^3}\Big] \\ &\geq c\,\varepsilon^3 - \frac{m}{\delta}\Big[|\bar{r}_a|^2_{L^2(o)^3} + |\bar{u}_a|^2_{L^2(o)^3}\Big]\,. \end{aligned}$$

Taking $a \to 0$, we get the contradiction

$$0 \geq c\,\varepsilon^3,$$

which ends the proof. □

Proof of Theorem 6.1. We note that the assumptions of Proposition 6.4 are fulfilled and that (6.28) is valid for $\{p_n\}$, for any $n \in \mathbb{N}$. Then, if we fix $(\bar{\mu}, \bar{\rho}) = \bar{y}_n = (\bar{u}_n, \bar{r}_n)$ in (5.11) with $p = p_n$, we get immediately that $\{\bar{y}_n\}$ is bounded in $V(o)^2$. We may assume that $\bar{u}_n \to \bar{u}\,,\ \bar{r}_n \to \bar{r}$, both weakly in $V(o)$, on a subsequence. Due to the uniform convergence of the coefficients, one may pass to the limit in (5.11) and see that $\bar{y} = (\bar{u}, \bar{r})$ is indeed the solution of (5.11) associated with p. As the solution of (5.11) is unique, $\bar{y}$ is the weak limit of the whole sequence.

Now, we have to show that the convergence is valid in the strong topology of $V(o)^2$. We subtract the equations corresponding to $\bar{y}_n\,,\ \bar{y}$; we intercalate advantageous terms (see the last step in the proof of Theorem 3.1) and; finally, we take test functions of the form $\bar{y}_n - \bar{y} \in V(o)^2$. As the difference of the corresponding right-hand sides converges to 0 (by the above weak convergence property), a detailed calculus gives that

$$\lim_{n\to\infty} \mathcal{B}_p(\bar{y}_n - \bar{y},\, \bar{y}_n - \bar{y}) = 0\,. \tag{6.32}$$

By (6.28), (6.32), the proof is finished. □

Corollary 6.5. *If $\mathcal{K} \subset C^2(\bar{o})$ is compact and $j : V(o)^2 \times C^2(\bar{o}) \to R$ is lower semicontinuous, then the shape optimization problem* (P′) *admits at least one optimal solution $p \in \mathcal{K}$.*

7. Sensitivity analysis for shells. We investigate some differentiability properties of the mapping $p \in C^2(\bar{o}) \mapsto \bar{y} \in V(o)^2$ defined by (5.11). We consider $p + \lambda\,q\,,\ \lambda \in \mathbb{R}_+$, and $q \in C^2(\bar{o})$, a small perturbation of $p \in C^2(\bar{o})$, and we denote by $\bar{y}_\lambda = (\bar{u}^\lambda, \bar{r}^\lambda) \in V(o)^2$ the corresponding solution of (5.11). Similarly, we denote by $\bar{n}_\lambda \in C^1(\bar{o})^3\,,\ F_\lambda \in C^1(\bar{\Omega})^3\,,\ J_\lambda \in C(\bar{\Omega})^9\,,\ h^\lambda_{ij} \in C(\bar{\Omega})\,,\ g^{ij}_\lambda \in C(\bar{\Omega})\,,\ \mathcal{B}_\lambda$, etc., all the quantities defined in section 5, starting from $p_\lambda = p + \lambda\,q$. We shall simply write $\mathcal{B}$ for $\mathcal{B}_p$.

It is elementary, though tedious, to check that the below listed limits and linear and bounded operators exist in the indicated spaces:

$$\lim_{\lambda\to 0} \frac{\bar{n}_\lambda - \bar{n}}{\lambda} = \tilde{n}(q)\,,\ \tilde{n} : C^2(\bar{o}) \to C^1(\bar{o})^3\,, \tag{7.1}$$

$$\lim_{\lambda\to 0}\frac{J_\lambda - J}{\lambda} = \tilde J(q)\,,\ \tilde J\,:\,C^2(\bar o)\,\to C(\bar\Omega)^9\,, \tag{7.2}$$

$$\lim_{\lambda\to 0}\frac{J_\lambda^{-1} - J^{-1}}{\lambda} = \tilde I(q)\,,\ \tilde I\,:\,C^2(\bar o)\,\to C(\bar\Omega)^9\,, \tag{7.3}$$

$$\lim_{\lambda\to 0}\frac{h_{ij}^\lambda - h_{ij}}{\lambda} = \tilde h_{ij}(q)\,,\ \tilde h_{ij}\,:\,C^2(\bar o)\,\to C(\bar\Omega)\,, \tag{7.4}$$

$$\lim_{\lambda\to 0}\frac{\det J_\lambda - \det J}{\lambda} = \mathcal{D}(q)\,,\ \mathcal{D}\,:\,C^2(\bar o)\,\to C(\bar\Omega)\,, \tag{7.5}$$

$$\lim_{\lambda\to 0}\frac{g_\lambda^{ij} - g^{ij}}{\lambda} \,=\, \tilde g^{ij}(q)\,,\ \tilde g^{ij}\,:\,C^2(\bar o)\,\to C(\bar\Omega)\,. \tag{7.6}$$

By Theorem 6.1, we also know that

$$\bar y_\lambda \,\longrightarrow\, \bar y \quad \text{strongly in } V(o)^2\,. \tag{7.7}$$

Now, we subtract the equations for $\bar y_\lambda$ and for $\bar y$, we divide by λ, and we prove that it is possible to take $\lambda\to 0$. In the right-hand side, we have

$$\begin{aligned}
&\lim_{\lambda\to 0}\left\{\int_\Omega\sum_{l=1}^3 f_l(\mu_l + x_3\,\rho_l)\frac{\det J_\lambda - \det J}{\lambda}\,d\bar x\right.\\
&\left.+\int_{\Gamma_1}\sum_{l=1}^3\sum_{i,j=1}^3 g_l(\mu_l + x_3\,\rho_l)\frac{\det J_\lambda\sqrt{\nu_i\,g_\lambda^{ij}\,\nu_j} - \det J\sqrt{\nu_i\,g^{ij}\,\nu_j}}{\lambda}\,d\tau\right\}\\
&=\sum_{l=1}^3\int_\Omega f_l(\mu_l + x_3\,\rho_l)\mathcal{D}(q)\,d\bar x\\
&\quad+\sum_{l=1}^3\sum_{i,j=1}^3\int_{\Gamma_1} g_l(\mu_l + x_3\,\rho_l)\left[\mathcal{D}(q)\sqrt{\nu_i\,g^{ij}\,\nu_j} + \det J\frac{\nu_i\,\tilde g^{ij}(q)\,\nu_j}{2\sqrt{\nu_i\,g^{ij}\,\nu_j}}\right]d\tau\,.
\end{aligned} \tag{7.8}$$

Here $\bar v = (\bar\mu,\bar\rho)\in V(o)^2$ is an arbitrary test function.

As the computation of $\frac{1}{\lambda}[\mathcal{B}_\lambda - \mathcal{B}]$ is quite lengthy, we write in detail just the terms from the bilinear functionals associated with the coefficient $2\,\tilde\mu$, namely,

$$\begin{aligned}
&\frac{1}{\lambda}\left\{\int_\Omega\sum_{i=1}^3\left[\left(\frac{\partial u_i^\lambda}{\partial x_1} + x_3\,\frac{\partial r_i^\lambda}{\partial x_1}\right)h_{1i}^\lambda + \left(\frac{\partial u_i^\lambda}{\partial x_2} + x_3\,\frac{\partial r_i^\lambda}{\partial x_2}\right)h_{2i}^\lambda + r_i^\lambda\,h_{3i}^\lambda\right]\right.\\
&\times\left[\left(\frac{\partial \mu_i}{\partial x_1} + x_3\,\frac{\partial \rho_i}{\partial x_1}\right)h_{1i}^\lambda + \left(\frac{\partial \mu_i}{\partial x_2} + x_3\,\frac{\partial \rho_i}{\partial x_2}\right)h_{2i}^\lambda + \rho_i\,h_{3i}^\lambda\right]|\det J_\lambda|d\bar x\\
&-\int_\Omega\left[\left(\frac{\partial u_i}{\partial x_1} + x_3\frac{\partial r_i}{\partial x_1}\right)h_{1i} + \left(\frac{\partial u_i}{\partial x_2} + x_3\frac{\partial r_i}{\partial x_2}\right)h_{2i} + r_i h_{3i}\right]\\
&\left.\times\left[\left(\frac{\partial \mu_i}{\partial x_1} + x_3\frac{\partial \rho_i}{\partial x_1}\right)h_{1i} + \left(\frac{\partial \mu_i}{\partial x_2} + x_3\frac{\partial \rho_i}{\partial x_2}\right)h_{2i} + \rho_i h_{3i}\right]|\det J|d\bar x\right\}\\
&=\int_\Omega\sum_{i=1}^3\left[\left(\frac{\partial\frac{u_i^\lambda - u_i}{\lambda}}{\partial x_1} + x_3\frac{\partial\frac{r_i^\lambda - r_i}{\lambda}}{\partial x_1}\right)h_{1i} + \left(\frac{\partial\frac{u_i^\lambda - u_i}{\lambda}}{\partial x_2} + x_3\frac{\partial\frac{r_i^\lambda - r_i}{\lambda}}{\partial x_2}\right)h_{2i} + \frac{r_i^\lambda - r_i}{\lambda}h_{3i}\right]
\end{aligned}$$

$$
\begin{aligned}
&\times\left[\left(\frac{\partial\mu_i}{\partial x_1}+x_3\frac{\partial\rho_i}{\partial x_1}\right)h_{1i}+\left(\frac{\partial\mu_i}{\partial x_2}+x_3\frac{\partial\rho_i}{\partial x_2}\right)h_{2i}+\rho_i h_{3i}\right]|\det J|d\bar{x}\\
&+\int_\Omega\sum_{i=1}^{3}\left[\left(\frac{\partial u_i^\lambda}{\partial x_1}+x_3\frac{\partial r_i^\lambda}{\partial x_1}\right)\frac{h_{1i}^\lambda\det J_\lambda-h_{1i}\det J}{\lambda}\right.\\
&\left.+\left(\frac{\partial u_i^\lambda}{\partial x_2}+x_3\frac{\partial r_i^\lambda}{\partial x_2}\right)\frac{h_{2i}^\lambda\det J_\lambda-h_{2i}\det J}{\lambda}+r_i^\lambda\frac{h_{3i}^\lambda\det J_\lambda-h_{3i}\det J}{\lambda}\right]\\
&\times\left[\left(\frac{\partial\mu_i}{\partial x_1}+x_3\frac{\partial\rho_i}{\partial x_1}\right)h_{1i}^\lambda+\left(\frac{\partial\mu_i}{\partial x_2}+x_3\frac{\partial\rho_i}{\partial x_2}\right)h_{2i}^\lambda+\rho_i h_{3i}^\lambda\right]d\bar{x}\\
&+\int_\Omega\sum_{i=1}^{3}\left[\left(\frac{\partial u_i^\lambda}{\partial x_1}+x_3\frac{\partial r_i^\lambda}{\partial x_1}\right)h_{1i}+\left(\frac{\partial u_i^\lambda}{\partial x_2}+x_3\frac{\partial r_i^\lambda}{\partial x_2}\right)h_{2i}+r_i^\lambda h_{3i}\right]\\
&\times\left[\left(\frac{\partial\mu_i}{\partial x_1}+x_3\frac{\partial\rho_i}{\partial x_1}\right)\frac{h_{1i}^\lambda-h_{1i}}{\lambda}+\left(\frac{\partial\mu_i}{\partial x_2}+x_3\frac{\partial\rho_i}{\partial x_2}\right)\frac{h_{2i}^\lambda-h_{2i}}{\lambda}+\rho_i\frac{h_{3i}^\lambda-h_{3i}}{\lambda}\right]|\det J|d\bar{x}\,.
\end{aligned}
\tag{7.9}
$$

According to (7.4), (7.5), and (6.10) the last two integrals are of the form $Z_\lambda(\bar{y}_\lambda\,,\,\bar{v})$, and there is a constant independent of $\lambda>0$ such that the bilinear forms Z_λ satisfy

$$
|Z_\lambda(\bar{y}_\lambda\,,\,\bar{v}|\le C|\bar{y}_\lambda|_{V(o)^2}|\bar{v}|_{V(o)^2}\,. \tag{7.10}
$$

Applying the same technique to all of the terms of $\mathcal{B}_\lambda-\mathcal{B}$, (7.8)–(7.10) give

$$
\mathcal{B}\Big(\frac{\bar{y}_\lambda-\bar{y}}{\lambda}\,,\,\bar{v}\Big)\;=\;\tilde{Z}_\lambda(\bar{y}_\lambda,\bar{v})\quad\forall\;\bar{v}\in V(o)^2\,, \tag{7.11}
$$

where $\tilde{Z}_\lambda$ is obtained by adding together all the terms from (7.8)–(7.10).

By fixing $\bar{v}=\frac{\bar{y}_\lambda-\bar{y}}{\lambda}$ in (7.11), and taking into account (7.10) and (7.7), we see that $\{\frac{\bar{y}_\lambda-\bar{y}}{\lambda}\}$ is bounded in $V(o)^2$, due to Proposition 6.4. We may take a weakly convergent subsequence,

$$
\frac{\bar{y}_\lambda-\bar{y}}{\lambda}\to\hat{y}\quad\text{weakly in}\quad V(o)^2\,, \tag{7.12}
$$

and we can pass to the limit in (7.11). The obtained equation in variations has the form

$$
\mathcal{B}(\hat{y},\bar{v})=Z(\bar{v})\quad\forall\;\bar{v}\in V(o)^2\,, \tag{7.13}
$$

where $Z(\bar{v})=\lim_{\lambda\to0}\tilde{Z}_\lambda(\bar{y}_\lambda,\,\bar{v})$ and $Z:V(o)^2\to\mathbb{R}$ is a linear bounded functional. Notice that (7.13) has a unique solution $\hat{y}\in V(o)^2$, due to (6.28). We thus have proved the following proposition.

PROPOSITION 7.1. *The mapping $p\in C^2(\bar{o})\mapsto\bar{y}\in V(o)^2$ given by* (5.11) *is Gâteaux differentiable, and the directional derivative $\hat{y}$ satisfies* (7.13).

We introduce now the so-called adjoint system with unknowns $\bar{s}=(\bar{a}\,,\,\bar{b})\in V(o)^2$,

$$
\mathcal{B}(\bar{s}\,,\,\bar{v})\;=\;\nabla_1 j(\bar{y}\,,\,p)(\bar{v})\;\;\forall\;\bar{v}\in V(o)^2\,. \tag{7.14}
$$

The existence and the uniqueness of the solution to (7.14) are clear due to the properties of $\mathcal{B}$. We have assumed that j is Fréchet differentiable on $V(o)^2\times C^2(\bar{o})$, and $\nabla_1 j$, $\nabla_2 j$ denote the partial differentials with respect to $\bar{y}\,,\,p$.

PROPOSITION 7.2. *If j is Fréchet differentiable, then the directional derivative of the cost functional Π in problem (P′), at the point $p \in C^2(\bar{o})$ and in the direction $q \in C^2(\bar{o})$, is given by*

$$\nabla\Pi(p)q \;=\; \nabla_2 j(\bar{y}\,, p)q + Z(\bar{s})\,. \tag{7.15}$$

Proof.

$$\lim_{\lambda\to 0} \frac{\Pi(p+\lambda q) - \Pi(p)}{\lambda} \;=\; \nabla_2 j(\bar{y}\,, p)q + \nabla_1 j(\bar{y}\,, p)\hat{y}\,,$$

by the chain rule and Proposition 7.1. Moreover, by (7.14), (7.13), we have

$$\nabla_1 j(\bar{y}\,, p)\hat{y} \;=\; \mathcal{B}(\bar{s}\,, \hat{y}) \;=\; \mathcal{B}(\hat{y}\,, \bar{s}) \;=\; Z(\bar{s})\,. \qquad \square$$

Remark 7.3. In order to compute (7.15) from $p\,,\ q \in C^2(\bar{o})$, one has to compute $\bar{y}$ by (5.11), $\bar{s}$ by (7.14), and Z by (7.13). The computation of Z is standard (see (7.9), (7.8)) but tedious, and we do not detail it here.

COROLLARY 7.4. *Assume that p^* is a (local) optimal shape for (P′), that $\bar{y}^*$ is the associated deformation, and that all the above assumptions are fulfilled. Then*

(i) *If $\mathcal{K} \subset C^2(\bar{o})$ is convex, we have*

$$\nabla_2 j(\bar{y}^*\,, p^*)q + Z(\bar{s}) \geq 0 \quad \forall\ q \in \mathcal{K} - p^*\,.$$

(ii) *If $\mathcal{K}$ is not convex, we have*

$$\nabla_2 j(\bar{y}^*\,, p^*)q + Z(\bar{s}) \geq 0 \quad \forall\ q \in T(\mathcal{K}\,, p^*)\,.$$

Remark 7.5. Corollary 7.4 gives the standard optimality conditions for problem (P′). The directional derivative obtained in Proposition 7.2 may be used, in principle, in the numerical computations, as in the case of the curved rods. However, the coercivity properties of the bilinear functional $\mathcal{B}_p$ are valid just for small thickness ε, and the coercivity constant depends in a very bad manner, namely like ε^3 (see Proposition 6.2 or Sprekels and Tiba [18]). This shows that instabilities (the locking problem) may appear in the numerical experiments and special numerical schemes are to be used. The interested reader may consult Chenais and Paumier [9] and Pitkäranta and Leino [15] for a discussion on the approximation of the state equation (5.11).

8. Numerical experiments. In the papers of Ignat, Sprekels, and Tiba [12], [13], many numerical examples concerning the deformation of 3D curved rods and the optimization of planar arches are reported. Here, we concentrate on the problem discussed in sections 2–4. Namely, we assume that a certain field of forces acting on "any possible curved rod" is given (see (8.4)–(8.7) below), and we search for the geometry which produces the minimum value for some cost functional. The cost considered in the examples is related to various components of the deformation $\bar{\tau}$ of the line of centroids of the curved rod. The "locking phenomenon," specific to numerical computations involving thin structures (see [9], [15]), is avoided in our experiments by allowing the thickness of the curved rod to be "larger" than the division that we consider for the interval $[0, L]\,,\ L = 4\pi\sqrt{2}$. Namely, we have divided the interval $[0, L]$ into 100 equal parts and we have taken the cross section of the curved rod to always be given by a disk with radius $R = 0.3$. For the integrals over the cross section, the usual change of variables to polar coordinates leads to the integration over the

rectangle $[0,R]\times[0,2\pi]$, which allows the use of simple numerical integration formulae corresponding to the discrete grids. We have divided it into 8, respectively, 80, parts and we have used Simpson's iterative formula.

In general, as initial iteration to the optimization algorithm, we have considered the spiral, lying on the cylinder $x_1^2+x_2^2=1$, given by

$$\varphi^0(x_3)=\frac{\pi}{4}\,,\quad \psi^0(x_3)=\frac{\pi}{2}+\frac{x_3}{\sqrt{2}}\,,\quad x_3\in[0,L]\,. \tag{8.1}$$

A simple calculus shows that the rod parametrization corresponding to (8.1) is

$$\bar\theta(x_3)=\left(\cos\frac{x_3}{\sqrt{2}}\,,\ \sin\frac{x_3}{\sqrt{2}}\,,\ \frac{x_3}{\sqrt{2}}\right),\quad x_3\in[0,L]\,. \tag{8.2}$$

Deformations for this example of a curved rod, under the action of various body forces, have been computed in Ignat, Sprekels, and Tiba [13]. The Lamé constants taken into account are $\lambda=50\,,\ \mu=100$. The solution of the state system (2.13) in the Sobolev space $H_0^1(0,L)^9$ is approximated by linear splines from V_h^9, where $h=10^{-2}L$ is the division norm of $[0,L]$, and where

$$V_h=\{v_h\in C[0,L]\,;\ v_h(0)=v_h(L)=0\,,\ v_h \text{ is piecewise linear in } [0,L]\}\,. \tag{8.3}$$

The same matrix governs the discrete equations for both (2.13) and the adjoint system (4.20). We underline that finding the matrix (which has to be recomputed in each optimization iteration) is the most time-consuming step of the algorithm. This is due to the 3D character of the objects that we are studying. The model (2.13) provides a dimension reduction up to ODEs, and this is reflected in that the coefficients involve the computation of many integrals over the cross section. One can compute the gradient of the cost functional and use projected gradient methods for the optimization of the geometry of the 3D rods, as explained in section 4. We have used the Uzawa algorithm combined with the Armijo line search rule for the minimization of the cost.

A first class of examples is obtained when the force $\bar f=(0,0,f_3)$ with the variants

$$f_3(x_3)=\begin{cases} 10, & x_3\in\left[0,\ \dfrac{L}{2}\right],\\[2ex] -10, & x_3\in\left]\dfrac{L}{2}\,,L\right],\end{cases} \tag{8.4}$$

$$f_3(x_3)\equiv 10\quad\text{in }\ [0,L]\,, \tag{8.5}$$

$$f_3(x_3)=\begin{cases} 10, & x_3\in\left[0,\ \dfrac{L}{2}\right],\\[2ex] 0, & x_3\in\left]\dfrac{L}{2},\ L\right],\end{cases} \tag{8.6}$$

$$f_3(x_3)=\begin{cases} 0, & x_3\in\left[0,\ \dfrac{L}{2}\right],\\[2ex] 10, & x_3\in\left]\dfrac{L}{2},\ L\right].\end{cases} \tag{8.7}$$

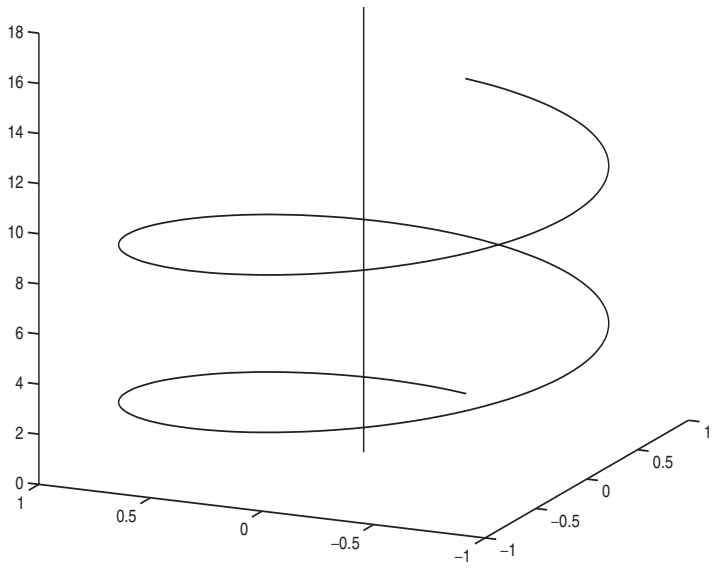


FIG. 1.

The control problem has the cost functional $\Pi = \frac{1}{2}|\tau_i|^2_{L^2(0,L)}$ with $i = 2, 3$ (compare with (2.14)) and state equation (2.13) with various $\bar{f}$ chosen as above. We have also imposed the constraint (2.15), with $\varepsilon = \frac{\pi}{8}$, to avoid the appearance of self-intersecting curves. We have neglected (2.11), but it may be checked a posteriori that $\det J \neq 0$.

In all the cases (8.4)–(8.7), the vertical column, which corresponds to $\varphi \equiv 0$, was the geometric solution of the given problem. Indeed, the vertical column is the most resistant structure with respect to vertical forces as in (8.4)–(8.7). In this case also, the lateral displacements τ_1 , τ_2 are several orders of magnitude smaller than the vertical displacement.

Figure 1 shows the initial and the final geometries, obtained in one or two iterations. In Figures 2–5, the values of τ_3 (in the final iteration) are shown, and one can see their dependence on the forces (8.4)–(8.7), respectively.

The fact that these examples have a clear physical interpretation provides a validation of the model and of the approximation and optimization procedures that we are using.

In another set of numerical tests, we have considered $\bar{f} = 10\bar{b}$ (recall (2.3)). Again the initial iteration was given by (8.1) (or (8.2)) or by the following perturbation of it,

$$\varphi^0(x_3) = \begin{cases} \dfrac{\pi}{4} + 0,1, & x_3 \in \left[0, \dfrac{L}{2}\right], \\ \dfrac{\pi}{4} - 0,1, & x_3 \in \left]\dfrac{L}{2}, L\right], \end{cases} \tag{8.8}$$

and the objective functional was the same as above.

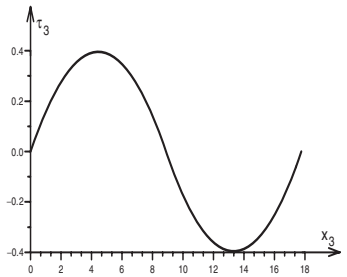


FIG. 2.

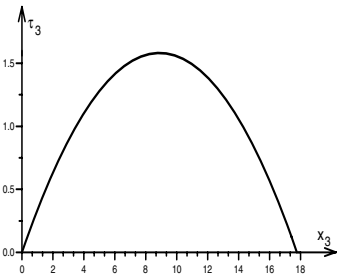


FIG. 3.

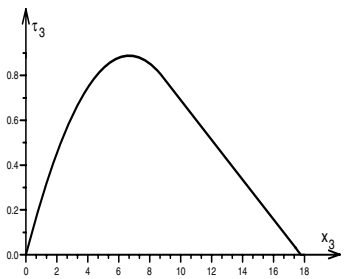


FIG. 4.

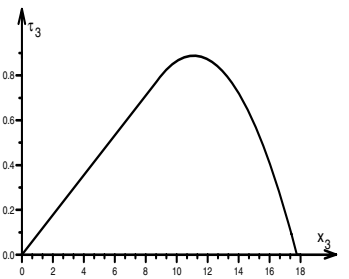


FIG. 5.

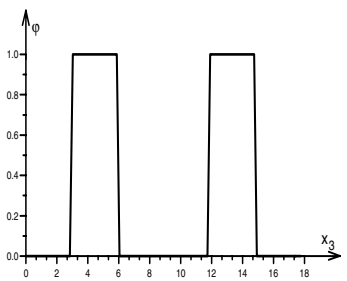


FIG. 6.

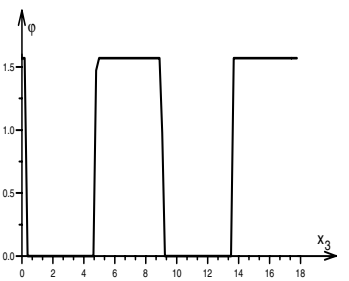


FIG. 7.

Notice that, under our parametrization, it is very simple to change the initial iteration, which is an important advantage in nonconvex optimization problems. The main property of this choice of $\bar{f}$ is that it always acts in the horizontal plane, although in various directions. It is also very easy to construct, under our approach. For the constraints, we have taken $\varepsilon = 0$ in (2.15). This allows horizontal curves as well, but self-intersections may appear (which indeed was the case). That is, in this set of experiments (2.11) is violated. In the examples that we have computed, a clear decrease in the cost was observed and the tendency was to produce a horizontal curve as the solution. Although self-intersections are present, horizontal curves will deform just in the horizontal plane under the action of $\bar{f} = 10\bar{b}$. That is, a mechanical interpretation is still possible (and due to this, it was necessary to allow $\varepsilon = 0$ in (2.15)).

An interesting feature of this type of experiment was that the optimal φ has a bang-bang structure. Figures 6 and 7 show this for when the initial iteration was given by (8.8) and (8.1), respectively.

In general, one experiment took between two and three hours, on a powerful Compaq GS80 workstation.

Acknowledgment. The authors thank both referees for their suggestions which led to an improvement of the presentation.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] S. S. Antman, *Nonlinear Problems of Elasticity*, Springer-Verlag, Berlin, 1995.
[3] K. Arunakirinathar and B. D. Reddy, *Mixed finite element methods for elastic rods of arbitrary geometry*, Numer. Math., 64 (1993), pp. 13–43.
[4] V. Barbu and Th. Precupanu, *Convexity and Optimization in Banach Spaces*, D. Reidel, Dordrecht, The Netherlands, 1986.
[5] E. Bloch, *A First Course in Geometric Topology and Differential Geometry*, Birkhäuser-Verlag, Basel, Switzerland, 1997.
[6] A. Blouza, *Existence et unicité pour le modèle de Naghdi pour une coque peu régulière*, C. R. Acad. Sci. Paris Sér. I Math., 324 (1997), pp. 839–844.
[7] H. Cartan, *Formes Differentielles*, Hermann, Paris, 1967.
[8] D. Chapelle, *A locking-free approximation of curved rods by straight beam elements*, Numer. Math., 77 (1997), pp. 299–322.
[9] D. Chenais and J. C. Paumier, *On the locking phenomenon for a class of elliptic problems*, Numer. Math., 67 (1994), pp. 427–440.
[10] D. Chenais and B. Rousselet, *Dependence of the buckling load of a nonshallow arch with respect to the shape of its midcurve*, RAIRO Modél. Math. Anal. Numér., 24 (1990), pp. 307–341.
[11] Ph. Ciarlet, *Mathematical Elasticity* III: *Theory of Shells*, North–Holland, Amsterdam, 2000.
[12] A. Ignat, J. Sprekels, and D. Tiba, *Analysis and optimization of nonsmooth arches*, SIAM J. Control Optim., 40 (2001), pp. 1107–1133.
[13] A. Ignat, J. Sprekels, and D. Tiba, *A model of a general elastic curved rod*, Math. Methods Appl. Sci., 25 (2002), pp. 835–854.
[14] A. Myslinski, J. Piekarski, and B. Rousselet, *Design sensitivity for a hyperelastic rod in large displacement with respect to its midcurve shape*, J. Optim. Theory Appl., 96 (1998), pp. 683–708.
[15] J. Pitkäranta and Y. Leino, *On the membrane locking of h-p finite elements in a cylindrical shell problem*, Internat. J. Numer. Methods Engrg., 37 (1994), pp. 1053–1070.
[16] B. Rousselet, *A finite strain rod model and its design sensitivity*, Mech. Structures Mach., 20 (1992), pp. 415–432.
[17] J. Sprekels and D. Tiba, *Sur les arches lipschitziennes*, C. R. Acad. Sci. Paris Sér. I Math., 331 (2000), pp. 179–184.
[18] J. Sprekels and D. Tiba, *An analytic approach to a generalized Naghdi shell model*, Adv. Math. Sci. Appl., 12 (2002), pp. 175–190.
[19] L. Trabucho and J. M. Viaño, *Mathematical modelling of rods*, in Handbook of Numerical Analysis, Vol. IV, Handb. Numer. Anal. IV, P. G. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 1996, pp. 487–974.
[20] F. Tröltzsch, *Optimality Conditions for Parabolic Control Problems and Applications*, Teubner-Texte Math. 62, Teubner, Leipzig, Germany, 1984.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 3, pp. 777–800

ON THE ROBUSTNESS OF $\mathcal{KL}$-STABILITY FOR DIFFERENCE INCLUSIONS: SMOOTH DISCRETE-TIME LYAPUNOV FUNCTIONS*

CHRISTOPHER M. KELLETT† AND ANDREW R. TEEL‡

Abstract. We consider stability with respect to two measures of a difference inclusion, i.e., of a discrete-time dynamical system with the push-forward map being set-valued. We demonstrate that robust stability is equivalent to the existence of a smooth Lyapunov function and that, in fact, a continuous Lyapunov function implies robust stability. We also present a sufficient condition for robust stability that is independent of a Lyapunov function. Toward this end, we develop several new results on the behavior of solutions of difference inclusions. In addition, we provide a novel result for generating a smooth function from one that is merely upper semicontinuous.

Key words. difference inclusions, stability with respect to two measures, Lyapunov functions, robustness, smoothing upper-semicontinuous functions

AMS subject classifications. 39A12, 93C55, 93D09, 93D30

DOI. 10.1137/S0363012903435862

1. Introduction.

The close connection between robustness of stability properties for differential equations and the existence of Lyapunov functions has been implicit in the literature since the result of Kurzweil [13]. In particular, Kurzweil exploited the inherent robustness of asymptotic stability of the origin for differential equations defined by a continuous right-hand side in order to demonstrate the existence of a smooth Lyapunov function. Since Kurzweil, robustness of the assumed stability property has played a key role in deriving Lyapunov functions. Results on the existence of Lyapunov functions for asymptotically stable closed sets became available in the 1960s in the works of Hoppensteadt [7] and Wilson [23]. These results were extended by Lin, Sontag, and Wang [15] to consider asymptotic stability of closed sets for locally Lipschitz differential equations subject to disturbances. Recently, Clarke, Ledyaev, and Stern [4] demonstrated the existence of a smooth Lyapunov function for upper-semicontinuous differential inclusions with an asymptotically stable origin.

Rather than considering differential inclusions, we will consider the difference inclusion

$$x^+ \in F(x), \quad x \in \mathcal{G}, \tag{1.1}$$

where $\mathcal{G} \subseteq \mathbb{R}^n$ is open. Difference inclusions are a natural way to consider difference equations subject to disturbances or controlled difference equations. One may consider a set-valued map as

$$x^+ \in F(x) := f(x, \mathcal{V}),$$

where $\mathcal{V}$ is a set of disturbances or the admissible control set. We use $\phi \in \mathcal{S}(x)$ to denote a solution of the difference inclusion (1.1) from initial condition $x \in \mathcal{G}$, i.e., a

*Received by the editors October 10, 2003; accepted for publication (in revised form) March 6, 2005; published electronically September 15, 2005. This work was partially supported by AFOSR grants F49620-00-1-0106 and -03-1-0203 and by NSF grant ECS-9988813.
http://www.siam.org/journals/sicon/44-3/43586.html
†The Hamilton Institute, National University of Ireland, Maynooth, Ireland (chris.kellett@nuim.ie).
‡Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 (teel@ece.ucsb.edu).

function satisfying $\phi(0,x) = x$ and

$$\phi(k+1,x) \in F(\phi(k,x)) \quad \forall k \in \mathbb{Z}_{\geq 0}.$$

Whereas in the continuous-time case a solution was an absolutely continuous function, in the discrete-time case solutions are sequences of points. Solutions are defined for all $k \in \mathbb{Z}_{\geq 0}$ when $F(\cdot)$ maps $\mathcal{G}$ to subsets of $\mathcal{G}$, which is the discrete-time counterpart to forward completeness for continuous-time systems.

In the 1970s, Lakshmikantham and Salvadori [14] demonstrated a locally Lipschitz Lyapunov function for a differential equation under the assumption of stability with respect to two measures, a concept first introduced by Movchan [17]. Stability with respect to two measures can be seen to cover uniform global or local asymptotic stability of a point, prescribed motion, or closed set. In fact, Teel and Praly [22, Proposition 1] (following [15, Proposition 2.5]) demonstrated that $\mathcal{KL}$-stability with respect to two measures is equivalent to uniform stability and global boundedness coupled with uniform global attractivity (both properties being defined in an appropriate two-measure sense). In Proposition 2.2 we show that this property carries over to the discrete-time case.

A smooth Lyapunov function for output stability, a special case of stability with respect to two measures where one of the measures is the norm of the output function, was presented by Sontag and Wang [20, Theorem 2]. Teel and Praly [22] extended these results to consider the existence of a smooth Lyapunov function under the assumption of $\mathcal{KL}$-stability with respect to two measures for differential inclusions. It is this last result by Teel and Praly [22, Theorem 1] that we propose to develop in the discrete-time case, namely, the equivalence of robustness of $\mathcal{KL}$-stability with respect to two measures for a difference inclusion and the existence of a smooth Lyapunov function. This is the result of Theorem 2.7.

In Theorem 2.8 we present a result stating that when the set-valued map defining the difference inclusion (1.1) is compact and nonempty, a continuous Lyapunov function is sufficient to demonstrate robustness. This result has important implications in robustness analysis. The authors used this fact in [11, Theorem 14] to demonstrate robustness for a (discontinuous) difference equation. Frequently, in model predictive control, a continuous Lyapunov function is assumed (see Mayne et al. [16]) which guarantees robustness of stability. Recently, Grimm et al. [6] presented several examples where model predictive control is nonrobust. Intuitively, these results follow from the lack of a continuous Lyapunov function.

A question of great interest over many years is the so-called converse Lyapunov question, namely, what stability requirements guarantee the existence of a Lyapunov function? We see from Theorem 2.7 that, for $\mathcal{KL}$-stability with respect to two measures, this question is reduced to that of finding sufficient conditions for robustness. The result of Theorem 2.10 states that if the difference inclusion $x^+ \in F(x)$ is $\mathcal{KL}$-stable, the set-valued map $F(x)$ is nonempty and compact for each $x \in \mathcal{G}$, and $F(\cdot)$ is continuous, then the $\mathcal{KL}$-stability is robust. In [9] and [10], other sufficient conditions were presented for robustness of $\mathcal{KL}$-stability. For example, $\mathcal{KL}$-stability is robust when using a single measurement function that is a proper indicator function for a compact attractor. Each of these sufficient conditions then allows us to state a converse Lyapunov theorem.

Previous converse Lyapunov theorems for discrete-time systems appeared in books by Agarwal [1, Theorem 5.12.5] and Stuart and Humphries [21, Theorem 1.7.6], where uniform global asymptotic stability of the origin or a compact attractor for a locally

Lipschitz single-valued mapping yields a locally Lipschitz Lyapunov function. Nešić, Teel, and Kokotović [18] demonstrated the equivalence of uniform global asymptotic stability of the origin for a difference equation (with no regularity) and the existence of a Lyapunov function (with no regularity).

Jiang and Wang [8, Theorem 1] showed that uniform global asymptotic stability to a closed set $\mathcal{A}$ for a difference equation with disturbances is equivalent to the existence of a smooth Lyapunov function under the assumption that the difference equation is continuous. The assumption of continuity on the difference equation (and compactness of the set of allowable disturbances) gives rise to a continuous set-valued map. This result can then be seen to be a special case of Theorem 2.7 and Theorem 2.10 with $\omega_1(\cdot) = \omega_2(\cdot) = |\cdot|_{\mathcal{A}}$, where $|x|_{\mathcal{A}} := \inf_{a\in\mathcal{A}} |x - a|$.

The authors [11] demonstrated that global asymptotic stability of a point for an upper-semicontinuous difference inclusion implied the existence of a smooth Lyapunov function. This result follows from the results presented here and in [10] (see also [9]).

We will require two sets of technical results, heretofore unknown in the literature. In section 5 we develop results pertaining to difference inclusions which parallel those found in the work of Filippov [5] for differential inclusions. Specifically, we prove results on closeness of solutions under perturbations (Lemmas 5.1 and 5.2) as well as on uniform convergence of sequences of solutions (Lemma 5.3). The second novel technical result involves smoothing nonsmooth functions on a given open domain. As in much previous work (e.g., [13], [15], [22], and [23]), we first construct a Lyapunov function satisfying the desired decrease condition, but with a rather weak regularity property, and then apply a smoothing result to obtain the smooth Lyapunov function without destroying the decrease property. In the past, these smoothing results applied to continuous functions. In section 3, we present a novel smoothing theorem which obtains a smooth function from one that is upper semicontinuous.

2. Smooth Lyapunov functions and robustness. We now turn to precise statements of our results. Recall that a function $\alpha : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is of class-$\mathcal{K}$ if it is continuous, zero at zero, and strictly increasing. A function is of class-$\mathcal{K}_\infty$ if, in addition to being class-$\mathcal{K}$, it is unbounded. A function $\beta : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is said to belong to class-$\mathcal{KL}$ if, for each $t \geq 0$, $\beta(\cdot, t)$ is of class-$\mathcal{K}$ and, for each $s \geq 0$, $\beta(s, \cdot)$ is nonincreasing and $\lim_{t\to\infty} \beta(s, t) = 0$.

DEFINITION 2.1. *Let $\omega_i : \mathcal{G} \to \mathbb{R}_{\geq 0}$, $i = 1, 2$, be continuous functions. Let $F(\cdot)$ be a set-valued map from $\mathcal{G}$ to subsets of $\mathcal{G}$. We say that the difference inclusion $x^+ \in F(x)$ is $\mathcal{KL}$-stable with respect to (ω_1, ω_2) on $\mathcal{G}$ if there exists a function $\beta \in \mathcal{KL}$ such that for every initial condition $x \in \mathcal{G}$ all solutions $\phi \in \mathcal{S}(x)$ satisfy*

$$\omega_1(\phi(k, x)) \leq \beta(\omega_2(x), k) \quad \forall k \in \mathbb{Z}_{\geq 0}. \tag{2.1}$$

Note that appropriate choices for the measurement functions $\omega_1(\cdot)$ and $\omega_2(\cdot)$ as well as the domain $\mathcal{G}$ allow us to recover several classical stability notions. For instance, global asymptotic stability of the origin (for a given difference inclusion evolving in $\mathbb{R}^n$) corresponds to taking $\mathcal{G} = \mathbb{R}^n$ and the measurement functions $\omega_1(x) = \omega_2(x) = |x|$ for all $x \in \mathbb{R}^n$. Other stability notions, such as local asymptotic stability or partial state stability, can be covered by appropriately choosing the domain and measurement functions.

Lin et al. [15, Proposition 2.5] demonstrated that $\mathcal{KL}$-stability with respect to $(|\cdot|_{\mathcal{A}}, |\cdot|_{\mathcal{A}})$ (where $\mathcal{A}$ is a closed set) is equivalent to uniform stability and uniform attractivity of the set $\mathcal{A}$ (i.e., $\mathcal{KL}$-stability is equivalent to uniform global asymptotic stability of the set $\mathcal{A}$). Teel and Praly [22, Proposition1] extended this result to the

consideration of the general two-measure case; that is, $\mathcal{KL}$-stability with respect to two measures is equivalent to uniform stability and global boundedness coupled with uniform global attractivity, where these properties are defined in an appropriate two-measure sense. This result also holds in the discrete time. The details are similar to the continuous-time result and may be found in [9].

PROPOSITION 2.2. *Let $\omega_i : \mathcal{G} \to \mathbb{R}_{\geq 0}$, $i = 1, 2$, be continuous and let $F(\cdot)$ be a set-valued map from $\mathcal{G}$ to subsets of $\mathcal{G}$. The following are equivalent:*

1. *The difference inclusion $x^+ \in F(x)$ is $\mathcal{KL}$-stable with respect to (ω_1, ω_2) on $\mathcal{G}$.*
2. *The following hold:*
   (a) *(Uniform stability and global boundedness): There exists a function $\gamma \in \mathcal{K}_\infty$ such that, for each $x \in \mathcal{G}$, all solutions $\phi \in \mathcal{S}(x)$ satisfy*

$$\omega_1(\phi(k, x)) \leq \gamma(\omega_2(x)) \quad \forall k \in \mathbb{Z}_{\geq 0}.$$

   (b) *(Uniform global attractivity): For each $r, \varepsilon > 0$, there exists $K(r, \varepsilon) > 0$ such that, for each $x \in \mathcal{G}$, all solutions $\phi \in \mathcal{S}(x)$ satisfy*

$$\omega_2(x) \leq r, \quad k \geq \mathbb{Z}_{\geq K} \quad \Longrightarrow \quad \omega_1(\phi(k, x)) \leq \varepsilon.$$

For a continuous function $\sigma : \mathcal{G} \to \mathbb{R}_{\geq 0}$ we define the σ-perturbation of $F(\cdot)$ as

$$F_\sigma(x) := \left\{ v \in \mathbb{R}^n : v \in \{\eta\} + \sigma(\eta)\overline{\mathcal{B}}, \eta \in F\left(x + \sigma(x)\overline{\mathcal{B}}\right) \right\}. \tag{2.2}$$

We denote the solution set of the difference inclusion $x^+ \in F_\sigma(x)$ starting from an initial condition $x \in \mathcal{G}$ by $\mathcal{S}_\sigma(x)$. We will use $\mathcal{B}$ (or $\overline{\mathcal{B}}$) to denote the open (or closed) unit ball in $\mathbb{R}^n$. For two sets $\mathcal{O}_1$ and $\mathcal{O}_2$, we denote the intersection of $\mathcal{O}_1$ and the complement of $\mathcal{O}_2$ by $\mathcal{O}_1 \backslash \mathcal{O}_2$.

The following set will be used in what follows:

$$\mathcal{A} := \left\{ \xi \in \mathcal{G} : \sup_{k \in \mathbb{Z}_{\geq 0}, \phi \in \mathcal{S}(\xi)} \omega_1(\phi(k, \xi)) = 0 \right\}. \tag{2.3}$$

In most cases the set $\mathcal{A}$ will be nonempty, but we observe that this is not necessary for the following results to hold. When $\mathcal{A}$ is empty, we define $|x|_{\mathcal{A}} = \inf_{a \in \mathcal{A}} |x - a|$ to be infinite.

For stability with respect to (ω, ω) ($\omega : \mathcal{G} \to \mathbb{R}_{\geq 0}$ continuous) the closed set $\mathcal{A}$ is

$$\mathcal{A} := \{ x \in \mathcal{G} : \omega(x) = 0 \}.$$

This follows from the previous definition (2.3) by examining the $\mathcal{KL}$-estimate defining stability. Specifically, if

$$\omega(\phi(k, x)) \leq \beta(\omega(x), k) \quad \forall x \in \mathcal{G},\ \phi \in \mathcal{S}(x),\ k \in \mathbb{Z}_{\geq 0},$$

then, for $\xi \in \mathcal{G}$, $\omega(\xi) = 0$ if and only if $\sup_{k \in \mathbb{Z}_{\geq 0}, \phi \in \mathcal{S}(\xi)} \omega(\phi(k, \xi)) = 0$.

Our robustness definition is defined relative to the above σ-perturbation.

DEFINITION 2.3. *Let $F(\cdot)$ be a set-valued map from $\mathcal{G}$ to subsets of $\mathcal{G}$. We say that the difference inclusion $x^+ \in F(x)$ is robustly $\mathcal{KL}$-stable with respect to (ω_1, ω_2) on $\mathcal{G}$ if there exists a continuous function $\sigma : \mathcal{G} \to \mathbb{R}_{\geq 0}$ such that*

1. *for all $x \in \mathcal{G}$, $\{x\} + \sigma(x)\overline{\mathcal{B}} \subset \mathcal{G}$;*
2. *for all $x \in \mathcal{G} \backslash \mathcal{A}$, $\sigma(x) > 0$;*

3.

$$\mathcal{A} = \mathcal{A}_\sigma := \left\{ \xi \in \mathcal{G} : \sup_{k \in \mathbb{Z}_{\geq 0},\ \phi \in \mathcal{S}_\sigma(\xi)} \omega_1(\phi(k,\xi)) = 0 \right\}; \quad \text{and} \tag{2.4}$$

4. *the difference inclusion $x^+ \in F_\sigma(x)$ is $\mathcal{KL}$-stable with respect to (ω_1, ω_2) on $\mathcal{G}$.*

In what follows, we denote the exponential function by e.

DEFINITION 2.4. *A function $V : \mathcal{G} \to \mathbb{R}_{\geq 0}$ is said to be a Lyapunov function with respect to (ω_1, ω_2) on $\mathcal{G}$ for the difference inclusion $x^+ \in F(x)$ if there exist $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ such that for all $x \in \mathcal{G}$,*

$$\alpha_1(\omega_1(x)) \leq V(x) \leq \alpha_2(\omega_2(x)), \tag{2.5}$$

$$\sup_{f \in F(x)} V(f) \leq V(x)e^{-1}, \quad \text{and} \tag{2.6}$$

$$V(x) = 0 \iff x \in \mathcal{A}. \tag{2.7}$$

We claim that the above decrease condition (2.6) can be stated as

$$\sup_{f \in F(x)} V(f) \leq V(x) - \alpha(V(x)) \quad \forall x \in \mathcal{G}, \tag{2.8}$$

where $\alpha : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is continuous and positive definite. However, we prefer (2.6) because of the symmetry with the continuous time decrease condition

$$\sup_{\omega \in F(x)} \langle \nabla V(x), \omega \rangle \leq -V(x),$$

which yields an exponential decrease of the Lyapunov function along trajectories. The following claim is proved in section 8.

CLAIM 1. *Suppose we are given functions $V : \mathcal{G} \to \mathbb{R}_{\geq 0}$, $\alpha : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, and $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ satisfying* (2.5), (2.7), *and* (2.8). *Then there exist $W : \mathcal{G} \to \mathbb{R}_{\geq 0}$ and functions $\hat{\alpha}_1$, $\hat{\alpha}_2 \in \mathcal{K}_\infty$ such that for all $x \in \mathcal{G}$*

$$\hat{\alpha}_1(\omega_1(x)) \leq W(x) \leq \hat{\alpha}_2(\omega_2(x)), \tag{2.9}$$

$$\sup_{f \in F(x)} W(f) \leq e^{-1} W(x), \quad \text{and} \tag{2.10}$$

$$W(x) = 0 \iff x \in \mathcal{A}. \tag{2.11}$$

Prior to stating our first result we require two definitions related to set-valued maps.

DEFINITION 2.5. *The set-valued map $F(\cdot)$ is said to be upper semicontinuous on (the open set) $\mathcal{O}$ if for each $x \in \mathcal{O}$ and $\varepsilon > 0$ there exists $\delta > 0$ such that, for all $\xi \in \mathcal{O}$ satisfying $|x - \xi| < \delta$, we have $F(\xi) \subseteq F(x) + \varepsilon \mathcal{B}$.*

We point out that the concept of upper semicontinuity for a set-valued map is not the same as that for extended real-valued functions. In fact, for $f : \mathbb{R}^n \to \mathbb{R}^n$, the set-valued map $x \mapsto \{f(x)\}$ is upper semicontinuous if and only if $x \mapsto f(x)$ is continuous.

DEFINITION 2.6. *We say that the set-valued map $F(\cdot)$ satisfies the basic conditions on $\mathcal{G}$ if $F(\cdot)$ is upper semicontinuous on $\mathcal{G}$ and, for each $x \in \mathcal{G}$, $F(x)$ is nonempty and compact.*

In continuous time the "basic conditions" also require convexity of $F(x)$ for each $x \in \mathcal{G}$. This is necessary to guarantee solutions of the differential inclusion $\dot{x} \in F(x)$ (see [5]). Obviously, solutions to the difference inclusion $x^+ \in F(x)$ will exist so long as $F(x)$ is nonempty.

With all the necessary definitions in hand, we may state under what conditions robust stability is equivalent to the existence of a Lyapunov function. The following is the discrete-time analogue of [22, Theorem 1] and is proved in section 6.

THEOREM 2.7. *Let $F(\cdot)$ mapping $\mathcal{G}$ to subsets of $\mathcal{G}$ satisfy the basic conditions on $\mathcal{G}$. Then, for the difference inclusion $x^+ \in F(x)$, there exists a smooth Lyapunov function with respect to (ω_1, ω_2) on $\mathcal{G}$ if and only if the inclusion is robustly $\mathcal{KL}$-stable with respect to (ω_1, ω_2) on $\mathcal{G}$.*

Remark 1. We note that (2.4) and (2.7) were not required in the corresponding definitions of robustness and a Lyapunov function in [22]. The addition of (2.4) to the definition of robustness significantly simplifies the proof. In order to maintain the equivalence of robustness and the existence of a Lyapunov function, one would then expect that an extra property is required of $V(\cdot)$. This property is (2.7). This is not unreasonable as, in the case of a single measure, we see that the upper and lower bounds (2.5) actually imply (2.7). □

It is possible to weaken the conditions of Theorem 2.7 and still maintain the necessity. This means that, in order to demonstrate robustness, it is only necessary to exhibit a continuous Lyapunov function (rather than a smooth one). Furthermore, note that we can drop the regularity requirement on the set-valued map. This allows application of the theorem, for example, to the consideration of discontinuous difference equations.

THEOREM 2.8. *Let $F(\cdot)$ mapping $\mathcal{G}$ to subsets of $\mathcal{G}$ be compact and nonempty, and suppose we have a continuous Lyapunov function. Then $x^+ \in F(x)$ is robustly $\mathcal{KL}$-stable with respect to (ω_1, ω_2) on $\mathcal{G}$.*

Since Lyapunov functions can sometimes be difficult to find, we would like a sufficient condition for robustness that is independent of having a Lyapunov function. Intuitively, if the set-valued map $F(\cdot)$ of (1.1) is sufficiently regular, robustness should follow since small perturbations will lead to small deviations. In fact, continuity of $F(\cdot)$ outside of the set $\mathcal{A}$ is sufficient.

DEFINITION 2.9. *We say the set-valued map $F(\cdot)$ is continuous on (the open set) $\mathcal{O}$ if, in addition to being upper semicontinuous on $\mathcal{O}$, for each $x \in \mathcal{O}$ and $\varepsilon > 0$ there exists $\delta > 0$ such that, for $z \in \mathcal{O}$ satisfying $|z - x| < \delta$, we have $F(x) \subseteq F(z) + \varepsilon\mathcal{B}$.*

The following theorem is the discrete-time counterpart of [22, Theorem 2] and is proved in section 7.

THEOREM 2.10. *Let $F(\cdot)$ be a set-valued map from $\mathcal{G}$ to subsets of $\mathcal{G}$ satisfying the basic conditions on $\mathcal{G}$ and continuous on an open set containing $\mathcal{G}\backslash\mathcal{A}$. Under these conditions, if $x^+ \in F(x)$ is $\mathcal{KL}$-stable with respect to (ω_1, ω_2) on $\mathcal{G}$, then the inclusion is robustly $\mathcal{KL}$-stable with respect to (ω_1, ω_2) on $\mathcal{G}$.*

3. Smoothing functions. Frequently one wishes to prove that certain assumptions such as asymptotic stability of a set or asymptotic controllability to a set imply the existence of a function satisfying certain boundedness and decrease properties, as well as a given regularity property. Typically, one constructs a function which satisfies all the given properties (i.e., boundedness and decrease properties) excepting the desired regularity property. One may then take the additional step of "smoothing" the constructed function without destroying the boundedness or decrease properties. Such techniques were first used by Kurzweil [13] and Wilson [23]. Throughout this

section, we take $\mathcal{O}$ to be an open set.

We will smooth nonsmooth functions via an integration which involves a change of variables. We will require the following assumption on the function we wish to smooth.

ASSUMPTION 1. *The function $V : \mathcal{O} \to \mathbb{R}_{\geq 0}$ is such that*

1. *$V(\cdot)$ is upper semicontinuous and locally bounded on $\mathcal{O}$,*
2. *$V(x) > 0$ implies that there exists $\delta > 0$ such that $|z-x| < \delta$ implies $V(z) > 0$.*

Define

$$\mathcal{A} := \{x \in \mathcal{O} : V(x) = 0\}. \tag{3.1}$$

We observe that, under the above assumption on $V(\cdot)$, the set $\mathcal{O}\backslash\mathcal{A}$ is open. Note that we need not assume that $\mathcal{A}$ is nonempty.

We will also require an assumption on the "smoothing perturbation."

ASSUMPTION 2. *The smooth function $\sigma : \mathcal{O}\backslash\mathcal{A} \to \mathbb{R}_{>0}$ satisfies the following:*

1. *for each $x^* \in \mathcal{A}$ and $\varepsilon > 0$ there exists $\delta > 0$ such that*

$$x \in \mathcal{O}\backslash\mathcal{A}\ , \quad |x - x^*| \leq \delta \qquad \Longrightarrow \qquad \sigma(x) \leq \varepsilon, \tag{3.2}$$

2.

$$x \in \mathcal{O}\backslash\mathcal{A} \qquad \Longrightarrow \qquad \{x\} + \sigma(x)\overline{\mathcal{B}} \subset \mathcal{O}. \tag{3.3}$$

Item 1 implies that the function $\sigma(\cdot)$ can be continuously extended to the set $\mathcal{A}$ by defining it to be identically zero on $\mathcal{A}$. For the case where $\mathcal{A}$ is empty, item 1 is trivially satisfied.

We define $V_s : \mathcal{O} \to \mathbb{R}_{\geq 0}$ as

$$\begin{aligned} V_s(x) &:= 0, && x \in \mathcal{A}, \\ V_s(x) &:= \int V(x + \sigma(x)\xi)\psi(\xi)\, d\xi, && x \in \mathcal{O}\backslash\mathcal{A}, \end{aligned} \tag{3.4}$$

where $\psi : \mathbb{R}^n \to [0, 1]$ is smooth, vanishes on $\mathbb{R}^n\backslash\overline{\mathcal{B}}$, and satisfies $\int \psi(\xi)d\xi = 1$.

The following theorem is a generalization of [11, Theorem 20], where the smoothing was carried out on $\mathbb{R}^n\backslash\{0\}$.

THEOREM 3.1. *Under Assumptions* 1 *and* 2*, the function $V_s : \mathcal{O} \to \mathbb{R}_{\geq 0}$ defined by* (3.4) *is well defined, continuous on $\mathcal{O}$, and smooth on $\mathcal{O}\backslash\mathcal{A}$.*

Proof. The properties of $\sigma(\cdot)$ and $\psi(\cdot)$ and the upper semicontinuity of $V(\cdot)$ guarantee that the (Lebesgue) integral in (3.4) is well defined.

Continuity at $\mathcal{A}$. Since $V_s(x) \equiv 0$ for $x \in \mathcal{A}$, the function is clearly continuous in the interior of $\mathcal{A}$. It remains to check continuity at the boundary of $\mathcal{A}$. Let x^* belong to the boundary of $\mathcal{A}$ so that $V_s(x^*) = 0$. Let $\varepsilon > 0$ be given. Since $V(\cdot)$ is upper semicontinuous, there exists $\delta_2 > 0$ such that $V(z) \leq \varepsilon$ for all $z \in \mathcal{O}$ satisfying $|z-x^*| \leq \delta_2$. Since (3.2) holds, there exists $\delta > 0$ such that $\{x\}+\sigma(x)\overline{\mathcal{B}} \subseteq \{x^*\}+\delta_2\overline{\mathcal{B}}$ for all $|x-x^*| \leq \delta$. Consequently, with the fact that $\int \psi(\xi)d\xi = 1$, if $|x-x^*| \leq \delta$ and $x \in \mathcal{O}\backslash\mathcal{A}$, then

$$V_s(x) = \int V(x + \sigma(x)\xi)\psi(\xi)d\xi \leq \sup_{z \in \{x^*\}+\delta_2\overline{\mathcal{B}}} V(z) \leq \varepsilon;$$

i.e., $V_s(x)$ is continuous for x in the boundary of $\mathcal{A}$.

Finally, if we can establish that V_s is smooth on $\mathcal{O}\backslash\mathcal{A}$, then it will be continuous on $\mathcal{O}$.

Smoothness on $\mathcal{O}\backslash\mathcal{A}$. For each $x \in \mathcal{O}\backslash\mathcal{A}$, we perform a change of variables under the integration with $z = x + \sigma(x)\xi$ so that

$$V_s(x) = \sigma(x)^{-n} \int V(z)\psi(\sigma(x)^{-1}(z-x))\, dz.$$

For notational purposes, we define $h(x,z) := \psi(\sigma(x)^{-1}(z-x))$ so that, for each $x \in \mathcal{O}\backslash\mathcal{A}$,

$$V_s(x) = \sigma(x)^{-n} \int V(z)h(x,z)\, dz.$$

For $x, z \in \mathcal{O}\backslash\mathcal{A}$ such that $|z-x| > \sigma(x)$ we note that $h(x,z)$ and all of its higher order partial derivatives with respect to x vanish. From this, (3.4), and the fact that $\psi(\cdot)$ and $\sigma(\cdot)$ are smooth (the latter on $\mathcal{O}\backslash\mathcal{A}$) it follows that each of these partial derivatives is continuous in x uniformly in z.

Because of the properties of $\sigma(\cdot)$, to establish smoothness of $V_s(\cdot)$ on $\mathcal{O}\backslash\mathcal{A}$ it is enough to establish smoothness of

$$W_s(x) := \int V(z)h(x,z)\, dz.$$

We note that, using the mean value theorem, for each $x \in \mathcal{O}\backslash\mathcal{A}$, $\varepsilon > 0$, and $v \in \mathbb{R}^n$, there exists $\lambda \in [0,1]$ such that

$$\begin{aligned}\frac{W_s(x+\varepsilon v) - W_s(x)}{\varepsilon} &= \int V(z)\frac{h(x+\varepsilon v, z) - h(x,z)}{\varepsilon}\, dz \\ &= \int V(z)\langle \nabla h(x+\varepsilon\lambda v, z), v\rangle\, dz \\ &= r(x,\varepsilon,v) + \int V(z)\langle \nabla h(x,z), v\rangle\, dz,\end{aligned}$$

where $r(x,\varepsilon,v) := \int V(z)\langle \nabla h(x+\varepsilon\lambda v, z) - \nabla h(x,z), v\rangle dz$.

Using that $V(\cdot)$ is locally bounded on $\mathcal{O}$, (3.3), the fact that $\nabla h(x,z) = 0$ when $|z-x| > \sigma(x)$, and the continuity of $\nabla h(\cdot,z)$, which is uniform in z, for each $\rho > 0$ and $M > 0$ there exists $\varepsilon^* > 0$ such that if $\varepsilon \in (0,\varepsilon^*]$ and $|v| \le M$, then $|r(x,\varepsilon,v)| \le \rho$. It follows that $W_s(\cdot)$ is (Fréchet) differentiable (hence continuous) and

$$\langle \nabla W_s(x), v\rangle = \int V(z)\langle \nabla h(x,z), v\rangle\, dz.$$

Repeating this argument for higher order derivatives, we conclude that $W_s(\cdot)$ is smooth on $\mathcal{O}\backslash\mathcal{A}$. □

The following lemma can be applied to the function $V_s(\cdot)$ obtained from Theorem 3.1 in order to obtain a function that is smooth on the entire domain $\mathcal{O}$. The lemma appeared as [22, Lemma 17], which derives from [15, Lemma 4.3] and [13, Theorem 6].

LEMMA 3.2. *Let $\mathcal{A} \subset \mathcal{O}$ be a closed set, and assume that $V_s : \mathcal{O} \to \mathbb{R}_{\ge 0}$ is continuous, the restriction of V_s to $\mathcal{O}\backslash\mathcal{A}$ is smooth, $V_s(x) = 0$ for all $x \in \mathcal{A}$, and $V_s(x) > 0$ for all $x \in \mathcal{O}\backslash\mathcal{A}$. Then there exists a strictly convex function $\rho \in \mathcal{K}_\infty$, smooth on $(0,\infty)$, such that $V := \rho \circ V_s$ is smooth on $\mathcal{O}$.*

Frequently, to construct the function $\sigma(\cdot)$ used in the integral smoothing of Theorem 3.1, we will first specify a constraint which $\sigma(\cdot)$ must satisfy and then take a smaller smooth function.

LEMMA 3.3. *Given a function $\sigma_2 : \mathcal{O} \to \mathbb{R}_{>0}$ bounded away from zero on compact subsets of $\mathcal{O}$ there exists a smooth function $\sigma_1 : \mathcal{O} \to \mathbb{R}_{>0}$, also bounded away from zero on compact subsets of $\mathcal{O}$, such that, for all $x \in \mathcal{O}$, $\sigma_1(x) \leq \sigma_2(x)$.*

Proof. We let $\{\mathcal{U}_i\}_{i=1}^{\infty}$ be a locally finite open cover of $\mathcal{O}$ with $\overline{\mathcal{U}}_i$ a compact subset of $\mathcal{O}$ and let $\{\kappa_i\}_{i=1}^{\infty}$ be a smooth partition of unity on $\mathcal{O}$ subordinate to $\{\mathcal{U}_i\}$. Define $\varepsilon_i := \inf_{\xi \in \mathcal{U}_i} \sigma_2(\xi)$,

$$\sigma_1(x) := \sum_{i=1}^{\infty} \kappa_i(x)\varepsilon_i,$$

and, for each $x \in \mathcal{O}$, $\mathcal{I}_x := \{j : x \in \mathcal{U}_j\}$. The set $\mathcal{I}_x$ is finite for each $x \in \mathcal{O}$. We also note that

$$\max_{j \in \mathcal{I}_x} \varepsilon_j = \max_{j \in \mathcal{I}_x} \inf_{\xi \in \mathcal{U}_j} \sigma_2(\xi) \leq \sigma_2(x).$$

Since $\overline{\mathcal{U}}_i$ is a compact subset of $\mathcal{O}$ for each i and $\sigma_2(\cdot)$ is bounded away from zero on compact subsets of $\mathcal{O}$, we have $\varepsilon_i > 0$ for each i. Thus $\sigma_1(x) > 0$ for all $x \in \mathcal{O}$. Also,

$$\sigma_1(x) \leq \max_{j \in \mathcal{I}_x} \varepsilon_j \leq \sigma_2(x).$$

Finally, σ_1 is smooth on $\mathcal{O}$, inheriting this property from the κ_i. ∎

4. Set-valued maps. Prior to stating our novel results for difference inclusions, we require certain facts from set-valued analysis. Our primary sources for set-valued analysis include the books by Aubin and Cellina [2], Aubin and Frankowska [3], Filippov [5], and Kisielewicz [12].

Given a set-valued map $F(\cdot)$ from an open set $\mathcal{O} \subset \mathbb{R}^n$ to subsets of $\mathbb{R}^n$, we define the mapping of a compact set M by

$$F(M) := \bigcup_{\xi \in M} F(\xi).$$

We also define the composition of two set-valued maps $F(\cdot)$ and $G(\cdot)$ to be

$$F(G(x)) := \bigcup_{w \in G(x)} F(w),$$

and we denote the n-times composition of $F(\cdot)$ with itself by $F^n(\cdot)$ (e.g., $F(F(x)) = F^2(x)$).

The following is well known. See, for example, [12, Proposition 2.3].

CLAIM 2. *Let $F(\cdot)$ be an upper-semicontinuous set-valued map from $\mathcal{O}$ to subsets of $\mathbb{R}^n$, let $M \subset \mathcal{O}$ be compact, and let $F(x)$ be compact for all $x \in \mathcal{O}$. Then the set $F(M)$ is compact.*

For $\delta \geq 0$, we define the δ-perturbation of the set-valued map $F(\cdot)$ by

$$F_\delta(x) := F\left(\{x\} + \delta\overline{\mathcal{B}}\right) + \delta\overline{\mathcal{B}}$$

and the δ-inflation of a set M by

$$M_\delta := M + \delta\overline{\mathcal{B}}.$$

The following claim, which is not difficult to prove, extends the concept of upper semicontinuity to the consideration of compact sets rather than merely points.

CLAIM 3. *Let $F(\cdot)$ be an upper-semicontinuous set-valued map from $\mathcal{O}$ to subsets of $\mathbb{R}^n$ and let $M \subset \mathcal{O}$ be compact. Then for every $\varepsilon > 0$ there exists $\delta > 0$ such that*

$$F_\delta(M_\delta) \subseteq F(M) + \varepsilon\overline{\mathcal{B}} \ .$$

CLAIM 4. *Let $F(\cdot)$ be an upper-semicontinuous set-valued map from $\mathcal{O}$ to subsets of $\mathbb{R}^n$. Let $k \in \mathbb{Z}_{>0}$, $i \in \{1, 2, \ldots, k\}$, and S_i compact subsets of $\mathcal{O}$. Then there exist $\rho \in \mathcal{K}_\infty$ and $c > 0$ such that for every $\delta \in (0, c]$*

$$F_\delta(S_{i_\delta}) \subseteq F(S_i) + \rho(\delta)\overline{\mathcal{B}}.$$

Proof. For a particular S_i, let $\varepsilon > 0$. Then, from the result of Claim 3, there exists $\delta_i > 0$ such that $F_{\delta_i}(S_{i_\delta}) \subseteq F(S_i) + \varepsilon\overline{\mathcal{B}}$. For fixed $\varepsilon > 0$, let $\bar{\delta}_i(\varepsilon)$ be the supremum of all applicable $\delta_i(\varepsilon)$. Therefore,

$$F_{\bar{\delta}_i(\varepsilon)}(S_{i_{\bar{\delta}_i(\varepsilon)}}) \subseteq F(S_i) + \varepsilon\overline{\mathcal{B}}.$$

We note that $\bar{\delta}_i(\varepsilon)$ is positive and nondecreasing, but not necessarily continuous. Choose $\alpha_i \in \mathcal{K}$ such that $\alpha_i(r) \leq k\bar{\delta}_i(r)$ for all $r \in \mathbb{R}_{\geq 0}$ with $k \in (0, 1)$. Let $c_i := \lim_{r\to\infty} \alpha_i(r)$ and $\rho_i(r) := \alpha_i^{-1}(r)$ for all $r \in [0, c_i)$. Then ρ_i is continuous, zero at zero, strictly increasing, and is defined on $[0, c_i)$. Given $\delta_i < c_i$, let $\varepsilon = \rho_i(\delta_i)$. Then $\delta_i < \bar{\delta}_i(\varepsilon)$ and

$$F_{\delta_i}(S_{i_\delta}) \subseteq F(S_i) + \rho_i(\delta_i)\overline{\mathcal{B}}.$$

Let $c^* := \min_{i\in\{1,\ldots,k\}}\{c_i\}$ and, for each $r \in [0, c^*)$, let $\hat{\rho}(r) := \max_{i\in\{1,\ldots,k\}} \rho_i(r)$. Then, for each $i \in \{1, 2, \ldots, k\}$, $F_\delta(S_{i_\delta}) \subseteq F(S_i) + \hat{\rho}(\delta)\overline{\mathcal{B}}$ for all $\delta \in (0, c^*)$. Finally, let $\rho \in \mathcal{K}_\infty$ be such that $\rho(r) \geq \hat{\rho}(r)$ for all $r \in [0, \frac{1}{2}c^*]$. Therefore, with $c := \frac{1}{2}c^*$,

$$F_\delta(S_{i_\delta}) \subseteq F(S_i) + \rho(\delta)\overline{\mathcal{B}} \quad \forall \delta \in (0, c]. \qquad \square$$

CLAIM 5. *Suppose $F(\cdot)$ is an upper-semicontinuous set-valued map from $\mathcal{O}$ to subsets of $\mathbb{R}^n$ and that, for each $x \in \mathcal{O}$, $F(x)$ is nonempty and compact. Let M be a compact set in $\mathcal{O}$ and $K \in \mathbb{Z}_{>0}$. Assume $F^k(M) \subset \mathcal{O}$ for all $k \in \{1, \ldots, K\}$. Then there exist $\tilde{\rho} \in \mathcal{K}_\infty$ and $\tilde{c} > 0$ such that, for every $\delta \in (0, \tilde{c}]$ and $k \in \{1, \ldots, K\}$,*

$$F_\delta^k(M_\delta) \subseteq F^k(M) + \tilde{\rho}(\delta)\overline{\mathcal{B}}.$$

Proof. Define the compact sets $S_0 := M$, $S_1 := F(M), \ldots, S_k := F^k(M)$. Let $\rho \in \mathcal{K}_\infty$ and $c > 0$ come from Claim 4. Without loss of generality, assume $\rho(s) \geq s$ for all $s \in \mathbb{R}_{\geq 0}$. Let $\tilde{c} > 0$ be such that $\rho^{k-1}(\tilde{c}) < c$ and define $\tilde{\rho}(r) := \rho^k(r)$ for all $r \in [0, c)$ (where $\rho^k(\cdot)$ is the k-times composition of $\rho(\cdot)$ with itself). From Claim 4 we may write

$$F_\delta(M_\delta) = F_\delta(S_0 + \delta\overline{\mathcal{B}}) \subseteq F(M) + \rho(\delta)\overline{\mathcal{B}}.$$

Since $\delta < \tilde{c}$, we have that $\rho(\delta) < c$.

Assume the result holds for $k-1$; i.e., $F_\delta^{k-1}(M_\delta) \subseteq F^{k-1}(M) + \rho^{k-1}(\delta)\overline{\mathcal{B}}$. Noting that $\delta \leq \rho^{k-1}(\delta) < c$ we may write

$$\begin{aligned} F_\delta^k(M_\delta) &= F_\delta\left(F_\delta^{k-1}(M_\delta)\right) \ \subseteq \ F_\delta\left(F^{k-1}(M) + \rho^{k-1}(\delta)\overline{\mathcal{B}}\right) \\ &\subseteq F_{\rho^{k-1}(\delta)}\left(F^{k-1}(M) + \rho^{k-1}(\delta)\overline{\mathcal{B}}\right) \ \subseteq \ F^k(M) + \rho^k(\delta)\overline{\mathcal{B}} \\ &= F^k(M) + \tilde{\rho}(\delta)\overline{\mathcal{B}}, \end{aligned}$$

where the final subset is obtained by appealing to Claim 4. □

We will need to apply the lemmas in the following section to the difference inclusion defined by $x^+ \in F_\sigma(x)$ with $F_\sigma(\cdot)$ as in (2.2). To do this, we need to know that $F_\sigma(\cdot)$ satisfies the basic conditions.

CLAIM 6. *If $F(\cdot)$ is a set-valued map from $\mathcal{G}$ to subsets of $\mathcal{G}$ satisfying the basic conditions on $\mathcal{G}$, and $\sigma : \mathcal{G} \to \mathbb{R}_{\geq 0}$ satisfies item 1 of Definition 2.3, then $F_\sigma(\cdot)$ satisfies the basic conditions on $\mathcal{G}$.*

Proof. That $F_\sigma(x)$ is nonempty follows from $F(x)$ nonempty. Similarly, $F_\sigma(x)$ being compact follows from $F(x)$ compact, the compactness of the closed unit ball, $F(\cdot)$ upper semicontinuous, and the fact that upper-semicontinuous maps send compacts to compacts (see Claim 2).

Appealing to Claim 3 with $M := \{x\} + \sigma(x)\overline{\mathcal{B}}$ and any $\varepsilon > 0$ there exists $\delta > 0$ such that $F(M_\delta) = F\left(\{x\} + (\sigma(x) + \delta)\overline{\mathcal{B}}\right) \subseteq F\left(\{x\} + \sigma(x)\overline{\mathcal{B}}\right) + \varepsilon\overline{\mathcal{B}}$. Let $\varepsilon_\sigma = \frac{\delta}{2} > 0$. Then, since $\sigma(\cdot)$ is continuous, there exists $\delta_\sigma \in (0, \frac{\delta}{2}]$ such that if $|x - z| < \delta_\sigma$, then $|\sigma(x) - \sigma(z)| < \varepsilon_\sigma = \frac{\delta}{2}$. Therefore, $\{z\} + \sigma(z)\overline{\mathcal{B}} \subseteq \{x\} + (\sigma(x) + \delta)\overline{\mathcal{B}}$ and

$$F\left(\{z\} + \sigma(z)\overline{\mathcal{B}}\right) \subseteq F\left(\{x\} + \sigma(x)\overline{\mathcal{B}}\right) + \varepsilon\overline{\mathcal{B}};$$

i.e., $F_\sigma(\cdot)$ is upper semicontinuous on $\mathcal{G}$. □

5. Difference inclusions. In this section we present three new results for difference inclusions which will be necessary for the proofs of the results presented in section 2.

The first result makes use of a perturbed difference inclusion. Let $F(\cdot)$ map an open set $\mathcal{O} \subset \mathbb{R}^n$ to subsets of $\mathbb{R}^n$, let $\delta \geq 0$, and consider

$$x^+ \in F_\delta(x) := F(x + \delta\overline{\mathcal{B}}) + \delta\overline{\mathcal{B}}, \quad x \in \mathcal{O}. \tag{5.1}$$

We denote the solution set of (5.1) from the point $x \in \mathbb{R}^n$ by $\mathcal{S}_\delta(x)$. This result is similar to a result on closeness of solutions for differential inclusions (see, for example, [5, section 8, Corollary 2]).

LEMMA 5.1. *Let $\mathcal{O}$ be open, $\omega : \mathcal{O} \to \mathbb{R}_{\geq 0}$ be continuous, and $F(\cdot)$ map $\mathcal{O}$ to subsets of $\mathbb{R}^n$ satisfy the basic conditions on $\mathcal{O}$. Let the triple (K, ε, M) be such that $K \in \mathbb{Z}_{>0}$, $\varepsilon > 0$, $M \subset \mathcal{O}$ compact, and $F^k(M) \subset \mathcal{O}$ for all $k \in \{1, \ldots, K\}$. Under these conditions, there exists $\delta > 0$ such that for every $x \in M_\delta$*

1. *every solution $\psi \in \mathcal{S}_\delta(x)$ satisfies $\psi(k, x) \in \mathcal{O}$ for $k \in \{0, \ldots, K\}$, and*
2. *for every $\psi \in \mathcal{S}_\delta(x)$ there exists $\bar{x} \in M$ and $\phi \in \mathcal{S}(x)$ such that for all $k \in \{0, \ldots, K\}$ we have*

$$|\omega(\psi(k, x)) - \omega(\phi(k, \bar{x}))| \leq \varepsilon. \tag{5.2}$$

Proof. The first item follows from Claim 5 and the fact that $F^k(M)$ is compact for each $k \in \mathbb{Z}_{>0}$.

If the second item is not true, then no matter how small we pick δ, there is an initial condition in M_δ and a solution to $x^+ \in F_\delta(x)$ starting at this initial condition such that, no matter which initial condition in M and solution of $x^+ \in F(x)$ we pick, the condition (5.2) is violated for some $k \in \{0, \ldots, K\}$. In particular, there exist sequences $x_i \in M_{1/i}$ and $\psi_i \in \mathcal{S}_{1/i}(x_i)$ such that, no matter which initial condition in M and solution of $x^+ \in F(x)$ we pick, the condition (5.2) is violated for some $k \in \{0, \ldots, K\}$. The sequence x_i has a subsequence, which we will not relabel, converging to a point $f_0^* \in M$. Associated with this subsequence is a sequence of

points $\psi_i(1, x_i) \in F_{1/i}(x_i)$. This sequence has a converging subsequence and, from the upper semicontinuity of F and compactness of $F(f_0^*)$, its accumulation point, denoted f_1^*, belongs to $F(f_0^*)$. Continuing in this way we get a sequence of initial conditions $x_i \in M_{1/i}$ and a sequence of solutions $\psi_i \in \mathcal{S}_{1/i}(x_i)$ such that $x_i \to f_0^* \in M$ and $\psi_i(k, x_i) \to f_k^* \in F(f_{k-1}^*)$. Now with the solution $\phi \in \mathcal{S}(f_0^*)$ given by $\phi(k, f_0^*) = f_k^* \in F(\phi(k-1, f_0^*))$ for all $k \in \{1, \dots, K\}$, and using the continuity of ω, condition (5.2) holds for all i sufficiently large. This is a contradiction and thus proves the lemma. ∎

We will require the following lemma on closeness of solutions for difference inclusions defined by continuous set-valued maps.

LEMMA 5.2. *Suppose $F(\cdot)$ is a set-valued map from $\mathcal{O}$ to subsets of $\mathbb{R}^n$ continuous on an open set $\mathcal{O}_1 \subseteq \mathcal{O}$ and that, for each $x \in \mathcal{O}$, $F(x)$ is compact and nonempty. Furthermore, suppose $\omega : \mathcal{O} \to \mathbb{R}_{\geq 0}$ is continuous. For each triple (K, ε, x_0) such that $K \in \mathbb{Z}_{>0}$, $\varepsilon > 0$, and $x_0 \in \mathcal{O}$, and for each solution $\phi \in \mathcal{S}(x_0)$ such that $\phi(k, x_0) \in \mathcal{O}_1$ for all $k \in \{0, \dots, K\}$ there exists a $\delta > 0$ such that, for every $x \in \{x_0\} + \delta\overline{\mathcal{B}}$, there exists a solution $\psi \in \mathcal{S}(x)$ such that, for all $k \in \{0, \dots, K+1\}$,*

$$|\omega(\phi(k, x_0)) - \omega(\psi(k, x))| \leq \varepsilon.$$

Proof. Define the compact set

$$\mathcal{C} := \{\phi(k, x_0)\} \subset \mathcal{O}_1 \subseteq \mathcal{O} \quad \forall k \in \{0, \dots, K\}.$$

For the given $\varepsilon > 0$, since $\omega(\cdot)$ is continuous, there exists $\delta_\omega > 0$ such that $r \in \mathcal{C}$ and $|s - r| \leq \delta_\omega$ imply $|\omega(s) - \omega(r)| \leq \varepsilon$. Without loss of generality, we also impose $\delta_\omega \leq \varepsilon$ and $\mathcal{C} + \delta_\omega\overline{\mathcal{B}} \subset \mathcal{O}_1$.

From the continuity of $F(\cdot)$ at $\phi(K, x_0)$, there exists $\delta_K \in (0, \delta_\omega]$ such that for all $z \in \mathcal{O}$ satisfying $|z - \phi(K, x_0)| \leq \delta_K$ we have $F(\phi(K, x_0)) \subseteq F(z) + \delta_\omega\overline{\mathcal{B}}$. Similarly, from the continuity of $F(\cdot)$ at $\phi(K-1, x_0)$, there exists $\delta_{K-1} \in (0, \delta_\omega]$ such that for all $z \in \mathcal{O}$ satisfying $|z - \phi(K-1, x_0)| \leq \delta_{K-1}$ we have $F(\phi(K-1, x_0)) \subseteq F(z) + \delta_K\overline{\mathcal{B}}$. We repeat this procedure until we reach the initial point x_0. From the previous step we will have a $\delta_1 \in (0, \delta_\omega]$. Then, from the continuity of $F(\cdot)$ at x_0, there exists a $\delta_0 \in (0, \delta_\omega]$ such that, for all $z \in \mathcal{O}$,

$$|z - x_0| \leq \delta_0 \quad \Longrightarrow \quad F(x_0) \subseteq F(z) + \delta_1\overline{\mathcal{B}}. \tag{5.3}$$

From (5.3), for any $x \in \{x_0\} + \delta_0\overline{\mathcal{B}}$ there exists a point $\psi(1, x) \in F(x)$ such that

$$|\phi(1, x_0) - \psi(1, x)| \leq \delta_1. \tag{5.4}$$

This follows from (5.3) since $|x - x_0| \leq \delta_0$, so that $\phi(1, x_0) \in F(x_0) \subseteq F(x) + \delta_1\overline{\mathcal{B}}$.

Since (5.4) holds, we see that there exists a point $\psi(2, x) \in F(\psi(1, x))$ such that

$$|\phi(2, x_0) - \psi(2, x)| \leq \delta_2.$$

This follows from (5.4) since $\phi(2, x_0) \in F(\phi(1, x_0)) \subseteq F(\psi(1, x)) + \delta_2\overline{\mathcal{B}}$. That is, for the point $\phi(2, x_0)$, there exists an element in $F(\psi(1, x))$ (which we have called $\psi(2, x)$) that is no more than δ_2 away from $\phi(2, x_0)$.

We can repeat this procedure at each step until we get to $\phi(K+1, x_0)$.

Since, for each $\ell \in \{0, \dots, K\}$, we imposed $\delta_\ell \leq \delta_\omega$ we see that, with $\psi \in \mathcal{S}(x)$ constructed as above,

$$|\omega(\phi(k, x_0)) - \omega(\psi(k, x))| \leq \varepsilon \quad \forall k \in \{0, \dots, K+1\}. \qquad \blacksquare$$

We present a lemma regarding sequences of solutions. This lemma is similar to the continuous-time results found in [22, Lemmas 4 and 5], which derived from [5, section 7, Theorem 3].

LEMMA 5.3. *Let $F(\cdot)$ mapping $\mathcal{O}$ to subsets of $\mathbb{R}^n$ satisfy the basic conditions on $\mathcal{O}$. Let $x \in \mathcal{O}$ be given and suppose that all solutions $\phi \in \mathcal{S}(x)$ are defined and belong to $\mathcal{O}$ for all $k \geq 0$. Then each sequence $\{\phi_n\}_{n=1}^{\infty}$ of solutions in $\mathcal{S}(x)$ has a subsequence converging to a function $\phi \in \mathcal{S}(x)$ and the convergence is uniform on each finite time interval.*

Proof. From Claim 2 we know that for each $k \in \mathbb{Z}_{\geq 0}$ the set $F^k(x)$ is a compact set. Since for all n and k, $\phi_n(k,x) \in F^k(x)$, it follows that $\{\phi_n\}_{n=1}^{\infty}$ has a converging subsequence $\{\phi_{1m}\}_{m=1}^{\infty}$ such that $\phi_{1m}(1,x) \to f_1^* =: \phi(1,x)$. Similarly, $\{\phi_{1m}\}_{m=1}^{\infty}$ has a converging subsequence $\{\phi_{2m}\}_{m=1}^{\infty}$ such that $\phi_{2m}(2,x) \to f_2^* =: \phi(2,x)$, and so on. In this way, we construct a subsequence which converges to a solution $\phi \in \mathcal{S}(x)$, and, for a finite time interval, this convergence is uniform. ∎

6. Proof of Theorem 2.7. We demonstrate that robust $\mathcal{KL}$-stability is a necessary and sufficient condition for the existence of a smooth Lyapunov function with respect to (ω_1, ω_2).

6.1. Sufficiency. One of the most useful lemmas regarding comparison functions is frequently referred to as Sontag's lemma on $\mathcal{KL}$-estimates [19, Proposition 7]. This lemma allows us to view asymptotic stability as exponential stability via a suitable nonlinear scaling. The following lemma is a slight refinement of Sontag's original lemma wherein we specify the required regularity property of one of the $\mathcal{K}_\infty$ functions.

LEMMA 6.1. *For each $\beta \in \mathcal{KL}$ and $\lambda > 0$, there exist $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ such that $\alpha_1(\cdot)$ is Lipschitz on its domain, continuously differentiable on $(0,\infty)$, $\alpha_1(s) \leq s\alpha_1'(s)$ for all $s > 0$, and $\alpha_1(\beta(s,t)) \leq \alpha_2(s)e^{-\lambda t}$ for all $(s,t) \in \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$.*

6.1.1. Bounds. Given $\beta_\sigma \in \mathcal{KL}$ from the assumption of $\mathcal{KL}$-stability with respect to (ω_1, ω_2) for $x^+ \in F_\sigma(x)$, Lemma 6.1 yields $\hat{\alpha}_1, \hat{\alpha}_2 \in \mathcal{K}_\infty$ such that

$$\hat{\alpha}_1\left(\beta_\sigma(s,k)\right) \leq \hat{\alpha}_2(s)e^{-2k} \quad \forall k \in \mathbb{Z}_{\geq 0},\ \forall s \geq 0. \tag{6.1}$$

For each $x \in \mathcal{G}$ we define the function $V_1 : \mathcal{G} \to \mathbb{R}_{\geq 0}$ by

$$V_1(x) := \sup_{k \in \mathbb{Z}_{\geq 0}, \phi \in \mathcal{S}_\sigma(x)} \hat{\alpha}_1(\omega_1(\phi(k,x)))e^k. \tag{6.2}$$

We claim that

$$V_1(x) = 0 \quad \Longleftrightarrow \quad x \in \mathcal{A}_\sigma = \mathcal{A}. \tag{6.3}$$

To see this, note that if $x \in \mathcal{A}_\sigma$, then, by definition (2.4), $V_1(x) = 0$. Furthermore, $V_1(x) = 0$ implies that

$$0 = \sup_{k \in \mathbb{Z}_{\geq 0}, \phi \in \mathcal{S}_\sigma(x)} \hat{\alpha}_1(\omega_1(\phi(k,x))),$$

and, since $\hat{\alpha}_1 \in \mathcal{K}_\infty$, $x \in \mathcal{A}_\sigma = \mathcal{A}$.

It is easy to see that, for all $x \in \mathcal{G}$,

$$V_1(x) \geq \sup_{\phi \in \mathcal{S}_\sigma(x)} \hat{\alpha}_1(\omega_1(\phi(0,x)))e^0 = \hat{\alpha}_1(\omega_1(x)), \quad \text{and} \tag{6.4}$$

$$V_1(x) \leq \sup_{k\in\mathbb{Z}_{\geq 0}} \hat{\alpha}_1(\beta_\sigma(\omega_2(x),k))e^k$$
$$\leq \sup_{k\in\mathbb{Z}_{\geq 0}} \hat{\alpha}_2(\omega_2(x))e^{-k} = \hat{\alpha}_2(\omega_2(x)). \tag{6.5}$$

For each $x \in \mathcal{G}$ and $\phi \in \mathcal{S}_\sigma(x)$ we may write

$$\begin{aligned} V_1(\phi(j,x)) &= \sup_{k\in\mathbb{Z}_{\geq 0},\ \psi\in\mathcal{S}_\sigma(\phi(j,x))} \hat{\alpha}_1\left(\omega_1(\psi(k,\phi(j,x)))\right)e^k \\ &\leq \sup_{k\in\mathbb{Z}_{\geq j},\ \psi\in\mathcal{S}_\sigma(x)} \hat{\alpha}_1(\omega_1(\psi(k,x)))e^{k-j} \\ &\leq \sup_{k\in\mathbb{Z}_{\geq 0},\ \psi\in\mathcal{S}_\sigma(x)} \hat{\alpha}_1(\omega_1(\psi(k,x)))e^k e^{-j} \\ &= V_1(x)e^{-j}\ . \end{aligned}$$

We note that $w \in F_\sigma(x)$ implies the existence of $\phi \in \mathcal{S}_\sigma(x)$ such that $\phi(1,x) = w$. Therefore, we may write

$$\sup_{f\in F_\sigma(x)} V_1(f) \leq e^{-1}V_1(x) \quad \forall x \in \mathcal{G}. \tag{6.6}$$

6.1.2. Smoothing V_1. We proceed to smooth the function $V_1(\cdot)$ without destroying the nature of the upper and lower bounds (6.4) and (6.5) and the decrease condition (6.6). To do this, we will appeal to Theorem 3.1, which requires Assumptions 1 and 2 and uses a set $\mathcal{A}$ defined in (3.1) that, as a consequence of (6.3), is the same as the set $\mathcal{A}$ of (2.3). Assumption 1 requires that $V_1(\cdot)$ be upper semicontinuous and that if $V_1(x) > 0$, then $V_1(z) > 0$ for z near x.

$V_1(\cdot)$ is upper semicontinuous: We first show that for each $x \in \mathcal{G}\backslash\mathcal{A}$ there exists a solution such that the supremum defining $V_1(\cdot)$ is attained for some solution and over a finite time interval.

Claim 7. *Let $x \in \mathcal{G}$ be such that $V_1(x) > 0$. Define*

$$K(x) := -\left\lfloor \ln\left(\frac{V_1(x)}{\hat{\alpha}_2(\omega_2(x))}\right)\right\rfloor + 1. \tag{6.7}$$

Then there exists $\hat{\phi} \in \mathcal{S}_\sigma(x)$ such that, for every $\kappa \geq K(x)$,

$$V_1(x) = \max_{k\in\{0,\ldots,\kappa\}} \hat{\alpha}_1\left(\omega_1(\hat{\phi}(k,x))\right)e^k. \tag{6.8}$$

Proof. It is obvious that

$$\sup_{k\in\{0,\ldots,\kappa\},\phi\in\mathcal{S}_\sigma(x)} \hat{\alpha}_1\left(\omega_1(\phi(k,x))\right)e^k \leq V_1(x). \tag{6.9}$$

We note that, for all $x \in \mathcal{G}\backslash\mathcal{A}$,

$$e^{-\kappa} \leq e^{-K(x)} \leq \frac{V_1(x)}{\hat{\alpha}_2(\omega_2(x))}e^{-1}.$$

Therefore, with (2.1) and (6.1), we may write

$$\begin{aligned} V_1(x) &= \max\left\{ \sup_{k\in\{0,\dots,\kappa\},\phi\in\mathcal{S}_\sigma(x)} \hat{\alpha}_1\left(\omega_1(\phi(k,x))\right)e^k, \right. \\ &\qquad\qquad \left. \sup_{k\in\mathbb{Z}_{\geq\kappa},\phi\in\mathcal{S}_\sigma(x)} \hat{\alpha}_1\left(\omega_1(\phi(k,x))\right)e^k \right\} \\ &\leq \max\left\{ \sup_{k\in\{0,\dots,\kappa\},\phi\in\mathcal{S}_\sigma(x)} \hat{\alpha}_1\left(\omega_1(\phi(k,x))\right)e^k, \quad \hat{\alpha}_2(\omega_2(x))e^{-\kappa} \right\} \\ &\leq \max\left\{ \sup_{k\in\{0,\dots,\kappa\},\phi\in\mathcal{S}_\sigma(x)} \hat{\alpha}_1\left(\omega_1(\phi(k,x))\right)e^k, \quad V_1(x)e^{-1} \right\}, \end{aligned}$$

which, together with (6.9), implies

$$\begin{aligned} V_1(x) &= \sup_{k\in\{0,\dots,\kappa\},\phi\in\mathcal{S}_\sigma(x)} \hat{\alpha}_1\left(\omega_1(\phi(k,x))\right)e^k \\ &= \sup_{\phi\in\mathcal{S}_\sigma(x)} \max_{k\in\{0,\dots,\kappa\}} \hat{\alpha}_1\left(\omega_1(\phi(k,x))\right)e^k, \end{aligned}$$

where we can pass to the "max" since the supremum is taken over a finite number of elements. Now let $\{\phi_\ell\}_{\ell=1}^\infty$ be a maximizing sequence of solutions in $\mathcal{S}_\sigma(x)$; i.e.,

$$V_1(x) = \lim_{\ell\to\infty} \max_{k\in\{0,\dots,\kappa\}} \hat{\alpha}_1\left(\omega_1(\phi_\ell(k,x))\right)e^k.$$

Since $F_\sigma(\cdot)$ satisfies the basic conditions (see Claim 6), we may appeal to Lemma 5.3 to see that a subsequence of $\{\phi_\ell(\cdot,x)\}_{\ell=1}^\infty$ converges uniformly on $\{0,\dots,\kappa\}$ to some solution $\hat{\phi}\in\mathcal{S}_\sigma(x)$. Since the functions $\hat{\alpha}_1(\cdot)$ and $\omega_1(\cdot)$ are continuous, we may write

$$V_1(x) = \max_{k\in\{0,\dots,\kappa\}} \hat{\alpha}_1\left(\omega_1(\hat{\phi}(k,x))\right)e^k. \qquad \square$$

We now prove that $V_1(\cdot)$ is upper semicontinuous. In order to obtain a contradiction, suppose that there exist $x\in\mathcal{G}$ and a sequence $\{x_\ell\}_{\ell=1}^\infty$ of points in $\mathcal{G}$ converging to $x\in\mathcal{G}$ such that

$$\limsup_{\ell\to\infty} V_1(x_\ell) > V_1(x) \geq 0.$$

Without loss of generality, for all ℓ and some $\eta>0$

$$V_1(x_\ell) \geq \eta. \tag{6.10}$$

Define $\kappa := \sup_\ell K(x_\ell)$. From (6.10), the continuity of $\hat{\alpha}_2\circ\omega_2(\cdot)$, and the definition of $K(\cdot)$ in (6.7), we see that $\kappa<\infty$. Let $\hat{\phi}\in\mathcal{S}_\sigma(x_\ell)$ come from Claim 7 so that

$$V_1(x_\ell) = \max_{k\in\{0,\dots,\kappa\}} \hat{\alpha}_1\left(\omega_1(\hat{\phi}(k,x_\ell))\right)e^k.$$

Let $\varepsilon>0$. Since $F_\sigma(\cdot)$ satisfies the basic conditions, we appeal to Lemma 5.1 with the triple (κ,ε,x) and the continuity of $\hat{\alpha}_1\circ\omega_1(\cdot)$ to assert the existence of ℓ_ε so that, for all $\ell\geq\ell_\varepsilon$, there exists $\psi_\ell\in\mathcal{S}_\sigma(x)$ such that

$$\begin{aligned} V_1(x_\ell) = \max_{k\in\{0,\dots,\kappa\}} \hat{\alpha}_1\left(\omega_1(\hat{\phi}(k,x_\ell))\right)e^k &\leq \varepsilon + \max_{k\in\{0,\dots,\kappa\}} \hat{\alpha}_1\left(\omega_1(\psi_\ell(k,x))\right)e^k \\ &\leq \varepsilon + V_1(x). \end{aligned}$$

This implies $\limsup_{\ell\to\infty} V_1(x_\ell) \leq V_1(x)$, which is a contradiction. In addition, it also establishes continuity of $V_1(\cdot)$ at each point $x \in \{\xi \in \mathcal{G} : V_1(\xi) = 0\}$ since, for each such x, we may write

$$0 \leq \limsup_{z\to x} V_1(z) \leq V_1(x) = 0.$$

Next we establish item 1 of Assumption 1.

CLAIM 8. *If $V_1(x) > 0$, then there exists $\delta > 0$ such that $|z - x| < \delta$ implies $V_1(z) > 0$.*

With this claim, we see that the set $\mathcal{G}\backslash\mathcal{A}$ is open.

Proof. If $x \in \mathcal{G}\backslash\mathcal{A}$ is such that $\omega_1(x) > 0$, then the result follows from the continuity of $\omega_1(\cdot)$, the lower bound (6.4), and the fact that $\hat{\alpha}_1 \in \mathcal{K}_\infty$. So we just need to consider points $x \in \mathcal{G}\backslash\mathcal{A}$ such that $\omega_1(x) = 0$. We first assert that

$$\sup_{f\in F(x)} V_1(f) > 0. \tag{6.11}$$

If this were not the case, then with the lower bound (6.4) and the fact that $\hat{\alpha}_1 \in \mathcal{K}_\infty$ we would have $\max_{f\in F(x)} \omega_1(f) = 0$. Furthermore, with the decrease condition (6.6) and $F(x) \subseteq F_\sigma(x)$ we would have, for all $f \in F(x)$,

$$\sup_{g\in F(f)} V_1(g) \leq e^{-1} V_1(f) = 0.$$

Iterating and using the condition $\omega_1(x) = 0$, we would have

$$\sup_{k\in\mathbb{Z}_{\geq 0}, \phi\in\mathcal{S}(x)} \omega_1(\phi(k, x)) = 0;$$

i.e., $x \in \mathcal{A}$, which is a contradiction.

We also have, according to the definition of robust $\mathcal{KL}$ stability, $\sigma(x) > 0$ for $x \in \mathcal{G}\backslash\mathcal{A}$. Using the continuity of $\sigma(\cdot)$, there exists $\delta > 0$ such that $\delta \leq \min_{q\in\delta\overline{\mathcal{B}}} \sigma(x + q)$. It follows that

$$0 \in \bigcap_{z\in\delta\overline{\mathcal{B}}} \{z\} + \sigma(x + z)\overline{\mathcal{B}}$$

and thus, for any $z \in \delta\overline{\mathcal{B}}$, we see that $F(x) \subseteq F(x + z + \sigma(x + z)\overline{\mathcal{B}})$. Now, using (6.6) and (6.11), we have, for $z \in \delta\overline{\mathcal{B}}$,

$$e^{-1} V_1(x + z) \geq \sup_{f\in F_\sigma(x+z)} V_1(f) \geq \sup_{f\in F(x)} V_1(f) > 0,$$

which establishes the claim. ∎

Finally, we need to construct a function $\sigma_2 : \mathcal{G}\backslash\mathcal{A} \to \mathbb{R}_{>0}$ satisfying Assumption 2 and such that the smooth function $V(\cdot)$ of Theorem 3.1 retains bounds like (6.4) and (6.5) and the decrease condition (6.6).

Construction of σ_2:

CLAIM 9. *There exists a smooth function $\sigma_1 : \mathcal{G}\backslash\mathcal{A} \to \mathbb{R}_{>0}$ such that, for all $x \in \mathcal{G}\backslash\mathcal{A}$,*

$$\sigma_1(x) \leq \sigma(x), \qquad \textit{and} \tag{6.12}$$

$$\sup_{f\in F_{\sigma_1}(x), f\in\mathcal{G}\backslash\mathcal{A}} V_1(f) \leq e^{-1} \inf_{z\in\sigma_1(x)\overline{\mathcal{B}}} V_1(x + z)\ . \tag{6.13}$$

Proof. We define a function $\sigma_1 : \mathcal{G}\backslash\mathcal{A} \to \mathbb{R}_{>0}$ by associating to each $x \in \mathcal{G}\backslash\mathcal{A}$ one-half the supremum over all values $\tilde{\sigma}_1$ satisfying

$$2\tilde{\sigma}_1 \leq \min_{q \in \tilde{\sigma}_1 \overline{\mathcal{B}}} \sigma(x+q). \tag{6.14}$$

The existence of $\sigma_1(\cdot)$ follows from continuity of $\sigma(\cdot)$ and the fact that $\sigma(x) > 0$ for all $x \in \mathcal{G}\backslash\mathcal{A}$. These properties for $\sigma(\cdot)$ also guarantee that $\sigma_1(\cdot)$ is bounded away from zero on compact subsets of $\mathcal{G}\backslash\mathcal{A}$.

It follows from (6.14) that

$$\sigma_1(x) \leq \sigma(x) \quad \forall x \in \mathcal{G}\backslash\mathcal{A}, \tag{6.15}$$

i.e., (6.12) holds. We further see that

$$\sigma_1(x)\overline{\mathcal{B}} \subseteq \bigcap_{z \in \sigma_1(x)\overline{\mathcal{B}}} \{z\} + \sigma(x+z)\overline{\mathcal{B}},$$

so that

$$F(x + \sigma_1(x)\overline{\mathcal{B}}) \subseteq \bigcap_{z \in \sigma_1(x)\overline{\mathcal{B}}} F(x + z + \sigma(x+z)\overline{\mathcal{B}}). \tag{6.16}$$

With the definition of $F_\sigma(\cdot)$, (6.15), and (6.16) we see that, for any $z \in \sigma_1(x)\overline{\mathcal{B}}$,

$$F_{\sigma_1}(x) \subseteq F_\sigma(x+z). \tag{6.17}$$

From the decrease condition (6.6), we have

$$e^{-1}V_1(x+z) \geq \sup_{f \in F_\sigma(x+z)} V_1(f).$$

Taking the infimum on both sides and appealing to (6.17) we have

$$\inf_{z \in \sigma_1(x)\overline{\mathcal{B}}} e^{-1}V_1(x+z) \geq \inf_{z \in \sigma_1(x)\overline{\mathcal{B}}} \left[\sup_{f \in F_{\sigma(x+z)}(x+z)} V_1(f) \right] \geq \sup_{f \in F_{\sigma_1(x)}(x)} V_1(f).$$

It is clear that this inequality and (6.15) hold for any function smaller than $\sigma_1(\cdot)$, and so we can smooth $\sigma_1(\cdot)$ using Lemma 3.3 to prove the claim. ∎

Let the function $\sigma_a : \mathcal{G}\backslash\mathcal{A} \to \mathbb{R}_{\geq 0}$ be given by

$$\sigma_a(x) := \min\left\{1, \frac{1}{2}\sup\left\{\eta \ : \ |z-x| \leq \eta \ \Rightarrow \ |\omega_2(z) - \omega_2(x)| \leq \frac{1}{2}\omega_2(x)\right\}\right\} \tag{6.18}$$

and the function $\sigma_b : \mathcal{G}\backslash\mathcal{A} \to \mathbb{R}_{\geq 0}$ be given by

$$\sigma_b(x) := \min\left\{1, \frac{1}{2}\sup\left\{\eta \ : \ V_1(x + \eta\overline{\mathcal{B}}) \geq \hat{\alpha}_1\left(\frac{1}{2}\omega_1(x)\right)\right\}\right\}. \tag{6.19}$$

We then define, for each $x \in \mathcal{G}\backslash\mathcal{A}$,

$$\bar{\sigma}_2(x) := \min\{\sigma_a(x), \sigma_b(x), \sigma_1(x), |x|_{\mathcal{A}}\}. \tag{6.20}$$

Before proceeding, we demonstrate that the function $\bar{\sigma}_2(\cdot)$ is bounded away from zero on compact subsets of $\mathcal{G}\backslash\mathcal{A}$. Since $\sigma_1(\cdot)$ and $|\cdot|_{\mathcal{A}}$ are continuous and positive on $\mathcal{G}\backslash\mathcal{A}$, we need to establish this property only for $\sigma_a(\cdot)$ and $\sigma_b(\cdot)$.

We first note that $\omega_2(x) > 0$ for $x \in \mathcal{G}\backslash\mathcal{A}$. Suppose not. If $x \notin \mathcal{A}$, then there exists a solution $\phi \in \mathcal{S}_\sigma(x)$ and a time $k \in \mathbb{Z}_{\geq 0}$ such that $\omega_1(\phi(k,x)) > 0$. However, from the stability estimate we see that $0 < \omega_1(\phi(k,x)) \leq \beta_\sigma(\omega_2(x), 0) = 0$, which is a contradiction.

First we demonstrate that $\sigma_a(\cdot)$ is bounded away from zero on compact subsets of $\mathcal{G}\backslash\mathcal{A}$. Suppose not. Then there exists a compact set $D \subset \mathcal{G}\backslash\mathcal{A}$, a sequence $\{x_i\}_{i=1}^{\infty}$ in D, and a sequence $z_i \in \{x_i\} + \sigma_a(x_i)\overline{\mathcal{B}} \subset \{x_i\} + \frac{1}{i}\overline{\mathcal{B}}$ such that

$$|\omega_2(z_i) - \omega_2(x_i)| > \frac{1}{2}\omega_2(x_i). \tag{6.21}$$

The sequence x_i has an accumulation point $x^* \in D$. Now, since $x^* \in D$, we have $\omega_2(x^*) > 0$. Since $\omega_2(\cdot)$ is continuous we have, as $i \to \infty$, $|\omega_2(z_i) - \omega_2(x_i)| \to 0$ while $\frac{1}{2}\omega_2(x_i) \to \frac{1}{2}\omega_2(x^*) > 0$. This contradicts (6.21).

Next we demonstrate that $\sigma_b(\cdot)$ is bounded away from zero on compact subsets of $\mathcal{G}\backslash\mathcal{A}$. Suppose not. Then there exist a compact set $D \subset \mathcal{G}\backslash\mathcal{A}$, a sequence $x_i \in D$, and a sequence z_i with $|x_i - z_i| \leq 1/i$ such that

$$V_1(z_i) < \hat{\alpha}_1\left(\frac{1}{2}\omega_1(x_i)\right). \tag{6.22}$$

The sequence x_i has an accumulation point $x^* \in D$. Henceforth we use x_i to denote the converging subsequence, and z_i the associated subsequence. Suppose $\omega_1(x^*) > 0$. Using the continuity of $\omega_1(\cdot)$, there exists i^* such that $\omega_1(z_i) \geq \frac{1}{2}\omega_1(x_i)$ for all $i \geq i^*$ and thus

$$V_1(z_i) \geq \hat{\alpha}_1\left(\omega_1(z_i)\right) \geq \hat{\alpha}_1\left(\frac{1}{2}\omega_1(x_i)\right),$$

which contradicts (6.22).

Alternatively, suppose $\omega_1(x^*) = 0$. We make the following claim.

Claim 10. *There exists $c > 0$ such that*

$$V_1(x^* + \sigma_1(x^*)\overline{\mathcal{B}}) \geq c. \tag{6.23}$$

Proof. Suppose not. Then, for all $c > 0$

$$\inf_{z \in \sigma_1(x^*)\overline{\mathcal{B}}} V_1(x^* + z) < c.$$

We note that this implies that if $f \in F(x^*)$, then $f \in \mathcal{A}$. Suppose not. Then, since $x^* \in D \subset \mathcal{G}\backslash\mathcal{A}$, (6.13) implies that $\sup_{f \in F(x^*)} V_1(f) = 0$. However, appealing to (6.3), we see that $V_1(f) = 0$ if and only if $f \in \mathcal{A}$.

From (6.6), we see that if $f \in \mathcal{A}$, then

$$\sup_{f_1 \in F(f)} V_1(f_1) \leq \sup_{f_1 \in F_\sigma(f)} V_1(f_1) \leq e^{-1} V_1(f) = 0.$$

In other words, any solution starting from a point $f \in F(x^*)$ is such that $V_1(\cdot)$ remains identically zero and, from (6.4), we see that $\omega_1(\cdot)$ also remains identically

zero. Furthermore, with $\omega_1(x^*) = 0$, it follows that any solution starting at x^* is such that $\omega_1(\cdot)$ remains identically zero so that $x^* \in \mathcal{A}$, which contradicts $x^* \in \mathcal{G}\backslash\mathcal{A}$. ∎

For sufficiently large i we have $z_i \in \{x^*\} + \sigma_1(x^*)\overline{\mathcal{B}}$ and, since $\omega_1(x^*) = 0$, we have $\hat{\alpha}_1\left(\frac{1}{2}\omega_1(x_i)\right) \leq c$ so that, with (6.23),

$$V(z_i) \geq \hat{\alpha}_1\left(\frac{1}{2}\omega_1(x_i)\right),$$

which contradicts (6.22).

Let the function $\sigma_2 : \mathcal{G}\backslash\mathcal{A} \to \mathbb{R}_{>0}$ come from the application of Lemma 3.3 to the function $\bar{\sigma}_2(\cdot)$ defined in (6.20) so that $\sigma_2(\cdot)$ is smooth and positive on $\mathcal{G}\backslash\mathcal{A}$. We see that $\sigma_2(\cdot)$ is such that, for a sequence of points $x_i \in \mathcal{G}\backslash\mathcal{A}$ such that $x_i \to x^* \in \mathcal{A}$, we have $\sigma_2(x_i) \to 0$, since $\sigma_2(x) \leq |x|_{\mathcal{A}}$. We also note that $x \in \mathcal{G}\backslash\mathcal{A}$ implies $\{x\} + \sigma_2(x)\overline{\mathcal{B}} \subset \mathcal{G}$. This follows from the fact that $x \in \mathcal{G}\backslash\mathcal{A}$ implies $\{x\} + \sigma_1(x)\overline{\mathcal{B}} \subset \mathcal{G}$ (which stems from the fact that $\sigma_2(x) \leq \sigma_1(x) \leq \sigma(x)$ and the definition of robust $\mathcal{KL}$ stability). Consequently, $\sigma_2(\cdot)$ satisfies Assumption 2.

Let $\psi : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ be smooth, vanish outside $\mathbb{R}^n\backslash\overline{\mathcal{B}}$, and satisfy $\int \psi(\xi)d\xi = 1$. For $x \in \mathcal{G}\backslash\mathcal{A}$ we define

$$V_s(x) := \int V_1(x + \sigma_2(x)\xi)\psi(\xi)\, d\xi.$$

For $x \in \mathcal{A}$ we let $V_s(x) = 0$. That $V_s(\cdot)$ is smooth on $\mathcal{G}\backslash\mathcal{A}$ and continuous on $\mathcal{G}$ follows from Theorem 3.1.

Using (6.18) and $\sigma_2(x) \leq \sigma_a(x)$ for all $x \in \mathcal{G}\backslash\mathcal{A}$, we see that

$$V_s(x) \leq \max_{z \in \{x\} + \sigma_2(x)\overline{\mathcal{B}}} \hat{\alpha}_2(\omega_2(z)) \leq \hat{\alpha}_2\left(\frac{3}{2}\omega_2(x)\right).$$

From (6.19) and $\sigma_2(x) \leq \sigma_b(x)$ for all $x \in \mathcal{G}\backslash\mathcal{A}$, we have

$$V_s(x) \geq \hat{\alpha}_1\left(\frac{1}{2}\omega_1(x)\right).$$

We now check that an appropriate decrease condition holds for $V_s(\cdot)$. Suppose $x \notin \mathcal{A}$ and $f \in F(x)$ is such that $f \in \mathcal{A}$. Then it is obvious that $V_s(f) \leq e^{-1}V_s(x)$. If $x \in \mathcal{A}$, by definition of $\mathcal{A}$ this implies that $f \in \mathcal{A}$ for all $f \in F(x)$. Therefore, $V_s(f) = 0 = e^{-1}V_s(x)$. It remains to check the decrease condition for $x, f \notin \mathcal{A}$.

Making use of (6.6), the result of Claim 9, and the fact that $\sigma_2(\cdot) \leq \sigma_1(\cdot)$, we may write

$$\begin{aligned}
\max_{f \in F(x)} V_s(f) &= \max_{f \in F(x)} \int V_1(f + \sigma_2(f)\xi)\psi(\xi)\, d\xi \;\leq\; \int \max_{f \in F_{\sigma_2}(x)} V_1(f)\psi(\xi)\, d\xi \\
&\leq e^{-1} \int \min_{z \in \sigma_2(x)\overline{\mathcal{B}}} V_1(x + z)\psi(\xi)\, d\xi \;\leq\; e^{-1} \int V_1(x + \sigma_2(x)\xi)\psi(\xi)\, d\xi \\
&= e^{-1}V_s(x). \qquad (6.24)
\end{aligned}$$

Let $\rho \in \mathcal{K}_\infty$ come from Lemma 3.2 and define $V(x) := \rho \circ V_s(x)$. Then we may write

$$\begin{aligned}
V(x) &\leq \rho \circ \hat{\alpha}_2\left(\frac{3}{2}\omega_2(x)\right) =: \alpha_2(\omega_2(x)), \qquad \text{and} \\
V(x) &\geq \rho \circ \hat{\alpha}_1\left(\frac{1}{2}\omega_1(x)\right) =: \alpha_1(\omega_1(x)).
\end{aligned}$$

Since $\rho \in \mathcal{K}_\infty$ is convex, $\rho(e^{-1}s) \leq e^{-1}\rho(s)$ for all $s \in \mathbb{R}_{\geq 0}$. Consequently, following (6.24), we may write

$$\begin{aligned}\max_{f\in F(x)} V(f) &= \rho\left(\max_{f\in F(x)} V_s(f)\right) \leq \rho(V_s(x)e^{-1}) \leq \rho(V_s(x))e^{-1} \\ &= V(x)e^{-1}. \qquad \square\end{aligned}$$

6.2. Necessity. We note that in order to demonstrate that robust $\mathcal{KL}$-stability follows from a smooth Lyapunov function, we actually only make use of a continuous Lyapunov function. Furthermore, the upper semicontinuity of the set-valued map $F(\cdot)$ is not used. This, then, is the result of Theorem 2.8 as well as the necessity of Theorem 2.7

We assume we have a continuous function $V : \mathcal{G} \to \mathbb{R}_{\geq 0}$ and functions $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ such that, for every $x \in \mathcal{G}$,

$$\alpha_1(\omega_1(x)) \leq V(x) \leq \alpha_2(\omega_2(x)), \tag{6.25}$$

$$\max_{f\in F(x)} V(f) \leq e^{-1}V(x), \tag{6.26}$$

and $V(x) = 0$ if and only if $x \in \mathcal{A}$. Since $V(\cdot)$ is continuous on $\mathcal{G}$ and bounded away from zero on compact subsets of $\mathcal{G}\backslash\mathcal{A}$, we see that $\mathcal{G}\backslash\mathcal{A}$ is open.

Let $\varepsilon > 0$ satisfy $(1+\varepsilon)^2 e^{-1} < 1$. Since $V(\cdot)$ is continuous, for each $x \in \mathcal{G}\backslash\mathcal{A}$ there exists $\tilde{\delta}_1 > 0$ such that for all $\xi \in \mathcal{G}$

$$|x - \xi| \leq \tilde{\delta}_1 \quad \Longrightarrow \quad |V(x) - V(\xi)| \leq \varepsilon V(x). \tag{6.27}$$

For each $x \in \mathcal{G}\backslash\mathcal{A}$ let $\delta_1(x)$ be one-half the supremum over all $\tilde{\delta}_1 \leq 1$ such that (6.27) holds and, for $x \in \mathcal{A}$, let $\delta_1(x) = 0$. Then $\delta_1(\cdot)$ is bounded away from zero on compact subsets of $\mathcal{G}\backslash\mathcal{A}$. Suppose not. Then there exists a compact set $D \subset \mathcal{G}\backslash\mathcal{A}$, a sequence of points $x_i \in D$, and an accumulation point $x^* \in D$ such that $\delta_1(x_i) \to 0$ as $x_i \to x^*$ and

$$|V(z_i) - V(x_i)| > \varepsilon V(x_i), \tag{6.28}$$

where $z_i \in \{x_i\} + \delta(x_i)\overline{\mathcal{B}}$. Since $\delta(x_i) \to 0$ we may pick a subsequence (which we do not relabel) such that $\delta(x_i) < \frac{1}{i}$, which implies that $|z_i - x_i| \leq \frac{1}{i}$. Since $x^* \in D$, $V(x^*) > 0$. With the continuity of $V(\cdot)$, as $i \to \infty$, $|V(z_i) - V(x_i)| \to 0$ while

$$V(x_i) \to V(x^*) > 0,$$

which contradicts (6.28).

For every $x \in \mathcal{G}$ let $\delta_2(x)$ be the supremum over all $\hat{\delta} \leq 1$ such that $\{x\} + 2\hat{\delta}\overline{\mathcal{B}} \subset \mathcal{G}$. Since $\mathcal{G}$ is open, $\hat{\delta}$ always exists and satisfies $\hat{\delta} > 0$. Moreover $\delta_2(\cdot)$ is bounded away from zero on compact subsets of $\mathcal{G}$. Suppose not. Then there exists $D \subset \mathcal{G}$ compact and a sequence $\{x_i\}_{i=1}^\infty$ such that $x_i \in D$ and the sequence has an accumulation point $x^* \in D$ such that $\delta_2(x_i) \to 0$, and for each i there exists $z_i \in \{x_i\} + \frac{1}{2}\delta_2(x_i)\overline{\mathcal{B}}$ so that $z_i \notin \mathcal{G}$. We pick a subsequence (without relabeling) such that $\delta(x_i) < \frac{1}{2i}$. Since $\mathcal{G}$ is open and $D \subset \mathcal{G}$ is compact, x^* is in the interior of $\mathcal{G}$, and consequently, for i sufficiently large, $\{x^*\} + \frac{1}{i}\overline{\mathcal{B}} \subset \mathcal{G}$. Therefore, again for i sufficiently large, we see that

$$z_i \in \{x_i\} + \frac{1}{2i}\overline{\mathcal{B}} \subset \{x^*\} + \frac{1}{i}\overline{\mathcal{B}},$$

which is a contradiction.

For each $x \in \mathcal{G}$ we define

$$\delta(x) := \min\{\delta_1(x), \delta_2(x), |x|_{\mathcal{A}}\}$$

and note that $x \in \mathcal{A}$ implies $\delta(x) = 0$. That $\delta(\cdot)$ is bounded away from zero on compact subsets of $\mathcal{G}\backslash\mathcal{A}$ follows from the fact that $\delta_2(x)$ is bounded away from zero on compact subsets of $\mathcal{G}$ and that $\delta_1(x)$ and $|x|_{\mathcal{A}}$ are bounded away from zero on compact subsets of $\mathcal{G}\backslash\mathcal{A}$.

Let the function $\sigma : \mathcal{G} \to \mathbb{R}_{\geq 0}$ come from the application of Lemma 3.3 to the function $\delta(\cdot)$ on $\mathcal{G}\backslash\mathcal{A}$ and define $\sigma(x) = 0$ for $x \in \mathcal{A}$. The restriction of $\sigma(\cdot)$ to $\mathcal{G}\backslash\mathcal{A}$ is smooth and, for all $x \in \mathcal{G}\backslash\mathcal{A}$, $\sigma(x) > 0$, satisfying item 2 in Definition 2.3. Since $\delta(x) \leq |x|_{\mathcal{A}}$ for all $x \in \mathcal{G}$, it follows that $\sigma(\cdot)$ is continuous on $\mathcal{G}$. Since $\delta(x) \leq \delta_2(x)$, $\{x\} + \sigma(x)\overline{\mathcal{B}} \subset \mathcal{G}$ for all $x \in \mathcal{G}$, satisfying item 1 of Definition 2.3. It remains to check items 3 and 4 of Definition 2.3.

Since $\sigma(x) \leq \delta_1(x)$ for all $x \in \mathcal{G}\backslash\mathcal{A}$, we may write

$$V(\xi + \sigma(\xi)\overline{\mathcal{B}}) \leq (1+\varepsilon)V(\xi)\ . \tag{6.29}$$

Then, using the definition of $F_\sigma(x)$ in (2.2), the bound (6.29), the decrease (6.26), and (6.27) we may write, for $x \in \mathcal{G}\backslash\mathcal{A}$,

$$\begin{aligned}\max_{\substack{f\in F_\sigma(x)\\ f\in\mathcal{G}\backslash\mathcal{A}}} V(f) \;&\leq\; (1+\varepsilon) \max_{\substack{f\in F(x+\sigma(x)\overline{\mathcal{B}})\\ f\in\mathcal{G}\backslash\mathcal{A}}} V(f) \;\leq\; (1+\varepsilon)e^{-1} \max_{z\in\{x\}+\sigma(x)\overline{\mathcal{B}}} V(z)\\ &\leq\; (1+\varepsilon)^2 e^{-1} V(x). \qquad (6.30)\end{aligned}$$

We can see that (6.30) holds for all $x \in \mathcal{G}$ by considering the two remaining cases. First, suppose $x \in \mathcal{A}$; then, by the definition of $\mathcal{A}$ and the fact that $\sigma(x) = 0$ for $x \in \mathcal{A}$, $f \in \mathcal{A}$, (6.30) is trivially satisfied. Second, suppose $x \notin \mathcal{A}$ and $f \in \mathcal{A}$, then (6.30) is again satisfied since $0 \leq (1+\varepsilon)^2 e^{-1} V(x)$. Let $\lambda := (1+\varepsilon)^2 e^{-1} < 1$. From (6.30), we see that, for any $x \in \mathcal{G}$ and $\phi \in \mathcal{S}_\sigma(x)$, we may write

$$V(\phi(k,x)) \leq \lambda^k V(x) \quad \forall k \in \mathbb{Z}_{\geq 0}.$$

If $x \in \mathcal{A}$, then $V(x) = 0$ and, using (6.25), we may write, for any $\phi \in \mathcal{S}_\sigma(x)$,

$$\alpha_1(\omega_1(\phi(k,x))) \leq V(\phi(k,x)) \leq V(x)\lambda^k = 0 \quad \forall k \in \mathbb{Z}_{\geq 0},$$

which implies that $x \in \mathcal{A}_\sigma$; i.e.,

$$\sup_{k\in\mathbb{Z}_{\geq 0}, \phi\in\mathcal{S}_\sigma(x)} \omega_1(\phi(k,x)) = 0 \quad \forall x \in \mathcal{A}.$$

Furthermore, if $x \in \mathcal{A}_\sigma$, then $x \in \mathcal{A}$ as a consequence of $\mathcal{S}(x) \subseteq \mathcal{S}_\sigma(x)$. Consequently, $\mathcal{A} = \mathcal{A}_\sigma$ and item 3 of Definition 2.3 is satisfied.

Now, using the upper and lower $\mathcal{K}_\infty$ bounds (6.25) on the Lyapunov function we may write, for all $x \in \mathcal{G}$, $\phi \in \mathcal{S}(x)$, and $k \in \mathbb{Z}_{\geq 0}$,

$$\alpha_1(\omega_1(\phi(k,x))) \leq V(\phi(k,x)) \leq \lambda^k V(x) \leq \lambda^k \alpha_2(\omega_2(x)).$$

Inverting $\alpha_1(\cdot)$ we obtain

$$\omega_1(\phi(k,x)) \leq \alpha_1^{-1}\left(\alpha_2(\omega_2(x))\lambda^k\right) =: \beta_\sigma(\omega_2(x), k);$$

i.e., $x^+ \in F_\sigma(x)$ is $\mathcal{KL}$-stable with respect to (ω_1, ω_2) on $\mathcal{G}$, satisfying item 4 of Definition 2.3. Therefore $x^+ \in F(x)$ is robustly $\mathcal{KL}$-stable with respect to (ω_1, ω_2) on $\mathcal{G}$. ∎

7. Proof of Theorem 2.10. If we can demonstrate that there exists a continuous Lyapunov function, the result of Theorem 2.8 yields that the $\mathcal{KL}$-stability is robust. Toward this end, we will define a (Lyapunov) function that is similar to (6.2), with the only difference being that the solution set under consideration in (6.2) is for the perturbed difference inclusion $x^+ \in F_\sigma(x)$. Here, however, we are not assuming robust $\mathcal{KL}$-stability. Rather, we are assuming $\mathcal{KL}$-stability of $x^+ \in F(x)$ and continuity of $F(\cdot)$ on $\mathcal{G}\backslash\mathcal{A}$.

In particular, we apply Lemma 6.1, with $\lambda = 2$, to the function $\beta \in \mathcal{KL}$ defining the stability estimate in order to obtain functions $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$ such that $\alpha_1(\beta(s,k)) \leq \alpha_2(s)e^{-2k}$ for all $(s,k) \in \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$. We define our Lyapunov function as

$$V(x) := \sup_{k \in \mathbb{Z}_{\geq 0}, \phi \in \mathcal{S}(x)} \alpha_1(\omega_1(\phi(k,x)))e^k \quad \forall x \in \mathcal{G}. \tag{7.1}$$

We can then obtain appropriate upper and lower bounds, the required decrease condition, upper semicontinuity of $V(\cdot)$ on $\mathcal{G}$, and continuity of $V(\cdot)$ on $\mathcal{A}$ by following the proof given in section 6.1. We note that the result of Claim 7 also holds. Therefore, in order to appeal to Theorem 2.8 it remains to show that $V(\cdot)$ as defined by (7.1) is lower semicontinuous on $\mathcal{G}\backslash\mathcal{A}$.

Lower semicontinuity of $V(\cdot)$ on $\mathcal{G}\backslash\mathcal{A}$: Let $x \in \mathcal{G}\backslash\mathcal{A}$. Appealing to Claim 7, there exists $\hat{\phi} \in \mathcal{S}(x)$ and $K(x) \in \mathbb{Z}_{\geq 0}$ such that

$$V(x) = \max_{k \in \{0,\ldots,K(x)\}} \alpha_1\left(\omega_1(\hat{\phi}(k,x))\right)e^k.$$

Let $\kappa \in \{0,\ldots,K(x)-1\}$ be the smallest integer such that $\hat{\phi}(\kappa+1,x) \in \mathcal{A}$ or, if $\hat{\phi}(k,x) \in \mathcal{G}\backslash\mathcal{A}$ for all $k \in \{0,\ldots,K(x)\}$, let $\kappa = K(x)$. We see that

$$\max_{k \in \{0,\ldots,K(x)\}} \alpha_1\left(\omega_1(\hat{\phi}(k,x))\right)e^k = \max_{k \in \{0,\ldots,\kappa\}} \alpha_1\left(\omega_1(\hat{\phi}(k,x))\right)e^k$$

since, if $K(x) \neq \kappa$, then $\hat{\phi}(k,x) \in \mathcal{A}$ for $k \in \{\kappa,\ldots,K(x)\}$, which implies that, for those k, $\omega_1(\hat{\phi}(k,x)) = 0$.

Since $F(\cdot)$ is continuous on $\mathcal{G}\backslash\mathcal{A}$ and since $\hat{\phi}(k,x) \in \mathcal{G}\backslash\mathcal{A}$ for all $k \in \{0,\ldots,\kappa\}$, given any $\varepsilon > 0$, Lemma 5.2 yields a $\delta > 0$ such that, for any $\bar{x} \in \{x\} + \delta\overline{\mathcal{B}}$, there exists a solution $\psi \in \mathcal{S}(\bar{x})$ such that we may write

$$\begin{aligned}
V(x) &= \max_{k \in \{0,\ldots,\kappa\}} \alpha_1\left(\omega_1(\hat{\phi}(k,x))\right)e^k \\
&\leq \max_{k \in \{0,\ldots,\kappa\}} \alpha_1\left(\omega_1(\psi(k,\bar{x}))\right)e^k \\
&\quad + \max_{k \in \{0,\ldots,\kappa\}} \left|\alpha_1\left(\omega_1(\hat{\phi}(k,x))\right) - \alpha_1\left(\omega_1(\psi(k,\bar{x}))\right)\right|e^k \\
&\leq \sup_{k \in \mathbb{Z}_{\geq 0}} \alpha_1\left(\omega_1(\psi(k,\bar{x}))\right)e^k + \varepsilon \\
&\leq V(\bar{x}) + \varepsilon.
\end{aligned}$$

Therefore, $V(\cdot)$ is lower semicontinuous at x; i.e., $\liminf_{z \to x} V(z) \geq V(x)$. ∎

8. Proof of Claim 1. In order to simplify the presentation, we define

$$V^+(x) := \sup_{f \in F(x)} V(f) \quad \forall x \in \mathcal{G}.$$

We will use W_1^+ and W^+ for the same purpose.

Let $g \in \mathcal{K}_\infty$ be such that $g'(\cdot)$ is nondecreasing, $g'(s) \geq 1$ for all $s \geq 0$, and such that there exists $\gamma \in \mathcal{K}_\infty$ such that

$$\alpha(s)\,(\exp(g(s)) - 1) \geq \gamma(s) \quad \forall s \geq 0.$$

We define $\rho(s) := \exp(g(s)) - 1$ and note that

$$\rho'(s) = (\rho(s) + 1)\, g'(s) \geq \rho(s)\ .$$

The equality shows that $\rho'(\cdot)$ is nondecreasing, so that, by the mean value theorem,

$$\rho(V^+(x)) - \rho(V(x)) \leq \rho'(V^+(x))[V^+(x) - V(x)] \quad \forall x \in \mathcal{G}.$$

For all $x \in \mathcal{G}$ we define $W_1(x) := \rho(V(x))$. It is obvious that $W_1(x) = 0$ if and only if $x \in \mathcal{A}$. For all $s \geq 0$ let $\mu(s) := \frac{1}{2}\min\left\{s, \gamma \circ \rho^{-1}(s)\right\}$ so that $\mu \in \mathcal{K}_\infty$. With this definition we observe that $(\mathrm{Id} - \mu) \in \mathcal{K}_\infty$. Now, for every $x \in \mathcal{G}$ either $\rho(V^+(x)) \leq \frac{1}{2}\rho(V(x))$ or $\rho(V^+(x)) \geq \frac{1}{2}\rho(V(x))$. In the first case

$$W_1^+(x) := \rho(V^+(x)) \leq \frac{1}{2}\rho(V(x)) = \frac{1}{2}W_1(x) \leq W_1(x) - \mu(W_1(x)), \tag{8.1}$$

while, in the latter case, we may write

$$\begin{aligned} W_1^+(x) - W_1(x) &= \rho(V^+(x)) - \rho(V(x)) \leq \rho'(V^+(x))\left[V^+(x) - V(x)\right] \\ &\leq -\rho'(V^+(x))\alpha(V(x)) \leq -\rho(V^+(x))\alpha(V(x)) \\ &\leq -\frac{1}{2}\rho(V(x))\alpha(V(x)) \leq -\frac{1}{2}\gamma(V(x)) \\ &= -\frac{1}{2}\gamma \circ \rho^{-1}(W_1(x)) \leq -\mu(W_1(x))\ . \end{aligned} \tag{8.2}$$

Combining (8.1) and (8.2) we have $W_1^+(x) \leq W_1(x) - \mu(W_1(x))$ for all $x \in \mathcal{G}$.

We require the following lemma, which appeared as [9, Lemma 2.4].

LEMMA 8.1. *If $\ell > 1$ and $\varphi \in \mathcal{K}_\infty$ satisfies $(\varphi - \mathrm{Id}) \in \mathcal{K}_\infty$, then there exists $\tilde{\alpha} \in \mathcal{K}_\infty$ such that $\tilde{\alpha} \circ \varphi(s) = \ell\tilde{\alpha}(s)$ for all $s \geq 0$.*

Define $\varphi \in \mathcal{K}_\infty$ as $\varphi(s) := (\mathrm{Id} - \mu)^{-1}(s)$ for all $s \geq 0$. We note that $\varphi(\cdot)$ is well defined by virtue of $(\mathrm{Id} - \mu) \in \mathcal{K}_\infty$. From the definition of $\varphi(\cdot)$ we see that, for all $s \geq 0$, $s - \varphi^{-1}(s) = \mu(s)$ or, equivalently, $\varphi(s) - s = \mu \circ \varphi(s)$. Therefore, $(\varphi - \mathrm{Id}) \in \mathcal{K}_\infty$. Let $\ell = e^1 > 1$ and let $\tilde{\alpha} \in \mathcal{K}_\infty$ come from Lemma 8.1. For all $x \in \mathcal{G}$ we define $W(x) := \tilde{\alpha}(W_1(x))$. We may then write

$$\begin{aligned} W^+(x) &= \tilde{\alpha}(W_1^+(x)) \leq \tilde{\alpha}\left(W_1(x) - \mu(W_1(x))\right) \\ &= \tilde{\alpha}(\varphi^{-1}(W_1(x))) = e^{-1}\tilde{\alpha}(W_1(x)) = e^{-1}W(x) \quad \forall x \in \mathcal{G}. \end{aligned}$$

Finally, we define the functions $\hat{\alpha}_1, \hat{\alpha}_2 \in \mathcal{K}_\infty$ by $\hat{\alpha}_1 := \tilde{\alpha} \circ \rho \circ \alpha_1$ and $\hat{\alpha}_2 := \tilde{\alpha} \circ \rho \circ \alpha_2$ so that (2.9) holds. ∎

REFERENCES

[1] R. P. AGARWAL, *Difference Equations and Inequalities: Theory, Methods, and Applications*, 2nd ed., Marcel Dekker, New York, 2000.

[2] J.-P. AUBIN AND A. CELLINA, *Differential Inclusions: Set-Valued Maps and Viability Theory*, Springer-Verlag, New York, 1984.

[3] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser Boston, Boston, 1990.
[4] F. H. Clarke, Y. S. Ledyaev, and R. J. Stern, *Asymptotic stability and smooth Lyapunov functions*, J. Differential Equations, 149 (1998), pp. 69–114.
[5] A. F. Filippov, *Differential Equations with Discontinuous Righthand Sides*, Kluwer Academic, Norwell, MA, 1988.
[6] G. Grimm, M. Messina, A. R. Teel, and S. Tuna, *Examples when model predictive control is nonrobust*, Automatica, submitted.
[7] F. C. Hoppensteadt, *Singular perturbations on the infinite interval*, Trans. Amer. Math. Soc., 123 (1966), pp. 521–535.
[8] Z. Jiang and Y. Wang, *A converse Lyapunov theorem for discrete-time systems with disturbances*, Systems Control Lett., 45 (2002), pp. 49–58.
[9] C. M. Kellett, *Advances in Converse and Control Lyapunov Functions*, Ph.D. thesis, University of California, Santa Barbara, 2002.
[10] C. M. Kellett and A. R. Teel, *Results on converse Lyapunov theorems for difference inclusions*, in Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, HI, 2003.
[11] C. M. Kellett and A. R. Teel, *Smooth Lyapunov functions and robustness of stability for difference inclusions*, Systems Control Lett., 52 (2004), pp. 395–405.
[12] M. Kisielewicz, *Differential Inclusions and Optimal Control*, Kluwer Academic, Dordrecht, The Netherlands, 1991.
[13] J. Kurzweil, *On the inversion of Ljapunov's second theorem on stability of motion*, Amer. Math. Soc. Transl., Ser. 2, 24 (1956), pp. 19–77.
[14] V. Lakshmikantham and L. Salvadori, *On Massera type converse theorem in terms of two different measures*, Boll. Unione Mat. Ital., 13 (1976), pp. 293–301.
[15] Y. Lin, E. D. Sontag, and Y. Wang, *A smooth converse Lyapunov theorem for robust stability*, SIAM J. Control Optim., 34 (1996), pp. 124–160.
[16] D. Mayne, J. Rawlings, C. Rao, and P. Scokaert, *Constrained model predictive control: Stability and optimality*, Automatica, 36 (2000), pp. 789–814.
[17] A. A. Movchan, *Stability of processes with respect to two metrics*, J. Appl. Math. Mech., 24 (1960), pp. 1475–1740 (translation of Prikl. Mat. Mekh.).
[18] D. Nešić, A. R. Teel, and P. V. Kokotović, *Sufficient conditions for stabilization of sampled-data nonlinear systems via discrete-time approximations*, Systems Control Lett., 38 (1999), pp. 259–270.
[19] E. D. Sontag, *Comments on integral variants of ISS*, Systems Control Lett., 34 (1998), pp. 93–100.
[20] E. D. Sontag and Y. Wang, *Lyapunov characterizations of input to output stability*, SIAM J. Control Optim., 39 (2001), pp. 226–249.
[21] A. Stuart and A. Humphries, *Dynamical Systems and Numerical Analysis*, Cambridge University Press, Cambridge, UK, 1996.
[22] A. R. Teel and L. Praly, *A smooth Lyapunov function from a class-$\mathcal{KL}$ estimate involving two positive semidefinite functions*, ESAIM Control Optim. Calc. Var., 5 (2000), pp. 313–367.
[23] F. W. Wilson, *Smoothing derivatives of functions and applications*, Trans. Amer. Math. Soc., 139 (1969), pp. 413–428.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 3, pp. 801–815

STATIONARY FILTER FOR CONTINUOUS-TIME MARKOVIAN JUMP LINEAR SYSTEMS*

MARCELO D. FRAGOSO† AND NEI C. S. ROCHA‡

Abstract. We derive a stationary filter for the best linear mean square filter (BLMSF) of continuous-time Markovian jump linear systems (MJLS). It amounts here to obtain the convergence of the error covariance matrix of the BLMSF to a stationary value under the assumption of mean square stability of the MJLS and ergodicity of the associated Markov chain θ_t. It is shown that there exists a unique solution for the stationary Riccati filter equation, and this solution is the limit of the error covariance matrix of the BLMSF. The advantage of this scheme is that it is easy to implement since the filter gain can be performed offline, leading to a linear time-invariant filter.

Key words. Kalman filter, Riccati equation, jump systems, Markov parameters

AMS subject classifications. 93E11, 93C05, 93C60, 60J75, 60J27

DOI. 10.1137/S0363012903436259

1. Introduction.

Markovian jump linear systems (MJLS) have been the subject of extensive research over the last few years, and the associated literature is now fairly extensive (see, e.g., [2], [4], [5], [10], [11], [12], [13], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [35], [36], [43], [44], [50], and references therein). Since its inception, this class of models has an intimate connection with systems which are vulnerable to abrupt changes in their structure, which includes, for instance, safety-critical and high-integrity systems (e.g., aircraft, chemical plants, nuclear power station, robotic manipulator systems, large scale flexible structures for space stations such as antenna, solar arrays, etc.). This, in turn, has led to applications in a variety of fields (see, e.g., [3], [12], [21], [22], [33], [34], [39], [42], [44], [1], [48], [49], [51], and references therein), which illustrate the breadth of possibilities of applications of MJLS. For instance, it is said in [48] that the results achieved by MJLS, when applied to the synthesis problem of wing deployment of an uncrewed air vehicle, were quite encouraging.

Filtering theory has been widely celebrated as a great achievement in stochastic systems theory and is of fundamental importance in application. The appearance of seminal papers such as [8], [9], [32], [40], and [53] gave an enormous impetus to the theory. Although the theoretical machinery available to deal with nonlinear estimation problems is by now rather considerable (see, e.g., [37] for an overview of the classical results and [15] for a nice introduction), there are yet many challenging questions in this area. One of these is the fact that the description of the optimal nonlinear filter can rarely be given in terms of a closed finite system of stochastic differential equations, i.e., the so-called finite filters; the exceptions are, for instance, the classical

*Received by the editors October 17, 2003; accepted for publication (in revised form) March 24, 2005; published electronically September 15, 2005. Research supported in part by the Brazilian National Research Council-CNPq grants 472920/03-0 and 302587/2004-7, by the Research Council of the State of Sâo Paulo-FAPESP grant 03/06736-7, by PRONEX grant 015/98, and IM-AGIMB.
http://www.siam.org/journals/sicon/44-3/43625.html

†National Laboratory for Scientific Computing—LNCC/CNPq, Av. Getulio Vargas 333, Petrópolis, Rio de Janeiro, CEP 25651-070, Brazil (frag@lncc.br).

‡Universidade Federal do Rio de Janeiro—UFRJ, Instituto de Matemática—IM, Av. Brig. Trompowski, s/n., Cidade Universitária, Ilha do Fundão, Rio de Janeiro, RJ, Brazil (rocha@im.ufrj.br).

Kalman filter, those described in [52], and Benes' class. More recently, the concept of estimation algebra, proposed initially in [7], has been used to enlarge this class in [55]. Unfortunately, this is what happens with the optimal nonlinear filter for our MJLS model (see, e.g., [45]). In view of this, the best linear mean square filter (BLMSF) for our MJLS model has been derived in [30]. This filter provides some of the desirable features of the Kalman filter.

With some simplifications in the MJLS model, finite filters have been derived. For instance, with the simplification that the observation process is not fed by the state, in [2] a finite filter is derived. Another finite-dimensional related paper is [53], but here the variable to be estimated is just the Markov chain (the state variable is assumed accessible). In the nonlinear setting, this problem has been studied in [56] for a small noise observation scenario, without requiring knowledge of the generator of the Markov chain, and asymptotic optimality is proved. See also, e.g., [19], [20], [23], [24], [51], and references therein, for some recent related results. For the case in which the associated Markov chain is assumed to be accessible, i.e., the operation mode is known for every $t \geq 0$, H_∞ filtering and the robust Kalman filtering for uncertain MJLS (including the time-lag case) has been studied, for instance, in [17], [18], [41], and [47].

In this paper, we make a further foray into the problem of linear estimation for the class of MJLS. The BLMSF derived in [30] is a function of the error covariance matrix whose dynamics is governed by two matrix differential equations: one associated with the second moment of the state variable and the other associated with the second moment of the estimator. Our aim here is to work out a certain matrix Riccati differential equation for the error covariance matrix of the BLMSF and show that the unique solution of this matrix Riccati differential equation converges to the unique solution of an algebraic Riccati equation, under the hypotheses of mean square stability of the system and ergodicity of the Markovian process θ_t. In the spirit of the Kalman filter theory, when using the algebraic Riccati equation, instead of the differential Riccati equation, for the BLMSF, we call this filter the stationary BLMSF. In addition to interest in its own right, it is a well-known fact that stationary filters have, *prima facie*, the desirable advantage of being easier to implement. Finally, the discrete-time counterpart of our problem was already contemplated in [13] and has inspired our work.

A brief outline of the content of this paper is as follows. In section 2, we fix the notation and recall a few notions and facts which can be found in [14]. The MJLS model is described in section 3. A Riccati-like differential equation for the BLMSF is derived in section 4. Finally, an asymptotic analysis is carried out in section 5.

2. Notation and preliminaries. We shall denote by $\mathbb{R}^n$ the n-dimensional Euclidean space and by $\mathbb{B}(\mathbb{R}^n, \mathbb{R}^m)$ the normed bounded linear space of all $m \times n$ matrices with $\mathbb{B}(\mathbb{R}^n) := \mathbb{B}(\mathbb{R}^n, \mathbb{R}^n)$. For $L \in \mathbb{B}(\mathbb{R}^n)$, L' will indicate the transpose of L. As usual, $L \geq 0\,(L > 0)$ will mean that the symmetric matrix $L \in \mathbb{B}(\mathbb{R}^n)$ is positive semidefinite (positive definite), respectively. In addition, we set $\mathbb{B}(\mathbb{R}^n)^+ := \{L \in \mathbb{B}(\mathbb{R}^n); L = L' \geq 0\}$. We use $\mathbb{R}^+$ to denote the interval $[0, \infty)$, and by $L \otimes K \in \mathbb{B}(\mathbb{R}^{sn}, \mathbb{R}^{rm})$ we mean the Kronecker product for any $L \in \mathbb{B}(\mathbb{R}^s, \mathbb{R}^r)$ and $K \in \mathbb{B}(\mathbb{R}^n, \mathbb{R}^m)$. We recall also that for $L \in \mathbb{B}(\mathbb{R}^n)$ and $K \in \mathbb{B}(\mathbb{R}^m)$ the Kronecker sum is defined as $L \oplus K := L \otimes I_m + I_n \otimes K \in \mathbb{B}(\mathbb{R}^{nm})$ (see, e.g., [6]). For $D_i \in \mathbb{B}(\mathbb{R}^n)$, $i = 1, \ldots, N$, $\mathrm{diag}(D_i)$ stands for an $Nn \times Nn$ matrix, where the matrices D_i are put together corner-to-corner diagonally, with all the other entries being zero, and $1_{\{\cdot\}}$ for the Dirac measure. In addition, we denote by $\mathrm{Re}\{\lambda_i(\mathcal{T})\}$ the real part of the eigenvalue

$\lambda_i(\mathcal{T})$ of the operator $\mathcal{T}$ and write generically $\mathrm{Re}\{\lambda(\mathcal{T})\} < 0$ if all its eigenvalues have real part less than zero. Furthermore, $\mathbb{H}^{n,m}$ represents the linear space composed of all sequences of N matrices $V = (V_1, \ldots, V_N)$ with $V_i \in \mathbb{B}(\mathbb{R}^n, \mathbb{R}^m)$, for $i = 1, \ldots, N$, and, for simplicity, we define $\mathbb{H}^n := \mathbb{H}^{n,n}$. Also, we shall define $\mathbb{H}^{n+} := \{V = (V_1, \ldots, V_N);$ $V_i \geq 0$ for $i = 1, \ldots, N\}$. Finally, we shall write, for $V = (V_1, \ldots, V_N) \in \mathbb{H}^n$ and $S = (S_1, \ldots, S_N) \in \mathbb{H}^n$, that $V \geq S$ if $V - S = (V_1 - S_1, \ldots, V_N - S_N) \in \mathbb{H}^{n^+}$, and that $V > S$ if $V_i - S_i > 0$ for $i = 1, \ldots, N$.

As usual, we define $\mathcal{H} := L_2(\Omega, \mathcal{F}, \mathbb{P})$, the Hilbert space of all square integrable r.v.'s in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, equipped with the inner product $\langle x, y\rangle = Ex'y$. Convergence here will be in the *quadratic mean* (q.m.) sense, i.e., a sequence $\{x(n)\}$ converges to x if $\|x(n) - x\|_2 \to 0$. We define also $\mathcal{H}_0 = \{x \in \mathcal{H} | Ex = 0\}$, the closed subspace of all centered r.v.'s of $\mathcal{H}$ and therefore a Hilbert space. In addition, $x \in \mathcal{H}$ and $y \in \mathcal{H}$ are said to be *orthogonal* (from now on $x \perp y$) if $\langle x, y\rangle = 0$.

For any stochastic process $y(t) \in \mathcal{H}_0$, we define $\mathcal{H}_t^y := \mathcal{L}\{y(s),\ 0 \leq s \leq t\}$, the space of all linear combinations $\sum_i \alpha_i y(t_i)$, where $t_i < t$ and q.m. limits of these combinations (a closed subspace) such that $\mathcal{H}_s^y \subset \mathcal{H}_t^y \subset \mathcal{H}^y := \mathcal{H}_\infty^y$ for $s \leq t$. We recall that if $\{y(t)\}$ is q.m. continuous, then $\mathcal{H}^y$ is a separable Hilbert space, and as a fundamental property of a Hilbert space, any $z \in \mathcal{H}_0$ has a unique decomposition (cf. [14, p. 45]), $z = \hat{z} + \tilde{z}$, where $\hat{z} = \mathfrak{P}_t^y z \in \mathcal{H}_t^y$ and $\tilde{z} \perp \mathcal{H}_t^y$, where $\mathfrak{P}_t^y$ denotes the *projection operator* which projects each element of $\mathcal{H}_0$ onto $\mathcal{H}_t^y$. Moreover, we have the following properties (cf. [14, p. 45]): (i) $\|z - \mathfrak{P}_t^y z\| = \min_{y \in \mathcal{H}_t^y} \|z - y\|$, and therefore $\hat{z} = \mathfrak{P}_t^y z$ is the linear minimum mean square error estimator of z given $\mathcal{H}_t^y$, i.e., the best linear estimator is the projection of z onto $\mathcal{H}_t^y$; (ii) $\tilde{z} = z - \hat{z} \perp \mathcal{H}_t^y$.

3. The model. Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \in \mathbb{R}^+}, \mathbb{P})$ be a stochastic basis (a complete probability space, carrying its natural filtration $\{\mathcal{F}_t\}_{t \in \mathbb{R}^+}$) on which all the processes in this work are defined. Let us consider the class of hybrid dynamical systems modeled by the following MJLS:

$$dx(t) = A_{\theta_t} x(t) dt + C_{\theta_t} dw_0(t), \quad x(0) = x, \quad t \in \mathbb{R}^+, \tag{3.1}$$

$$dy(t) = H_{\theta_t} x(t) dt + G_{\theta_t} dw(t), \quad y(0) = 0, \tag{3.2}$$

where $\{x(t)\}$ denotes the state vector in $\mathbb{R}^n$ (*signal process*), and $\{y(t)\}$ denotes the output process in $\mathbb{R}^m$, which generates the observational information that is available at time t, i.e., $\mathcal{H}_t^y$. Furthermore, we assume the following:

(A1) $\theta = \{(\theta_t,\ \mathcal{F}_t),\ t \in \mathbb{R}^+\}$ is a nonobserved homogeneous Markov process with right continuous trajectories and taking values on the finite set $\mathcal{S} := \{1, 2, \ldots, N\}$. We assume the following:

$$P(\theta_{t+h} = j \mid \theta_t = i) = \begin{cases} \lambda_{ij} h + o(h), & i \neq j, \\ 1 + \lambda_{ii} h + o(h), & i = j, \end{cases}$$

where $\Lambda := [(\lambda_{ij})]$ is the stationary $N \times N$ transition rate matrix of $\{\theta_t\}$ with $\lambda_{ij} > 0$, $i \neq j$, and $\lambda_i = -\lambda_{ii} = \sum_{j\,:\,j \neq i} \lambda_{ij} < \infty$, i.e., the process is supposed to be conservative (see, e.g., [38]). In addition, defining $p_{ij}(t) = P(\theta_{t+s} = j \mid \theta_s = i)$ and $p_i(t) = P(\theta_t = i)$ for $i, j = 1, \ldots, N$ and denoting $\mathbf{P}(t) := [p_{ij}(t)]$ and $\mathbf{p}(t) := (p_1(t), \ldots, p_N(t))'$, it is well known, under the hypothesis of the homogeneity of the process $\theta = \{(\theta_t,\ \mathcal{F}_t),\ t \in \mathbb{R}^+\}$, that $\frac{d}{dt}\mathbf{P}(t) = \mathbf{P}(t)\Lambda$, $\mathbf{P}(0) = I$, and $\frac{d}{dt}\mathbf{p}(t) = \Lambda'\mathbf{p}(t)$. Moreover, we assume that $\{(\theta_t, \mathcal{F}_t), t \in \mathbb{R}^+\}$ has initial distribution $\{v(i); i = 1, \ldots, N\}$.

(A2) $\mathcal{W}_0 = \{(w_0(t), \mathcal{F}_t),\ t \in \mathbb{R}^+\}$ and $\mathcal{W} = \{(w(t), \mathcal{F}_t),\ t \in \mathbb{R}^+\}$ are statistically mutually independent Wiener processes in $\mathbb{R}^r$ and $\mathbb{R}^p$, respectively.

(A3) $x1_{\{\theta_0=i\}}$, $i = 1, \dots, N$, are second order *r.v.*'s with $E[x1_{\{\theta_0=i\}}] = \mu_i$ and $E[xx'1_{\{\theta_0=i\}}] = V_i$.

(A4) $x(0)$ and $\{\theta_t\}$ are independent of $\{w_0(t)\}$ and $\{w(t)\}$.

(A5) $G_iG_i' > 0$ for $i = 1, \dots, N$.

In addition, notice that A_{θ_t}, C_{θ_t}, H_{θ_t}, and G_{θ_t} are random matrices such that, for $\theta_t = i$, $i \in \mathcal{S}$, they assume the values $A_i \in \mathbb{B}(\mathbb{R}^n)$, $C_i \in \mathbb{B}(\mathbb{R}^r, \mathbb{R}^n)$, $H_i \in \mathbb{B}(\mathbb{R}^n, \mathbb{R}^m)$, and $G_i \in \mathbb{B}(\mathbb{R}^p, \mathbb{R}^m)$, respectively.

4. The Riccati differential equation (RDE). In order to derive the stationary filter for (3.1)–(3.2), we need first to work out a Riccati differential equation for the main result concerning the best linear mean square filter obtained in [30]. First, define $z_i(t) := x(t)1_{\{\theta_t=i\}} \in \mathbb{R}^n$, $z(t) := (z_1(t)', \dots, z_N(t)')' \in \mathbb{R}^{Nn}$, $\hat{z}_i(t) := \mathfrak{P}_t^y z_i(t)$, $\hat{z}(t) := (\hat{z}_1(t)', \dots, \hat{z}_N(t)')'$, $\tilde{z}_i(t) := z_i(t) - \hat{z}_i(t)$, $\tilde{z}(t) := (\tilde{z}_1(t)', \dots, \tilde{z}_N(t)')' = z(t) - \hat{z}(t)$, and $\mathcal{P}(t) := E[\tilde{z}(t)\tilde{z}(t)']$. Furthermore, $Z(t) := \operatorname{diag}(Z_i(t))$ with $Z_i(t) := E[z_i(t)z_i(t)'] \in \mathbb{B}(\mathbb{R}^n)$, $\mathcal{Z}_t := (Z_1(t), \dots, Z_N(t)) \in \mathbb{H}^n$, and $\hat{Z}(t) := E[\widehat{z}(t)\widehat{z}(t)']$. With these definitions, we clearly have $x(t) = \sum_{i=1}^N z_i(t)$, $\hat{x}(t) := \mathfrak{P}_t^y x(t) = \sum_{i=1}^N \hat{z}_i(t)$, and $\tilde{x}(t) = x(t) - \hat{x}(t) = \sum_{i=1}^N [z_i(t) - \hat{z}_i(t)] = \sum_{i=1}^N \tilde{z}_i(t)$.

In addition, we shall be using the notation

$$\mathcal{A} := \Lambda' \otimes I_n + \operatorname{diag}(A_i) \quad \in \quad \mathbb{B}(\mathbb{R}^{Nn}), \tag{4.1}$$

$$H := [H_1, \dots, H_N] \quad \in \quad \mathbb{B}\big(\mathbb{R}^{Nn}, \mathbb{R}^m\big), \tag{4.2}$$

$$G_t^p := [\sqrt{p_1(t)}G_1, \dots, \sqrt{p_N(t)}G_N] \quad \in \quad \mathbb{B}\big(\mathbb{R}^{Np}, \mathbb{R}^m\big) \tag{4.3}$$

so that

$$0 < G_t^p G_t^{p\prime} = \sum_{j=1}^N G_j G_j' p_j(t) \in \mathbb{B}(\mathbb{R}^m).$$

Define also the innovations process $\{\nu(t)\}$ as

$$\nu(t) := y(t) - \int_0^t \hat{m}(s)\, ds,$$

where $\hat{m}(t) = \mathfrak{P}_t^y[H_{\theta_t}x(t)] = \sum_{i=1}^N H_i\hat{z}_i(t)$, or

$$d\nu(t) = dy(t) - H\hat{z}(t)\, dt.$$

THEOREM 4.1. *For system* (3.1)–(3.2)*, the best linear mean square estimator* $\hat{x}(t)$ *is given by the following filter:*

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

where

$$d\widehat{z}(t) = \mathcal{A}\widehat{z}(t)\, dt + \mathcal{P}(t)H'(G_t^pG_t^{p\prime})^{-1}d\nu(t), \tag{4.4}$$

$$\widehat{z}(0) = \mu := (\mu_1', \dots, \mu_N')', \tag{4.5}$$

with

$$\mathcal{P}(t) := Z(t) - \hat{Z}(t) \in \mathbb{B}(\mathbb{R}^{Nn})^+,$$

where

$$\dot{Z}_i(t) = A_i Z_i(t) + Z_i(t)A_i' + \sum_{j=1}^{N} Z_j(t)\lambda_{ji} + C_i C_i' p_i(t), \tag{4.6}$$

$$Z_i(0) = V_i \geq 0, \quad i = 1, \ldots, N, \tag{4.7}$$

and

$$\begin{aligned}\dot{\hat{Z}}(t) = {} & \overline{\mathcal{A}}\hat{Z}(t) + \hat{Z}(t)\overline{\mathcal{A}}' + \hat{Z}(t)H'(G_t^p G_t^{p\prime})^{-1}H\hat{Z}(t) \\ & + Z(t)H'(G_t^p G_t^{p\prime})^{-1}HZ(t)\end{aligned} \tag{4.8}$$

with $\hat{Z}(0) = \mu\mu'$, *and*

$$\overline{\mathcal{A}} = \mathcal{A} - Z(t)H'(G_t^p G_t^{p\prime})^{-1}H. \tag{4.9}$$

Proof. See Theorem 4.1 in [30]. □

Notice that in (4.4) the term $\mathcal{P}(t)$ is obtained as the difference of two terms $(Z(t) - \hat{Z}(t))$ which are derived via (4.6) and (4.8). Our first step then is to obtain a Riccati differential equation for $\mathcal{P}(t)$ as follows.

LEMMA 4.2. *$\mathcal{P}(t)$ satisfies the following matrix Riccati differential equation:*

$$\begin{aligned}\dot{\mathcal{P}}(t) &= \mathcal{A}\mathcal{P}(t) + \mathcal{P}(t)\mathcal{A}' - \mathcal{P}(t)H'(G_t^p G_t^{p\prime})^{-1}H\mathcal{P}(t) + \mathcal{C}_t\mathcal{C}_t' + \mathcal{V}(\mathcal{Z}_t), \\ \mathcal{P}(0) &= \mathcal{P}_0 \geq 0,\end{aligned} \tag{4.10}$$

with $\mathcal{C}_t := \operatorname{diag}(\sqrt{p_i(t)}C_i)$, $\mathcal{Z}_t := (Z_1(t), \ldots, Z_N(t))$, *where* $Z_i(t)$ *is solution of* (4.6), *and* $\mathcal{V}(\mathcal{Z}_t)$ *is defined by*

$$\begin{aligned}\mathcal{V}(\mathcal{Z}_t) := {} & \begin{bmatrix} \sum_{j=1}^N Z_j(t)\lambda_{j1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sum_{j=1}^N Z_j(t)\lambda_{jN} \end{bmatrix} \\ & - (\Lambda' \otimes I_n)Z(t) - Z(t)(\Lambda' \otimes I_n)'.\end{aligned} \tag{4.11}$$

Moreover, $\mathcal{V}(\mathcal{Z}_t) \geq 0$ *and is a linear operator.*

Proof. Let us first prove that $\mathcal{V}(.)$ is a linear operator and $\mathcal{V}(\mathcal{Z}_t) \geq 0$. For any $\mathcal{Q} = (Q_1, \ldots, Q_N) \in \mathbb{H}^{n+}$, $\mathcal{R} = (R_1, \ldots, R_N) \in \mathbb{H}^{n+}$, α and $\beta \in \mathbb{R}$, it is straightforward to show that $\mathcal{V}(\alpha\mathcal{Q} + \beta\mathcal{R}) = \alpha\mathcal{V}(\mathcal{Q}) + \beta\mathcal{V}(\mathcal{R})$. Now, defining $d\mathcal{D}_t = \left[x(t)'d(1_{\{\theta_t=1\}}) \cdots x(t)'d(1_{\{\theta_t=N\}})\right]' \in \mathbb{R}^{Nn}$, one can show that $E[d\mathcal{D}_t d\mathcal{D}_t'] = \mathcal{V}(\mathcal{Z}_t)\,dt$, by using Lemma 4.3 in [29]. Then, for any constant vector $v \in \mathbb{R}^{Nn}$,

$$\begin{aligned}v'\mathcal{V}(\mathcal{Z}_t)v\,dt &= v'E\left[d\mathcal{D}_t d\mathcal{D}_t'\right]v \\ &= E\left[\left\|d\mathcal{D}_t'v\right\|^2\right] \\ &\geq 0\end{aligned}$$

and therefore $\mathcal{V}(\mathcal{Z}_t) \geq 0$. Now, by the definition of $Z(t)$, we have

$$\begin{aligned}\dot{Z}(t) = {} & [\operatorname{diag}(A_i)]Z(t) + Z(t)[\operatorname{diag}(A_i)]' + \mathcal{C}_t\mathcal{C}_t' \\ & + \begin{bmatrix} \sum_{j=1}^N Z_j(t)\lambda_{j1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sum_{j=1}^N Z_j(t)\lambda_{jN}, \end{bmatrix},\end{aligned}$$

and bearing in mind the definition of $\mathcal{V}(\mathcal{Z}_t)$ and $\mathcal{A}$, we have

$$\dot{Z}(t) = \mathcal{A}Z(t) + Z(t)\mathcal{A}' + \mathcal{C}_t\mathcal{C}_t' + \mathcal{V}(\mathcal{Z}_t).$$

Now from (4.8) and (4.9), we have

$$\dot{\hat{Z}}(t) = \mathcal{A}\hat{Z}(t) + \hat{Z}(t)\mathcal{A}' + \big[Z(t) - \hat{Z}(t)\big]H'(G_t^pG_t^{p\prime})^{-1}H\big[Z(t) - \hat{Z}(t)\big].$$

Finally, bearing in mind that $\mathcal{P}(t) = Z(t) - \hat{Z}(t)$, we get

$$\dot{\mathcal{P}}(t) = \mathcal{A}\mathcal{P}(t) + \mathcal{P}(t)\mathcal{A}' - \mathcal{P}(t)H'(G_t^pG_t^{p\prime})^{-1}H\mathcal{P}(t) + \mathcal{C}_t\mathcal{C}_t' + \mathcal{V}(\mathcal{Z}_t),$$

which proves the lemma. □

In the next section we shall use the following definition of mean square stability.

DEFINITION 4.3. *A linear system with a Markovian jump parameter is mean square stable (MSS) if for any initial condition x_0 and initial distribution $\{v(i), i = 1, \dots, N\}$, there exist $q \in \mathbb{R}^n$ and $Q \in \mathbb{B}(\mathbb{R}^n)^+$ independent of x_0 such that*

(a) $\|E[x(t)] - q\| \longrightarrow 0$ *as* $t \longrightarrow \infty$;

(b) $\|E[x(t)x(t)'] - Q\| \longrightarrow 0$ *as* $t \longrightarrow \infty$.

Finally, we shall also need the following notation. Define the operators φ and $\hat{\varphi}$ in the following way: for $V = (V_1, \dots, V_N) \in \mathbb{H}^{n,m}$, with $V_i = (v_{i1}, \dots, v_{in})$, $v_{ij} \in \mathbb{R}^m$,

$$\varphi(V_i) = \begin{bmatrix} v_{i1} \\ \vdots \\ v_{in} \end{bmatrix} \in \mathbb{R}^{mn} \quad \text{and} \quad \hat{\varphi}(V) = \begin{bmatrix} \varphi(V_1) \\ \vdots \\ \varphi(V_N) \end{bmatrix} \in \mathbb{R}^{Nmn}.$$

Furthermore, for $v_i := \varphi(V_i)$ and $\mathfrak{V} := \hat{\varphi}(V)$, $i = 1, \dots, N$, we define

$$\hat{\varphi}^{-1}\begin{bmatrix} v_1 \\ \vdots \\ v_N \end{bmatrix} := \begin{bmatrix} \hat{\varphi}_1^{-1}(\mathfrak{V}) & \cdots & \hat{\varphi}_N^{-1}(\mathfrak{V}) \end{bmatrix} = \begin{bmatrix} \varphi^{-1}(v_1) & \cdots & \varphi^{-1}(v_N) \end{bmatrix}.$$

That is, φ is a mapping that stacks up the columns of a matrix from left to right and makes a long vector out of the matrix.

5. Asymptotic analysis of the RDE. In this section, we obtain our main result concerning the asymptotic behavior of the matrix Riccati differential equation (4.10). We shall show that, under the assumption of ergodicity of the process $\{\theta_t\}$ and of mean square stability of (3.1), the unique solution of the matrix Riccati differential equation (4.10) converges to the unique solution of an algebraic Riccati equation.

The main result reads as follows.

THEOREM 5.1. *Assume that system* (3.1) *is MSS, according to Definition* 4.3*, and $\{\theta_t\}$ is ergodic. Then for any $\mathcal{Z}_0 = (Z_1(0), \dots, Z_N(0)) \in \mathbb{H}^n$, with $Z_i(0) \geq 0$, $i = 1, \dots, N$, and $\mathcal{P}_0 \geq 0$ we have that $\mathcal{P}(t) \to \mathcal{P}$ exponentially fast with $\mathcal{P}$ the unique positive semidefinite solution of the algebraic Riccati equation (ARE):*

$$\mathcal{A}\mathcal{P} + \mathcal{P}\mathcal{A}' - \mathcal{P}H'(\mathcal{G}\mathcal{G}')^{-1}H\mathcal{P} + \mathcal{C}\mathcal{C}' + \mathcal{V}(\mathcal{Q}) = 0, \tag{5.1}$$

where $\mathcal{A} - \mathcal{P}H'(\mathcal{G}\mathcal{G}')^{-1}H$ *is a stable matrix,*

$$\mathcal{C} := \operatorname{diag}(\sqrt{\pi_i}C_i), \qquad \mathcal{G} := [\sqrt{\pi_1}G_1 \cdots \sqrt{\pi_N}G_N],$$

with $\{\pi_i; i = 1, \ldots, N\}$ *the limit distribution of* θ_t, $\mathcal{Q} := (Q_1, \ldots, Q_N)$, *with* Q_i *the limit of* $Z_i(t)$, *in the sense that* $\|Z_i(t) - Q_i\| \to 0$ *as* $t \to \infty$, *where* $Q_i = \hat{\varphi}_i^{-1}(-\mathcal{F}^{-1}\hat{\varphi}(\mathcal{R}_1, \ldots, \mathcal{R}_N))$ *with* $\mathcal{F} := \Lambda' \otimes I_{n^2} + \operatorname{diag}(A_i \oplus A_i)$ *and* $\mathcal{R}_i = C_i C_i' \pi_i$.

Proof. The idea of the proof runs as follows. First, notice that from Lemma 5.1 in [31], (4.10) has a unique positive semidefinite solution. We have to prove then, existence and uniqueness of a positive semidefinite solution, $\mathcal{P}$, for the algebraic Riccati equation (5.1). Now, proving that $\mathcal{P}(t) \to \mathcal{P}$ is tantamount to proving that there exist lower and upper bound functions $P_\star(t)$ and $P^\star(t)$ for $\mathcal{P}(t)$, i.e., $P_\star(t) \leq \mathcal{P}(t) \leq P^\star(t)$, and these functions squeeze asymptotically $\mathcal{P}(t)$ to $\mathcal{P}$.

From Proposition 5.7 of [29], if system (3.1) is MSS, then $\operatorname{Re}(\lambda(\mathcal{F})) < 0$, where $\mathcal{F} = \Lambda' \otimes I_{n^2} + \operatorname{diag}(A_i \oplus A_i)$. But from Proposition 4.3 of [29], if $\operatorname{Re}(\lambda(\mathcal{F})) < 0$, then $\operatorname{Re}(\lambda(\mathcal{A})) < 0$. Therefore, we conclude that matrix $\mathcal{A}$ in (4.10) is stable. Now, from Theorem 4.1 in [31] (see also Theorem 4.1 in [54]), it follows that there exists a unique positive semidefinite solution $\mathcal{P}$ to (5.1) and, furthermore, $\mathcal{A} - \mathcal{P}H'(\mathcal{G}\mathcal{G}')^{-1}H$ is stable.

Now, from Proposition 5.6 of [29], under the assumption of mean square stability of system (3.1), $\|Z_i(t) - Q_i\| \to 0$ as $t \to \infty$, where $Q_i = \hat{\varphi}_i^{-1}(-\mathcal{F}^{-1}\hat{\varphi}(\mathcal{R}_1, \ldots, \mathcal{R}_N))$ with $\mathcal{R}_i = C_i C_i' \pi_i$. In addition, notice that $\mathcal{V}(\mathcal{Z}_t) = \mathcal{V}(Z_1(t), \ldots, Z_N(t)) \to \mathcal{V}(Q_1, \ldots, Q_N) = \mathcal{V}(\mathcal{Q})$ as $t \to \infty$, with $\mathcal{V}(\mathcal{Z}_t)$ defined by (4.11), since $\mathcal{V}(\mathcal{Z}_t)$ is a linear bounded operator.

Finally, from Lemmas 6.1 and 6.4 in the appendix, there exist matrices $P_\star(t)$ and $P^\star(t)$ such that

$$P_\star(t) \leq \mathcal{P}(t) \leq P^\star(t)$$

and

$$\lim_{t\to\infty} P_\star(t) = \lim_{t\to\infty} P^\star(t) = \mathcal{P}$$

exponentially fast and, consequently, we have $\lim_{t\to\infty} \mathcal{P}(t) = \mathcal{P}$, which completes the proof. □

REMARK 1. *Bearing in mind that for the case in which there is no jump* ($N = 1$) *we have that* $\mathcal{V}(\mathcal{Q}) = 0$, *it is not difficult to show that in this case our filter reduces to the Kalman filter.*

REMARK 2. *Some preliminary simulations for the discrete-time case can be found in* [13]. *This includes some comparison with the IMM algorithm* (*the interacting multiple model algorithm*) *derived in* [4]. *However, for a better assessment of the full potentialities of our filter in applications, exhaustive simulations of adequate examples, and comparison with other approximation of the infinite-dimensional filter is required. For instance, it would be interesting to carry out exhaustive simulations in order to compare the nonstationary filter with the stationary one, and a certain PEM filter* (*polymorphic estimator filter*), *which can be found in* [51].

6. Appendix. We assume here the hypothesis of Theorem 5.1. In addition, we assume the results regarding existence and uniqueness of a positive semidefinite solution, $\mathcal{P}$, for the algebraic Riccati equation (5.1), including the fact that $\mathcal{A} - \mathcal{P}H'(\mathcal{G}\mathcal{G}')^{-1}H$ is stable.

In order to prove the following results we shall rewrite the matrix Riccati equation (4.10) in a more convenient way. Defining $T_t := \mathcal{P}(t)H'K_t^{-1}$ and $K_t := G_t^p G_t^{p\prime}$, (4.10) is given by

$$\dot{\mathcal{P}}(t) = (\mathcal{A} - T_t H)\,\mathcal{P}(t) + \mathcal{P}(t)\,(\mathcal{A} - T_t H)' + T_t K_t T_t' + \mathcal{C}_t\mathcal{C}_t' + \mathcal{V}(\mathcal{Z}_t),$$
$$\mathcal{P}(0) = E\big[\tilde{z}_0\tilde{z}_0'\big] \geq 0.$$

In addition, since the Markovian process $\{\theta_t\}$ is ergodic by the assumption, there exist limit probabilities $\{\pi_i;\ i = 1,\dots,N\}$, which do not depend upon the initial distribution, with $\sum_{i=1}^N \pi_i = 1$, satisfying the following inequalities:

$$\max_i |p_{ij}(t) - \pi_j| \leq \alpha e^{-\beta t} \quad \text{and} \quad \max_i |p_j(t) - \pi_j| \leq \alpha e^{-\beta t}$$

for some positive constants α and β; therefore, $p_{ij}(t) \to \pi_j$ and $p_j(t) \to \pi_j$, exponentially fast, as $t \to \infty$. Thus,

$$G_t^p = \Big[\sqrt{p_1(t)}G_1, \dots, \sqrt{p_N(t)}G_N\Big] \to [\sqrt{\pi_1}G_1, \dots, \sqrt{\pi_N}G_N] = \mathcal{G}$$

and

$$\mathcal{C}_t = \operatorname{diag}(\sqrt{p_i(t)}C_i) \to \operatorname{diag}(\sqrt{\pi_i}C_i) = \mathcal{C}.$$

So, as $t \to \infty$, we have $G_t^p G_t^{p\prime} \to \mathcal{G}\mathcal{G}'$ and $\mathcal{C}_t\mathcal{C}_t' + \mathcal{V}(\mathcal{Z}_t) \to \mathcal{C}\mathcal{C}' + \mathcal{V}(\mathcal{Q})$ exponentially fast.

LEMMA 6.1. *Let $P^\star(t)$ be the solution of the matrix differential equation given by*

$$\dot{P}^\star(t) = \mathfrak{A}P^\star(t) + P^\star(t)\mathfrak{A}' + T_\infty K_t T_\infty' + \mathcal{C}_t\mathcal{C}_t' + \mathcal{V}(\mathcal{Z}_t), \tag{6.1}$$
$$P^\star(0) = \mathcal{P}(0) = E\big[\tilde{z}_0\tilde{z}_0'\big] \geq 0,$$

where $\mathfrak{A} := \mathcal{A} - T_\infty H$, $T_\infty := \mathcal{P}H'K^{-1}$ with $\mathcal{P}$ a solution of (5.1), and $K := \mathcal{G}\mathcal{G}'$ with $\mathcal{G} := [\sqrt{\pi_1}G_1, \dots, \sqrt{\pi_N}G_N]$. Then $P^\star(t) \geq \mathcal{P}(t)$ for any $t \in [0,\infty)$ and $\lim_{t\to\infty} P^\star(t) = \mathcal{P}$.

Proof. Define $\tilde{P}^\star(t) := P^\star(t) - \mathcal{P}(t)$. Then

$$\dot{\tilde{P}}^\star(t) = \mathfrak{A}\tilde{P}^\star(t) + \tilde{P}^\star(t)\mathfrak{A}' + (T_t - T_\infty)\,K_t\,(T_t - T_\infty)', \tag{6.2}$$
$$\tilde{P}^\star(0) = 0.$$

Let $\Phi^\star(t,s)$ be the transition matrix associated with $\mathfrak{A}$, i.e., $\Phi^\star(t,s) = e^{\mathfrak{A}(t-s)}$. Then, the solution of (6.2) is given by

$$\tilde{P}^\star(t) = \int_0^t \Phi^\star(t,s)\,(T_s - T_\infty)\,K_s\,(T_s - T_\infty)'\,\Phi^\star(t,s)'\,ds.$$

Since $K_t > 0$ for all $t \in [0,\infty)$, we have $\tilde{P}^\star(t) \geq 0$ and, consequently, $P^\star(t) \geq \mathcal{P}(t)$. Now, let us consider the solution of (6.1). Defining $\mathcal{W}(t) := T_\infty K_t T_\infty' + \mathcal{C}_t\mathcal{C}_t' + \mathcal{V}(\mathcal{Z}_t)$, we get from (6.1)

$$P^\star(t) = \Phi^\star(t,0)\mathcal{P}(0)\Phi^\star(t,0)' + \int_0^t \Phi^\star(t,s)\mathcal{W}(s)\Phi^\star(t,s)'\,ds. \tag{6.3}$$

Since $\mathfrak{A}$ is stable, $\Phi^\star(t,0) = e^{\mathfrak{A}t} \to 0$ as $t \to \infty$. Defining

$$\mathcal{I}(t) := \int_0^t \Phi^\star(t,s)\mathcal{W}(s)\Phi^\star(t,s)'\,ds,$$

it is clear that $\lim_{t\to\infty} P^\star(t)$ exists if and only if $\lim_{t\to\infty}\mathcal{I}(t)$ exists. Now we shall show that this limit does exist. First observe that

$$\begin{aligned}\frac{\partial}{\partial s}\left[\Phi^\star(t,s)\mathcal{W}(s)\Phi^\star(t,s)'\right] &= -\mathfrak{A}\Phi^\star(t,s)\mathcal{W}(s)\Phi^\star(t,s)' \\ &\quad + \Phi^\star(t,s)\dot{\mathcal{W}}(s)\Phi^\star(t,s)' \\ &\quad - \Phi^\star(t,s)\mathcal{W}(s)\Phi^\star(t,s)'\mathfrak{A}'.\end{aligned} \tag{6.4}$$

Integrating both sides of (6.4), we have

$$\mathfrak{A}\mathcal{I}(t) + \mathcal{I}(t)\mathfrak{A}' = \Phi^\star(t,0)\mathcal{W}(0)\Phi^\star(t,0)' + \int_0^t \Phi^\star(t,s)\dot{\mathcal{W}}(s)\Phi^\star(t,s)'\,ds - \mathcal{W}(t). \tag{6.5}$$

Since $\Phi^\star(t,0)\to 0$ and $\mathcal{W}(t)\to T_\infty K T_\infty' + \mathcal{C}\mathcal{C}' + \mathcal{V}(\mathcal{Q})$ as $t\to\infty$, it follows that $\lim_{t\to\infty}\mathcal{I}(t)$ exists if and only if $\lim_{t\to\infty}\int_0^t \Phi^\star(t,s)\mathcal{W}(s)\Phi^\star(t,s)'\,ds$ exists. To show that this limit does exist, first observe that

$$\begin{aligned}\left\|\int_0^t e^{\mathfrak{A}(t-s)}\dot{\mathcal{W}}(s)e^{\mathfrak{A}'(t-s)}\,ds\right\| &\le \mu^2\int_0^t e^{-2\alpha(t-s)}\|\dot{\mathcal{W}}(s)\|\,ds \\ &\le \mu^2\int_0^\infty e^{-2\alpha(t-s)}\|\dot{\mathcal{W}}(s)\|ds.\end{aligned}$$

Define $f(t) := e^{-2\alpha t}$ and $g(t) := \|\dot{\mathcal{W}}(t)\|$. Then

$$\begin{aligned}\left\|\int_0^t e^{\mathfrak{A}(t-s)}\dot{\mathcal{W}}(s)e^{\mathfrak{A}'(t-s)}\,ds\right\| &\le \mu^2\int_0^\infty f(t-s)g(s)\,ds, \\ &= \mu^2(f*g)(t),\end{aligned}$$

where $(f*g)(t)$ stands for the convolution of f and g.

Clearly $f(t)\in L_1(\mathbb{R}^+)$. Let us now show that $g(t)\in L_1(\mathbb{R}^+)$,

$$\|\dot{\mathcal{W}}(t)\| \le \|T_\infty\|^2\left\|\frac{d}{dt}K_t\right\| + \left\|\frac{d}{dt}\left[\mathcal{C}_t\mathcal{C}_t'\right]\right\| + \left\|\frac{d}{dt}\mathcal{V}(\mathcal{Z}_t)\right\|.$$

We are going to show that $\|\frac{d}{dt}K_t\|$, $\|\frac{d}{dt}[\mathcal{C}_t\mathcal{C}_t']\|$, and $\|\frac{d}{dt}\mathcal{V}(\mathcal{Z}_t)\|$ belong to $L_1(\mathbb{R}^+)$, and, consequently, $\|\frac{d}{dt}\mathcal{W}(t)\|\in L_1(\mathbb{R}^+)$.

It is straightforward to show that $|\dot{p}_j(t)|\le \max_j|p_j(t)-\pi_j|\sum_{i=1}^N|\lambda_{ij}|$, and since $\max_j|p_j(t)-\pi_j|\le \alpha e^{-\beta t}$ for some constants $\alpha>0$ and $\beta>0$, we have that $|\dot{p}_j(t)|\in L_1(\mathbb{R}^+)$.

Now $\|\frac{d}{dt}K_t\| \le \sum_{j=1}^N \|G_jG_j'\||\dot{p}_j(t)|$ and $\|\frac{d}{dt}(\mathcal{C}_t\mathcal{C}_t')\| = \|\mathrm{diag}(C_iC_i'\dot{p}_i(t))\| \le \|\mathrm{diag}(\|C_iC_i'\||\dot{p}_i(t)|)\| = \max_{1\le i\le N}\{\|C_iC_i'\||\dot{p}_i(t)|\}$, so $\|\frac{d}{dt}K_t\|\in L_1(\mathbb{R}^+)$ and $\|\frac{d}{dt}(\mathcal{C}_t\mathcal{C}_t')\|\in L_1(\mathbb{R}^+)$. As for $\|\frac{d}{dt}\mathcal{V}(\mathcal{Z}_t)\|$, we have

$$\begin{aligned}\left\|\frac{d}{dt}\mathcal{V}(\mathcal{Z}_t)\right\| &\le \left\|\mathrm{diag}\left(\sum_{l=1}^N \dot{Z}_l(t)\lambda_{li}\right)\right\| + \|(\Lambda'\otimes I_n)\mathrm{diag}(\dot{Z}_i(t))\| \\ &\quad + \|\mathrm{diag}(\dot{Z}_i(t))(\Lambda'\otimes I_n)'\| \\ &\le \left\|\mathrm{diag}\left(\sum_{l=1}^N \|\dot{Z}_l(t)\|\,|\lambda_{li}|\right)\right\| + 2\,\|(\Lambda'\otimes I_n)\|\,\|\mathrm{diag}(\|\dot{Z}_i(t)\|)\| \\ &\le \max_i\left\{\sum_{l=1}^N \|\dot{Z}_l(t)\|\,|\lambda_{li}|\right\} + 2\,\|(\Lambda'\otimes I_n)\|\max_i\{\|\dot{Z}_i(t)\|\}.\end{aligned}$$

Define $|\lambda|_{\max} := \max_{i,j}\{|\lambda_{ij}|\}$. Then $\|\frac{d}{dt}\mathcal{V}(\mathcal{Z}_t)\| \leq \gamma \max_i\{\|\dot{Z}_i(t)\|\}$ with $\gamma := N|\lambda|_{\max} + 2\|(\Lambda' \otimes I_n)\| > 0$. So, it suffices to prove that $\|Z_i(t)\| \in L_1(\mathbb{R}^+)$ to have that $\|\frac{d}{dt}\mathcal{V}(\mathcal{Z}_t)\| \in L_1(\mathbb{R}^+)$.

But from Proposition 5.7 in [29], we have

$$\hat{\varphi}(\dot{\mathcal{Z}}_t) = \mathcal{F}\hat{\varphi}(\mathcal{Z}_t) + \hat{\varphi}(R(t)), \tag{6.6}$$

where $R(t) := [R_1(t), \ldots, R_N(t)]$ with $R_i(t) := C_i C_i' p_i(t)$. Now taking the derivative on both sides of (6.6), we have

$$\frac{d}{dt}\hat{\varphi}(\dot{\mathcal{Z}}_t) = \mathcal{F}\hat{\varphi}(\dot{\mathcal{Z}}_t) + \hat{\varphi}(\dot{R}(t)).$$

Now define $u(t) := \hat{\varphi}(\dot{\mathcal{Z}}_t)$, so

$$\dot{u}(t) = \mathcal{F}u(t) + \hat{\varphi}(\dot{R}(t)),$$

whose solution is

$$u(t) = e^{\mathcal{F}t}u(0) + \int_0^t e^{\mathcal{F}(t-s)}\hat{\varphi}(\dot{R}(s))\, ds.$$

Thus

$$\begin{aligned}\|u(t)\| &\leq \|e^{\mathcal{F}t}\|\,\|u(0)\| + \int_0^t \|e^{\mathcal{F}(t-s)}\|\|\hat{\varphi}(\dot{R}(s))\|\, ds \\ &\leq \|e^{\mathcal{F}t}\|\,\|u(0)\| + \int_0^\infty \|e^{\mathcal{F}(t-s)}\|\|\hat{\varphi}(\dot{R}(s))\|\, ds.\end{aligned}$$

Since $\mathcal{F}$ is stable, we have $\|e^{\mathcal{F}t}\| \leq \alpha_1 e^{-\beta_1 t}$ with $\alpha_1 > 0$ and $\beta_1 > 0$. Also, $\|\hat{\varphi}(\dot{R}(t))\| \in L_1(\mathbb{R}^+)$, because $\|\hat{\varphi}(\dot{R}(t))\| = (\sum_{i=1}^N k_i(\dot{p}_i(t))^2)^{\frac{1}{2}} \leq \max_j |\dot{p}_i(t)|(\sum_{i=1}^N k_i)^{\frac{1}{2}}$ for some constants $k_i > 0$. Therefore,

$$\|u(t)\| \leq \alpha_1\,\|u(0)\|\, e^{-\beta_1 t} + \alpha_1 \int_0^\infty e^{-\beta_1(t-s)}\|\hat{\varphi}(\dot{R}(s))\|\, ds. \tag{6.7}$$

Defining $h_1(t) := e^{-\beta_1 t} \in L_1(\mathbb{R}^+)$ and $h_2(t) := \|\hat{\varphi}(\dot{R}(t))\| \in L_1(\mathbb{R}^+)$ and bearing in mind that the integral on the right-hand side of (6.7) is $(h_1 * h_2)(t) \in L_1(\mathbb{R}^+)$ (cf. [46, Theorem 8.14, p. 170]) we show that $\|\hat{\varphi}[\dot{Z}_1(t), \ldots, \dot{Z}_N(t)]\| \in L_1(\mathbb{R}^+)$. Finally, we have $\|\dot{Z}_i(t)\| \in L_1(\mathbb{R}^+)$, for all i, and $\max_i\{\|\dot{Z}_i(t)\|\} \in L_1(\mathbb{R}^+)$ and, consequently, $\|\frac{d}{dt}\mathcal{V}(\mathcal{Z}_t)\| \in L_1(\mathbb{R}^+)$, which proves that $\|\frac{d}{dt}\mathcal{W}(t)\| \in L_1(\mathbb{R}^+)$.

Now, since $f(t) \in L_1(\mathbb{R}^+)$ and $g(t) \in L_1(\mathbb{R}^+)$, we have that $(f * g)(t) \in L_1(\mathbb{R}^+)$ (cf. [46, Theorem 8.14, p. 170]), and so

$$\lim_{t\to\infty}\left\|\int_0^t e^{\mathfrak{A}(t-s)}\dot{\mathcal{W}}(s)e^{\mathfrak{A}'(t-s)}\, ds\right\| = 0,$$

that is,

$$\int_0^t e^{\mathfrak{A}(t-s)}\dot{\mathcal{W}}(s)e^{\mathfrak{A}'(t-s)}\, ds \to 0 \text{ as } t \to \infty$$

which proves that $\lim_{t\to\infty}\mathcal{I}(t)$ exists. Defining $\lim_{t\to\infty}\mathcal{I}(t) = \lim_{t\to\infty}P^\star(t) := \overline{P}^\star$ and taking limits on both sides of (6.5), we obtain

$$\mathfrak{A}\overline{P}^\star + \overline{P}^\star\mathfrak{A}' = \lim_{t\to\infty}\int_0^t e^{\mathfrak{A}(t-s)}\dot{\mathcal{W}}(s)e^{\mathfrak{A}'(t-s)}ds - \left[T_\infty K T_\infty' + \mathcal{C}\mathcal{C}' + \mathcal{V}(\mathcal{Q})\right]. \tag{6.8}$$

Finally, (6.8) becomes

$$(\mathcal{A} - T_\infty H)\,\overline{P}^\star + \overline{P}^\star\,(\mathcal{A} - T_\infty H)' + T_\infty K T_\infty' + \mathcal{C}\mathcal{C}' + \mathcal{V}(\mathcal{Q}) = 0. \tag{6.9}$$

Notice that $\mathcal{P}$ is also a solution of (6.9) because, replacing $\overline{P}^\star$ by $\mathcal{P}$ in (6.9), and taking into account that $T_\infty = \mathcal{P}H'K^{-1}$, we have

$$\mathcal{A}\mathcal{P} + \mathcal{P}\mathcal{A}' - \mathcal{P}H'\left(\mathcal{G}\mathcal{G}'\right)^{-1} H\mathcal{P} + \mathcal{C}\mathcal{C}' + \mathcal{V}(\mathcal{Q}) = 0,$$

which is (5.1).

Since $\mathcal{A} - T_\infty H$ is stable, the algebraic Riccati equation (6.9) admits a unique solution. Therefore, we must have $\overline{P}^\star = \mathcal{P}$. In short, we have $\lim_{t\to\infty} P^\star(t) = \mathcal{P}$ and $P^\star(t) \geq \mathcal{P}(t)$. □

LEMMA 6.2. *Let $Z_i(t)$ and $\overline{Z_i}(t)$ be solutions of matrix differential equations given, respectively, by*

$$\begin{aligned} \dot{Z}_i(t) &= A_i Z_i(t) + Z_i(t)A_i' + \sum_{j=1}^{N} Z_j(t)\lambda_{ji} + C_i C_i' p_i(t), \\ Z_i(0) &= V_i \geq 0 \end{aligned} \tag{6.10}$$

and

$$\begin{aligned} \dot{\bar{Z}}_i(t) &= A_i \bar{Z}_i(t) + \bar{Z}_i(t)A_i' + \sum_{j=1}^{N} \bar{Z}_j(t)\lambda_{ji} + C_i C_i' \alpha_i(t), \\ \bar{Z}_i(0) &= V_i \geq 0, \end{aligned} \tag{6.11}$$

where $\alpha_i(t) = \inf_{s\in[0,\infty)}\{p_i(t+s)\}$. *Then, for all $t > 0$, we have $Z_i(t) \geq \bar{Z}_i(t)$.*

Proof. First observe that $p_i(t) \geq \alpha_i(t)$ and that for $0 \leq s \leq t$, $\alpha_i(t) \geq \alpha_i(s)$, i.e., $\alpha_i(t)$ is a nondecreasing function of t. Moreover, $\alpha_i(t) \to \pi_i$ exponentially fast, as $t \to \infty$, because $p_i(t) \to \pi_i$ exponentially fast. Next, from Lemma 4.1 in [29], under positive semidefinite initial conditions, (6.10) and (6.11) admit positive semidefinite solution for all t, which can be obtained by successive approximations. Then $Z_i(t) \geq 0$ and $\bar{Z}_i(t) \geq 0$ for all t.

To see this, notice that (6.11) can be rewritten as

$$\begin{aligned} \dot{\bar{Z}}_i(t) &= \left(A_i + \frac{1}{2}\lambda_{ii} I\right)\bar{Z}_i(t) + \bar{Z}_i(t)\left(A_i + \frac{1}{2}\lambda_{ii} I\right)' \\ &\quad + \sum_{j=1,\ j\neq i}^{N} \bar{Z}_j(t)\lambda_{ji} + C_i C_i' \alpha_i(t), \\ \bar{Z}_i(0) &= V_i \geq 0. \end{aligned}$$

Now let $\overline{\Phi}_i(t,s)$ be the transition matrix associated with $A_i + \frac{1}{2}\lambda_{ii} I$, i.e., $\overline{\Phi}_i(t,s) = e^{(A_i + \frac{1}{2}\lambda_{ii} I)(t-s)}$. Then the solution $\bar{Z}_i(t)$ will be given by

$$\begin{aligned} \bar{Z}_i(t) &= \overline{\Phi}_i(t,0) V_i \overline{\Phi}_i'(t,0) \\ &\quad + \int_0^t \overline{\Phi}_i(t,s)\left[\sum_{j=1,\ j\neq i}^{N} \bar{Z}_j(s)\lambda_{ji} + C_i C_i' \alpha_i(s)\right]\overline{\Phi}_i'(t,s)\,ds. \end{aligned}$$

Since $\alpha_i(t) \geq 0$ for all t, and since $\bar{Z}_j(t)\lambda_{ji} \geq 0$ for all t and $i \neq j$, we have that

$$\sum_{j=1,\ j\neq i}^{N} \bar{Z}_j(u)\lambda_{ji} + C_i C_i' \alpha_i(u) \geq 0 \quad \text{for all } u \geq 0.$$

We shall now show that $Z_i(t) \geq \bar{Z}_i(t)$. (6.10) can be rewritten as

$$\begin{aligned}\dot{Z}_i(t) &= \left(A_i + \frac{1}{2}\lambda_{ii}I\right) Z_i(t) + Z_i(t)\left(A_i + \frac{1}{2}\lambda_{ii}I\right)' \\ &\quad + \sum_{j=1,\ j\neq i}^{N} Z_j(t)\lambda_{ji} + C_iC_i'p_i(t), \\ Z_i(0) &= V_i \geq 0.\end{aligned}$$

Define $R_i(t) := Z_i(t) - \bar{Z}_i(t)$. Then, we get

$$\begin{aligned}\dot{R}_i(t) &= \left(A_i + \frac{1}{2}\lambda_{ii}I\right) R_i(t) + R_i(t)\left(A_i + \frac{1}{2}\lambda_{ii}I\right)' \\ &\quad + \sum_{j=1,\ j\neq i}^{N} R_j(t)\lambda_{ji} + C_iC_i'\left[p_i(t) - \alpha_i(t)\right], \\ R_i(0) &= 0.\end{aligned}$$

Since $C_iC_i'[p_i(t) - \alpha_i(t)] \geq 0$, because $p_i(t) \geq \alpha_i(t)$ for all t, the solution $R_i(t)$ is obtained in the same way as we did for the solution of $\bar{Z}_i(t)$ and possesses the same properties as those of $\bar{Z}_i(t)$. So, $R_i(t) \geq 0$, which proves that $Z_i(t) \geq \bar{Z}_i(t)$, completing the proof of lemma. □

LEMMA 6.3. *Let $\mathcal{Z}_t = (Z_1(t), \ldots, Z_N(t)) \in \mathbb{H}^{n+}$, and let $\overline{\mathcal{Z}}_t = (\bar{Z}_1(t), \ldots, \bar{Z}_N(t)) \in \mathbb{H}^{n+}$. Then $\mathcal{V}(\mathcal{Z}_t) \geq \mathcal{V}(\overline{\mathcal{Z}}_t)$ for all $t \geq 0$, where $\mathcal{V}(\mathcal{Z}_t)$ is defined by* (4.11).

Proof. We have shown that $\mathcal{V}(\mathcal{Z}_t)$ is a linear operator and that $\mathcal{V}(\mathcal{Z}_t) \geq 0$ for all $\mathcal{Z}_t = (Z_1(t), \ldots, Z_N(t)) \in \mathbb{H}^{n+}$. From the previous lemma, $\mathcal{Z}_t - \overline{\mathcal{Z}}_t \geq 0$. Then, $0 \leq \mathcal{V}(\mathcal{Z}_t - \overline{\mathcal{Z}}_t) = \mathcal{V}(\mathcal{Z}_t) - \mathcal{V}(\overline{\mathcal{Z}}_t)$, which is equivalent to saying that $\mathcal{V}(\mathcal{Z}_t) \geq \mathcal{V}(\overline{\mathcal{Z}}_t)$, completing the proof. □

LEMMA 6.4. *Let $P_\star(t)$ be the solution of the Riccati differential equation given by*

$$\begin{aligned}\dot{P}_\star(t) &= (\mathcal{A} - T_\star(t)H)\, P_\star(t) + P_\star(t)\,(\mathcal{A} - T_\star(t)H)' \\ &\quad + T_\star(t)\bar{K}_tT_\star'(t) + \bar{C}_t\bar{C}_t' + \mathcal{V}(\overline{\mathcal{Z}}_t), \\ P_\star(0) &= 0,\end{aligned} \tag{6.12}$$

where $T_\star(t) := P_\star(t)H'\bar{K}_t^{-1}$, $\bar{K}_t := \bar{G}_t\bar{G}_t'$, $\bar{G}_t := [\sqrt{\alpha_1(t)}G_1 \ldots \sqrt{\alpha_N(t)}G_N]$, $\bar{C}_t := \operatorname{diag}(\sqrt{\alpha_i(t)}C_i)$, $\mathcal{V}(\overline{\mathcal{Z}}_t)$ is the linear operator defined by (4.11) *applied to $\overline{\mathcal{Z}}_t = (\bar{Z}_1(t), \ldots, \bar{Z}_N(t))$, the solution of the matrix differential equation given by* (6.11)*, and $\alpha_i(t) = \inf_{s\in[0,\infty)}\{p_i(t+s)\}$. Then, for $0 \leq s \leq t$, $P_\star(s) \leq P_\star(t)$. In addition, $P_\star(t) \leq \mathcal{P}(t)$ for all $t \in [0,\infty)$ and $\lim_{t\to\infty} P_\star(t) = \mathcal{P}$, where $\mathcal{P}$ is the solution of* (5.1).

Proof. First, observe that, for all t, $\mathcal{C}_t\mathcal{C}_t' \geq \bar{C}_t\bar{C}_t' \geq 0$ and $K_t \geq \bar{K}_t \geq 0$. Also from the previous lemma, $\mathcal{V}(\mathcal{Z}_t) \geq \mathcal{V}(\overline{\mathcal{Z}}_t) \geq 0$. In addition, from the exponential speed of convergence of $\alpha_i(t)$ to π_i, we have that $\bar{C}_t\bar{C}_t' \to \mathcal{C}\mathcal{C}'$, $\mathcal{V}(\overline{\mathcal{Z}}_t) \to \mathcal{V}(\mathcal{Q})$, and $\bar{K}_t \to K = \mathcal{G}\mathcal{G}'$ exponentially fast as $t \to \infty$.

Also, for $0 \leq s \leq t$, $P_\star(s) \leq P_\star(t)$, from Lemma 5.2 in [54].

Let us now prove that $P_\star(t) \leq \mathcal{P}(t)$ for all $t \in [0,\infty)$. In order to do that we rewrite (6.13) in the following way:

$$\begin{aligned}\dot{P}_\star(t) &= (\mathcal{A} - T_tH)\, P_\star(t) + P_\star(t)\,(\mathcal{A} - T_tH)' \\ &\quad + \bar{C}_t\bar{C}_t' + \mathcal{V}(\overline{\mathcal{Z}}_t) + T_t\bar{K}_tT_t' - (T_t - T_\star(t))\,\bar{K}_t\,(T_t - T_\star(t))'.\end{aligned}$$

Defining $\tilde{P}_\star(t) := \mathcal{P}(t) - P_\star(t)$, we have

$$\begin{aligned}\dot{\tilde{P}}_\star(t) &= (\mathcal{A} - T_t H)\,\tilde{P}_\star(t) + \tilde{P}_\star(t)\,(\mathcal{A} - T_t H)' \\ &\quad + T_t\left[K_t - \bar{K}_t\right] T_t' + \left[\mathcal{C}_t\mathcal{C}_t' - \bar{C}_t\bar{C}_t'\right] + \left[\mathcal{V}(\mathcal{Z}_t) - \mathcal{V}(\overline{\mathcal{Z}}_t)\right] \\ &\quad + (T_t - T_\star(t))\,\bar{K}_t\,(T_t - T_\star(t))', \\ \tilde{P}_\star(0) &= \mathcal{P}(0) \geq 0.\end{aligned}$$

Let $\Phi(t,s)$ be the transition matrix associated with $\mathcal{A} - T_t H$. Then

$$\begin{aligned}\tilde{P}_\star(t) &= \Phi(t,0)\mathcal{P}(0)\Phi'(t,0) \\ &\quad + \int_0^t \Phi(t,s)\left\{T_s\left[K_s - \bar{K}_s\right] T_s' + \left[\mathcal{C}_s\mathcal{C}_s' - \bar{C}_s\bar{C}_s'\right] + \left[\mathcal{V}(\mathcal{Z}_s) - \mathcal{V}(\overline{\mathcal{Z}}_s)\right]\right. \\ &\quad \left. + (T_s - T_\star(s))\,\bar{K}_s\,(T_s - T_\star(s))'\right\}\Phi'(t,s)ds.\end{aligned}$$

Since $\mathcal{P}(0) \geq 0$, $K_t - \bar{K}_t \geq 0$, $\mathcal{C}_t\mathcal{C}_t' - \bar{C}_t\bar{C}_t' \geq 0$, $\mathcal{V}(\mathcal{Z}_t) - \mathcal{V}(\overline{\mathcal{Z}}_t) \geq 0$, and $(T_t - T_\star(t))\bar{K}_t(T_t - T_\star(t))' \geq 0$ because $\bar{K}_t \geq 0$ for all $t \geq 0$, we have $\tilde{P}_\star(t) \geq 0$, which proves that $\mathcal{P}(t) \geq P_\star(t)$.

Now, since $P_\star(t)$ is a nondecreasing function of t and is bounded above by $\mathcal{P}$, because $P_\star(t) \leq \mathcal{P}(t) \leq P^\star(t)$ and $\lim_{t\to\infty} P^\star(t) = \mathcal{P}$, we have that $\lim_{t\to\infty} P_\star(t)$ exists. So, there exists a matrix $\overline{P}_\star$, such that $P_\star(t) \to \overline{P}_\star$ as $t \to \infty$. Now, due to the monotonicity and convergence of $P_\star(t)$, it is a classical result that $\dot{P}_\star(t) \to 0$ as $t \to \infty$. Therefore, $\overline{P}_\star$ is the solution of the algebraic Riccati equation given by

$$\left(\mathcal{A} - \overline{T}_\star H\right)\overline{P}_\star + \overline{P}_\star\left(\mathcal{A} - \overline{T}_\star H\right)' + \overline{T}_\star K \overline{T}_\star' + \mathcal{C}\mathcal{C}' + \mathcal{V}(\mathcal{Q}) = 0,$$

where $\overline{T}_\star = \overline{P}_\star H'(\mathcal{G}\mathcal{G}')^{-1}$. But the above equation can be rewritten as

$$\mathcal{A}\overline{P}_\star + \overline{P}_\star\mathcal{A}' - \overline{P}_\star H'\left(\mathcal{G}\mathcal{G}'\right)^{-1} H\overline{P}_\star + \mathcal{C}\mathcal{C}' + \mathcal{V}(\mathcal{Q}) = 0. \tag{6.13}$$

Since (6.13) is equivalent to (5.1), we have, from the unicity of solution of (5.1), $\overline{P}_\star = \mathcal{P}$. Therefore, $\lim_{t\to\infty} P_\star(t) = \mathcal{P}$. □

Acknowledgment. The authors would like to express their gratitude to the referees for their suggestions and helpful comments.

REFERENCES

[1] Y. Bar-Shalom and X.R. Li, *Estimation and Tracking. Principles, Techniques and Software,* Artech House, Boston, MA, 1993.

[2] T. Björk, *Finite dimensional optimal filters for a class of Itô-processes with jumping parameters,* Stochastics, 4 (1980), pp. 167–183.

[3] W.P. Blair Jr. and D.D. Sworder, *Continuous-time regulation of a class of econometric models,* IEEE Trans. Syst. Man Cyber., 5 (1975), pp. 341–346.

[4] H.A.P. Blom and Y. Bar-Shalom, *The interacting multiple model algorithm for systems with Markovian switching coefficients,* IEEE Trans. Automat. Control, 33 (1988), pp. 780–783.

[5] E. Boukas and P. Shi, *Stochastic stability and guaranteed cost control of discrete-time uncertain systems with Markovian jumping parameters,* Internat. J. Robust Nonlinear Control, 8 (1998), pp. 1155–1167.

[6] J.W. Brewer, *Kronecker products and matrix calculus in system theory,* IEEE Trans. Circuits and System, 25 (1978), pp. 772–781.

[7] R.W. Brockett and J.M.C. Clark, *The geometry of the conditional density functions,* in Analysis and Optimization of Stochastic Systems, O.L.R. Jacobs, et al., eds., Academic Press, New York, 1980, pp. 299–309.

[8] J.M.C. Clark, *Conditions for One-to-One Correspondence between an Observation Process and Its Innovation,* Technical Report, Centre for Computing and Automation, Imperial College, London, 1969.

[9] J.M.C. Clark, *The design of robust approximations to the stochastic differential equations of nonlinear filtering,* in Communication Systems and Random Process Theory, NATO Adv. Study Inst. Ser., J.K. Skwirzynski, ed., Sijthoff and Noordhoff, Alphen aan den Rijn, 1978.

[10] O.L.V. Costa and M.D. Fragoso, *Stability results for discrete-time linear systems with Markovian jumping parameters,* J. Math. Anal. Appl., 179 (1993), pp. 154–178.

[11] O.L.V. Costa and M.D. Fragoso, *Discrete-time LQ-optimal control problems for infinite markov jump parameter systems*, IEEE Trans. Automat. Control, 40 (1995), pp. 2076–2088.

[12] O.L.V. Costa, M.D. Fragoso, and R.P. Marques, *Discrete-Time Markov Jump Linear Systems,* Probab. Appl. (N.Y.), Springer-Verlag, New York, 2004.

[13] O.L.V. Costa and S. Guerra, *Stationary filter for linear minimum mean square error estimator of discrete-time Markovian jump systems,* IEEE Trans. Automat. Control, 47 (2002), pp. 1351–1356.

[14] M.H.A. Davis, *Linear Estimation and Stochastic Control,* Chapman and Hall, London, 1977.

[15] M.H.A. Davis and S.I. Marcus, *An introduction to nonlinear filtering,* in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, NATO Adv. Study Inst. Ser., M. Hazewinkel and J. C. Willems, eds., D. Reidel, Dordrecht, 1981, pp. 53–75.

[16] C.E. de Souza and M.D. Fragoso, H^∞ *control for linear systems with Markovian jumping parameters,* Control Theory Adv. Technol., 9 (1993), pp. 457–466.

[17] C.E. de Souza and M.D. Fragoso, H^∞ *filtering for Markovian jump linear systems,* Internat. J. Systems Sci., 33 (2002), pp. 909–915.

[18] C.E. de Souza and M.D. Fragoso, *Robust* H^∞ *filtering for uncertain Markovian jump linear systems,* Internat. J. Robust Nonlinear Control, 12 (2002), pp. 435–446.

[19] A. Doucet and C. Andrieu, *Iterative algorithms for state estimation of jump Markov linear systems,* IEEE Trans. Automat. Control, 49 (2000), pp. 1216–1227.

[20] A. Doucet and A. Logothetis, and V. Krishnamurthy, *Stochastic sampling algorithms for state estimation of jump Markov linear systems,* IEEE Trans. Automat. Control, 45 (2000), pp. 188–202.

[21] J.B.R. do Val and T. Başar, *Receding horizon control of jump linear systems and a macroeconomic policy problem,* J. Econom. Dynam. Control, 23 (1999), pp. 1099–1131.

[22] F. Dufour and P. Bertrand, *An image based filter for discrete-time Markovian jump linear systems,* Automatica, 32 (1996), pp. 241–247.

[23] R.J. Elliot, L. Aggoun, and J.B. Moore, *Hidden Markov Models: Estimation and Control,* Springer-Verlag, New York, 1995.

[24] R. Elliot, F. Dufour, and F. Sworder, *Exact hybrid filters in discrete-time,* IEEE Trans. Automat. Control, 41 (1996), pp. 1807–1810.

[25] Y. Fang, *A new general sufficient condition for almost sure stability of jump linear systems,* IEEE Trans. Automat. Control, 42 (1997), pp. 378–382.

[26] X. Feng, K.A. Loparo, Y. Ji, and H.J. Chizeck, *Stochastic stability properties of jump linear systems,* IEEE Trans. Automat. Control, 37 (1992), pp. 1884–1892.

[27] M.D. Fragoso, *On a partially observable LQG problem for systems with Markovian jumping parameters,* Systems Control Lett., 10 (1988), pp. 349–356.

[28] M.D. Fragoso and J. Baczynski, *Optimal control for continuous-time linear quadratic problems with infinite Markov jump parameters,* SIAM J. Control Optim., 40 (2001), pp. 270–297.

[29] M.D. Fragoso and O.L.V. Costa, *A unified approach for stochastic and mean square stability of continuous-time linear systems with Markovian jumping parameters and additive disturbances,* SIAM J. Control Optim., to appear.

[30] M.D. Fragoso, O.L.V. Costa, and J. Baczynski, *The minimum linear mean square filter for a class of hybrid systems,* IEEE Trans. Automat. Control, to appear.

[31] M.D. Fragoso, O.L.V. Costa, and C.E. de Souza, *A new approach to linearly perturbed Riccati equations arising in stochastic control,* Appl. Math. Optim., 37 (1998), pp. 99–126.

[32] M. Fujisaki, G. Kallianpur, and H. Kunita, *Stochastic differential equations for the nonlinear filtering problem,* Osaka J. Math., 9 (1972), pp. 19–40.

[33] W.S. Gray and O. Gonzalez, *Modelling electromagnetic disturbances in closed-loop computer controlled flight systems,* in IEEE 37th Conference on Decision and Control, Philadelphia, PA, 1998, pp. 359–364.

[34] W.S. Gray, O.R. González, and M. Doğan, *Stability analysis of digital linear flight controllers subject to electromagnetic disturbances,* IEEE Trans. Aerosp. Electron. Syst., 36 (2000), pp. 1204–1218.
[35] Y. Ji and H.J. Chizeck, *Controllability, stabilizability, and continuous-time Markovian jumping linear quadratic control,* IEEE Trans. Automat. Control, 35 (1990), pp. 777–788.
[36] Y. Ji, H.J. Chizek, X. Feng, and K.A. Loparo, *Stability and control of discrete-time jump linear systems,* Control-Theory Adv. Techol., 7 (1991), pp. 247–270.
[37] G. Kallianpur, *Stochastic Filtering Theory,* Springer-Verlag, New York, 1980.
[38] S. Karlin and H.M. Taylor, *A Second Course in Stochastic Processes,* Academic Press, New York, 1981.
[39] M. Khambaghi, R. Malhamé, and M. Perrier, *White water and broke recirculation policies in paper mills via Markovian jump linear quadratic control,* in Proceedings of the American Control Conference, Philadelphia, PA, 1998, pp. 738–743.
[40] H.J. Kushner, *Dynamical equations for nonlinear filtering,* J. Differential Equations, 3 (1967), pp. 179–190.
[41] M. Mahmoud and P. Shi, *Robust Kalman filtering for continuous time-lag systems with Markovian jump parameters,* IEEE Trans. Circuits and System, 50 (2003), pp. 98–105.
[42] R. Malhame and C.Y. Chong, *Electric load model synthesis by diffusion approximation in a high order hybrid state stochastic systems,* IEEE Trans. Automat. Control, 30 (1985), pp. 854–860.
[43] M. Mariton, *Almost sure and moments stability of Jump linear systems,* Systems Control Lett., 11 (1988), pp. 393–397.
[44] M. Mariton, *Jump Linear Systems in Automatic Control,* Marcel Dekker, New York, 1990.
[45] N.C.S. Rocha, *Filtering for Continuous-Time Linear Systems with Markovian Jumps (Portuguese),* Ph.D. thesis, Federal University of Rio de Janeiro, UFRJ/COPPE, 2004.
[46] W. Rudin, *Real and Complex Analysis,* 3rd ed., McGraw-Hill, New York, 1987.
[47] P. Shi, E. Boukas, and R. Agarwal, *Kalman filtering for continuous-time uncertain systems with Markovian jump parameters,* IEEE Trans. Automat. Control, 44 (1999), pp. 1592–1597.
[48] A. Stoica and I. Yaesh, *Jump Markovian-based control of wing deployment for an uncrewed air vehicle,* J. Guidance, 25 (2002), pp. 407–411.
[49] D.D. Sworder and R.O. Rogers, *An LQ solution to a control problem associated with a solar thermal central receiver,* IEEE Trans. Automat. Control, AC-28 (1983), pp. 971–978.
[50] D.D. Sworder, *Feedback control for a class of linear systems with jump parameters,* IEEE Trans. Automat. Control, AC-14 (1969), pp. 9–14.
[51] D.D. Sworder and J.E. Boyd, *Estimation Problems in Hybrid Systems,* Cambridge University Press, Cambridge, UK, 1999.
[52] J. Van Schuppen, *Stochastic filtering theory: A discussion of concepts methods and results,* in Stochastic Control Theory and Stochastic Differential Systems, Lecture Notes in Control and Inform. Sci. 16, M. Kohlman and W. Vogel, eds., Springer-Verlag, New York, 1979, pp. 209–226.
[53] W.H. Wonham, *Some applications of stochastic differential equations to optimal nonlinear filtering,* SIAM J. Control Optim., 3 (1965), pp. 347–369.
[54] W.H. Wonham, *On a matrix Riccati equation of stochastic control Optim.*, SIAM J. Control, 6 (1968), pp. 681–697.
[55] S.S.-T Yau, *Finite dimensional filters with nonlinear drift I: A class of filters including both Kalman-Bucy filters and Benes filters,* J. Math. System Estim. Control, 4 (1994), pp. 181–203.
[56] Q. Zhang, *Nonlinear filtering and control of a switching diffusion with small observation noise,* SIAM J. Control Optim., 36 (1998), pp. 1638–1668.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 3, pp. 816–866

A SYMBOLIC APPROACH TO PERFORMANCE ANALYSIS OF QUANTIZED FEEDBACK SYSTEMS: THE SCALAR CASE*

FABIO FAGNANI† AND SANDRO ZAMPIERI‡

Abstract. When dealing with the control of a large number of interacting systems, the fact that the flow of information has to be limited becomes an essential feature of the control design. The first consequence of the limited information flow constraint is that the signals that the controllers and the systems exchange have to be quantized. Though quantization has already been extensively considered in the control literature, its analysis from the point of view of the information flow demand has been considered only recently.

Limiting the information flow between a plant and a controller will necessarily lead to a performance degradation of the feedback loop, and we expect a trade-off between the achievable performance and the amount of information exchange allowed in the loop.

Most of the success of modern digital communication theory in the last 50 years is due to the contributions of information theory, which proposed a symbolically based analysis of the communication channel performance. The same goal is much more difficult to reach in digital control theory.

This paper proposes an attempt toward this direction. The main contribution of this paper is to provide a complete analysis of the trade-off between performance and information flow in the simple case of the stabilization of a scalar linear system by means of a memoryless quantized feedback map.

Key words. stability, stabilization, communication constraint, quantized feedback, chaotic control, symbolic dynamics, Markov chains, entropy

AMS subject classifications. 93D15, 37B10, 37E05

DOI. 10.1137/S0363012903434315

1. Introduction.

Stabilization by quantized feedback controllers has been widely investigated in the last few years (see [2, 3, 6, 7, 9, 16, 20, 21, 22, 23, 25, 26] and references therein). There are two different situations in which quantization appears to be a central feature in the control design. The first is related to control systems in which either the sensor's or the actuator's limitations impose the condition that their measures or their commands can take a limited number of different values. In this case, the number of quantization levels provides a measure of the sensor or of the actuator complexity. Another situation in which quantization plays an important role is when plants and controllers exchange information through digital communication channels with a limited capacity. In this last case, the measures and the commands need to be quantized before being communicated and the number of quantization levels is strictly related to the information flow between the components of the control system and so to the capacity required to transmit the control information.

Two different approaches have been proposed in the literature for solving the control problem with a quantized feedback. The first approach considers memoryless feedback quantizers. In particular in [6] there is a first mathematical analysis of control systems with uniform quantized feedback, while in [26, 2] a first bound of the number of quantization intervals needed to stabilize a linear system is proposed. In

*Received by the editors September 11, 2003; accepted for publication (in revised form) March 9, 2005; published electronically September 15, 2005.
http://www.siam.org/journals/sicon/44-3/43431.html
†Dipartimento di Matematica, Politecnico di Torino, C.so Duca degli Abruzzi, 24, 10129 Torino, Italy (fabio.fagnani@polito.it).
‡Dipartimento di Ingegneria dell'Informazione, Università di Padova, via Gradenigo, 6/a, 35131 Padova, Italy (zampi@dei.unipd.it).

[7] logarithmic quantizers are shown to yield the Lyapunov stability. In [9] a chaos-based quantized controller was proposed and a first comparison between uniform, logarithmic, and chaotic quantized feedback controllers was presented in the scalar case. In [23] performance of uniform quantized feedback controllers is analyzed for general linear systems.

The second approach considers quantized feedback controllers with an internal state. In particular [3] proposes a stabilization technique in which the quantizer is scaled according to the state growth. In [25] this technique is used to show the relation between the degree of instability of the system to be controlled and the number of quantization levels of the feedback quantizer. The same relation was found independently in [20] in a different context.

In general the analysis of memoryless quantized feedback controllers is difficult, while the results become quite neat for quantized feedback controllers with infinite memory. Notice that, while it is reasonable to allow a memory structure on sensors and actuators when designing control systems under communication constraint [25, 20], in situations in which quantization is due to the poor quality of sensors or actuators, only the memoryless quantized feedback controller becomes a reasonable model.

This paper considers memoryless quantized controllers for which, as we mentioned, a mathematical analysis is more complicated. The relation between controller complexity and controller performance is investigated by using theoretical information and combinatorial techniques. One of the main contributions of this paper is to show that the controller performance has to be described by two conflicting parameters, one evaluating the steady state and the other evaluating the transient of the controlled system. Roughly speaking we proved that, for a fixed controller complexity, a good steady state implies a bad transient and vice versa.

More precisely, in this paper we consider the stabilization problem for discrete time linear systems with a one-dimensional state, namely, a system described by the equation

$$x_{t+1} = ax_t + u_t.$$

While in the classical control setting this stabilization problem is completely trivial and there is little to be said, in the memoryless quantized feedback setting nontrivial issues already come up in this simple situation. In this setup, a memoryless quantized feedback is a control law $u_t = k(x_t)$, where $k(\cdot)$ is a quantized (i.e., piecewise constant) map. Let N be the number of distinct values which $k(\cdot)$ is allowed to take. The number N will provide a measure of the information flow in the feedback loop. In the literature referenced above several different quantized stabilizing strategies have been proposed in this context. Moreover in [2, 26] the minimum value of N (as a function of $|a|$) has been found, ensuring the existence of a memoryless quantized controller yielding stability (but not convergence) of the system.

The aim of this paper is to compare different quantized control strategies proposed in the literature in terms of complexity and performance and to establish a number of results showing fundamental limitations of quantized control. To be more precise about performance, notice first that, if the original system is unstable, a state feedback with finitely many quantization intervals can only yield so-called practical stabilization, namely, the convergence of any initial state belonging to a bigger bounded region I into another smaller target region of the state space J. The ratio C between the measure of the starting region and the target region is called contraction, and it provides a description of the steady state properties of the closed loop system.

Beyond C, the expected time T needed to shrink the state of the plant from the starting set to the target set will measure the transient controller performance. Notice that these two parameters represent a particular way of evaluating the steady state and the transient performance of the controller. There are other possible choices. For instance, it is possible to evaluate the transient by means of a quadratic-like index. Some preliminary investigations show that the techniques proposed in this paper can also be applied in this setup and yield similar trade-off results.

We will evaluate the relations between the parameters N, C, T, and a in a series of different stabilization strategies. In all cases we will see that, for fixed a, as C grows, either N has to grow or T has to grow. However, different strategies exhibit different growth rates of the two parameters N and T. In all cases an increasing value of $|a|$ either requires an increase in N or yields a degradation of C and T. These results extend the relations between N and $|a|$ proposed in [2, 26] and complete the analysis started in [9], where, however, the parameter T was interpreted as the sup norm of the entrance time and where a stronger notion of stability was considered. The relations between the parameters N, C, T pointed out in the examples are in accordance with some fundamental bounds which are proved in the second part of the paper, proving in this way the optimality of the proposed quantized controller synthesis techniques.

Now we present an outline of the contents of this paper and of our main results. In section 2, we present all basic definitions and notation. In particular we introduce the concepts of stability and almost stability, and we state precisely the problems we want to solve. Moreover, we introduce some basic tools from the ergodic theory of piecewise affine maps. Using these we show that the expected entrance time T is always finite if we have almost stability.

Section 3 is devoted to the introduction and the discussion of a general stabilization strategy based on nesting an initial given quantized stabilizer.

Section 4 is devoted to the analysis of some examples. We show that, by nesting the quantized deadbeat controller in a suitable way, we can obtain a variety of different quantized stabilizers, which can be analyzed in terms of the parameters N and T as functions of a and C. There are three particularly significant cases. The first is the quantized deadbeat control which is obtained by using uniform quantized feedback. In this case N grows linearly in C and $|a|$ and T tend to the constant 1. The second is the logarithmic quantized feedback strategy. In this case, instead, both N and T grow logarithmically in C. The latter is the chaotic quantized feedback strategy. In this last case only almost stability can be achieved and N tends to the constant $\lceil |a| \rceil$ while T grows linearly in C. Notice that the first and the last strategies present dual characteristics of N and T as functions of C. It is interesting to observe that, if we take any linear feedback $u_t = kx_t$, with $k \in \mathbb{R}$, such that the linear closed loop system $x_{t+1} = (a+k)x_t$ is asymptotically stable, then the expected entrance time T of this controlled system is such that $T/\log C$ tends to a constant which is a decreasing function of $|a + k|$. Hence the logarithmic regime corresponds to the performance which can be obtained through the allocation of the eigenvalue inside the unit circle and the absolute value of this eigenvalue determines the logarithmic rate.

In section 5, we obtain universal bounds relating T, N, and C for fixed $|a|$. The main results are presented in Theorems 3 and 4 and Corollaries 3 and 2. All these results, except Theorem 4, need the assumption $|a| > 2$. Corollary 3 says two things: First, in order to obtain expected entrance time T growing at most logarithmically with respect to C, we need a number of quantization intervals N growing at least logarithmically with respect to C. Second, if we use a number of quantization intervals

N growing at most logarithmically with respect to C, we obtain expected entrance times T growing at least logarithmically with respect to C. Moreover, the corollary furnishes a quantitative trade-off between the two ratios $T/\log C$ and $N/\log C$ which turns out to be interesting if related to the previous comment on the logarithmic regime which can be obtained in the linear feedback case. Another consequence of the results presented in this section is that the chaos-based stabilization strategy is somehow optimal since its performance cannot be improved without paying this with a greater information flow. Finally, Theorem 5 shows that any stabilization strategy yielding stability has the ratio $N/\log C$ bounded from below.

To prove the results in section 5 we need to use the tools of combinatorial analysis of the symbolic dynamics associated with piecewise affine maps. This is developed in section 6, which contains the deeper mathematical result of this paper, which is Theorem 6. This theorem provides a new bound on the number of the paths on a graph with possibly infinite uncountable edges, when this graph has some specific properties. This theorem is very general and has potential applications in other situations such as the analysis of quantized feedback systems when the state is multidimensional [10].

We conclude this introduction with a few remarks to emphasize the reasons why we limited our analysis to scalar state space systems. From an application viewpoint, these may be seen as a relatively uninteresting family of systems to be considered. However, this simple case already contains all the interesting issues of the coupling between control and information and mathematically leads to nontrivial problems. The completeness of the results obtained in this paper, because of the simplified setup we choose, will provide the guidelines for future investigations on more general situations (see [10]). Observe finally that first order systems can be considered as simplified models of more general systems and that one important case in which control under communication constraint is relevant is just when many simple systems have to be controlled by a unique centralized controller.

Notation. We present here some notation which will be used in this paper. If A and B are two sets, then $A \setminus B := \{a \in A : a \notin B\}$. Given a map $f : A \to B$ and $B_1 \subseteq B$ we define

$$f^{-1}(B_1) := \{a \in A : f(a) \in B_1\}.$$

The symbol $A^{\mathbb{N}}$ denotes the set of all sequences taking values on the set A, while the symbol A^* denotes the set of all finite words over the alphabet A. The symbol $\#A$ denotes the cardinality of A.

The symbol $\mathbb{R}_+$ denotes the set of all positive real numbers. If $a \in \mathbb{R}_+$, then $\lceil a \rceil$ means the minimum integer greater than or equal to a, and $\log a$ is the natural logarithm of a. Given $a, b \in \mathbb{R}$, $a \wedge b$ and $a \vee b$ denote the minimum and the maximum between a and b, respectively. Given $K \subseteq \mathbb{R}$, $\overline{K}$ denotes the closure of K, while ∂K denotes the boundary set of K.

Let I be an interval in $\mathbb{R}$. Given any function $f : I \to \mathbb{R}$ we define

$$\text{supp } (f) := \{x \in I : f(x) \neq 0\}.$$

For any $J \subseteq I$ we denote by $\mathbf{1}_J$ the function defined on I which is 1 in J and 0 on $I \setminus J$, and it is called the indicator function of J. With the symbol $L^1(I)$ we mean the set of absolutely integrable functions which is a normed space with the norm

$$||f||_1 := \int_I |f(x)|dx \qquad \forall f \in L^1(I).$$

If $\mathcal{P} : L^1(I) \to L^1(I)$ is a linear continuous operator, then the symbol $||\mathcal{P}||_1$ denotes the induced norm of $\mathcal{P}$. The symbol $L^\infty(I)$ means the set of bounded functions on I which is a normed space as well. A function $f \in L^1(I)$ such that $f(x) \geq 0$ for all $x \in I$ and such that $||f||_1 = 1$ is called a density function on I. It induces a probability measure on I which will be denoted by $\mathbb{P}_f$, while the symbol $\mathbb{E}_f$ will denote the expected value with respect to $\mathbb{P}_f$. The probability measure and the expected value with respect to a uniform Lebesgue measure on I will be simply denoted by the symbols $\mathbb{P}$ and $\mathbb{E}$, respectively.

2. Problem statement. Consider the following discrete-time, one-dimensional linear model:

$$x_{t+1} = ax_t + u_t, \tag{1}$$

where $a \in \mathbb{R}$. Most of the paper is devoted to the stabilization problem, and so it is assumed that $|a| > 1$. Some results, however, hold true also for stable systems and so for $|a| \leq 1$.

Let $k : \mathbb{R} \to \mathbb{R}$ be a piecewise constant function with only finitely many discontinuities. If we use k as a static feedback in system (1), namely, we let $u_t = k(x_t)$, we obtain the closed loop system

$$x_{t+1} = \Gamma(x_t), \tag{2}$$

where $\Gamma(x) := ax + k(x)$ is a piecewise affine map with a fixed slope a. Autonomous systems such as (2) in which Γ is piecewise affine can exhibit very wild behavior. Their dynamical properties were extensively studied in the past [15, 18, 5, 24].

Remark. In fact, the definition we gave is not precise if we do not define what happens at the boundary points of the intervals. We assume there is a finite family of disjoint open intervals I_h such that $D := \cup_h I_h$ is dense in $\mathbb{R}$ and that $k(x) = u_h$ for every $x \in I_h$. In this way the associated closed loop map is defined as a map

$$\Gamma : D \to \mathbb{R}, \tag{3}$$

$$\Gamma(x) = ax + u_h \quad \text{if } x \in I_h.$$

In order to consider iterations of Γ we need to restrict the domain by considering

$$\Omega = \bigcap_{n=0}^{\infty} \Gamma^{-n}(D). \tag{4}$$

It is clear that $\Gamma(\Omega) \subseteq \Omega$. Notice that $\mathbb{R} \setminus \Omega$ is a countable subset of $\mathbb{R}$, and since most of the questions considered in this paper are related to mean properties, it will be sufficient to consider Γ as a map defined on Ω, disregarding all the orbits which will eventually get to a discontinuity point.

However, in those situations in which it is necessary to understand how the dynamics is defined at the boundaries, it is necessary to define the dynamics of Γ on all $\mathbb{R}$. This is done by considering, for any $x_0 \in \mathbb{R}$, the left and right limits of $\Gamma(x)$ for $x \to x_0$ denoted by $\Gamma(x_0-)$ and $\Gamma(x_0+)$, and by defining the enlarged set of orbits as

$$X_\Gamma = \big\{(x_t) \in \mathbb{R}^{\mathbb{N}} \mid x_{t+1} = \Gamma(x_t+) \text{ or } x_{t+1} = \Gamma(x_t+) \quad \forall t \in \mathbb{N}\big\}. \tag{5}$$

The subset $X_\Gamma \cap \Omega^{\mathbb{N}}$ consists of the orbits of Γ on Ω, and it is in bijection with Ω through the initial condition.

It is obvious that, by using quantized feedback controllers, only a "practical stability" can be obtained as detailed in the following definitions.

DEFINITION (invariance and almost invariance). *Given a closed interval I, we say that I is Γ-invariant if every orbit (x_t) of Γ with $x_0 \in I$ is such that $x_t \in I$ for every t. It is almost Γ-invariant if the assertion above is true for almost every initial condition x_0 with respect to the Lebesgue measure. When an interval I is invariant or almost invariant we will use in any case the notation $\Gamma : I \to I$.*

DEFINITION (stability and almost stability). *Given two closed intervals $J \subseteq I$, we say that Γ is (I, J)-stable if I and J are invariant by Γ and if for every orbit (x_t) of Γ with $x_0 \in I$, there exists an integer $t \geq 0$ such that $x_t \in J$. We say that Γ is almost (I, J)-stable if I and J are almost invariant and the convergence to J as defined above occurs for almost all initial conditions in the orbit $x_0 \in I$ with respect to the Lebesgue measure. A quantized feedback map $k : \mathbb{R} \to \mathbb{R}$ is said to be (almost) (I, J)-stabilizing if the corresponding closed loop map Γ is (almost) (I, J)-stable.*

Remark. With regard to almost invariance and almost stability, it is sufficient to work with Γ on the set Ω as defined in (4). The concepts of invariance and stability also depend on the dynamics on boundary points, and so the orbits have to be considered as defined in (5).

Assume that Γ is almost (I, J)-stable. The first entrance time function

$$T_{(I,J)} : I \cap \Omega \to \mathbb{N} \cup \{+\infty\}$$

is defined by

$$T_{(I,J)}(x) = \inf \left\{ n \in \mathbb{N} \mid \Gamma^n x \in J \right\} = \sum_{n=1}^{\infty} \mathbf{1}_{I \setminus J}(\Gamma^n x). \tag{6}$$

We put $T_{(I,J)}(x) := +\infty$ if $\Gamma^t x \notin J$ for all t. Notice that the map $T_{(I,J)}$ is always finite exactly when we have stability, while it is almost surely finite when we have almost stability.

Remark. Notice that, if we want to extend the function $T_{(I,J)}$ to all I, we cannot use definition (6). Indeed, there is a possible ambiguity for orbits touching discontinuity points since, given $x \in I$, there can be infinitely many orbits having x as an initial condition and, therefore, $\Gamma^n x$ is not uniquely defined. In this case definition (6) should be replaced as follows: we say that $T_{(I,J)}(x) = n$ if every orbit $(x_t) \in X_\Gamma$ such that $x_0 = x$ is such that $x_t \in J$ for any $t \geq n$, and if there exists an orbit $(x_t) \in X_\Gamma$ such that $x_0 = x$ and such that $x_{n-1} \notin J$.

The expected value of the entrance time with respect to a density function f on I is given by

$$\mathbb{E}_f(T_{(I,J)}) = \int_I T_{(I,J)}(x) f(x) dx.$$

It is clear that

$$\mathbb{E}_f(T_{(I,J)}) = \int_I \left[\sum_{n=1}^{\infty} \mathbf{1}_{I \setminus J}(\Gamma^n x) f(x) \right] dx = \sum_{n=1}^{\infty} n \mathbb{P}_f[T_{(I,J)} = n] = \sum_{n=0}^{\infty} \mathbb{P}_f[T_{(I,J)} > n].$$

In what follows, for any given (almost) (I, J)-stabilizing quantized feedback k yielding an (almost) (I, J)-stable piecewise affine closed loop map Γ, we will use the symbol $\mathbf{T}(k)$ or $\mathbf{T}(\Gamma)$ to denote the relative expected entrance time $\mathbf{E}(T_{(I,J)})$ with respect

to the uniform density function on I. Notice that this quantity depends only on the restriction of Γ to $I \setminus J$, and so we can assume that Γ is defined only on $I \setminus J$. For this reason the right parameter measuring the information flow will be the number of quantization intervals in $I \setminus J$, which will be denoted by the symbol $\mathbf{N}(k)$ or $\mathbf{N}(\Gamma)$. Finally the ratio between the length of I and the length of J will be called the contraction rate and will be denoted by $C(k)$ or $C(\Gamma)$.

The performance analysis of the quantized stabilization consists of determining, for a given $C > 1$, $N \in \mathbb{N}$, and $T > 0$, whether there exists or not a (almost) stabilizing quantized feedback k such that $C(k) = C$, $\mathbf{N}(k) = N$, and $\mathbf{T}(k) = T$, or, in other words, estimating the set

$$\mathcal{A} := \{(C, N, T) \ : \ \text{there exists } k \text{ such that } C(k) = C, \quad \mathbf{N}(k) = N, \quad \mathbf{T}(k) = T\}.$$

Remark. The analysis proposed in this paper can be extended to a family of more general performance measures. Let

$$V : I \to \mathbb{R}$$

be such that $0 \leq V(x) \leq 1$ for every $x \in I$, and $V(x) = 0$ for every $x \in J$. Another measure of the transient properties of the closed loop system is the following number:

$$\mathbb{E}\left(\sum_{n=0}^{\infty} V(\Gamma^n x)\right).$$

It is clear that, if $V(x) = \mathbf{1}_{I\setminus J}(x)$, then the previous cost coincides with the expected entrance time in J. If $V(x)$ is a general continuous function, then for any $\alpha \in [0, 1]$ we have that

$$\alpha \mathbf{1}_{I\setminus J(\alpha)}(x) \leq V(x) \leq \mathbf{1}_{I\setminus J}(x),$$

where $J(\alpha) := \{x \in I \ : \ V(x) \leq \alpha\}$. This fact implies that

$$\alpha \mathbb{E}(T_{J(\alpha)}) \leq \mathbb{E}\left(\sum_{n=0}^{\infty} V(\Gamma^n x)\right) \leq \mathbb{E}(T_{(I,J)}).$$

This shows that the dependence of this generalized performance index and of the expected entrance time on the parameters $C(\Gamma)$ and $\mathbf{N}(\Gamma)$ will be similar.

2.1. The Perron–Frobenius operator for piecewise affine maps. In this subsection we recall some standard results on the ergodic theory of piecewise affine maps, and we will present a first preliminary result asserting that the expected entrance time is always finite for almost (I, J)-stable piecewise affine maps.

Let $\Gamma : I \to I$ be a piecewise affine map with a fixed slope a and assume here that $|a| > 1$. It is a standard fact that Γ induces a linear transformation

$$\mathcal{P}_\Gamma : L^1(I) \to L^1(I)$$

called the Perron–Frobenius operator associated with Γ which is uniquely defined by the following duality relation:

$$\int_I (g \circ \Gamma)(x) f(x) dx = \int_I g(x)(\mathcal{P}_\Gamma f)(x) dx \tag{7}$$

for all $g \in L^\infty(I), f \in L^1(I)$. It can be shown that the operator $\mathcal{P}_\Gamma$ is bounded with $||\mathcal{P}_\Gamma||_1 \leq 1$, and it maps probability densities onto probability densities. An important interpretation of $\mathcal{P}_\Gamma$ is as follows. If we have a continuous random variable X defined on I whose density is f, then the density of the transformed random variable $X \circ \Gamma$ is $\mathcal{P}_\Gamma f$. A final important property of the Perron–Frobenius operator $\mathcal{P}_\Gamma$ is that $\mathcal{P}_{\Gamma^n} = \mathcal{P}_\Gamma^n$.

The relevance of the Perron–Frobenius operator in our investigations is due to the fact that

$$\mathbb{P}_f\big[T_{(I,J)} > n\big] = \int_{I\setminus J} \mathcal{P}_\Gamma^n f(x)dx,$$

which follows by iterating (7) and by taking $g(x) = \mathbf{1}_{I\setminus J}(x)$. This shows that the asymptotics of this operator and so its spectral properties will be relevant for our purposes.

We have the following result.

LEMMA 1. *Let Γ be almost (I, J)-stable. If $h(x) \in L^1(I)$ is an invariant density of $\mathcal{P}_\Gamma$, then*

$$\text{supp } h \subseteq J.$$

Proof. First we show that, since J is invariant by Γ, the fact that supp $f \subseteq J$ implies that supp $\mathcal{P}_\Gamma^k f \subseteq J$. Indeed, if $K \subseteq I \setminus J$, then $\Gamma^{-1}(K) \subseteq I \setminus J$, and so

$$\int_K (\mathcal{P}_\Gamma f)(x)dx = \int_{\Gamma^{-1}(K)} f(x)dx = 0.$$

We show now that, if h is invariant by $\mathcal{P}_\Gamma$, then $h\mathbf{1}_J$ and $h\mathbf{1}_{I\setminus J}$ are also invariant by $\mathcal{P}_\Gamma$. Indeed for any $g \in L^\infty(I), f \in L^1(I)$ we have that

$$\begin{aligned}\int_I g(x)(\mathcal{P}_\Gamma h\mathbf{1}_J)(x)dx &= \int_J g(x)(\mathcal{P}_\Gamma h\mathbf{1}_J)(x)dx = \int_J g(x)(\mathcal{P}_\Gamma h)(x)dx \\ &= \int_I g(x)\mathbf{1}_J(x)(\mathcal{P}_\Gamma h)(x)dx = \int_I g(x)(h\mathbf{1}_J)(x)dx,\end{aligned}$$

where in the first equality we used the fact that supp $h\mathbf{1}_J \subseteq J$. This shows that $h\mathbf{1}_J$ is invariant. Since both h and $h\mathbf{1}_J$ are invariant, so is $h\mathbf{1}_{I\setminus J}$, as well.

Finally, if we assume by contradiction that there exists a nonzero invariant density of $\mathcal{P}_\Gamma$ not supported inside J, then for the above considerations, there also exists a nonzero invariant density supported inside $I \setminus J$. Let us call it h_0. We can find $\delta > 0$ and a subset $K \subseteq I \setminus J$ of nonzero Lebesgue measure such that $h_0(x) > \delta$ for every $x \in K$. Consequently, $h_0 - \delta\mathbf{1}_K$ is a nonnegative function. Therefore, $\mathcal{P}_\Gamma^n(h_0 - \delta\mathbf{1}_K) = h_0 - \mathcal{P}_\Gamma^n(\delta\mathbf{1}_K)$ is also nonnegative for all $n \geq 0$. Since h_0 is 0 on J, it follows that $\mathcal{P}_\Gamma^n\mathbf{1}_K$ is 0 on J for every n. This implies that

$$\int_{\Gamma^{-n}(J)} \mathbf{1}_K(x)dx = \int_J (\mathcal{P}_\Gamma^n\mathbf{1}_K)(x)dx = 0,$$

which implies that $K \cap \Gamma^{-n}(J)$ has zero Lebesgue measure for all $n \geq 0$, which contradicts the almost (I, J)-stability of Γ. ∎

To obtain a good characterization of the spectral properties of $\mathcal{P}_\Gamma$ we need to restrict the type of densities to be considered. Let $\mathrm{BV}(I) \subseteq L^1(I)$ be the subspace of $L^1(I)$ constituted by the bounded variation functions on the interval I. More precisely, if we define the variation of a function f as

$$\bigvee f := \sup\left\{\sum_{i=1}^{n-1} |f(x_{i+1}) - f(x_i)| \; x_i \in I, \quad x_1 < x_2 < \cdots < x_n\right\},$$

then

$$\mathrm{BV}(I) := \Big\{f : I \to \mathbb{R} : \bigvee f < \infty\Big\}.$$

Now equip the space $\mathrm{BV}(I)$ with a new norm

$$|||f||| := \bigvee f + ||f||_1.$$

It is a classical fact that $\mathcal{P}_\Gamma(\mathrm{BV}(I)) \subseteq \mathrm{BV}(I)$ and that $\mathcal{P}_\Gamma|_{\mathrm{BV}(I)}$ is bounded with respect to the norm $|||\cdot|||$. Using the Lasota–Yorke inequality [15] and the spectral theorem of Ionescu-Tulcea and Marinescu [12], the following facts can be shown to hold true.

(i) Let σ_1 be the set of eigenvalues of modulus 1 of $\mathcal{P}_\Gamma$ seen as an operator on $L^1(I)$. Then this set is a finite multiplicative group. Moreover, each of these eigenvalues has a finite dimensional eigenspace contained in $\mathrm{BV}(I)$.

(ii) The Perron–Frobenius operator $\mathcal{P}_\Gamma$ on $\mathrm{BV}(I)$ admits the following decomposition:

$$\mathcal{P}_\Gamma = \sum_{\lambda\in\sigma_1} \lambda Q_\lambda + R, \tag{8}$$

where Q_λ are finite rank operators on $\mathrm{BV}(I)$, and R is a bounded operator on $\mathrm{BV}(I)$ such that

(a) $Q_\lambda \circ R = R \circ Q_\lambda = 0$ for all $\lambda \in \sigma_1$;
(b) $Q_\lambda \circ Q_{\lambda'} = 0$ for all $\lambda, \lambda' \in \sigma_1$ such that $\lambda \neq \lambda'$;
(c) $Q_\lambda \circ Q_\lambda = Q_\lambda$ for all $\lambda \in \sigma_1$;
(d) $|||R^n||| \leq c\gamma^n$ for all $n \in \mathbb{N}$, where c is a positive constant and $0 < \gamma < 1$.

An important consequence of the above results is that the spectrum of $\mathcal{P}_\Gamma$ in $\mathrm{BV}(I)$ is composed of a finite set of eigenvalues on the unit circle (with finite dimensional eigenspaces) and of another part contained in a disk of radius strictly smaller than 1.

We now state and prove the main result of this section.

PROPOSITION 1. *Let Γ be an almost (I, J)-stable piecewise affine map. Then, there exists a constant $K > 0$ such that*

$$\mathbb{E}_f(T_{(I,J)}) \leq K|||f|||$$

for every probability density $f \in \mathrm{BV}(I)$.

Proof. Notice preliminarily that there exists $\nu \in \mathbb{N}$ such that $\lambda^\nu = 1$ for every $\lambda \in \sigma_1$. This implies that

$$\mathcal{P}_\Gamma^\nu = \sum_{\lambda\in\sigma_1} Q_\lambda + R^\nu.$$

This implies that for any density $f \in \mathrm{BV}(I)$ we have that $Q_\lambda f$ is invariant by $\mathcal{P}_\Gamma^\nu$. Since $\mathcal{P}_\Gamma^\nu$ is the Perron–Frobenius operator for the map Γ^ν which is almost (I,J)-stable, then, by Lemma 1, we have that supp $Q_\lambda f \subseteq J$. Using this fact and formula (8), we obtain

$$\mathbb{P}_f[T_{(I,J)} > n] = \int_{I\setminus J} (\mathcal{P}_\Gamma^n f)(x)dx = \int_{I\setminus J} (R^n f)(x)dx \le c\gamma^n |||f|||$$

and hence

$$\mathbb{E}_f(T_{(I,J)}) = \sum_{n=0}^{+\infty} \mathbb{P}_f[T_{(I,J)} > n] \le \frac{c}{1-\gamma}|||f|||. \qquad \square$$

3. Nested quantized feedback strategies. Consider the linear discrete time system (1), where $|a| > 1$, and consider two intervals $J \subseteq I$. We want to stabilize it through a quantized state feedback, i.e., we want to find a quantized feedback map k such that the closed loop system (2) drives (almost) any initial state $x_0 \in I$ into a state evolution which, after a transient, enters the interval J. Several solutions to this problem can be proposed. In fact, we will show that, starting from a base quantized feedback, it is possible to construct a family of quantized feedbacks by iterating the base one.

More precisely, suppose that we have found an (I,J)-stabilizing quantized feedback $k_1(x)$ and a (J,K)-stabilizing quantized feedback $k_2(x)$. Then it is clear that the quantized feedback

$$k(x) = \begin{cases} k_1(x) & \text{if } x \in I \setminus J, \\ k_2(x) & \text{if } x \in J \setminus K \end{cases} \tag{9}$$

will be (I,K)-stabilizing. The analogous conclusion is less straightforward in case we start from almost stabilizing quantized feedbacks. In what follows we will show that this is indeed the case, namely, if $k_1(x)$ is almost (I,J)-stabilizing and $k_2(x)$ is almost (J,K)-stabilizing, then $k(x)$ is almost (I,K)-stabilizing.

Let $\Gamma : I \to I$ be an almost (I,J)-stable piecewise affine map with a fixed slope a such that $|a| > 1$, and let $\mathcal{P}_\Gamma$ be the Perron–Frobenius operator associated with Γ. From any density function $f \in L^1(I)$ it is possible to define a probability measure μ on J as the image of the measure $\mathbb{P}_f$ through the map

$$\psi(x) := \Gamma^{T_{(I,J)}(x)}(x),$$

where $T_{(I,J)}(x)$ is the first entrance time function of Γ. More precisely, if $A \subseteq J$ is a measurable set, then

$$\mu(A) := \mathbb{P}_f\big[\psi^{-1}(A)\big]. \tag{10}$$

The following result gives important information on the measure μ.

PROPOSITION 2. *For any density $f \in L^1(I)$, the measure μ defined in* (10) *is absolutely continuous with respect to the Lebesgue measure and its corresponding density h is given by*

$$h = \mathbf{1}_J f + \sum_{j=1}^{+\infty} \mathbf{1}_J \mathcal{P}_\Gamma\big(\mathbf{1}_{I\setminus J}\mathcal{P}_\Gamma^{j-1} f\big). \tag{11}$$

Moreover, there exists a constant $H > 0$ only depending on Γ such that

$$|||h||| \le H|||f||| \quad \forall f \in \mathrm{BV}(I).$$

Proof. Let $A \subseteq J$ be a measurable set. Then

$$\begin{aligned}\mu(A) = \mathbb{P}_f[\psi^{-1}(A)] &= \sum_{j=0}^{+\infty} \mathbb{P}_f\big[\psi^{-1}(A) \cap \{T_{(I,J)}(x) = j\}\big] \\ &= \sum_{j=0}^{+\infty} \mathbb{P}_f\big[\Gamma^{-j}(A) \cap \{T_{(I,J)}(x) = j\}\big] \\ &= \mathbb{P}_f[A] + \sum_{j=1}^{+\infty} \mathbb{P}_f\big[\Gamma^{-j}(A) \cap \Gamma^{-j+1}(I \setminus J)\big].\end{aligned} \tag{12}$$

Notice that

$$\begin{aligned}\mathbb{P}_f\big[\Gamma^{-j}(A) \cap \Gamma^{-j+1}(I \setminus J)\big] &= \int_I \mathbf{1}_{\Gamma^{-j}(A)}(x)\mathbf{1}_{\Gamma^{-j+1}(I\setminus J)}(x)f(x)dx \\ &= \int_I \mathbf{1}_{\Gamma^{-1}(A)}(x)\Big[\mathbf{1}_{I\setminus J}(x)\mathcal{P}_\Gamma^{j-1}f(x)\Big]dx \\ &= \int_A \mathcal{P}_\Gamma\big(\mathbf{1}_{I\setminus J}\mathcal{P}_\Gamma^{j-1}f\big)(x)dx = \int_A h_j(x)dx,\end{aligned}$$

where $h_j(x) := \mathbf{1}_J(x)\mathcal{P}_\Gamma(\mathbf{1}_{I\setminus J}\mathcal{P}_\Gamma^{j-1}f)(x)$. Using this relation in (12) and the fact that $h_j(x)$ are nonnegative we obtain, by Fatou's lemma, that

$$\mu(A) = \mathbb{P}_f[A] + \sum_{j=1}^{+\infty}\int_A h_j(x)dx = \int_A \left[\mathbf{1}_J(x)f(x) + \sum_{j=1}^{+\infty} h_j(x)\right]dx,$$

which shows that the series $\sum_{j=1}^{+\infty} h_j(x)$ converges in the L^1 sense. Hence, the function h, defined in (11), is in L^1 and μ is absolutely continuous with respect to the Lebesgue measure with density h.

We now show that there is also a convergence in the norm $|||\cdot|||$ if $f \in \mathrm{BV}(I)$. First notice that, by the Yorke inequality [15, Formula (6.1.12)], for all $g \in \mathrm{BV}(I)$ we have

$$\bigvee(g\mathbf{1}_J) \le 2\bigvee g + \frac{2}{|I|}||g||_1,$$

which implies that

$$|||g\mathbf{1}_J||| \le 2\bigvee g + \left(1 + \frac{2}{|I|}\right)||g||_1 \le \left(2 + \frac{2}{|I|}\right)|||g|||.$$

Using the previous inequality we can argue that

$$|||\mathbf{1}_J\mathcal{P}_\Gamma(\mathbf{1}_{I\setminus J}\mathcal{P}_\Gamma^{j-1}f)||| \le \left(2 + \frac{2}{|I|}\right)|||\mathcal{P}_\Gamma||| \cdot |||\mathbf{1}_{I\setminus J}\mathcal{P}_\Gamma^{j-1}f|||. \tag{13}$$

Using the spectral decomposition for $\mathcal{P}_\Gamma$ we can estimate this last term as

$$|||\mathbf{1}_{I\setminus J}\mathcal{P}_\Gamma^{j-1}f\,||| = |||R^{j-1}f||| \le c\gamma^{j-1}|||f|||, \tag{14}$$

where we used the same arguments used in Proposition 1. Putting together estimates (13) and (14), we finally obtain that sum (11) indeed converges in the norm $|||\cdot|||$ and, moreover, we have that

$$\left|\left|\left|\sum_{j=1}^{+\infty} \mathbf{1}_J \mathcal{P}_\Gamma\left(\mathbf{1}_{I\setminus J}\mathcal{P}_\Gamma^{j-1} f\right)\right|\right|\right| \leq \left(2+\frac{2}{|I|}\right)\frac{|||\mathcal{P}_\Gamma|||c}{1-\gamma}|||f|||\,,$$

which yields the thesis. □

From the previous proposition and from Proposition 1 we can argue that the composed quantized feedback $k(x)$ defined in (9) is always almost (I,K)-stabilizing. The previous result can also be used to obtain an estimate of the expected entrance time $\mathbf{T}(k)$. Let $T_{(I,J)}(x)$ for k_1, and let $T_{(J,K)}(x)$ be the first entrance time function for k_2. It is clear that the first entrance time function $T_{(I,K)}(x)$ of the quantized feedback k is given by

$$T_{(I,K)}(x) = T_{(I,J)}(x) + T_{(J,K)}\left(\Gamma_1^{T_{(I,J)}(x)}(x)\right).$$

This implies that

$$\begin{aligned}\mathbb{E}_f(T) &= \int_J T_{(I,J)}(x)f(x)dx + \int_J T_{(J,K)}\left(\Gamma_1^{T_{(I,J)}(x)}(x)\right)f(x)dx\\ &= \mathbb{E}_f(T_{(I,J)}) + \mathbb{E}_h(T_{(J,K)}),\end{aligned}$$

where h is the probability density on J obtained form f as shown in the previous proposition.

This shows a way to estimate the expected entrance time of $k(x)$. As far as the number of quantization intervals is concerned, it is clear that we have simply that $\mathbf{N}(k) = \mathbf{N}(k_1) + \mathbf{N}(k_2)$. Finally, the contraction rate of the overall quantized feedback is the product of the contraction rates of the component quantized feedbacks, i.e., $C(k) = C(k_1)C(k_2)$.

The previous considerations can be used to obtain a class of (almost) stabilizing quantized feedbacks starting from one. Indeed, assume that $k(x)$ is an (almost) (I,J)-stabilizing quantized feedback with contraction rate $C(k) = C$, $\mathbf{N}(k)$ quantization intervals, and expected entrance time $\mathbf{T}(k)$. Let

$$F(x) := \frac{x}{C} + \beta$$

be an affine map such that $J = F(I)$. It is clear that the quantized feedback

$$F \circ k \circ F^{-1} \;:\; F(I) \to F(I)$$

is (almost) $(F(I), F^2(I))$-stabilizing. Observe that the corresponding closed loop map is $F \circ \Gamma \circ F^{-1}$. The same construction can be iterated, obtaining for every $i = 0, 1, \ldots, \tau - 1$ the quantized feedback $F^i \circ k \circ F^{-i}$ which is (almost) $(F^i(I), F^{i+1}(I))$-stabilizing. The quantized feedback defined by

$$k^{(\tau)}(x) := F^i \circ k \circ F^{-i}(x) \quad \text{if} \quad x \in F^i(I) \setminus F^{i+1}(I)$$

will be (almost) $(I, F^\tau(I))$-stabilizing with the contraction rate $C(k^{(\tau)}) = C(k)^\tau$ and $\mathbf{N}(k^{(\tau)}) = \tau\mathbf{N}(k)$ quantization intervals. As far as the expected entrance time $\mathbf{T}(k^{(\tau)})$

is concerned, it is difficult in general to estimate its dependence on the number τ of iterations.

Consider the map

$$\Psi : I \to I : x \mapsto F^{-1} \circ \Gamma^{T_{(I,J)}(x)}(x), \tag{15}$$

where $T_{(I,J)}(x)$ is the first entrance time function for k. It follows from Proposition 2 that Ψ transforms absolutely continuous measures into themselves so that also in this case we can consider the associated Perron–Frobenius operator

$$\mathcal{P}_\Psi : L^1(I) \to L^1(I).$$

It is easy to see that

$$\mathcal{P}_\Psi f = C^{-1}(h \circ F),$$

where h is the density which is obtained from f as shown in (11).

It is clear from the previous considerations that

$$\mathbf{T}\big(k^{(\tau)}\big) = \sum_{i=0}^{\tau-1} \mathbb{E}_{\mathcal{P}_\Psi^i f}\big(T_{(I,J)}\big), \tag{16}$$

where f is the uniform probability density on I. From Propositions 2 and 1 we obtain

$$\mathbf{T}\big(k^{(\tau)}\big) \le \sum_{i=0}^{\tau-1} K|||\mathcal{P}_\Psi^i f||| \le \sum_{i=0}^{\tau-1} K H^i |||f||| \le \frac{K}{H-1} H^\tau |||f|||. \tag{17}$$

This is not a very good estimate, since we expect that in many situations the growth should be linear in τ. For instance, if the uniform density on I is invariant, then we have that $\mathbf{T}(k^{(\tau)}) = \tau \mathbf{T}(k)$. In this case from a triple $(C, N, T) \in \mathcal{A}$ we can obtain a sequence of triples $(C^\tau, \tau N, \tau T) \in \mathcal{A}$ for all $\tau \in \mathbb{N}$. This method will be used in the following subsections to obtain three specific quantized feedback strategies.

In general we cannot guarantee that Ψ will possess invariant probability densities, and it seems to be very difficult to obtain estimates which are better than (17). Notice that indeed Ψ is also a piecewise affine map but in general with an infinite number of continuity intervals. For this type of maps the theory is quite weak: invariant probability densities are not guaranteed to exist, and we may lose the spectral structure of the corresponding Perron–Frobenius operator we had in the finite case (see [4] for more details). There is, however, a case in which things go smooth, namely, when $T(x)$ is bounded. In this case Ψ is an expanding piecewise affine map with only a finite number of continuity intervals, and in this case invariant densities do exist and the Perron–Frobenius operator $\mathcal{P}_\Psi$ admits the usual spectral decomposition (8). In this case we have the following result.

PROPOSITION 3. *Assume that $T(x)$ is bounded. Then, there exist a probability density $\overline{f}$ and a bounded sequence $\{a_\tau\}$ such that*

$$\mathbf{T}\big(k^{(\tau)}\big) = \tau \mathbb{E}_{\overline{f}}(T) + a_\tau. \tag{18}$$

Proof. Let f be the uniform probability density on I. Moreover, let $\nu \in \mathbb{N}$ be such that $\lambda^\nu = 1$ for every $\lambda \in \sigma_1$, and let $Q = \sum_\lambda Q_\lambda$. Observe that for all $j \in \mathbb{N}$

we have $Q\mathcal{P}_\Psi^j = \mathcal{P}_\Psi^j - R^j$ and that $Q\mathcal{P}_\Psi^\nu = Q$. This implies that if we decompose $j = l\nu + r$, with $l \in \mathbb{N}$ and $r \in \{0, \dots, \nu - 1\}$, we have that

$$\mathcal{P}_\Psi^j = Q\mathcal{P}_\Psi^r + R^j.$$

Define

$$\overline{f} = Q\left(\frac{1}{\nu}\sum_{j=0}^{\nu-1}\mathcal{P}_\Psi^j f\right).$$

Then, if $\tau - 1 = l\nu + r$, we have that

$$\begin{aligned}\sum_{j=0}^{\tau-1}\mathcal{P}_\Psi^j f - \tau\overline{f} &= lQ\left(\sum_{j=0}^{\nu-1}\mathcal{P}_\Psi^j f\right) + Q\left(\sum_{j=0}^{r}\mathcal{P}_\Psi^j f\right) + \sum_{j=0}^{\tau-1}R^j f - \tau\overline{f}\\ &= (l\nu - \tau)\overline{f} + Q\left(\sum_{j=0}^{r}\mathcal{P}_\Psi^j f\right) + \sum_{j=0}^{\tau-1}R^j f.\end{aligned}$$

Notice that

$$\begin{aligned}\left|\left|\left|(l\nu - \tau)\overline{f} + Q\left(\sum_{j=0}^{r}\mathcal{P}_\Psi^j f\right) + \sum_{j=0}^{\tau-1}R^j f\right|\right|\right| &\le \nu|||\overline{f}|||\\ + |||Q|||\left(\sum_{j=0}^{\nu-1}|||\mathcal{P}_\Psi^j|||\right)|||f||| &+ \frac{C}{1-\gamma}|||f|||\end{aligned}$$

is bounded in τ. Observe finally that

$$\begin{aligned}|\mathbf{T}(k^{(\tau)}) - \tau\mathbb{E}_{\overline{f}}(T)| &= \left|\int_I T(x)\left(\sum_{j=0}^{\tau-1}\mathcal{P}_\Psi^j f(x) - \tau\overline{f}(x)\right)dx\right|\\ &\le \int_I T(x)\left|\sum_{j=0}^{\tau-1}\mathcal{P}_\Psi^j f(x) - \tau\overline{f}(x)\right|dx.\end{aligned}$$

The result now follows by applying Proposition 1. □

This has the following consequence. If the triple (C, N, T) is in $\mathcal{A}$ and corresponds to a situation in which the entrance time function is bounded, then we can obtain a sequence of triples $(C^\tau, \tau N, \tau\overline{T} + a_\tau) \in \mathcal{A}$ for all $\tau \in \mathbb{N}$, where $\overline{T}$ is the expected entrance time with respect to a suitable probability density, and $\{a_\tau\}$ is a bounded sequence.

4. Three stabilizing quantized feedback strategies. The method presented in the previous section will be used in the following subsections to obtain three specific quantized feedback strategies. In what follows we assume for simplicity that $I = [-1, 1]$ and $J = [\epsilon, \epsilon]$ with $\epsilon \le 1$, and so we have that $C = 1/\epsilon$. In this section we will simply write $C, \mathbf{N}, \mathbf{T}$ dropping the explicit dependence from k. All probabilistic considerations in this section will be carried on with respect to the uniform probability on I.

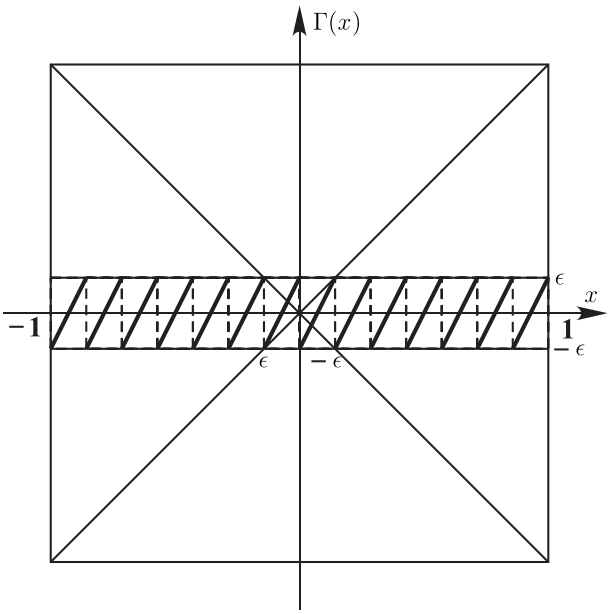


FIG. 1. *The map* Γ *associated with the quantized feedback defined in* (19).

4.1. Deadbeat quantized feedback strategy. The first strategy, which has been analyzed in detail by Delchamps in [6], consists of approximating the one-step deadbeat controller $k(x) := -ax$ by its quantized version, i.e., by a uniform quantized function $k(x)$ such that $-ax - \epsilon \le k(x) \le -ax + \epsilon$. One possibility is to take

$$k(x) := -(2h+1)\epsilon \quad \text{for} \quad h\frac{2\epsilon}{a} < x \le (h+1)\frac{2\epsilon}{a}, \tag{19}$$

which yields the closed loop map $\Gamma(x)$ illustrated in Figure 1.

This controller drives any state belonging to I into J in one step. In this case we have that

$$\mathbf{N} = 2\left\lceil |a|\frac{C-1}{2} \right\rceil \sim |a|C$$

and that

$$\mathbf{T} = \sum_{n=1}^{\infty} \mathbb{P}[T_J \ge n] = \mathbb{P}[T_J \ge 1] = 1 - \mathbb{P}[J] = 1 - 1/C,$$

where $f(C) \sim g(C)$ means that $f(C)/g(C)$ tends to 1 as $C \to \infty$.

Using the nesting strategy presented above we can construct a τ-step deadbeat quantized feedback simply iterating the one-step deadbeat quantized feedback. We only need to pay attention to the fact that the uniform density in I is invariant with respect to the map Ψ defined in (15). This happens if $|a|(C-1)/2$ is an integer. Assume that this is the case, and denote it by n. We obtain a triple contraction rate, quantization intervals, and expected entrance time equal to

$$\left(\frac{2n+|a|}{|a|}, 2n, \frac{2n}{2n+|a|}\right) \in \mathcal{A}.$$

Using the strategy presented above, we can iterate the construction τ times, obtaining in this way a sequence of triples

$$\left(\left(\frac{2n+|a|}{|a|}\right)^{\tau}, 2\tau n, \tau\frac{2n}{2n+|a|}\right) \in \mathcal{A}, \qquad n,\tau \in \mathbb{N},$$

which provides a family of quantized feedbacks parameterized by the two integers τ and n. We are mainly interested in understanding what asymptotic behavior can be obtained of $\mathbf{N}$ and $\mathbf{T}$ as $C \to \infty$. To this aim observe that

$$\frac{\mathbf{N}/|a|}{\mathbf{T}C^{1/\mathbf{T}}} = \left(\frac{2n+|a|}{|a|}\right)^{-\frac{|a|}{2n}} \in [1/e, 1].$$

Making the change of variable

$$C = \left(\frac{2n+|a|}{|a|}\right)^{\tau}, \qquad n = \frac{|a|}{2}\big(C^{\frac{1}{\tau}} - 1\big), \tag{20}$$

we obtain

$$\begin{aligned} \mathbf{N}/|a| &= \tau\big(C^{\frac{1}{\tau}} - 1\big), \\ \mathbf{T} &= \tau\big(1 - C^{-\frac{1}{\tau}}\big), \end{aligned}$$

where τ is any function of C that by (20) can be chosen arbitrarily subject to the fact that $\tau(C)/\log C$ is bounded from above. If in particular τ is fixed, we obtain

$$\begin{aligned} \mathbf{N}/|a| &\sim \tau C^{\frac{1}{\tau}}, \\ \mathbf{T} &\sim \tau. \end{aligned}$$

If instead we think of τ as a possible function of C, we can distinguish two different patterns of behavior: the case when $\tau(C)/\log C \to 0$ and the case when $\tau(C) \sim K\log C$. In the first case we have that

$$\mathbf{N}/|a| \sim \mathbf{T}C^{1/\mathbf{T}}, \tag{21}$$

and moreover $\mathbf{N}/\log C \to \infty$, namely, we have a superlogarithmic growth of the number of quantization intervals, while the expected entrance time has a sublogarithmic growth $\mathbf{T}/\log C \to 0$. In the second situation when $\tau(C) \sim K\log C$ we have that both $\mathbf{N}$ and $\mathbf{T}$ grow logarithmically in C. More precisely, we have that

$$\begin{aligned} \mathbf{N}/|a| &\sim K\big(e^{1/K} - 1\big)\log C, \\ \mathbf{T} &\sim K\big(1 - e^{-1/K}\big)\log C. \end{aligned} \tag{22}$$

4.2. Logarithmic quantized feedback strategy. The second strategy is based on the quantized feedback (we assume $a > 0$, the case $a < 0$ being completely analogous)

$$k(x) = \begin{cases} -a+1 & \text{if } \epsilon \le x \le 1, \\ +a-1 & \text{if } -1 \le x \le -\epsilon, \end{cases} \tag{23}$$

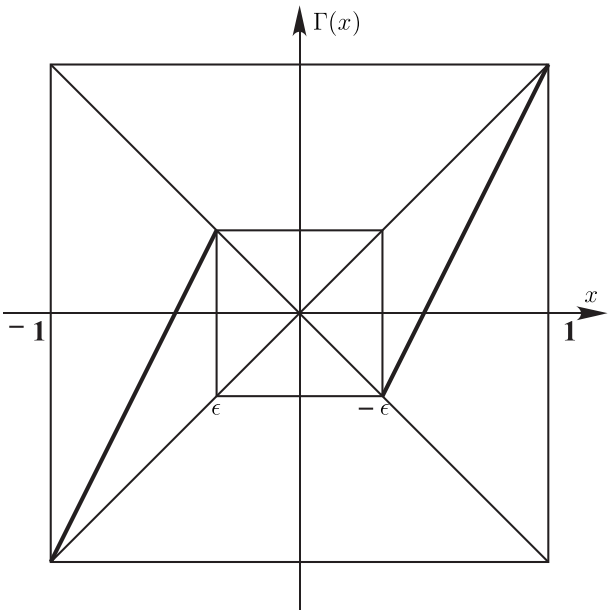


FIG. 2. *The map* Γ *associated with the quantized feedback defined in* (23).

where

$$\epsilon = \frac{a-1}{a+1}.$$

In this way we obtain an almost (I, J)-stabilizing quantized feedback, where $I = [-1, 1]$ and $J = [-\epsilon, \epsilon]$. The graph of closed loop map $\Gamma(x)$ is illustrated in Figure 2.

In this case we have a contraction rate $1/\epsilon$ and two quantization intervals. The expected entrance time can be found by noticing that

$$\Gamma^{-n}(I \setminus J) = [-1, -\epsilon_n] \cup [\epsilon_n, 1],$$

where $\epsilon_n = 1 - 2/(a+1)a^n$, which implies that the expected entrance time is

$$\sum_{n=0}^{\infty} \mathbb{P}[T_{(I,J)} > n] = \sum_{n=0}^{\infty} \mathbb{P}[\Gamma^{-n}(I \setminus J)] = \frac{2}{a+1} \sum_{n=0}^{\infty} a^{-n} = \frac{2a}{a^2-1}.$$

In general, when we do not restrict to positive a, we obtain a triple contraction rate, quantization intervals, the expected entrance time equal to

$$\left(\frac{|a|-1}{|a|+1}, 2, \frac{2|a|}{|a|^2-1} \right) \in \mathcal{A}.$$

Using the strategy presented above, we can iterate the construction τ times. In this case it is less obvious to show that the Lebesgue measure is invariant with respect to the map Ψ defined from Γ as in (15). To show this, observe preliminarily that if we assume that $\Gamma(x) = x$ for all $x \in J$, then

$$\lim_{n\to\infty} \Gamma^n(x) = \Gamma^{T_{(I,J)}(x)}(x) \qquad \text{for almost all } x \in I,$$

which implies that $\Gamma^n(x)$ converges to $\Gamma^{T_{(I,J)}(x)}(x)$ in distribution. Observe moreover that if the density function f_n of the random variable $\Gamma^n(x)$ is of the form

$$f_n(a) = \begin{cases} \alpha_n & \text{if } a \in J, \\ \beta_n & \text{if } a \in I \setminus J, \end{cases}$$

then also f_{n+1} has the same structure with $\alpha_{n+1} = 2\beta_n/|a| + \alpha_n$, and $\beta_{n+1} = \beta_n/|a|$. This implies that

$$\lim_{n\to\infty} f_n(a) = \begin{cases} 1/\epsilon & \text{if } a \in I_1, \\ 0 & \text{if } a \in I_0 \setminus I_1 \end{cases}$$

from which we can argue that the Lebesgue measure is invariant with respect to the map Ψ.

These facts allow us to obtain a sequence of triples

$$\left(\left(\frac{|a|+1}{|a|-1} \right)^\tau, 2\tau, \frac{2|a|}{|a|^2-1}\tau \right) \in \mathcal{A}, \qquad \tau \in \mathbb{N}.$$

Making the change of variable

$$C = \left(\frac{|a|+1}{|a|-1} \right)^\tau, \qquad \tau = \frac{\log C}{\log(|a|+1) - \log(|a|-1)},$$

we obtain

$$\mathbf{N}/|a| = \frac{2}{|a|} \; \frac{\log C}{\log(|a|+1) - \log(|a|-1)},$$

$$\mathbf{T} = \frac{2|a|}{|a|^2-1} \; \frac{\log C}{\log(|a|+1) - \log(|a|-1)}.$$

These expressions motivate the fact that this quantized feedback is called a logarithmic quantizer. The strategy obtained in this way coincides with the one proposed in [7, 9] which yields a Lyapunov stability.

4.3. Chaotic quantized feedback strategy. In [9] another possible quantized feedback yielding almost stability has been proposed. This control strategy exploits the chaotic behavior of the state evolution inside $I = [-1, 1]$ produced by the feedback map

$$k_0(x) := -(2h+1) \quad \text{for} \quad \frac{2}{a}h < x \le \frac{2}{a}(h+1), \tag{24}$$

when we have that $|a| \ge 2$. In this way we have that, for almost every initial condition x_0, the state evolution x_t is maintained inside the interval I and is dense in this interval. For this reason x_t will visit the interval $J = [-\epsilon, \epsilon]$. Therefore, if we modify this feedback map in J as follows:

$$k(x) = \begin{cases} k_0(x) & \text{if } x \in I \setminus J, \\ k_1(x) & \text{if } x \in J, \end{cases} \tag{25}$$

where $k_1(x)$ is any quantized feedback making J invariant (take for instance $k_1(x) = \epsilon k_0(x/\epsilon)$), we obtain that the state will move chaotically inside I until it enters the

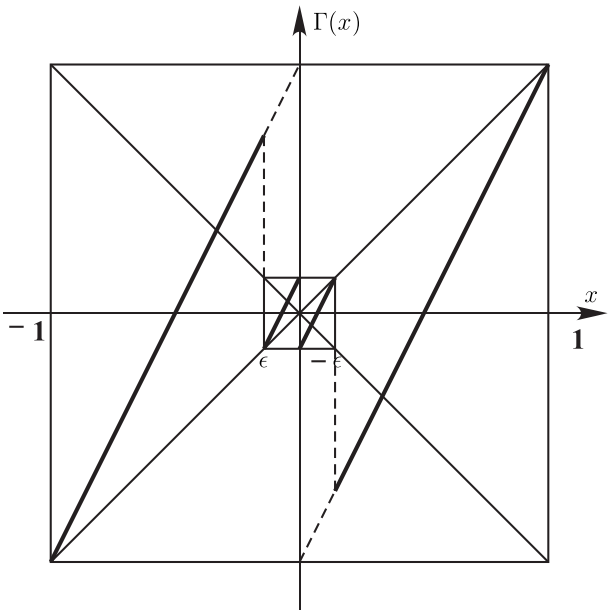


FIG. 3. *The map* Γ *associated with the quantized feedback defined in* (25).

interval J, and there it will be entrapped. In this way we obtain a feedback map requiring

$$\mathbf{N} = \lceil |a| \rceil$$

quantization intervals. The closed loop map $\Gamma(x)$ is shown in Figure 3 in the case $a = 2$.

In this case the evaluation of the expected entrance time can be done using Markov chain techniques. Assume that $\epsilon = 2^{-n}$. It is clear that, for evaluating the expected entrance time, we can refer to the system with feedback $k_0(x)$. Define the sets $I_i := [-i2^{-n}, -(i-1)2^{-n}] \cup [(i-1)2^{-n}, i2^{-n}]$, $i = 1, \ldots, 2^n$. In this way we have that $J = I_1$. Assuming that the initial state x_0 is uniformly distributed in I, we can argue that

$$\mathbb{P}[x_0 \in I_i] = 2^{-n}.$$

The initial distribution is described by the row vector

$$\pi := 2^{-n} \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \end{bmatrix} \in \mathbb{R}^{1 \times 2^n}.$$

Assuming that the iterated state x_t is uniformly distributed in each quantization interval I_i, the structure of the closed loop map $\Gamma_0(x) = ax + k_0(x)$ ensures also that the updated state $x_{t+1} = \Gamma_0(x_t)$ will be uniformly distributed in each quantization interval. Moreover, we have that

$$\mathbb{P}[x_{t+1} \in I_j | x_t \in I_i] = \Pi_{ij},$$

where Π_{ij} is the i,j-element of the matrix

$$\Pi = \frac{1}{2}\begin{bmatrix} 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & \cdots & 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & 1 & \cdots & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & 1 \end{bmatrix} \in \mathbb{R}^{2^n \times 2^n}.$$

Then (see [13]) the expected first entrance time in the state 1 is given by the formula

$$\mathbf{T} = \mathbb{E}\big(T_{(I,J)}\big) = \frac{d}{dz} w(z)_{|z=1},$$

where

$$w(z) := \frac{\pi\Pi(z)e_1}{e_1^T\Pi(z)e_1}$$

and where $\Pi(z) := \sum_{n\geq 0} \Pi^n z^n$ and $e_1 := [1\ 0\ \cdots\ 0]^T$. Since $\pi\Pi = \pi$, then

$$\pi\Pi(z) := \frac{1}{1-z}\pi.$$

It can be seen that

$$e_1^T\Pi(z)e_1 = 1 + 2^{-n}\frac{z^n}{1-z},$$

obtaining in this way

$$w(z) = \frac{1}{z^n + (1-z)2^n}$$

and

$$\mathbf{T} = \frac{d}{dz} w(z)_{|z=1} = 2^n - n.$$

In this way we obtained the triple

$$(2^n, 2, 2^n - n) \in \mathcal{A}.$$

Using the strategy presented above we can iterate this construction τ times. It can be shown that in this case also the Lebesgue measure is invariant with respect to the closed map Ψ defined from Γ as in (15). To show this we use the same kind of reasoning used in the previous subsection. Again, by defining Γ in such a way that $\Gamma(x) = x$ for all $x \in J$, we have that the random variable $\Gamma^n(x)$ converges to $\Gamma^{T_{(I,J)}(x)}(x)$ in distribution. Observe moreover that if the density function f_n of the random variable $\Gamma^n(x)$ is constant in each quantization interval I_i, then it can be

shown that f_{n+1} has also the same property. This implies that the limit density will also be a function which is constant in each set I_i and particularly in J. From this we can argue that the Lebesgue measure is invariant with respect to the map Ψ. These facts allow us to obtain a sequence of triples

$$(2^{\tau n}, \tau 2, \tau 2^n - \tau n) \in \mathcal{A}, \qquad n, \tau \in \mathbb{N}.$$

The previous reasoning can be extended to any situation in which $|a|$ is an integer. In this case we can obtain the sequence of triples

$$(|a|^{\tau n}, \tau |a|, \tau |a|^n - \tau n) \in \mathcal{A}, \qquad n, \tau \in \mathbb{N},$$

which provides a family of quantized feedbacks parameterized by the two integers τ, n. We are mainly interested in understanding what asymptotic behavior can be obtained for $\mathbf{N}$ and $\mathbf{T}$ as $C \to \infty$. To this aim observe that

$$\frac{\mathbf{T}}{\frac{\mathbf{N}}{|a|} C^{\frac{|a|}{\mathbf{N}}}} = 1 - \frac{n}{|a|^n} \in \left[1 - \frac{1}{e \log |a|}, 1\right].$$

Making the change of variable

$$C = |a|^{\tau n}, \qquad n = \frac{\log C}{\tau \log |a|}, \tag{26}$$

we obtain that

$$\begin{aligned} \mathbf{N}/|a| &= \tau, \\ \mathbf{T} &= \tau C^{\frac{1}{\tau}} - \frac{\log C}{\log |a|}, \end{aligned}$$

where τ is any function of C that, by (26), can be chosen arbitrarily subject to the fact that $\tau(C)/\log C$ is bounded from above. If in particular τ is fixed, we obtain

$$\begin{aligned} \mathbf{N}/|a| &= \tau, \\ \mathbf{T} &\sim \tau C^{\frac{1}{\tau}}. \end{aligned}$$

If instead we think of τ as a possible function of C, we can distinguish the case when $\tau(C)/\log C \to 0$ and the case when $\tau(C) \sim K \log C$. In the first case we have that

$$\mathbf{T} \sim \frac{\mathbf{N}}{|a|} C^{\frac{|a|}{\mathbf{N}}}, \tag{27}$$

and moreover $\mathbf{N}/\log C \to 0$, namely, a sublogarithmic growth of the number of quantization intervals, while the expected entrance time has a superlogarithmic growth $\mathbf{T}/\log C \to \infty$. In the second situation when $\tau(C) \sim K \log C$ we have that both $\mathbf{N}$ and $\mathbf{T}$ grow logarithmically in C. More precisely, we have that

$$\begin{aligned} \mathbf{N}/|a| &= K \log C, \\ \mathbf{T} &= \left(K e^{1/K} - \frac{1}{\log |a|}\right) \log C. \end{aligned} \tag{28}$$

Chaotic stabilizers can also be considered for nonintegers slopes a. Some preliminary results in this case have been obtained in [9]. In [8] the following more refined result is proved.

THEOREM 1. *Let a be such that $|a| > 2$, $I = [-1, 1]$, and $J = [-\epsilon, \epsilon]$, where $0 < \epsilon < 1$. There exists an almost (I, J)-stabilizing quantized feedback $k : I \to \mathbb{R}$ such that*

$$\mathbf{N} = \lceil |a| \rceil + 1,$$
$$\mathbf{T} \leq KC,$$

where K is a positive constant only depending on a.

Remark. The following table summarizes the properties of the different quantized feedback strategies.

	$\mathbf{N}/\lvert a\rvert$	$\mathbf{T}$
τ-step deadbeat quantizer	$\tau C^{\frac{1}{\tau}}$	τ
Logarithmic quantizer	$\frac{2}{\lvert a\rvert} \frac{\log C}{\log(\lvert a\rvert - 1) - \log(\lvert a\rvert + 1)}$	$\frac{2\lvert a\rvert}{\lvert a\rvert^2 - 1} \frac{\log C}{\log(\lvert a\rvert - 1) - \log(\lvert a\rvert + 1)}$
τ-step chaotic quantizer	τ	$\tau C^{\frac{1}{\tau}}$

This table highlights the relations between the parameters $|a|$, N, C, and T. In all cases it is possible to see that the steady state performance parameter C and transient performance parameter T are conflicting; namely, for fixed $|a|$ and N, an increasing value of C implies an increasing value of T and vice versa. Moreover, both the performance parameters are improved when N is increased and are worsened when $|a|$ is increased. A qualitative description of the relations between the parameters $|a|$, N, C, and T is given in Figure 4.

This suggests that looking for the stabilizing quantized feedback with minimal quantization intervals is a rather naive approach to the quantized control problem. Indeed, in case we do not consider the transient performance described by the parameter T, the optimal strategy would be clearly the chaos-based one. However, this provides only a partial view of the problem, since in fact the different strategies provide closed loop systems with different trade-off relations between the performance parameters T and C.

5. Bounds of the performance of a quantized feedback system. In this section we will present some general bounds involving the parameters $C(\Gamma)$, $\mathbf{N}(\Gamma)$, and $\mathbf{T}(\Gamma)$. These will be obtained by means of a symbolic representation of the dynamical system and using basic combinatorial arguments.

5.1. Symbolic descriptions of the dynamical system. Let $\Gamma : I \to I$ be a piecewise affine map with a fixed slope a. Let $J \subseteq I$ be another almost invariant

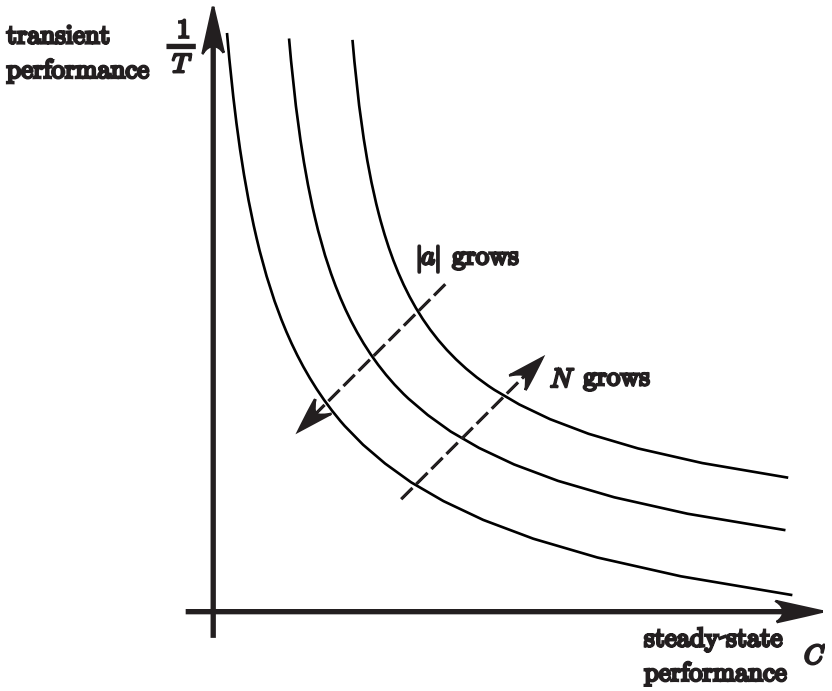


FIG. 4. *The qualitative relations between the parameters $|a|$, N, C, and T. The parameter C describes the steady state performance, the parameter $1/T$ describes the transient performance, the curves describe the trade-off between these two parameters for fixed N and $|a|$. Different curves refer to different values of N and $|a|$.*

interval. We can write

$$J = \overline{J_1 \cup J_2 \cup \cdots \cup J_M}, \qquad I = \overline{I_1 \cup I_2 \cup \cdots \cup I_N} \cup J,$$

where the I_h's and the J_l's are disjoint open intervals on which Γ is affine. In what follows, we will use the shorthand notation $C = C(\Gamma)$, $\mathbf{N} = \mathbf{N}(\Gamma)$, and $\mathbf{T} = \mathbf{T}(\Gamma)$. In this section, we will always consider Γ defined on the set Ω as defined in (4). Define the finite sets

$$\mathcal{I} = \{I_1, I_2, \ldots, I_{\mathbf{N}}\}, \quad \mathcal{J} = \{J_1, J_2, \ldots, J_M\},$$

and define a map $\psi : \Omega \to (\mathcal{I} \cup \mathcal{J})^{\mathbb{N}}$ by

$$\psi(x)_n = \omega_n \ \text{ if } \ \Gamma^n(x) \in \omega_n \,. \tag{29}$$

Notice that the above map is well defined by the way in which Ω has been defined. Consider the language $\Sigma_*(\Gamma)$ over the alphabet $\mathcal{I} \cup \mathcal{J}$ which is the subset of $(\mathcal{I} \cup \mathcal{J})^*$ consisting of all the finite words appearing in the infinite sequences in the range of ψ. If $|a| > 1$, then Γ locally expands and, as a consequence, the map ψ is injective. Indeed, it follows from (29) that $x \in \omega_0 \cap \cdots \cap \Gamma^{-n}\omega_n$ for every n. On the other hand it follows from the simple bound (33) that the length of this interval goes to 0 for $n \to +\infty$, which yields injectivity. This implies that all the dynamical and statistical properties of the map Γ can be read out from the language $\Sigma_*(\Gamma)$. Notice, for further use, the following properties of simple verification.

1. $\omega_0\omega_1\cdots\omega_n \in \Sigma_*(\Gamma)$ if and only if $\omega_0 \cap \Gamma^{-1}\omega_1 \cap \cdots \cap \Gamma^{-n}\omega_n \neq \emptyset$.
2. For all $\omega_0\omega_1\cdots\omega_n \in \Sigma_*(\Gamma)$ the map Γ^{n+1} is affine on the interval $\omega_0 \cap \Gamma^{-1}\omega_1 \cap \cdots \cap \Gamma^{-n}\omega_n$.
3. If $\omega_0\omega_1\cdots\omega_n$ and $\nu_0\nu_1\cdots\nu_m$ are in $\Sigma_*(\Gamma)$ and none of the two happens to be the initial subword of the other, then the two intervals $\omega_0 \cap \Gamma^{-1}\omega_1 \cap \cdots \cap \Gamma^{-n}\omega_n$ and $\nu_0 \cap \Gamma^{-1}\nu_1 \cap \cdots \cap \Gamma^{-m}\nu_m$ are disjoint.

As we mentioned above, language $\Sigma_*(\Gamma)$ contains all the dynamical and statistical properties of the map Γ. In particular this is true for the expected entrance time. Indeed, as the following lemma shows, the expected entrance time can be estimated by knowing how the number of words in $\Sigma_*(\Gamma)$ grows with respect to their length. More precisely, denote by γ_n the number of distinct words in sublanguage $\Sigma_*(\Gamma) \cap \mathcal{I}^*$ of length n, i.e.,

$$\gamma_n := \#\big\{\omega_0\omega_1\cdots\omega_{n-2}\omega_{n-1} \in \Sigma_*(\Gamma) \cap \mathcal{I}^*\big\}. \tag{30}$$

Then we have the following result.

LEMMA 2. *Given any $n \in \mathbb{N}$ we have that*

$$\mathbb{P}\big[T_{(I,J)} = n\big] \leq \mathbb{P}[J]\frac{\gamma_n}{|a|^n}, \tag{31}$$

$$\mathbb{P}\big[T_{(I,J)} \geq n\big] \geq \mathbb{P}[I \setminus J] - \mathbb{P}[J]\sum_{k=1}^{n-1}\frac{\gamma_k}{|a|^k}. \tag{32}$$

Proof. As mentioned above, the family of intervals of the form

$$\omega_0 \cap \Gamma^{-1}(\omega_1) \cap \cdots \cap \Gamma^{-(n-1)}(\omega_{n-1}) \cap \Gamma^{-n}(\omega_n), \qquad \omega_0, \ldots, \omega_{n-1} \in \mathcal{I}, \omega_n \in \mathcal{J}$$

constitutes a partition of the set of points of I which end inside J in exactly n steps. Moreover, since Γ^n is affine on each of these intervals, it follows that

$$\mathbb{P}\big[\omega_0 \cap \Gamma^{-1}(\omega_1) \cap \cdots \cap \Gamma^{-(n-1)}(\omega_{n-1}) \cap \Gamma^{-n}(\omega_n)\big] \leq \frac{\mathbb{P}[J]}{|a|^n}. \tag{33}$$

Therefore, if we let

$$\tilde{\gamma}_n := \#\{\omega_0\omega_1\cdots\omega_{n-2}\omega_{n-1} \in \Sigma_*(\Gamma) \mid \omega_0\omega_1\cdots\omega_{n-2}\omega_{n-1} \in \mathcal{I}^* \text{ and } \omega_{n-1} \in \mathcal{J}\}\ ,$$

we can argue that

$$\mathbb{P}[T_{(I,J)} = n] \leq \mathbb{P}[J]\frac{\tilde{\gamma}_{n+1}}{|a|^n} \leq \mathbb{P}[J]\frac{\gamma_n}{|a|^n}\ ,$$

where we used the fact that for all $n \geq 1$ we have that $\tilde{\gamma}_{n+1} \leq \gamma_n$.

We prove now the second assertion by induction on n. The assertion is trivial for $n = 1$. Assume by induction that the assertion holds for n, and let us prove it for $n + 1$. We can now write

$$\begin{aligned}\mathbb{P}\big[T_{(I,J)} \geq n+1\big] &= \mathbb{P}\big[T_{(I,J)} \geq n\big] - \mathbb{P}\big[T_{(I,J)} = n\big] \geq \mathbb{P}\big[T_{(I,J)} \geq n\big] - \mathbb{P}[J]\frac{\gamma_n}{|a|^n} \\ &\geq \mathbb{P}[I \setminus J] - \mathbb{P}[J]\sum_{k=1}^{n-1}\frac{\gamma_k}{|a|^k} - \mathbb{P}[J]\frac{\gamma_n}{|a|^n} \\ &= \mathbb{P}[I \setminus J] - \mathbb{P}[J]\sum_{k=1}^{n}\frac{\gamma_k}{|a|^k}. \qquad \square\end{aligned}$$

Notice that $\mathbb{P}[J] = C^{-1}$. This implies that formula (32) can be rewritten as

$$\mathbb{P}\big[T_{(I,J)} \geq n\big] \geq 1 - C^{-1} - C^{-1} \sum_{k=1}^{n-1} \frac{\gamma_k}{|a|^k} \tag{34}$$

from which we can argue that for any arbitrarily fixed $\overline{n} \in \mathbb{N}$ we have that

$$\begin{aligned} \mathbf{T} &= \mathbb{E}\big(T_{(I,J)}\big) \\ &= \sum_{n=1}^{+\infty} \mathbb{P}\big[T_{(I,J)} \geq n\big] \geq \sum_{n=1}^{\overline{n}} \mathbb{P}[T_{(I,J)} \geq n] \geq \overline{n}(1 - C^{-1}) - C^{-1} \sum_{n=1}^{\overline{n}} \sum_{k=1}^{n-1} \frac{\gamma_k}{|a|^k}. \end{aligned} \tag{35}$$

If we can establish upper bounds on γ_k, through (35), we can thus achieve lower bounds on $\mathbf{T}$. The following theorem provides the most relevant contribution of this paper, since it presents a bound on the growth of γ_k depending on the number of quantization intervals $\mathbf{N}$. The proof of this theorem is very long, and it will be presented in the last section.

THEOREM 2. *Assume that $|a| > 2$. Then*

$$\frac{\gamma_k}{|a|^k} \leq 2 \left[\sum_{s=1}^{r \wedge k} \binom{k-1}{s-1} \binom{r}{s} \left(\frac{s}{r}\right)^s \right] \left(\frac{\mathbf{N}M}{k \wedge \frac{\mathbf{N}M}{e}} \right)^{k \wedge \frac{\mathbf{N}M}{e}} \qquad \forall k \geq 1, \tag{36}$$

where $r \in \{1, \ldots, \mathbf{N}\}$ is independent of k, but may depend on the specific system, while M depends only on $|a|$.

Remark. In symbolic dynamics [17] the set $\overline{\Psi(\Omega)}$ (where the closure is to be intended in the product topology of $(\mathcal{I} \cup \mathcal{J})^{\mathbb{N}}$) is called shift. It can be shown that its topological entropy is $\log |a|$. As a consequence, for every $\epsilon > 0$, there exists $M_\epsilon > 0$ such that

$$\gamma_k \leq M_\epsilon (|a| + \epsilon)^k. \tag{37}$$

This type of estimate is of no use for our purposes for two reasons: first because the geometric growth rate $|a| + \epsilon$ causes a too quick growth in the double summation in (35), making impossible any useful estimate, and second because it is not clear how explicitly M_ϵ depends on the map Γ. In fact, estimate (36) is uniform with respect to all the possible piecewise affine maps having slope a and $\mathbf{N}$ quantization intervals. Notice, moreover, that for large k ($k \geq \max\{\mathbf{N}, \mathbf{N}/Me\}$), (36) can be written as

$$\gamma_k \leq (\overline{M}k)^{\mathbf{N}} |a|^k,$$

where $\overline{M}$ is a suitable constant depending only on a. This is clearly a much better estimate than (37).

Using Theorem 2 we obtain a lower bound estimate on $\mathbf{T}$.

COROLLARY 1. *For any $\overline{n} \in \mathbb{N}$ we have that*

$$\mathbf{T} \geq \overline{n}\big(1 - C^{-1}\big) - 2C^{-1} \left[\sum_{s=1}^{r \wedge \overline{n}-1} \binom{\overline{n}}{s+1} \binom{r}{s} \left(\frac{s}{r}\right)^s \right] \left(\frac{\mathbf{N}M}{\overline{n} - 1 \wedge \frac{\mathbf{N}M}{e}} \right)^{\overline{n}-1 \wedge \frac{\mathbf{N}M}{e}}. \tag{38}$$

Proof. From Theorem 2 we can argue that

$$
\begin{aligned}
(39)\qquad \sum_{k=1}^{n-1}\frac{\gamma_k}{|a|^k} &\le 2\sum_{k=1}^{n-1}\sum_{s=1}^{r\wedge k}\binom{k-1}{s-1}\binom{r}{s}\left(\frac{s}{r}\right)^s\left(\frac{\mathbf{N}M}{k\wedge\frac{\mathbf{N}M}{e}}\right)^{k\wedge\frac{\mathbf{N}M}{e}}\\
&= 2\sum_{s=1}^{r\wedge n-1}\sum_{k=s}^{n-1}\binom{k-1}{s-1}\binom{r}{s}\left(\frac{s}{r}\right)^s\left(\frac{\mathbf{N}M}{k\wedge\frac{\mathbf{N}M}{e}}\right)^{k\wedge\frac{\mathbf{N}M}{e}}\\
&\le 2\sum_{s=1}^{r\wedge n-1}\binom{r}{s}\left(\frac{s}{r}\right)^s\max_{k=s}^{n-1}\left\{\left(\frac{\mathbf{N}M}{k\wedge\frac{\mathbf{N}M}{e}}\right)^{k\wedge\frac{\mathbf{N}M}{e}}\right\}\sum_{k=s}^{n-1}\binom{k-1}{s-1}\\
&= 2\left[\sum_{s=1}^{r\wedge n-1}\binom{n-1}{s}\binom{r}{s}\left(\frac{s}{r}\right)^s\right]\left(\frac{\mathbf{N}M}{n-1\wedge\frac{\mathbf{N}M}{e}}\right)^{n-1\wedge\frac{\mathbf{N}M}{e}},
\end{aligned}
$$

where we used identity (88) and bound (93) of the appendix.

From (39) we can further obtain

$$
\begin{aligned}
\sum_{n=1}^{\overline{n}}\sum_{k=1}^{n-1}\frac{\gamma_k}{|a|^k} &\le 2\sum_{n=1}^{\overline{n}}\sum_{s=1}^{r\wedge n-1}\binom{n-1}{s}\binom{r}{s}\left(\frac{s}{r}\right)^s\left(\frac{\mathbf{N}M}{n-1\wedge\frac{\mathbf{N}M}{e}}\right)^{n-1\wedge\frac{\mathbf{N}M}{e}}\\
&= 2\sum_{s=1}^{r\wedge\overline{n}-1}\sum_{n=s+1}^{\overline{n}}\binom{n-1}{s}\binom{r}{s}\left(\frac{s}{r}\right)^s\left(\frac{\mathbf{N}M}{n-1\wedge\frac{\mathbf{N}M}{e}}\right)^{n-1\wedge\frac{\mathbf{N}M}{e}}\\
&\le 2\sum_{s=1}^{r\wedge\overline{n}-1}\binom{r}{s}\left(\frac{s}{r}\right)^s\max_{n=s+1}^{\overline{n}}\left\{\left(\frac{\mathbf{N}M}{n-1\wedge\frac{\mathbf{N}M}{e}}\right)^{n-1\wedge\frac{\mathbf{N}M}{e}}\right\}\\
&\quad\times\sum_{n=s+1}^{\overline{n}}\binom{n-1}{s}\\
&= 2\left[\sum_{s=1}^{r\wedge\overline{n}-1}\binom{\overline{n}}{s+1}\binom{r}{s}\left(\frac{s}{r}\right)^s\right]\left(\frac{\mathbf{N}M}{\overline{n}-1\wedge\frac{\mathbf{N}M}{e}}\right)^{\overline{n}-1\wedge\frac{\mathbf{N}M}{e}},
\end{aligned}
$$

where again we used identity (88) and bound (93). From this (38) follows by a simple substitution. □

In the following subsections we will exploit the previous result to obtain bounds describing the trade-off between the number of quantization intervals $\mathbf{N}$ and the expected entrance time $\mathbf{T}$ for a given almost (I, J)-stable piecewise affine map Γ with the contraction rate C. Three situations will be distinguished. First, we will consider the regime when $\mathbf{N}/\log C$ is sufficiently small. It contains the case when $\mathbf{N}/\log C \to 0$, namely, the regime of sublogarithmic growth of $\mathbf{N}$ in C. The corresponding expected entrance time $\mathbf{T}$ will exhibit a superlogarithmic growth in C. The second case considered will be a sort of a dual of the first one, since we will assume that $\mathbf{T}/\log C$ is sufficiently small. It contains the case when $\mathbf{T}/\log C \to 0$, namely, the regime of sublogarithmic growth of $\mathbf{T}$ in C. This time the corresponding number of quantization intervals $\mathbf{N}$ will exhibit a superlogarithmic growth in C. From these two cases

we will then be able to study in detail a third situation, the logarithmic regime, which is when both $\mathbf{N}$ and $\mathbf{T}$ exhibit a logarithmic growth. In this case, we will establish quantitative bounds relating the ratios $\mathbf{N}/\log C$ and $\mathbf{T}/\log C$.

5.2. The regime of sublogarithmic growth of N in C. In this subsection we will assume that $\mathbf{N}/\log C$ is small enough. In this case it is convenient to proceed the estimates in (38) as follows:

$$\sum_{s=1}^{r\wedge\overline{n}-1}\binom{\overline{n}}{s+1}\binom{r}{s}\left(\frac{s}{r}\right)^s \le \sum_{s=0}^{r\wedge\overline{n}-1}\binom{\overline{n}}{s+1}\binom{r}{s} = \binom{\overline{n}+r}{r+1} = \frac{\overline{n}}{r+1}\binom{\overline{n}+r}{r}$$

$$\le \frac{1}{\sqrt{\pi}}\frac{\overline{n}}{r+1}\left(1+\frac{\overline{n}}{r}\right)^r e^r \le \overline{n}\left(1+\frac{\overline{n}}{\mathbf{N}}\right)^{\mathbf{N}} e^{\mathbf{N}},$$

where we used bound (91), the fact that $\frac{2}{(r+1)\sqrt{\pi}} \le 1$ and that $\left(1+\frac{\overline{n}}{r}\right)^r e^r$ is an increasing function in r. We obtain in this way

$$\mathbf{T} \ge \overline{n}\left[1 - C^{-1} - \left(1+\frac{\overline{n}}{\mathbf{N}}\right)^{\mathbf{N}} A^{\mathbf{N}} C^{-1}\right], \tag{40}$$

where $A := e^{\left(\frac{M}{e}+1\right)}$. We are now ready to prove the following result.

THEOREM 3. *There exist $K_1 > 0$, $\beta_1 > 0$, and $C_1 > 1$ such that*

$$C \ge C_1 \quad \text{and} \quad \frac{\mathbf{N}}{\log C} \le \beta_1 \implies \mathbf{T} \ge K_1 \mathbf{N} C^{1/\mathbf{N}}. \tag{41}$$

Proof. If in (40) we choose $\overline{n} = \lceil D\mathbf{N}C^{1/\mathbf{N}}\rceil$ for some constant $D > 0$ which will be fixed later, we have that

$$\frac{\mathbf{T}}{\mathbf{N}C^{1/\mathbf{N}}} \ge \frac{\lceil D\mathbf{N}C^{1/\mathbf{N}}\rceil}{\mathbf{N}C^{1/\mathbf{N}}}\left[1 - C^{-1} - \left(1+\frac{\lceil D\mathbf{N}C^{1/\mathbf{N}}\rceil}{\mathbf{N}}\right)^{\mathbf{N}} A^{\mathbf{N}} C^{-1}\right] \tag{42}$$

$$\ge D\left[1 - C^{-1} - \left(1+\frac{D\mathbf{N}C^{1/\mathbf{N}}+\mathbf{N}}{\mathbf{N}}\right)^{\mathbf{N}} A^{\mathbf{N}} C^{-1}\right]$$

$$= D\left[1 - C^{-1} - \left(2C^{-1/\mathbf{N}} + D\right)^{\mathbf{N}} A^{\mathbf{N}}\right].$$

Assume now that $\mathbf{N} \le \beta \log C$ for some β which will be chosen later. This implies that

$$(2C^{-1/\mathbf{N}} + D)A \le (2e^{-1/\beta} + D)A.$$

By choosing β and D small enough, we obtain that $(2e^{-1/\beta}+D)A \le 1/2$. Let β_1 and D_1 be possible solutions of the this inequality. In this situation, we can argue that

$$\frac{\mathbf{T}}{\mathbf{N}C^{1/\mathbf{N}}} \ge D_1[1 - C^{-1} - (1/2)^{\mathbf{N}}] \ge D_1[1/2 - C^{-1}],$$

and so there exist $C_1 > 1$ and $K_1 > 0$ such that (41) holds true. ∎

Theorem 3 will be important for later results on the logarithmic regime. Notice, moreover, that the bound established in Theorem 3 resembles relation (27) between the expected entrance time and the number of quantization intervals that can be obtained when using the nested chaotic scheme proposed in subsection 4.3. However, there is a difference and in fact the bound provided by Theorem 3 is not tight in this case. Consider for simplicity the case in which $\tau = 1$, so that we have a simple chaotic quantized feedback. In this case we have $\mathbf{N} = \lceil |a| \rceil$ quantization intervals and this, by Theorem 3, yields the bound

$$\mathbf{T} \geq K_1 C^{1/\lceil |a| \rceil}.$$

However, this is not a good bound since we expect in this case that $\mathbf{T} \sim C$. In fact, this bound can be improved in this particular case by using Proposition 5 that is a modification of Theorem 2 in which r is fixed equal to 1.

COROLLARY 2. *There exist $K_1 > 0$ and $C_1 > 1$ such that*

$$C \geq C_1 \quad \text{and} \quad \mathbf{N} = \lceil |a| \rceil \implies \mathbf{T} \geq K_1 C.$$

Proof. By Proposition 5 we can argue that

$$\frac{\gamma_k}{|a|^k} \leq 2 \left(\frac{\mathbf{N}M}{k \wedge \frac{\mathbf{N}M}{e}} \right)^{k \wedge \frac{\mathbf{N}M}{e}} \leq 2e^{\frac{\mathbf{N}M}{e}} = 2e^{\frac{\lceil |a| \rceil M}{e}}.$$

Let $A := e^{\frac{\lceil |a| \rceil M}{e}}$. Then, by (35) this implies that

$$\mathbf{T} \geq \overline{n}\big(1 - C^{-1}\big) - C^{-1} 2 \binom{\overline{n}}{2} A = \overline{n}\big[1 - C^{-1} - C^{-1}(\overline{n} - 1)A\big].$$

Let $\overline{n} = \lceil DC \rceil$ for some constant $D > 0$ which will be fixed later. We have that

$$\frac{\mathbf{T}}{C} \geq D \left[1 - C^{-1} - (\lceil DC \rceil - 1) AC^{-1}\right] \geq D \left[1 - C^{-1} - DA\right],$$

and this implies the thesis. ∎

5.3. The regime of sublogarithmic growth of T in C. In this subsection, we will assume instead that $\mathbf{T}/\log C$ is small enough. In this case, it is convenient to proceed the estimates in (38) as follows:

$$\begin{aligned}
\sum_{s=1}^{r \wedge \overline{n}-1} \binom{\overline{n}}{s+1} \binom{r}{s} \left(\frac{s}{r}\right)^s &\leq \frac{1}{\sqrt{\pi}} \sum_{s=1}^{\overline{n}-1} \binom{\overline{n}}{s+1} \left(1 + \frac{r-s}{s}\right)^s e^s \left(\frac{s}{r}\right)^s \\
&= \frac{1}{\sqrt{\pi}} \sum_{s=1}^{\overline{n}-1} \binom{\overline{n}}{s+1} e^s \leq \frac{1}{\sqrt{\pi}} \sum_{s=0}^{\overline{n}} \binom{\overline{n}}{s} e^{s-1} \\
&= \tfrac{1}{\sqrt{e\pi}} (1+e)^{\overline{n}},
\end{aligned}$$

where again we used bound (91). We thus obtain

$$\mathbf{T} \geq \overline{n}\big(1 - C^{-1}\big) - C^{-1} \frac{2}{\sqrt{e\pi}} (1+e)^{\overline{n}} \left(\frac{\mathbf{N}M}{\overline{n} - 1 \wedge \frac{\mathbf{N}M}{e}} \right)^{\overline{n} - 1 \wedge \frac{\mathbf{N}M}{e}} \tag{43}$$

$$\geq \overline{n}\big(1-C^{-1}\big) - C^{-1}A^{\overline{n}-1}\left(\frac{\mathbf{N}M}{\overline{n}-1\wedge\frac{\mathbf{N}M}{e}}\right)^{\overline{n}-1\wedge\frac{\mathbf{N}M}{e}},$$

where $A := \frac{2(1+e)^2}{e\sqrt{\pi}}$, and where the last inequality holds if $\overline{n} \geq 2$. We are now ready to prove the following result.

THEOREM 4. *There exist $K_2 > 0$, $\gamma_2 > 0$, and $C_2 > 1$ such that*

$$C \geq C_2 \quad \text{and} \quad \frac{\lceil \mathbf{T} \rceil}{\log C} \leq \gamma_2 \implies \mathbf{N} \geq K_2 \lceil \mathbf{T} \rceil C^{\frac{1}{\lceil \mathbf{T} \rceil}}. \tag{44}$$

Proof. We first show that we can find $C' > 1$ and $\gamma > 0$ such that

$$C \geq C' \quad \text{and} \quad \frac{\lceil \mathbf{T} \rceil}{\log C} \leq \gamma \implies \lceil \mathbf{T} \rceil \leq \frac{\mathbf{N}M}{e}. \tag{45}$$

Assume by contradiction that $\lceil \mathbf{T} \rceil > \mathbf{N}M/e$. Then, choosing $\overline{n} := \lceil \mathbf{T} \rceil + 1$, it follows from (35) and (43) that

$$\mathbf{T} \geq (\lceil \mathbf{T} \rceil + 1)\big(1-C^{-1}\big) - C^{-1}A^{\lceil \mathbf{T} \rceil}e^{\frac{\mathbf{N}M}{e}} \geq (\lceil \mathbf{T} \rceil + 1)\big(1-C^{-1}\big) - C^{-1}\,(eA)^{\lceil \mathbf{T} \rceil}, \tag{46}$$

which implies that

$$\begin{aligned} 0 &\geq C(\lceil \mathbf{T} \rceil - \mathbf{T} + 1) - (eA)^{\lceil \mathbf{T} \rceil} - \lceil \mathbf{T} \rceil - 1 \\ &\geq C - (eA)^{\lceil \mathbf{T} \rceil} - \lceil \mathbf{T} \rceil - 1 \\ &\geq C - (eA)^{\gamma \log C} - \gamma \log C - 1 = C - C^{\gamma \log eA} - \gamma \log C - 1. \end{aligned} \tag{47}$$

If we choose $\gamma < (\log eA)^{-1}$, it is clear that there exists $C' > 1$ such that

$$C - C^{\gamma \log eA} - \gamma \log C - 1 > 0$$

for all $C \geq C'$. For such values of C (47) cannot hold. Hence (45) must hold.

Assume now that (45) holds true and choose again $\overline{n} := \lceil \mathbf{T} \rceil + 1$ in (43). Then we obtain

$$\mathbf{T} \geq (\lceil \mathbf{T} \rceil + 1)\big(1-C^{-1}\big) - \left(\frac{\mathbf{N}MA}{\lceil \mathbf{T} \rceil}\right)^{\lceil \mathbf{T} \rceil} C^{-1}. \tag{48}$$

Solving with respect to $\mathbf{N}$, we obtain

$$\begin{aligned} \mathbf{N} &\geq \frac{\lceil \mathbf{T} \rceil}{AM}[C(\lceil \mathbf{T} \rceil - \mathbf{T} + 1) - \lceil \mathbf{T} \rceil - 1]^{1/\lceil \mathbf{T} \rceil} \geq \frac{\lceil \mathbf{T} \rceil}{AM}[C - \lceil \mathbf{T} \rceil - 1]^{1/\lceil \mathbf{T} \rceil} \\ &\geq \frac{\lceil \mathbf{T} \rceil}{AM}[C - \gamma \log C - 1]^{1/\lceil \mathbf{T} \rceil}\ . \end{aligned} \tag{49}$$

Observe finally that

$$\lim_{C\to\infty} \frac{C - \gamma \log C - 1}{C} = 1,$$

which implies that for any $\epsilon > 0$ there exists $C'' > 0$ such that $C - \gamma \log C - 1 > (1-\epsilon)C$ for all $C > C''$. From this we can argue that

$$\mathbf{N} \geq \frac{\lceil \mathbf{T} \rceil}{AM}[(1-\epsilon)C]^{1/\lceil \mathbf{T} \rceil} \geq \frac{1-\epsilon}{AM}\lceil \mathbf{T} \rceil C^{1/\lceil \mathbf{T} \rceil}.$$

By letting $K_2 := \frac{1-\epsilon}{AM}$, $C_2 := C' \vee C''$, and $\gamma_2 := \gamma$, we have thus proved the thesis. □

Also in this case it is interesting to compare the bound provided by the previous theorem with relation (21) between the number of quantization intervals and the expected entrance time that can be obtained when using the nested strategy proposed in subsection 4.1. In this case this comparison shows that, up to a multiplication by a constant, the bound is tight.

5.4. The logarithmic regime. We have the following direct consequence of previous theorems.

COROLLARY 3. *There exist $C_0 > 1$ and two functions $F, G : \mathbb{R}_+ \to \mathbb{R}_+$ which are decreasing and converging to 0 at $+\infty$ such that for all $C > C_0$ we have that*

$$\frac{\mathbf{N}}{\log C} \geq F\left(\frac{\lceil \mathbf{T} \rceil}{\log C}\right) \quad \text{and} \quad \frac{\lceil \mathbf{T} \rceil}{\log C} \geq G\left(\frac{\mathbf{N}}{\log C}\right). \tag{50}$$

Proof. Notice first that the function $f :]0,1] \to \mathbb{R} : x \mapsto xe^{1/x}$ is strictly decreasing, and its image is $[e, +\infty)$. Let $C_0 := C_1 \vee C_2$, where C_1, C_2 are the constants introduced, respectively, in Theorems 3 and 4. Define the function

$$F(x) = \begin{cases} 1 \wedge \beta_1 & \text{if } 0 \leq x \leq K_1 f(1 \wedge \beta_1), \\ f^{-1}(x/K_1) & \text{if } x > K_1 f(1 \wedge \beta_1), \end{cases}$$

where K_1 and β_1 are the constants provided by Theorem 3. This function is decreasing such that $F(+\infty) = 0$. We want to show that if $C > C_0$, then

$$\frac{\mathbf{N}}{\log C} \geq F\left(\frac{\lceil \mathbf{T} \rceil}{\log C}\right).$$

If $\mathbf{N}/\log C > 1 \wedge \beta_1$, then

$$\frac{\mathbf{N}}{\log C} > \max_{x \in \mathbb{R}_+} F(x) \geq F\left(\frac{\lceil \mathbf{T} \rceil}{\log C}\right).$$

If instead $\mathbf{N}/\log C \leq 1 \wedge \beta_1$, then by Theorem 3 we can argue that

$$\frac{\mathbf{T}}{\log C} \geq K_1 \frac{\mathbf{N}}{\log C} C^{1/\mathbf{N}} = K_1 f\left(\frac{\lceil \mathbf{N} \rceil}{\log C}\right),$$

which implies that

$$\frac{\mathbf{N}}{\log C} \geq f^{-1}\left(\frac{\mathbf{T}}{K_1 \log C}\right) = F\left(\frac{\mathbf{T}}{\log C}\right).$$

In the same way it can be shown that

$$\frac{\lceil \mathbf{T} \rceil}{\log C} \geq G\left(\frac{\mathbf{N}}{\log C}\right),$$

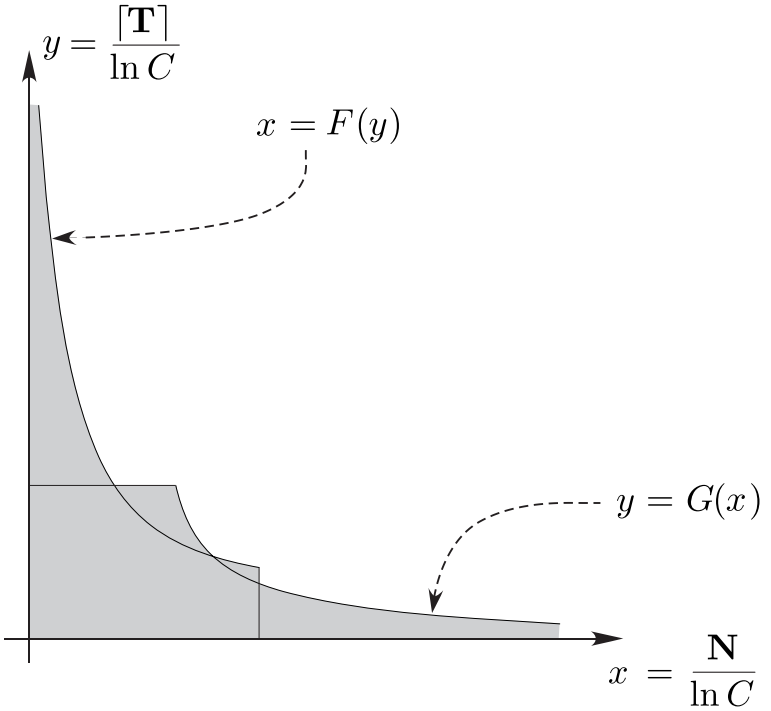


FIG. 5. *The grey region in this figure represents the set in which the pairs* $(\mathbf{N}/\log C, \mathbf{T}/\log C)$ *cannot belong.*

where

$$G(x) = \begin{cases} 1 \wedge \gamma_2 & \text{if } 0 \le x \le K_2 f(1 \wedge \beta_1), \\ f^{-1}(x/K_2) & \text{if } x > K_2 f(1 \wedge \gamma_2), \end{cases}$$

where K_2 and γ_2 are the constants provided by Theorem 4. □

Remark. The constraint provided by the previous corollary is illustrated in Figure 5 which shows explicitly the region in which the pairs $(\mathbf{N}/\log C, \mathbf{T}/\log C)$ cannot belong. Observe moreover that the functions $F(x)$ and $G(x)$ in the previous corollary which determine the boundary of this region tend to 0 as the function $f(x) = xe^{1/x}$. This is in agreement with the behavior of the logarithmic regime exhibited in the nesting of both deadbeat quantized feedbacks and chaotic quantized feedbacks (see (22) and (28)). This implies that, up to multiplicative constants, our bounds appear to be quite tight and that the examples presented in section 4 cannot be improved much.

5.5. The case when $|a| \le 2$. All previous results have been obtained under the assumption $|a| > 2$. In fact, part of the results presented in this subsection can be extended to the case $|a| \le 2$. Indeed, in this case, using the second part of Theorem 6, we obtain the estimate

$$\frac{\gamma_k}{2^k} \le \left[\sum_{s=1}^{r \wedge k} \binom{k+s-1}{2s-1} \binom{r}{s} \left(\frac{s}{r}\right)^s \right] \left(\frac{\mathbf{N}M}{k \wedge \frac{\mathbf{N}M}{e}} \right)^{k \wedge \frac{\mathbf{N}M}{e}}.$$

By similar arguments used to deal with the case $|a| > 2$, we obtain

$$\mathbf{T} \geq \overline{n}(1 - C^{-1}) - C^{-1} 2 \left[\sum_{s=1}^{r \wedge \overline{n}-1} \binom{\overline{n}+s}{2s+1} \binom{r}{s} \left(\frac{s}{r}\right)^s \right] \times \left(\frac{\mathbf{N}M}{\overline{n} - 1 \wedge \frac{\mathbf{N}M}{e}} \right)^{\overline{n}-1 \wedge \frac{\mathbf{N}M}{e}} \left(\frac{2}{|a|} \right)^{\overline{n}-1}$$

for all $\overline{n} \in \mathbb{N}$. Observing that

$$\sum_{s=1}^{r \wedge \overline{n}-1} \binom{\overline{n}+s}{2s+1} \binom{r}{s} \left(\frac{s}{r}\right)^s \leq \frac{1}{\sqrt{\pi}} \sum_{s=1}^{\overline{n}-1} \binom{2\overline{n}-1}{2s+1} e^s \leq \frac{1}{\sqrt{\pi}} \sum_{s=1}^{\overline{n}-1} \binom{2\overline{n}-1}{2s+1} e^{2s+1} \leq \frac{1}{\sqrt{\pi}} (1+e)^{2\overline{n}-1},$$

we thus obtain

$$\mathbf{T} \geq \overline{n}(1 - C^{-1}) - C^{-1} \frac{2}{\sqrt{\pi}} (1+e)^{2\overline{n}-1} \left(\frac{\mathbf{N}M}{\overline{n} - 1 \wedge \frac{\mathbf{N}M}{e}} \right)^{\overline{n}-1 \wedge \frac{\mathbf{N}M}{e}} \left(\frac{2}{|a|} \right)^{\overline{n}-1}$$

$$\geq \overline{n}(1 - C^{-1}) - C^{-1} A^{\overline{n}-1} \left(\frac{\mathbf{N}M}{\overline{n} - 1 \wedge \frac{\mathbf{N}M}{e}} \right)^{\overline{n}-1 \wedge \frac{\mathbf{N}M}{e}},$$

where $A := \frac{4(1+e)^3}{|a|\sqrt{\pi}}$ and where the last inequality holds if $\overline{n} \geq 2$. The previous inequality looks exactly like (43). This immediately implies that Theorem 4 also holds true for $|a| \leq 2$. We can instead only recover a part of Corollary 3: (50) remains true for small values of γ, as it is easy to see from the proof we gave.

5.6. Stabilizing quantized feedbacks. In this section, we will show that quantized control strategies yielding stability or even almost stability, but with only a countable subset of points never entering inside J, require a number of quantization intervals which grows at least logarithmically in C. The result is based on Theorem 7 which is given in the last section.

THEOREM 5. *If Γ is almost (I, J)-stable and if the set of points in I never entering inside J is at most countable, then there exists $\beta > 0$, only depending on a such that*

$$\mathbf{N} / \log C \geq \beta$$

for all $C > 1$.

Proof. Using (87) we can argue that

$$\sum_{k=1}^{n-1} \frac{\gamma_k}{|a|^k} \leq e^{\frac{\mathbf{N}}{e}} \sum_{k=0}^{+\infty} \binom{k + 2\mathbf{N} - 1}{2\mathbf{N} - 1} \left(\frac{2}{|a|} \right)^k = \frac{e^{\frac{\mathbf{N}}{e}}}{\left(1 - \frac{2}{|a|}\right)^{2\mathbf{N}}} = \left(\frac{e^{1/e}|a|^2}{(|a|-1)^2} \right)^{\mathbf{N}}.$$

By letting $A := \frac{e^{1/e}|a|^2}{(|a|-1)^2}$, from (34) we can argue that

$$\mathbb{P}\big[T_{(I,J)} \geq n\big] \geq 1 - C^{-1} - A^{\mathbf{N}} C^{-1} \geq 1 - C^{-1}(1+A)^{\mathbf{N}}. \tag{51}$$

Since Γ is almost stable, by Proposition 1 we have that $\mathbb{E}(T_{(I,J)}) < +\infty$, which implies that

$$\lim_{n\to\infty} \mathbb{P}\big[T_{(I,J)} \geq n\big] = 0.$$

From this and (51) we can argue that $1 - C^{-1}(1+A)^{\mathbf{N}} \leq 0$, which implies that

$$\mathbf{N} \geq \frac{\log C}{\log(1+A)}. \qquad \square$$

6. Estimation of paths in a class of weighted graphs. For proving our main result, namely Theorem 2, we introduce a class of weighted graphs and propose a method for bounding the number of paths on this graphs. In the last section, we will show how this bound can be used for proving Theorem 2. We use this strategy which considers this graph abstraction because the general result we are going to prove is useful to deal with more general situations, such as quantized controller with memory or the case in which the state of the system is multidimensional (see [10]).

Consider a direct graph $\mathcal{G}$ on a vertex set $\mathcal{X}$ (which is not necessarily finite or countable). For any choice of $\mathcal{X}_1, \dots, \mathcal{X}_k \subset \mathcal{X}$ we define $\mathcal{F}_k[\mathbf{x}_1 \in \mathcal{X}_1, \dots, \mathbf{x}_k \in \mathcal{X}_k]$ to be the set of paths $\mathbf{x}_1 \cdots \mathbf{x}_k \in \mathcal{X}^*$ on the graph $\mathcal{G}$ such that $\mathbf{x}_1 \in \mathcal{X}_1, \dots, \mathbf{x}_k \in \mathcal{X}_k$.

Assume the graph $\mathcal{G}$ has the following structure. We assume there exist a finite partition

$$\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2 \cup \cdots \cup \mathcal{X}_{\mathbf{N}},$$

a subset $\mathcal{X}_P \subseteq \mathcal{X}$, and a function $q : \mathcal{X} \to]0,1[$ with the following properties.

(A) There exist numbers $q_1, \dots, q_{\mathbf{N}} \in]0,1[$ such that

$$\begin{aligned} q(\mathbf{x}) &\leq q_i \quad \forall \mathbf{x} \in \mathcal{X}_i, \\ q(\mathbf{x}) &= q_i \quad \forall \mathbf{x} \in \mathcal{X}_{P,i} := \mathcal{X}_P \cap \mathcal{X}_i. \end{aligned}$$

(B) There exist positive numbers D_1 and α_1 such that, for every $\mathbf{x}' \in \mathcal{X}$, $\mathcal{X}'' \subseteq \mathcal{X}$, and $k \geq 2$,

$$\#\mathcal{F}_k[\mathbf{x}_1 = \mathbf{x}', \mathbf{x}_2, \dots, \mathbf{x}_{k-1} \in \mathcal{X}, \mathbf{x}_k \in \mathcal{X}''] \leq D_1 \frac{q(\mathbf{x}')}{\inf_{\mathbf{x}'' \in \mathcal{X}''} q(\mathbf{x}'')} \alpha_1^{k-2}.$$

(C) There exist positive numbers D_2 and α_2 such that, for every $\mathbf{x}' \in \mathcal{X}$, $i \in \{1, \dots, \mathbf{N}\}$, and $k \geq 2$,

$$\#\mathcal{F}_k[\mathbf{x}_1 = \mathbf{x}', \mathbf{x}_2, \dots, \mathbf{x}_{k-1} \in \mathcal{X} \setminus \mathcal{X}_P, \mathbf{x}_k \in \mathcal{X}_i] \leq D_2 \alpha_2^{k-2}.$$

Then, if we define

$$\gamma_{k,h} = \sup_{\mathbf{x}' \in \mathcal{X}_{P,h}} \#\mathcal{F}_k[\mathbf{x}_1 = \mathbf{x}', \mathbf{x}_2, \dots, \mathbf{x}_k \in \mathcal{X}], \qquad h = 1, \dots, \mathbf{N},$$

$$\gamma_k = \sum_{h=1}^{\mathbf{N}} \gamma_{k,h}, \tag{52}$$

we have the following result.

THEOREM 6. *We have the following bounds.*

(1) *If* $\alpha_1 > \alpha_2$, *then*

$$\frac{\gamma_k}{\alpha_1^k} \leq 2\left[\sum_{s=1}^{r\wedge k}\binom{k-1}{s-1}\binom{r}{s}\left(\frac{s}{r}\right)^s\right]\left(\frac{\mathbf{N}M}{k\wedge\frac{\mathbf{N}M}{e}}\right)^{k\wedge\frac{\mathbf{N}M}{e}} \qquad \forall k\geq 1. \tag{53}$$

(2) *If* $\alpha_1 \leq \alpha_2$, *then*

$$\frac{\gamma_k}{\alpha_2^k} \leq \left[\sum_{s=1}^{r\wedge k}\binom{k+s-1}{2s-1}\binom{r}{s}\left(\frac{s}{r}\right)^s\right]\left(\frac{\mathbf{N}M}{k\wedge\frac{\mathbf{N}M}{e}}\right)^{k\wedge\frac{\mathbf{N}M}{e}} \qquad \forall k\geq 1. \tag{54}$$

The constant $r \in \{1,\dots,\mathbf{N}\}$ *is independent of* k, *but may depend on the specific graph, while* M *depends only on the constants* $D_1, D_2, \alpha_1, \alpha_2$.

The proof of the previous theorem is quite lengthy. For this reason we prefer to divide it into various steps.

Remark. As specified in the previous theorem M depends only on the constants D_1, D_2, α_1, and α_2, and r depends on the specific graph. These conditions can be exchanged and the same bounds can be shown to hold true in which instead r depends only on the constants D_1, D_2, α_1, and α_2, and M depends on the specific graph. However, this exchange makes the bounds useless in general. Only in the specific situation considered in Proposition 5 does this point of view yield some advantages.

6.1. The proof of Theorem 6: Hierarchies of paths. Assume with no loss of generality that the subsets $\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_{\mathbf{N}}$ are ordered in such a way that

$$q_1 \geq q_2 \geq \cdots \geq q_{\mathbf{N}}.$$

For any choice of integers

$$0 = N_0 < N_1 < \cdots < N_{r-1} < N_r = \mathbf{N},$$

we can partition $\mathcal{X}_P$ into the subfamilies

$$\begin{aligned}\mathcal{X}_P^1 :&= \{\mathcal{X}_{P,N_0+1},\dots,\mathcal{X}_{P,N_1}\},\ \ \mathcal{X}_P^2 := \{\mathcal{X}_{P,N_1+1},\dots,\mathcal{X}_{P,N_2}\},\dots,\\ \mathcal{X}_P^r :&= \{\mathcal{X}_{P,N_{r-1}+1},\dots,\mathcal{X}_{P,N_r}\}\end{aligned} \tag{55}$$

and consider, moreover,

$$\begin{aligned}\mathcal{X}_P^{l+} &:= \bigcup_{j=l}^{r}\mathcal{X}_P^j,\\ \mathcal{X}^l &:= (\mathcal{X}\setminus\mathcal{X}_P)\cup\mathcal{X}_P^l,\\ \mathcal{X}^{l+} &:= (\mathcal{X}\setminus\mathcal{X}_P)\cup\mathcal{X}_P^{l+}.\end{aligned}$$

For each $k\in\mathbb{N}$, $h = 1,\dots,\mathbf{N}$, and $l = 1,\dots,r,r+1$ define

$$\gamma_{k,h,l} := \sup_{\mathbf{x}'\in\mathcal{X}_{P,h}} \#\mathcal{F}_k[\mathbf{x}_1 = \mathbf{x}', \mathbf{x}_2,\dots,\mathbf{x}_k \in \mathcal{X}^{l+}].$$

From these definitions it follows that $\mathcal{X}_P^{1+} = \mathcal{X}_P$, $\mathcal{X}^{1+} = \mathcal{X}$, and $\mathcal{X}^{(r+1)+} := \mathcal{X}\setminus\mathcal{X}_P$. This implies that $\gamma_{k,h,1} = \gamma_{k,h}$.

We present now two bounds on $\gamma_{k,h,l}$ which will be used in what follows. The first bound is based on the decomposition of the paths in $\mathcal{F}_k[\mathbf{x}_1 = \mathbf{x}', \mathbf{x}_2, \dots, \mathbf{x}_k \in \mathcal{X}^{l+}]$ according to the last exit from $\mathcal{X}_P^l$ among the indices $j = 2, \dots, k$

$$\mathcal{F}_k\big[\mathbf{x}_1 = \mathbf{x}', \mathbf{x}_2, \dots, \mathbf{x}_k \in \mathcal{X}^{l+}\big] =$$
$$\left\{ \bigcup_{j=2}^{k} \bigcup_{s=N_{l-1}+1}^{N_l} \mathcal{F}_k\big[\mathbf{x}_1 = \mathbf{x}', \mathbf{x}_2, \dots, \mathbf{x}_{j-1} \in \mathcal{X}^{l+}, \mathbf{x}_j \in \mathcal{X}_{P,s}, \mathbf{x}_{j+1}, \dots, \mathbf{x}_k \in \mathcal{X}^{(l+1)+}\big] \right\}$$
$$\times \bigcup \mathcal{F}_k\big[\mathbf{x}_1 = \mathbf{x}', \mathbf{x}_2, \dots, \mathbf{x}_k \in \mathcal{X}^{(l+1)+}\big].$$

Applying property (B) it follows that, for all $l = 1, \dots, r$,

$$\begin{aligned} \gamma_{k,h,l} \le & \sum_{j=2}^{k} \sum_{s=N_{l-1}+1}^{N_l} \sup_{\mathbf{x}' \in \mathcal{X}_{P,h}} \#\mathcal{F}_j\big[\mathbf{x}_1 = \mathbf{x}', \mathbf{x}_2, \dots, \mathbf{x}_{j-1} \in \mathcal{X}^{l+}, \mathbf{x}_j \in \mathcal{X}_{P,s}\big] \\ & \times \sup_{\mathbf{x}'' \in \mathcal{X}_{P,s}} \#\mathcal{F}_{k-j+1}\big[\mathbf{x}_j = \mathbf{x}'', \mathbf{x}_{j+1}, \dots, \mathbf{x}_k \in \mathcal{X}^{(l+1)+}\big] \\ & + \sup_{\mathbf{x}' \in \mathcal{X}_{P,h}} \#\mathcal{F}_k\big[\mathbf{x}_1 = \mathbf{x}', \mathbf{x}_2, \dots, \mathbf{x}_k \in \mathcal{X}^{(l+1)+}\big] \\ \le & \sum_{j=2}^{k} \sum_{s=N_{l-1}+1}^{N_l} D_1 \frac{q_h}{q_s} \alpha_1^{j-2} \gamma_{k-j+1,s,l+1} + \gamma_{k,h,l+1}. \end{aligned} \tag{56}$$

The second bound is based on the decomposition of the paths in $\mathcal{F}_k[\mathbf{x}_1 = \mathbf{x}', \mathbf{x}_2, \dots, \mathbf{x}_k \in \mathcal{X}^{l+}]$ according to the first entrance in $\mathcal{X}_P^{l+}$ among the indices $j = 2, \dots, k$:

$$\mathcal{F}_k\big[\mathbf{x}_1 = \mathbf{x}', \mathbf{x}_2, \dots, \mathbf{x}_k \in \mathcal{X}^{l+}\big] =$$
$$\left\{ \bigcup_{j=2}^{k} \bigcup_{s=N_{l-1}+1}^{\mathbf{N}} \mathcal{F}_k\big[\mathbf{x}_1 = \mathbf{x}', \mathbf{x}_2, \dots, \mathbf{x}_{j-1} \in \mathcal{X} \setminus \mathcal{X}_P, \mathbf{x}_j \in \mathcal{X}_{P,s}, \mathbf{x}_{j+1}, \dots, \mathbf{x}_k \in \mathcal{X}^{l+}\big] \right\}$$
$$\times \bigcup \mathcal{F}_k\big[\mathbf{x}_1 = \mathbf{x}', \mathbf{x}_2, \dots, \mathbf{x}_k \in \mathcal{X} \setminus \mathcal{X}_P\big].$$

Applying property (C) it follows that

$$\begin{aligned} \gamma_{k,h,l} \le & \sum_{j=2}^{k} \sum_{s=N_{l-1}+1}^{\mathbf{N}} \sup_{\mathbf{x}' \in \mathcal{X}_{P,h}} \mathcal{F}_j\big[\mathbf{x}_1 = \mathbf{x}', \mathbf{x}_2, \dots, \mathbf{x}_{j-1} \in \mathcal{X} \setminus \mathcal{X}_P, \mathbf{x}_j \in \mathcal{X}_{P,s}\big] \\ & \times \sup_{\mathbf{x}'' \in \mathcal{X}_{P,s}} \mathcal{F}_{k-j+1}\big[\mathbf{x}_j = \mathbf{x}'', \mathbf{x}_{j+1}, \dots, \mathbf{x}_k \in \mathcal{X}^{l+}\big] \\ & + \sup_{\mathbf{x}' \in \mathcal{X}_{P,h}} \mathcal{F}_k\big[\mathbf{x}_1 = \mathbf{x}', \mathbf{x}_2, \dots, \mathbf{x}_k \in \mathcal{X} \setminus \mathcal{X}_P\big] \\ \le & \sum_{j=2}^{k} \sum_{s=N_{l-1}+1}^{\mathbf{N}} D_2 \alpha_2^{j-2} \gamma_{k-j+1,s,l} + D_2 \alpha_2^{k-1}. \end{aligned} \tag{57}$$

Notice that the previous bound holds true for $l = 1, \dots, r, r+1$.

Define

$$\delta_{k,l} := \sum_{h=N_{l-1}+1}^{\mathbf{N}} \gamma_{k,h,l}, \quad \tilde{\delta}_{k,l} := \sum_{h=N_{l-2}+1}^{N_{l-1}} \gamma_{k,h,l} \qquad l = 1, \dots, r, r+1,$$

which imply that $\delta_{k,r+1} = 0$ for all $k \in \mathbb{N}$. Observe that from (56) we can argue that

$$\begin{aligned}\delta_{k,l} &\le \sum_{h=N_{l-1}+1}^{\mathbf{N}} \left(\sum_{j=2}^{k} \sum_{s=N_{l-1}+1}^{N_l} D_1 \frac{q_h}{q_s} \alpha_1^{j-2} \gamma_{k-j+1,s,l+1} + \gamma_{k,h,l+1} \right) \\ &\le \sum_{j=2}^{k} D_1 \frac{\sum_{h=N_{l-1}+1}^{\mathbf{N}} q_h}{q_{N_l}} \alpha_1^{j-2} \sum_{s=N_{l-1}+1}^{N_l} \gamma_{k-j+1,s,l+1} + \sum_{h=N_{l-1}+1}^{N_l} \gamma_{k,h,l+1} \\ &\quad + \sum_{h=N_l+1}^{\mathbf{N}} \gamma_{k,h,l+1} \le D_1 \beta_l \sum_{j=2}^{k} \alpha_1^{j-2} \tilde{\delta}_{k-j+1,l+1} + \tilde{\delta}_{k,l+1} + \delta_{k,l+1} \\ &= D_1 \beta_l \sum_{j=0}^{k-2} \alpha_1^{j} \tilde{\delta}_{k-j-1,l+1} + \tilde{\delta}_{k,l+1} + \delta_{k,l+1} \ ,\end{aligned}$$

where we define

$$\beta_l := \frac{\sum_{h=N_{l-1}+1}^{\mathbf{N}} q_h}{q_{N_l}}. \tag{58}$$

On the other hand (57) implies that

$$\tilde{\delta}_{k,l} \le D_2 (N_{l-1} - N_{l-2}) \left(\sum_{j=2}^{k} \alpha_2^{j-2} \delta_{k-j+1,l} + \alpha_2^{k-1} \right),$$

which using the convention

$$\delta_{0,l} = 1, \qquad l = 1, \ldots, r, r+1,$$

is equivalent to

$$\tilde{\delta}_{k,l} \le D_2 \Delta N_{l-1} \sum_{j=2}^{k+1} \alpha_2^{j-2} \delta_{k-j+1,l} = D_2 \Delta N_{l-1} \sum_{j=0}^{k-1} \alpha_2^{j} \delta_{k-j-1,l},$$

where we defined $\Delta N_l := N_l - N_{l-1}$.

Summarizing we have the following two inequalities holding for $k \ge 1$:

$$\begin{aligned}\delta_{k,l} &\le D_1 \beta_l \sum_{j=0}^{k-2} \alpha_1^{j} \tilde{\delta}_{k-j-1,l+1} + \tilde{\delta}_{k,l+1} + \delta_{k,l+1}, \qquad l = 1, \ldots, r, \\ \tilde{\delta}_{k,l} &\le D_2 \Delta N_{l-1} \sum_{j=0}^{k-1} \alpha_2^{j} \delta_{k-j-1,l}, \qquad l = 1, \ldots, r, r+1.\end{aligned} \tag{59}$$

Now define the sequences $\eta_{k,l}, \tilde{\eta}_{k,l}$ for $k = 0, 1, 2, \ldots$ and $l = 1, \ldots, r, r+1$ by letting $\eta_{k,r+1} = \delta_{k,r+1} = 0$ for $k = 0, 1, \ldots$, and satisfying, for every $k \ge 0$, the following recursive relations:

$$\begin{aligned}\eta_{k,l} &= D_1 \beta_l \sum_{j=0}^{k-2} \alpha_1^{j} \tilde{\eta}_{k-j-1,l+1} + \tilde{\eta}_{k,l+1} + \eta_{k,l+1}, \\ \tilde{\eta}_{k,l} &= D_2 \Delta N_{l-1} \sum_{j=0}^{k-1} \alpha_2^{j} \eta_{k-j-1,l}.\end{aligned} \tag{60}$$

Notice that from the above recursive relations it follows that $\eta_{0,l} = 1$ for every l. This implies in particular that $\delta_{k,l} \leq \eta_{k,l}$ for every k and l. In what follows we will estimate $\eta_{k,l}$ by using the zeta transforms formalism.

Let

$$\eta_l(z) := \sum_{k=0}^{+\infty} \eta_{k,l} z^k, \quad \tilde{\eta}_l(z) := \sum_{k=0}^{+\infty} \tilde{\eta}_{k,l} z^k.$$

Then by some standard manipulations from (60) we obtain

$$\begin{aligned} \eta_l(z) &= D_1 \beta_l \frac{z}{1-\alpha_1 z} \tilde{\eta}_{l+1}(z) + \tilde{\eta}_{l+1}(z) + \eta_{l+1}(z), \\ \tilde{\eta}_l(z) &= D_2 \Delta N_{l-1} \frac{z}{1-\alpha_2 z} \eta_l(z), \end{aligned} \tag{61}$$

which yields

$$\eta_l(z) = \left\{ \left[D_1 \beta_l \frac{z}{1-\alpha_1 z} + 1 \right] D_2 \Delta N_l \frac{z}{1-\alpha_2 z} + 1 \right\} \eta_{l+1}(z).$$

By iterating this formula we obtain

$$\eta_1(z) = \prod_{l=1}^{r} \left\{ \left[D_1 \beta_l \frac{z}{1-\alpha_1 z} + 1 \right] D_2 \Delta N_l \frac{z}{1-\alpha_2 z} + 1 \right\}, \tag{62}$$

where we used the fact that $\eta_{r+1}(z) = 1$.

6.2. The proof of Theorem 6: Combinatorial bounds. We now want to estimate the coefficients $\eta_{k,1}$ of $\eta_1(z)$. We recall that $\gamma_k = \delta_{k,1} \leq \eta_{k,1}$. In order to obtain such bounds we will first need to work out some combinatorics.

Bounds on the coefficients of elementary symmetric polynomials. Consider the following polynomial in the indeterminates x and y:

$$p(x,y) := \prod_{l=1}^{r} \left\{ \left[\beta_l x + 1 \right] \alpha_l y + 1 \right\} = \sum_{s=0}^{r} \sum_{\sigma=0}^{s} \bar{p}_{\sigma,s} x^\sigma y^s. \tag{63}$$

The aim of this part of the section is to determine bounds on the coefficients $\bar{p}_{\sigma,s}$ if we assume that

$$\sum_{l=0}^{r} \alpha_l \leq \alpha, \qquad \sum_{l=0}^{r} \beta_l \leq \beta.$$

Consider preliminarily the polynomial

$$\prod_{l=1}^{r} \{\alpha_l y + 1\} = \sum_{s=0}^{r} p_s^r(\alpha_1, \ldots, \alpha_r) y^s. \tag{64}$$

The polynomials $p_s^r(\alpha_1, \ldots, \alpha_r)$ are called elementary symmetric polynomials [11], and they can be expressed by the formula

$$p_s^r(\alpha_1, \ldots, \alpha_r) = \sum_{1 \leq l_1 < \cdots l_s \leq r} \prod_{j=1}^{s} \alpha_{l_j}.$$

We have the following first elementary result.

LEMMA 3. *Assume that* $\sum_{l=0}^{r} \alpha_l \leq \alpha$. *Then*

$$p_s^r(\alpha_1, \ldots, \alpha_r) \leq \binom{r}{s} \left(\frac{\alpha}{r}\right)^s. \tag{65}$$

Proof. We will actually prove that bound (65) holds true, and it is attained when $\alpha_i = \alpha/r$ for all $i = 1, \ldots, r$. For $r = 2$ it can be proven directly. For the general case, it is sufficient to notice that

$$\begin{aligned}
&p_s^r(\alpha_1, \alpha_2, \ldots, \alpha_r) \\
&= p_s^{r-2}(\alpha_3, \ldots, \alpha_r) + p_1^2(\alpha_1, \alpha_2) p_{s-1}^{r-2}(\alpha_3, \ldots, \alpha_r) + p_2^2(\alpha_1, \alpha_2) p_{s-2}^{r-2}(\alpha_3, \ldots, \alpha_r) \\
&\leq p_s^{r-2}(\alpha_3, \ldots, \alpha_r) + p_1^2\left(\tfrac{\alpha_1+\alpha_2}{2}, \tfrac{\alpha_1+\alpha_2}{2}\right) p_{s-1}^{r-2}(\alpha_3, \ldots, \alpha_r) \\
&\quad + p_2^2\left(\tfrac{\alpha_1+\alpha_2}{2}, \tfrac{\alpha_1+\alpha_2}{2}\right) p_{s-2}^{r-2}(\alpha_3, \ldots, \alpha_r) \\
&= p_s^r\left(\tfrac{\alpha_1+\alpha_2}{2}, \tfrac{\alpha_1+\alpha_2}{2}, \alpha_3, \ldots, \alpha_r\right). \qquad \square
\end{aligned}$$

We come back to the problem of finding bounds on the coefficients $\bar{p}_{\sigma,s}$ of polynomial (63).

LEMMA 4. *For every* $1 \leq s \leq r$ *and* $0 \leq \sigma \leq s$, *the following bound holds:*

$$\bar{p}_{\sigma,s} \leq \binom{s}{\sigma} \left(\frac{\beta}{s}\right)^\sigma \binom{r}{s} \left(\frac{\alpha}{r}\right)^s. \tag{66}$$

Proof. Observe first that

$$p(x, y) = \prod_{l=1}^{r} \left\{[\beta_l x + 1]\alpha_l y + 1\right\} = \sum_{s=0}^{r} p_s^r\big(\alpha_1(1 + \beta_1 x), \ldots, \alpha_r(1 + \beta_r x)\big) y^s.$$

Moreover, we have that

$$\begin{aligned}
p_s^r\big(\alpha_1(1 + \beta_1 x), \ldots, \alpha_r(1 + \beta_r x)\big) &= \sum_{1 \leq l_1 < \cdots l_s \leq r} \prod_{j=1}^{s} \alpha_{l_j} \prod_{j=1}^{s} (1 + \beta_{l_j} x) \\
&= \sum_{1 \leq l_1 < \cdots l_s \leq r} \prod_{j=1}^{s} \alpha_{l_j} \sum_{\sigma=0}^{s} p_\sigma^s(\beta_{l_1}, \ldots, \beta_{l_s}) x^\sigma,
\end{aligned}$$

from which we can argue that, using Lemma 3,

$$\begin{aligned}
\bar{p}_{\sigma,s} &= \sum_{1 \leq l_1 < \cdots l_s \leq r} p_\sigma^s(\beta_{l_1}, \ldots, \beta_{l_s}) \prod_{j=1}^{s} \alpha_{l_j} \leq \binom{s}{\sigma} \left(\frac{\beta}{s}\right)^\sigma \sum_{1 \leq l_1 < \cdots l_s \leq r} \prod_{j=1}^{s} \alpha_{l_j} \\
&= \binom{s}{\sigma} \left(\frac{\beta}{s}\right)^\sigma p_s^r(\alpha_1, \ldots, \alpha_r) \leq \binom{s}{\sigma} \left(\frac{\beta}{s}\right)^\sigma \binom{r}{s} \left(\frac{\alpha}{r}\right)^s. \qquad \square
\end{aligned}$$

To apply the result provided by the previous lemma to our problem, we need to have bounds on $\sum_{l=1}^{r} \Delta N_l$ and $\sum_{l=1}^{r} \beta_l$. While it is evident that

$$\sum_{l=1}^{r} \Delta N_l = \mathbf{N},$$

it is less clear how to bound the other sum. This will depend indeed on the way the subfamilies $\mathcal{X}_P^i$ are selected. It follows from (58) that

$$\beta_l = \sum_{k=0}^{r-l} \frac{\sum_{h=N_{l+k-1}+1}^{N_{l+k}} q_h}{q_l} \leq \sum_{k=0}^{r-l} \Delta N_{l+k} \frac{q_{N_{l+k-1}+1}}{q_{N_l}} = \frac{q_{N_{l-1}+1}}{q_{N_l}} \sum_{k=0}^{r-l} \Delta N_{l+k} \frac{q_{N_{l+k-1}+1}}{q_{N_{l-1}+1}}. \tag{67}$$

Choose inductively the numbers N_l as follows:

$$N_l = \max \left\{ k \geq N_{l-1} + 1 \mid q_k \geq \frac{1}{2} q_{N_{l-1}+1} \right\}. \tag{68}$$

In this way we have that

$$\frac{q_{N_{l-1}+1}}{q_{N_l}} \leq 2, \qquad \frac{q_{N_{l+k-1}+1}}{q_{N_{l-1}+1}} \leq 2^{-k}.$$

Inserting in (67), we thus obtain

$$\beta_l \leq 2 \sum_{k=0}^{r-l} \Delta N_{l+k} 2^{-k} \quad \forall l = 1, \ldots, r, \tag{69}$$

which implies that

$$\sum_{l=1}^{r} \beta_l \leq 2 \sum_{l=1}^{r} \sum_{k=0}^{r-l} \Delta N_{l+k} 2^{-k} = 2 \sum_{k=0}^{r-1} \left(\sum_{l=1}^{r-k} \Delta N_{l+k} \right) 2^{-k} \leq 2\mathbf{N} \sum_{k=0}^{r-1} 2^{-k} \leq 4\mathbf{N}.$$

Hence it follows from Lemma 3 that in our case the coefficients $\bar{p}_{\sigma,s}$ can be bounded as

$$\bar{p}_{\sigma,s} \leq \binom{r}{s} \binom{s}{\sigma} \left(\frac{\mathbf{N} D_2}{r} \right)^s \left(\frac{4\mathbf{N} D_1}{s} \right)^\sigma. \tag{70}$$

Bounds on the coefficients of the power series. Define the coefficients $a_k^{\sigma,s}$ by

$$\left(\frac{1}{1-\alpha_1 z} \right)^\sigma \left(\frac{1}{1-\alpha_2 z} \right)^s = \sum_{k=0}^{+\infty} a_k^{\sigma,s} z^k. \tag{71}$$

The aim of this part of the section is to determine bounds on the coefficients $a_k^{\sigma,s}$. Simple combinatorial manipulation shows that

$$a_k^{\sigma,0} = \binom{k+\sigma-1}{\sigma-1} \alpha_1^k \quad \forall \sigma \geq 1 \quad \forall k \geq 0. \tag{72}$$

In general we have the bound given by the following lemma.

LEMMA 5. *Assume that $\alpha_1 > \alpha_2$. Then, for every $s \geq 0$, $\sigma \geq 1$, and $k \geq 0$, we have*

$$0 \leq a_k^{\sigma,s} \leq \left(\frac{\alpha_1}{\alpha_1 - \alpha_2} \right)^s a_k^{\sigma,0}.$$

Proof. We start by proving that

$$a_k^{\sigma,1} \le \frac{\alpha_1}{\alpha_1-\alpha_2} a_k^{\sigma,0}$$

by induction on k. It is trivial if $k=0$. Assume it to be true for $k-1$ (with $k \ge 1$), and let us prove it for k. Then

$$a_k^{\sigma,1} = \sum_{h=0}^{k} a_h^{\sigma,0}\alpha_2^{k-h} = \sum_{h=0}^{k} \binom{h+\sigma-1}{\sigma-1} \alpha_1^h \alpha_2^{k-h}$$

$$= \sum_{h=0}^{k-1} \binom{h+\sigma-1}{\sigma-1} \alpha_1^h \alpha_2^{k-h} + \binom{k+\sigma-1}{\sigma-1} \alpha_1^k = \alpha_2 a_{k-1}^{\sigma,1} + \binom{k+\sigma-1}{\sigma-1} \alpha_1^k.$$

Using the induction we obtain

$$a_k^{\sigma,1} \le \frac{\alpha_2\alpha_1}{\alpha_1-\alpha_2} a_{k-1}^{\sigma,0} + \binom{k+\sigma-1}{\sigma-1} \alpha_1^k = \left[\frac{\alpha_2}{\alpha_1-\alpha_2} \binom{k+\sigma-2}{\sigma-1}\right.$$

$$\left. + \binom{k+\sigma-1}{\sigma-1}\right] \alpha_1^k = \left[\frac{\alpha_2}{\alpha_1-\alpha_2} \frac{k}{k+\sigma-1} + 1\right] \binom{k+\sigma-1}{\sigma-1} \alpha_1^k$$

$$\le \left[\frac{\alpha_2}{\alpha_1-\alpha_2} + 1\right] \binom{k+\sigma-1}{\sigma-1} \alpha_1^k = \frac{\alpha_1}{\alpha_1-\alpha_2} a_k^{\sigma,0}.$$

Finally assume that the assertion of the lemma holds true for $s-1$. Then

$$a_k^{\sigma,s} = \sum_{h=0}^{k} a_h^{\sigma,s-1}\alpha_2^{k-h} \le \left(\frac{\alpha_1}{\alpha_1-\alpha_2}\right)^{s-1} \sum_{h=0}^{k} a_h^{\sigma,0}\alpha_2^{k-h} = \left(\frac{\alpha_1}{\alpha_1-\alpha_2}\right)^{s-1} a_h^{\sigma,1}$$

$$\le \left(\frac{\alpha_1}{\alpha_1-\alpha_2}\right)^{s} a_h^{\sigma,0}. \qquad \square$$

6.3. The proof of Theorem 6: The final step. We now want to use the estimates obtained above for bounding the coefficients $\eta_{k,1}$.

From (62) we can argue that

(73)

$$\eta_1(z) = \sum_{s=0}^{r}\sum_{\sigma=0}^{s} \bar{p}_{\sigma,s} \left(\frac{1}{1-\alpha_1 z}\right)^{\sigma} \left(\frac{1}{1-\alpha_2 z}\right)^{s} z^{\sigma+s} = \sum_{s=0}^{r}\sum_{\sigma=0}^{s} \bar{p}_{\sigma,s} \sum_{h=0}^{+\infty} a_h^{\sigma,s} z^{h+\sigma+s}$$

$$= \sum_{s=0}^{r}\sum_{\sigma=0}^{s}\sum_{k=\sigma+s}^{+\infty} \bar{p}_{\sigma,s} a_{k-\sigma-s}^{\sigma,s} z^k = \sum_{k=0}^{+\infty}\sum_{s=0}^{r\wedge k}\sum_{\sigma=0}^{s\wedge k-s} \bar{p}_{\sigma,s} a_{k-\sigma-s}^{\sigma,s} z^k,$$

where $\bar{p}_{\sigma,s}$ was defined in (63) and $a_k^{\sigma,s}$ in (71). Hence we have

$$\eta_{k,1} = \sum_{s=0}^{r\wedge k}\sum_{\sigma=0}^{s\wedge k-s} \bar{p}_{\sigma,s} a_{k-\sigma-s}^{\sigma,s}.$$

Decompose $\eta_{k,1}$ as follows:

$$\eta_{k,1} = \eta'_{k,1} + \eta''_{k,1},$$

where

$$\eta'_{k,1} = \sum_{s=1}^{r\wedge k} \sum_{\sigma=1}^{s\wedge k-s} \bar{p}_{\sigma,s} a^{\sigma,s}_{k-\sigma-s}, \qquad \eta''_{k,1} = \sum_{s=0}^{r\wedge k} \bar{p}_{0,s} a^{0,s}_{k-s}. \tag{74}$$

Assume now that $\alpha_1 > \alpha_2$, and fix

$$M := \frac{8D_1}{\alpha_1} \vee \frac{2D_2}{\alpha_1 - \alpha_2} \vee \frac{D_2}{\alpha_2}. \tag{75}$$

Inserting bounds of Lemmas 4 and 5, we now obtain

$$\begin{aligned}
\frac{\eta'_{k,1}}{\alpha_1^k} &= \sum_{s=1}^{r\wedge k} \sum_{\sigma=1}^{s\wedge k-s} \bar{p}_{\sigma,s} \frac{a^{\sigma,s}_{k-\sigma-s}}{\alpha_1^k} \\
&\leq \sum_{s=1}^{r\wedge k} \sum_{\sigma=1}^{s\wedge k-s} \binom{s}{\sigma}\binom{r}{s}\left(\frac{4\mathbf{N}D_1}{s}\right)^{\sigma}\left(\frac{\mathbf{N}D_2}{r}\right)^{s}\left(\frac{\alpha_1}{\alpha_1-\alpha_2}\right)^{s}\binom{k-s-1}{\sigma-1}\frac{\alpha_1^{k-\sigma-s}}{\alpha_1^k} \\
&\leq \sum_{s=1}^{r\wedge k} \sum_{\sigma=1}^{s\wedge k-s} \binom{r}{s}\binom{s}{\sigma}\left(\frac{s}{r}\right)^{s}\left(\frac{\mathbf{N}M}{2s}\right)^{s+\sigma}\binom{k-s-1}{\sigma-1} \\
&\leq \left[\sum_{s=1}^{r\wedge k}\binom{r}{s}\left(\frac{s}{r}\right)^{s}\sum_{\sigma=1}^{s\wedge k-s}\binom{s}{\sigma}\binom{k-s-1}{\sigma-1}\right] \max_{s=1}^{r\wedge k}\max_{\sigma=0}^{s\wedge k-s}\left\{\left(\frac{\mathbf{N}M}{2s}\right)^{s+\sigma}\right\}.
\end{aligned}$$

Observe that

$$\max_{\sigma=0}^{s\wedge k-s}\left\{\left(\frac{\mathbf{N}M}{2s}\right)^{s+\sigma}\right\} = \left(\frac{\mathbf{N}M}{2s}\right)^{s} \bigvee \left(\frac{\mathbf{N}M}{2s}\right)^{2s\wedge k}$$

and that by (93),

$$\max_{s=1}^{r\wedge k}\left\{\left(\frac{\mathbf{N}M}{2s}\right)^{s}\right\} \leq \left(\frac{\mathbf{N}M/2}{k\wedge\frac{\mathbf{N}M}{2e}}\right)^{k\wedge\frac{\mathbf{N}M}{2e}} \quad \max_{s=1}^{r\wedge k}\left\{\left(\frac{\mathbf{N}M}{2s}\right)^{2\left(s\wedge\frac{k}{2}\right)}\right\} \leq \left(\frac{\mathbf{N}M}{k\wedge\frac{\mathbf{N}M}{e}}\right)^{k\wedge\frac{\mathbf{N}M}{e}},$$

which implies

$$\max_{s=1}^{r\wedge k},\ \max_{\sigma=0}^{s\wedge k-s}\left\{\left(\frac{\mathbf{N}M}{2s}\right)^{s+\sigma}\right\} \leq \left(\frac{\mathbf{N}M}{k\wedge\frac{\mathbf{N}M}{e}}\right)^{k\wedge\frac{\mathbf{N}M}{e}}.$$

From this fact and using the combinatorial identity (89), we obtain

$$\frac{\eta'_{k,1}}{\alpha_1^k} \leq \left[\sum_{s=1}^{r\wedge k}\binom{k-1}{s-1}\binom{r}{s}\left(\frac{s}{r}\right)^{s}\right]\left(\frac{\mathbf{N}M}{k\wedge\frac{\mathbf{N}M}{e}}\right)^{k\wedge\frac{\mathbf{N}M}{e}}. \tag{76}$$

On the other hand, assuming $k \geq 1$, similar computations show that

$$\frac{\eta''_{k,1}}{\alpha_1^k} = \sum_{s=1}^{r\wedge k} \bar{p}_{0,s} \frac{a_{k-s}^{0,s}}{\alpha_1^k} = \left[\sum_{s=1}^{r\wedge k} \binom{r}{s} \left(\frac{\mathbf{N}D_2}{r}\right)^s \binom{k-1}{s-1} \alpha_2^{-s}\right] \frac{\alpha_2^k}{\alpha_1^k}$$

$$\leq \left[\sum_{s=1}^{r\wedge k} \binom{r}{s} \left(\frac{\mathbf{N}M}{r}\right)^s \binom{k-1}{s-1}\right]$$

$$\leq \left[\sum_{s=1}^{r\wedge k} \binom{k-1}{s-1} \binom{r}{s} \left(\frac{s}{r}\right)^s\right] \max_{1\leq s\leq r\wedge k} \left\{\left(\frac{\mathbf{N}M}{s}\right)^s\right\},$$

which yields

$$\frac{\eta''_{k,1}}{\alpha_1^k} \leq \left[\sum_{s=1}^{r\wedge k} \binom{k-1}{s-1} \binom{r}{s} \left(\frac{s}{r}\right)^s\right] \left(\frac{\mathbf{N}M}{k \wedge \frac{\mathbf{N}M}{e}}\right)^{k\wedge \frac{\mathbf{N}M}{e}}. \tag{77}$$

Putting together (76) and (77), we obtain the final bound

$$\frac{\gamma_k}{\alpha_1^k} \leq \frac{\eta_{k,1}}{\alpha_1^k} \leq 2\left[\sum_{s=1}^{r\wedge k} \binom{k-1}{s-1} \binom{r}{s} \left(\frac{s}{r}\right)^s\right] \left(\frac{\mathbf{N}M}{k \wedge \frac{\mathbf{N}M}{e}}\right)^{k\wedge \frac{\mathbf{N}M}{e}},$$

which proves Theorem 6 in the case when $\alpha_1 > \alpha_2$. Observe that M depends only on the parameters α_1, α_2, D_1, and D_2.

In the case when $\alpha_2 \geq \alpha_1$, we replace the estimate in Lemma 5 with

$$0 \leq a_k^{\sigma,s} \leq \binom{k+s+\sigma-1}{s+\sigma-1} \alpha_2^k \quad \forall s+\sigma \geq 1 \quad \forall k \geq 0 \tag{78}$$

and fix

$$M = \frac{8D_1}{\alpha_2} \vee \frac{2D_2}{\alpha_2}.$$

Similar computations show that, for $k \geq 1$, we can estimate

$$\frac{\gamma_k}{\alpha_2^k} \leq \frac{\eta_{k,1}}{\alpha_2^k} \leq \left[\sum_{s=1}^{r\wedge k} \sum_{\sigma=0}^{s\wedge k-s} \binom{r}{s} \binom{s}{\sigma} \left(\frac{\mathbf{N}M}{2r}\right)^s \left(\frac{\mathbf{N}M}{2s}\right)^\sigma \binom{k-1}{s+\sigma-1}\right] \tag{79}$$

$$\leq \left[\sum_{s=1}^{r\wedge k} \binom{k+s-1}{2s-1} \binom{r}{s} \left(\frac{s}{r}\right)^s\right] \left(\frac{\mathbf{N}M}{k \wedge \frac{\mathbf{N}M}{e}}\right)^{k\wedge \frac{\mathbf{N}M}{e}}.$$

The proof of Theorem 6 is now complete. □

7. Proof of Theorem 2. The aim of this section is to obtain a representation of the language $\Sigma_*(\Gamma)$ by a finite state automaton or equivalently by a graph. This will be called a Markov representation of the language. Then we will show that this representation satisfies conditions (A), (B), and (C) of the previous section, and so we will be in a position to apply the estimates proposed there.

7.1. The Markov representation. Assume $\Gamma : I \to I$ is any piecewise affine map. The graph representation of the language $\Sigma_*(\Gamma)$ can be constructed as follows. We define as the set of vertices the set $\mathcal{V} := \Sigma_*(\Gamma)$ and as set of edges $\mathcal{E}$ the set given by

$$(\omega_0\omega_1\cdots\omega_{n-1} \to \omega_0\omega_1\cdots\omega_{n-1}\omega_n) \in \mathcal{E} \iff \omega_0\omega_1\cdots\omega_{n-1}\omega_n \in \Sigma_*(\Gamma). \tag{80}$$

Moreover, we introduce the following labeling $\xi : \mathcal{E} \to \mathcal{I} \cup \mathcal{J}$ on the edges:

$$\xi(\omega_0\omega_1\cdots\omega_{n-1} \to \omega_0\omega_1\cdots\omega_{n-1}\omega_n) = \omega_n.$$

Notice that $\Sigma_*(\Gamma)$ coincides with the set of all the labeled words associated with the finite paths on the graph starting from the empty word ϵ. This representation of $\Sigma_*(\Gamma)$ will be called a *Markov representation*. This can be simplified by considering an equivalence relation on the vertices. With each finite word $\omega_0\omega_1\cdots\omega_n \in \Sigma_*(\Gamma)$ we associate its *symbolic future*

$$\mathrm{fut}_\Sigma(\omega_0\omega_1\cdots\omega_n) = \big\{\overline{\omega}_0\overline{\omega}_1\cdots\overline{\omega}_k \mid \overline{\omega}_0 = \omega_n \text{ and } \omega_0\omega_1\cdots\omega_n\overline{\omega}_1\cdots\overline{\omega}_k \in \Sigma_*(\Gamma)\big\},$$

which is a subset of $\Sigma_*(\Gamma)$. More roughly, the symbolic future of a word $\omega_0\omega_1\cdots\omega_n$ is the set of words whose concatenation with $\omega_0\omega_1\cdots\omega_n$ is in $\Sigma_*(\Gamma)$.

Consider also the *geometric future* which is

$$\mathrm{fut}(\omega_0\omega_1\cdots\omega_n) = \Gamma^n(\omega_0 \cap \Gamma^{-1}\omega_1 \cap \ldots \cap \Gamma^{-n}\omega_n).$$

The following result is in [5].

PROPOSITION 4. *Let $\omega_0\omega_1\cdots\omega_n$ and $\nu_0\nu_1\cdots\nu_m$ be two words in $\Sigma_*(\Gamma)$. Then*

$$\mathrm{fut}(\omega_0\omega_1\cdots\omega_n) = \mathrm{fut}(\nu_0\nu_1\cdots\nu_m) \iff \mathrm{fut}_\Sigma(\omega_0\omega_1\cdots\omega_n) = \mathrm{fut}_\Sigma(\nu_0\nu_1\cdots\nu_m). \tag{81}$$

Now define $\overline{\mathcal{X}}$ to be the quotient of the set $\Sigma_*(\Gamma)$ by the equivalence relation

$$\omega'_0\cdots\omega'_n \equiv \omega''_0\cdots\omega''_m \Leftrightarrow \mathrm{fut}_\Sigma(\omega'_0\cdots\omega'_n) = \mathrm{fut}_\Sigma(\omega''_0\cdots\omega''_m). \tag{82}$$

The elements of $\overline{\mathcal{X}}$ will be called *states* and will be denoted by the symbol $\mathbf{x}$. The symbol $\langle\omega_0\omega_1\cdots\omega_n\rangle$ represents the state consisting of the equivalent class which contains the word $\omega_0\omega_1\cdots\omega_n$. States representable by words of length 1 will be called *principal states*. The equivalence relation defining $\overline{\mathcal{X}}$ ensures that any state $\mathbf{x} \in \overline{\mathcal{X}}$ has a well-defined geometric future $\mathrm{fut}(\mathbf{x})$. In fact, the geometric future $\mathrm{fut}(\mathbf{x})$ uniquely determines the state $\mathbf{x}$. Edges and labels can be naturally redefined on $\overline{\mathcal{X}}$ to obtain a new labeled graph denoted by $\overline{\mathcal{G}}$ which is still a Markov representation of $\Sigma_*(\Gamma)$ and so, with the property that the labeled sequences associated with the finite paths on $\overline{\mathcal{G}}$, starting from empty word, corresponds to all the possible sequences in $\Sigma_*(\Gamma)$.

Notice that there is an edge connecting a state $\mathbf{x}'$ to another state $\mathbf{x}''$ labeled with ω if and only if $\mathrm{fut}(\mathbf{x}'') = \Gamma(\mathrm{fut}(\mathbf{x}')) \cap \omega$. This shows that the Markov representation $\overline{\mathcal{G}}$ has the property that the terminal state of any edge is determined by its initial state and by its label. This means that $\overline{\mathcal{G}}$ is a deterministic automaton. This implies in particular that there is a one to one correspondence between paths $\mathbf{x}_1\mathbf{x}_2\cdots\mathbf{x}_k$ on the graph $\overline{\mathcal{G}}$ starting from a principal state and words in $\Sigma_*(\Gamma)$. In order to count the number of words in $\Sigma_*(\Gamma)$ of length k, it will thus be equivalent to count the paths in $\overline{\mathcal{G}}$ of the same length k.

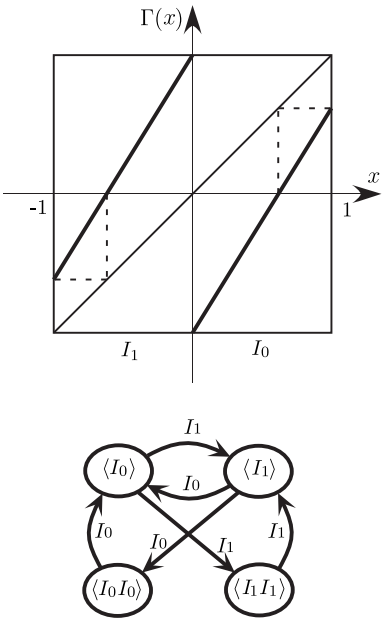


FIG. 6. *The map* Γ *of Example* 1 *and the graph* $\overline{\mathcal{G}}$ *describing the language associated with its dynamics.*

Example 1. We provide here a simple example which should clarify the concepts introduced so far. Consider the piecewise affine map $\Gamma : [-1,1] \to [-1,1]$ defined as follows:

$$\Gamma(x) := \begin{cases} ax+1 & \text{if } -1 < x < 0, \\ ax-1 & \text{if } 0 < x < 1, \end{cases}$$

where $a = \frac{1+\sqrt{5}}{2}$. The map $\Gamma(x)$ is shown in Figure 6. Let $I_0 :=]-1,0[$, and let $I_1 :=]0,1[$. For this particular choice of a we have that the set of states is finite:

$$\overline{\mathcal{X}} = \big\{\langle I_0\rangle, \langle I_1\rangle, \langle I_0 I_0\rangle, \langle I_1 I_1\rangle\big\}.$$

The graph $\overline{\mathcal{G}}$ is shown in Figure 6.

7.2. Properties of the Markov representation. Assume $\Gamma : I \to I$ is a piecewise affine map and that $J \subseteq I$ is another invariant interval as in the setting of section. We now want to show that the just introduced Markov representation restricted to $\Sigma_*(\Gamma) \cap \mathcal{I}^*$ (we are using the notation established in section 5.1) satisfies properties (A), (B), and (C) introduced in the previous section. To this aim we define

$$\mathcal{X}_P := \big\{\langle I_1\rangle, \langle I_2\rangle, \dots, \langle I_{\mathbf{N}}\rangle\big\},$$
$$\mathcal{X}_i := \big\{\langle \omega_0\omega_1\cdots\omega_k I_i\rangle \in \overline{\mathcal{X}} |\ \omega_0\omega_1\cdots\omega_k \in \Sigma_*(\Gamma)\cap\mathcal{I}^*\big\} = \big\{\mathbf{x}\in\overline{\mathcal{X}} \mid \mathrm{fut}(\mathbf{x}) \subseteq I_i\big\},$$
$$\mathcal{X} := \bigcup_{i=1}^{\mathbf{N}} \mathcal{X}_i = \big\{\mathbf{x}\in\overline{\mathcal{X}} \mid \mathrm{fut}(\mathbf{x}) \subseteq I\setminus J\big\},$$
$$q : \mathcal{X} \to]0,1[: \mathbf{x} \mapsto q(\mathbf{x}) := \mathbb{P}\big[\mathrm{fut}(\mathbf{x})\big],$$

and the graph $\mathcal{G}$ which coincides with the graph $\overline{\mathcal{G}}$ restricted to the set of states $\mathcal{X}$.

By taking $q_i = \mathbb{P}[I_i]$, we have that property (A) holds true. The next two lemmas will show that properties (B) and (C) also hold true with $\alpha_1 = |a|$, $\alpha_2 = 2$, $D_1 = |a|$, and $D_2 = 1$.

LEMMA 6. *Let $\mathbf{x}' \in \mathcal{X}$, $\mathcal{X}'' \subseteq \mathcal{X}$, and let $k \geq 2$. Then*

$$\#\mathcal{F}_k\big[\mathbf{x}_1 = \mathbf{x}', \mathbf{x}_2, \dots, \mathbf{x}_{k-1} \in \mathcal{X}, \mathbf{x}_k \in \mathcal{X}''\big] \leq \frac{\mathbb{P}\big[\mathrm{fut}(\mathbf{x}')\big]}{\inf_{\mathbf{x}'' \in \mathcal{X}''} \mathbb{P}\big[\mathrm{fut}(\mathbf{x}'')\big]} |a|^{k-1}.$$

Proof. Notice that the intervals of the form

$$\mathrm{fut}(\mathbf{x}') \cap \Gamma^{-1}\mathrm{fut}(\mathbf{x}_2) \cap \cdots \cap \Gamma^{-(k-2)}\mathrm{fut}(\mathbf{x}_{k-1}) \cap \Gamma^{-(k-1)}\mathrm{fut}(\mathbf{x}_k), \qquad \mathbf{x}_2, \dots, \mathbf{x}_k \in \mathcal{X},$$

constitute a family of disjoint subsets of $\mathrm{fut}(\mathbf{x}')$. This shows that

$$\begin{aligned}\mathbb{P}[\mathrm{fut}(\mathbf{x}')] \geq \sum_{\substack{\mathbf{x}_2, \dots, \mathbf{x}_{k-1} \in \mathcal{X} \\ \mathbf{x}_k \in \mathcal{X}''}} & \mathbb{P}\big[\mathrm{fut}(\mathbf{x}') \cap \Gamma^{-1}\mathrm{fut}(\mathbf{x}_2) \cap \cdots \\ & \cap \Gamma^{-(k-2)}\mathrm{fut}(\mathbf{x}_{k-1}) \cap \Gamma^{-(k-1)}\mathrm{fut}(\mathbf{x}_k)\big].\end{aligned}$$

Notice, moreover, that Γ^{k-1} is affine on each of these intervals and that

$$\Gamma^{k-1}\big(\mathrm{fut}(\mathbf{x}') \cap \Gamma^{-1}\mathrm{fut}(\mathbf{x}_2) \cap \cdots \cap \Gamma^{-(k-2)}\mathrm{fut}(\mathbf{x}_{k-1}) \cap \Gamma^{-(k-1)}\mathrm{fut}(\mathbf{x}_k)\big) = \mathrm{fut}(\mathbf{x}_k).$$

This implies that

$$\begin{aligned}&\mathbb{P}\big[\mathrm{fut}(\mathbf{x}') \cap \Gamma^{-1}\mathrm{fut}(\mathbf{x}_2) \cap \cdots \cap \Gamma^{-(k-2)}\mathrm{fut}(\mathbf{x}_{k-1}) \cap \Gamma^{-(k-1)}\mathrm{fut}(\mathbf{x}_k)\big] \\ &\quad \geq \frac{\inf_{\mathbf{x}'' \in \mathcal{X}''} \mathbb{P}[\mathrm{fut}(\mathbf{x}'')]}{|a|^{k-1}}\end{aligned}$$

if $\mathbf{x}'\mathbf{x}_2 \cdots \mathbf{x}_{k-1}\mathbf{x}_k \in \mathcal{F}_k[\mathbf{x}_1 = \mathbf{x}', \mathbf{x}_2, \dots, \mathbf{x}_{k-1} \in \mathcal{X}, \mathbf{x}_k \in \mathcal{X}'']$, and it is 0 otherwise. This yields the result. ∎

LEMMA 7. *Let $\mathbf{x}' \in \mathcal{X}$, and let $i = 1, \dots, \mathbf{N}$. Then*

$$\#\mathcal{F}_k[\mathbf{x}_1 = \mathbf{x}', \mathbf{x}_2, \dots, \mathbf{x}_{k-1} \in \mathcal{X} \setminus \mathcal{X}_P, \mathbf{x}_k \in \mathcal{X}_i] \leq 2^{k-2}.$$

Proof. As mentioned above, there is an edge connecting a state $\mathbf{x}'$ to another state $\mathbf{x}''$ with label ω if and only if $\mathrm{fut}(\mathbf{x}'') = \Gamma(\mathrm{fut}(\mathbf{x}')) \cap \omega$. Since the map Γ is affine on $\mathrm{fut}(\mathbf{x}')$, $\Gamma(\mathrm{fut}(\mathbf{x}'))$ is an interval, and so at most two followers of a state can be nonprincipal. The result follows by applying this argument. ∎

It follows from Lemmas 6 and 7 that the graph $\mathcal{G}$ satisfies properties (A), (B), and (C), and hence Theorem 6 holds true in this case. Notice that this yields Theorem 2, since γ_k defined in (30) coincides with γ_k defined in (30). Indeed, in this case we have that

$$\gamma_{k,h} = \#\mathcal{F}_k[\mathbf{x}_1 = \langle I_h \rangle, \mathbf{x}_2, \dots, \mathbf{x}_k \in \mathcal{X}],$$

so that $\gamma_k = \sum_h \gamma_{k,h}$ coincides with the number of paths of length k in the graph $\overline{\mathcal{G}}$, starting from a principal state and always remaining in $\mathcal{X}$. This, by the previous discussion, corresponds to the number of distinct subwords in $\Sigma_*(\Gamma) \cap \mathcal{I}^*$ of length k.

7.3. Estimation of the number of paths in the chaotic case. As mentioned in the remark after Theorem 6, in bound (53) we can fix r instead of the constant M. More precisely, instead of fixing the contraction factor equal to $1/2$ in (68) we can choose any $\delta \in]0,1[$. In this case, instead of (70), we obtain

$$\bar{p}_{\sigma,s} \leq \binom{r}{s}\binom{s}{\sigma}\left(\frac{\mathbf{N}D_2}{r}\right)^s\left(\frac{\mathbf{N}D_1}{s\delta(1-\delta)}\right)^\sigma. \tag{83}$$

In the case $\alpha_1 > \alpha_2$, the only consequence on the subsequent computations is that the factor $\frac{1}{\delta(1-\delta)}$ will enter in definition (75) of M. On the other hand, the number r also depends on the contraction factor δ. An important situation in which it is possible to take advantage of this degree of freedom is the following.

If we fix $\delta := q_{\mathbf{N}}/q_1 \wedge 1/2$, then $r = 1$, and in this way we obtain a simplified bound on γ_k in which, however, the constant M is a decreasing function of δ. In order to obtain an effective bound we need to have a bound from above on M, and so a bound from below on $q_{\mathbf{N}}/q_1$. In the context of piecewise affine maps this means that we need to have a bound from below on $\delta = \mathbb{P}[I_{\mathbf{N}}]/\mathbb{P}[I_1]$. An interesting situation in which this is possible is when $\mathbf{N} = \lceil |a| \rceil$, namely, for the chaotic quantized stabilizers.

PROPOSITION 5. *Let $|a| > 2$, and let $\mathbf{N} = \lceil |a| \rceil$. There exist constants $C_1 > 1$ and $M > 0$, depending only on $|a|$ such that if $C > C_1$, then*

$$\frac{\gamma_k}{|a|^k} \leq 2\left(\frac{\mathbf{N}M}{k \wedge \frac{\mathbf{N}M}{e}}\right)^{k \wedge \frac{\mathbf{N}M}{e}} \qquad \forall k \geq 1. \tag{84}$$

Proof. For the arguments presented above, we need only to prove that there exist constants $\delta_1 > 0$ and $C_1 > 1$, depending only on $|a|$ such that

$$C \geq C_1 \;\Rightarrow\; \frac{\mathbb{P}[I_{\mathbf{N}}]}{\mathbb{P}[I_1]} \geq \delta_1.$$

First notice that $1 \geq \mathbb{P}[\Gamma(I_1)] \geq |a|\,\mathbb{P}[I_1]$, from which we can argue that $\mathbb{P}[I_1] \leq 1/|a|$. Moreover,

$$\mathbb{P}[I_{\mathbf{N}}] = 1 - \mathbb{P}[J] - \sum_{h=1}^{\mathbf{N}-1}\mathbb{P}[I_h] \geq 1 - C^{-1} - (\mathbf{N}-1)\mathbb{P}[I_1] \geq 1 - C^{-1} - \frac{\lceil |a| \rceil - 1}{|a|},$$

and hence

$$\frac{\mathbb{P}[I_{\mathbf{N}}]}{\mathbb{P}[I_1]} \geq \frac{\mathbb{P}[I_{\mathbf{N}}]}{1/|a|} \geq |a| - |a|C^{-1} - \lceil |a| \rceil + 1 \overset{C\to\infty}{\longrightarrow} |a| - \lceil |a| \rceil + 1 > 0.$$

This proves the result. □

7.4. Estimation of the number of paths in the stable case. In this section, we will propose a bound on γ_k which holds true when Γ is (I,J)-stable or when Γ is almost (I,J)-stable but with only a countable subset of points in I never entering inside J. To obtain this bound we need the following lemma.

LEMMA 8. *Assume that there exists a state $\mathbf{x} \in \mathcal{X}$ such that there exist two distinct paths in the graph $\mathcal{G}$ both starting and ending in $\mathbf{x}$ and not passing by $\mathbf{x}$ in any intermediate step (simple loops through $\mathbf{x}$). Then there is an uncountable set of points in I never entering inside J.*

Proof. The proof is based on a general argument on the symbolic description of a one-dimensional expansive map as Γ which consists in constructing a sort of inverse of the map ψ defined in (29); see [5].

Given any loop $\nu = \mathbf{x}\mathbf{x}_1 \cdots \mathbf{x}_{k-1}\mathbf{x}$ in $\mathcal{G}$, if we consider the open interval

$$K_\nu = \mathrm{fut}(\mathbf{x}) \cap \Gamma^{-1}(\mathrm{fut}(\mathbf{x}_1)) \cap \cdots \cap \Gamma^{-(k-1)}(\mathrm{fut}(\mathbf{x}_{k-1})) \cap \Gamma^{-k}(\mathrm{fut}(\mathbf{x}))\,,$$

we have that Γ^k is affine on K_ν, and $\Gamma^k(K_\nu) = \mathrm{fut}(\mathbf{x})$. In particular it follows that

$$\mathbb{P}[K_\nu] = \mathbb{P}[\mathrm{fut}(\mathbf{x})]|a|^{-k}. \tag{85}$$

We now set some notation: if $\nu_1 = \mathbf{x}\mathbf{x}_1^1 \cdots \mathbf{x}_{k_1-1}^1\mathbf{x}$ and $\nu_2 = \mathbf{x}\mathbf{x}_1^2 \cdots \mathbf{x}_{k_2-1}^2\mathbf{x}$ are two loops through $\mathbf{x}$, we define the concatenation of ν_1 and ν_2 as the new loop

$$\nu = \nu_1 \wedge \nu_2 = \mathbf{x}\mathbf{x}_1^1 \cdots \mathbf{x}_{k_1-1}^1\mathbf{x}\mathbf{x}_1^2 \cdots \mathbf{x}_{k_2-1}^2\mathbf{x}.$$

Assume that there are two distinct simple loops ν_1 and ν_2 of length k_1 and k_2, respectively, through $\mathbf{x}$. The corresponding open intervals K_1 and K_2 as defined above are then disjoint. Define now a map $\Upsilon : \{1,2\}^{\mathbb{N}} \to I$ in the following way: given a sequence $(a_n) \in \{1,2\}^{\mathbb{N}}$, consider the set

$$\overline{K}_{a_1} \cap \Gamma^{-k_{a_1}}\left(\overline{K}_{a_2}\right) \cap \Gamma^{-k_{a_1}-k_{a_2}}\left(\overline{K}_{a_3}\right) \cap \cdots = \bigcap_{n=1}^{+\infty} \Gamma^{-\sum_{j=1}^{n-1} k_{a_j}}\left(\overline{K}_{a_n}\right). \tag{86}$$

Since

$$\bigcap_{n=1}^{q} \Gamma^{-\sum_{j=1}^{n-1} k_{a_j}}\left(\overline{K}_{a_n}\right)$$

is simply the closure of the open interval K associated with the loop $\nu_{a_1} \wedge \nu_{a_2} \wedge \cdots \wedge \nu_{a_q}$, it follows that it is nonempty and that, by (85), its size decreases by a factor

$$|a|^{-\sum_{j=1}^{n-1} k_{a_j}}.$$

Hence this implies that the set in (86) consists of exactly one point x. We then put $\Upsilon((a_n)) = x$. Call $\Delta = \Upsilon(\{1,2\}^{\mathbb{N}})$. A standard argument of symbolic dynamics of one-dimensional maps now show that there exists $\Delta_1 \subseteq \Delta$, at most countable, such that the counterimage set $\Upsilon^{-1}(x)$ is a singleton for every $x \in \Delta \backslash \Delta_1$. Indeed, it follows by the definition that the only points x which have more than one counterimage (and in fact exactly two) are those in the union of boundaries of the intervals

$$\bigcap_{n=1}^{q} \Gamma^{-\sum_{j=1}^{n-1} k_{a_j}}\left(\overline{K}_{a_n}\right),$$

namely, those in the subset

$$\Delta_1 = \bigcup_{q=1}^{+\infty} \bigcup_{a_1,\ldots a_q} \partial\left(\bigcap_{n=1}^{q} \Gamma^{-\sum_{j=1}^{n-1} k_{a_j}}\left(\overline{K}_{a_n}\right)\right),$$

which is clearly at most countable. Finally, the subset of points in Δ which will never enter inside Δ_1,

$$\Delta_2 = \bigcap_{k=0}^{+\infty} \Gamma^{-k}(\Delta \setminus \Delta_1),$$

is clearly uncountable.

We claim that no point in Δ_2 will ever enter inside J. Notice first that, by construction, $\Delta_2 \subseteq \Omega$. Now take $x \in \Delta_2$, and let $(a_n) \in \{1,2\}^{\mathbb{N}}$ be such that $\Upsilon(a_n) = x$. Then

$$\Gamma^{k_{a_1}}(\Upsilon(a_n)) = \Upsilon(\tilde{a}_n),$$

where $(\tilde{a}_n)$ is a sequence defined by $\tilde{a}_n = a_{n+1}$ for all $n \in \mathbb{N}$. This implies in particular that $\Gamma^{k_{a_1}}(x) \in \Delta_2$ by the way Δ_2 has been defined. Hence we have that for every $x \in \Delta_2$ either $\Gamma^{k_1}(x)$ or $\Gamma^{k_2}(x)$ is also in Δ_2. If, by contradiction, n_0 exists such that $\Gamma^n x \in J$ for every $n \geq n_0$, we could find for sure $n_1 \geq n_0$ such that $y = \Gamma^{n_1} x \in \Delta_2 \cap J$. Since $\Delta_2 \subseteq K_1 \cup K_2$ it would follow that $y \in \partial K_1 \cup \partial K_2$, which is absurd by the way Δ_2 has been defined. □

THEOREM 7. *Assume that Γ is almost (I,J)-stable with an at most countable subset of points in I never entering inside J. Then*

$$\frac{\gamma_k}{2^k} \leq \binom{k + 2\mathbf{N} - 1}{2\mathbf{N} - 1} e^{\frac{\mathbf{N}}{e}} \qquad \forall k \geq 1. \tag{87}$$

Proof. Decompose the set $\mathcal{X}_P$ into maximal subfamilies $\mathcal{X}_P^1, \mathcal{X}_P^2, \ldots, \mathcal{X}_P^m$ in such a way that two principal states belong to the same family if and only if there exists a loop in $\mathcal{G}$ connecting them. Also we can assume the families are ordered in such a way that if there exists a path from $\mathbf{x}_1 \in \mathcal{X}_P^i$ to $\mathbf{x}_2 \in \mathcal{X}_P^j$, then $i \leq j$. Let N_i be the cardinality of $\mathcal{X}_P^i$. We thus have $\mathbf{N} = \sum_{i=1}^m N_i$.

Given now any path ν of length k inside the graph $\mathcal{G}$ starting from a principal state, we can always split it as

$$\nu = \nu_1 \mu_1 \nu_2 \mu_2 \cdots \nu_m \mu_m,$$

where ν_i is a path connecting two principal states in $\mathcal{X}_P^i$, while ν_i is a path only consisting of nonprincipal states. Assume ν_i has length k_i^1 and that μ_i has length k_i^2. We thus have

$$k = \sum_{i=1}^m k_i^1 + \sum_{i=1}^m k_i^2.$$

The number of ways we can split k in the sum above is equal to

$$\binom{k + 2m - 1}{2m - 1}.$$

Once the numbers k_i^1 and k_i^2 have been fixed, we notice that the path ν_i can be chosen in N_i distinct ways corresponding to the ways we can choose the initial principal state. This follows from the fact that from any principal state in $\mathcal{X}_P^i$ there is exactly one path reaching another element in $\mathcal{X}_P^i$ because otherwise there would be two distinct

simple loops in $\mathcal{G}$ contradicting the result in Lemma 8. Notice that using the fact that $\sum_{i=1}^{m} N_i = \mathbf{N}$, by Lemma 3, the number of ways we can choose the family of paths $\nu_1, \nu_1, \ldots, \nu_m$ is bounded from above by

$$\prod_{i=1}^{m} N_i \leq \left(\frac{\mathbf{N}}{m}\right)^m \leq e^{\frac{\mathbf{N}}{e}}.$$

Once all the paths ν_i have been chosen, the remaining paths μ_i can be chosen in at most $2^{k_i^2}$ distinct ways. Hence the number of ways we can choose the family of paths $\mu_1, \mu_1, \ldots, \mu_m$ is bounded from above by 2^k. We thus have the thesis. □

8. Conclusions. In this paper, some stabilizing quantized feedback strategies are proposed and their different properties in terms of performance and communication requirements are compared. These strategies are based on nesting one base quantized feedback. The performance, defined as the expected time needed to get from a big initial state set into a smaller target state set, is analyzed by using the concept of the Perron–Frobenius operator associated with a nonlinear transformation.

The second part of the paper is devoted to the search of general bounds which could highlight the trade-off existing between performance and information flow required by a quantized control technique. This investigation is based on a symbolic representation of the closed loop nonlinear system. In this way the system is described by a Markov chain with possibly infinite states. Counting the paths on the graph which represents the Markov chain, it is possible to obtain bounds on the performance which yield to some interesting trade-off relations. This method is based on a technical result which is expressed in terms of general Markov chains and its proof, though quite long, is based on basic combinatorial relations.

It is our hope that, as information theory has been a successful symbolic technique to treat digital communication, a symbolic technique will be the right tool to deal with digital control as well. In fact, although this paper deals only with the static control of linear scalar systems, the symbolic method proposed here seems to be very promising for treating more general situations. In [10] the same method is applied for treating both the case in which a memory structure is allowed on the controller and the case in which the system is multidimensional. We hope that this method will be useful to solve other questions which remain open. In our opinion the most important ones are the following.

1. In most of the contributions on control with communication constraint proposed in the literature it is assumed that the channels are digital with a finite rate but are noiseless. In the future investigations it will be important to allow the presence of errors in the data exchange between the plant and the controller.
2. In our opinion more attention has to be devoted to the control problem with communication constraint in those situations in which there are more interacting agents to be controlled to achieve a joint control objective. In this case the communication constraint have to be imposed on the data which are exchanged by the differently located agents.
3. In this paper, we have been able to analyze the performance of some simple quantized feedback strategies. It remains to obtain an algorithm able to provide an approximate performance evaluation for any given specific quantized feedback. In our opinion a promising method could be based on the approximation of the Perron–Frobenius operator by a finite state Markov chain

which is connected with the so-called Ulam conjecture (see [19] and references therein).

Appendix: Some useful elementary combinatorics. In the paper, we use some elementary properties of the binomials. The first one is

$$\sum_{j=0}^{m} \binom{l+j}{j} = \binom{m+l+1}{m}, \tag{88}$$

which follows by iterating the elementary identity

$$\binom{m+l+1}{m} = \binom{m+l}{m} + \binom{m+l}{m-1}.$$

Another useful formula follows by comparing the binomial coefficients of the term z^k in the polynomial identity

$$(1+z)^{n_1}\left(1+z^{-1}\right)^{n_2} = (1+z)^{n_1+n_2} z^{-n_2},$$

which yields

$$\sum_{j=0}^{(n_1-k)\wedge n_2} \binom{n_1}{k+j}\binom{n_2}{j} = \binom{n_1+n_2}{k+n_2}. \tag{89}$$

Another useful formula is given by the following series of inequalities which holds true for all $n, m \geq 1$ [1, p. 113]:

$$\begin{aligned}\binom{n+m}{m} &\leq \sqrt{\frac{1}{2\pi}\left(\frac{1}{n}+\frac{1}{m}\right)}\left(1+\frac{n}{m}\right)^m\left(1+\frac{m}{n}\right)^n \\ &\leq \sqrt{\frac{1}{2\pi}\left(\frac{1}{n}+\frac{1}{m}\right)}\left(1+\frac{n}{m}\right)^m e^m \leq \sqrt{\frac{1}{2\pi}\left(\frac{1}{n}+\frac{1}{m}\right)}e^{n+m}.\end{aligned} \tag{90}$$

From (90) we can argue that for all $n \geq 0$ and $m \geq 1$,

$$\binom{n+m}{m} \leq \frac{1}{\sqrt{\pi}}\left(1+\frac{n}{m}\right)^m e^m. \tag{91}$$

Finally consider the function

$$f(x) := \left(\frac{A}{x}\right)^{Bx}.$$

This is a unimodal function having a unique maximum in $x_M = \frac{A}{e}$. This implies that for all $\bar{x} > 0$, we have

$$\max_{0<x\leq\bar{x}} f(x) = f(\bar{x} \wedge x_M) = \left(\frac{A}{\bar{x} \wedge \frac{A}{e}}\right)^{B\left(\bar{x}\wedge\frac{A}{e}\right)}. \tag{92}$$

Observe, moreover, that for all $\hat{x} > 0$, we have

$$\left(\frac{A}{x}\right)^{B(x\wedge\hat{x})} \leq \left(\frac{A}{x\wedge\hat{x}}\right)^{B(x\wedge\hat{x})},$$

which implies that

$$\max_{0<x\le\bar{x}}\left(\frac{A}{x}\right)^{B(x\wedge\hat{x})} \le \max_{0<x\le\bar{x}}\left(\frac{A}{x\wedge\hat{x}}\right)^{B(x\wedge\hat{x})} = \max_{0<x\le\bar{x}\wedge\hat{x}} f(x) = \left(\frac{A}{\bar{x}\wedge\hat{x}\wedge\frac{A}{e}}\right)^{B(\bar{x}\wedge\hat{x}\wedge\frac{A}{e})}. \tag{93}$$

REFERENCES

[1] R.B. Ash, *Information Theory,* Dover, New York, 1990 (corrected reprint of the 1965 original).

[2] J. Baillieul, *Feedback designs in information-based control,* in Proceedings of the Workshop on Stochastic Theory and Control, Kansas, Springer-Verlag, New York, 2001, pp. 35–57.

[3] R.W. Brockett and D. Liberzon, *Quantized feedback stabilization of linear systems,* IEEE Trans. Automat. Control, AC-45 (2000), pp. 1279–1289.

[4] A. Broise, F. Dal'bo, and M. Peigné, *Études spectrales d'opérateurs de transfert et application,* Astérisque, 238 (1996), pp. 1–177.

[5] J. Buzzi, *Intrinsic ergodicity of affine maps in* $[0,1]^d$*,* Monatsh. Math., 124 (1997), pp. 97–118.

[6] D.F. Delchamps, *Stabilizing a linear system with quantized state feedback,* IEEE Trans. Automat. Control, AC-35 (1990), pp. 916–924.

[7] N. Elia and S. K Mitter, *Stabilization of linear systems with limited information,* IEEE Trans. Automat. Control, AC-46 (2001), pp. 1384–1400.

[8] F. Fagnani, *Chaotic quantized feedback stabilizers: The scalar case,* Commun. Inf. Syst., 4 (2004), pp. 53–72.

[9] F. Fagnani and S. Zampieri, *Stability analysis and synthesis for scalar linear systems with a quantized feedback,* IEEE Trans. Automat. Control, AC-48 (2003), pp. 1569–1584.

[10] F. Fagnani and S. Zampieri, *Quantized stabilization of linear systems: Complexity versus performance,* IEEE Trans. Automat. Control, AC-49 (2004), pp. 1534–1548.

[11] T.W. Hungerford, *Algebra*, Springer-Verlag, New York, 1974.

[12] C.T. Ionescu-Tulcea and G. Marinescu, *Théorie ergodique pour des classes d'opérations non completement continues,* Ann. of Math., 52 (1950), pp. 140–147.

[13] J.G. Kemeny and J.L. Snell, *Finite Markov Chains,* Springer-Verlag, New York, 1976.

[14] A. Lasota and M.C. Mackey, *Chaos, Fractals, and Noise,* Springer-Verlag, New York, 1994.

[15] A. Lasota and J.A. Yorke, *On the existence of invariant measures for piecewise monotonic transformations,* Trans. Amer. Math. Soc., 186 (1973), pp. 481–488.

[16] D. Liberzon, *On stabilization of linear systems with limited information,* IEEE Trans. Automat. Control, AC-48 (2003), pp. 304–307.

[17] D. Lind and B. Marcus, *Symbolic Dynamics and Coding,* Cambridge University Press, Cambridge, UK, 1995.

[18] C. Liverani, *Decay of correlations in piecewise expanding maps,* J. Statist. Phys., 78 (1995), pp. 1111–1129.

[19] C. Liverani, *Rigorous numerical investigation of the statistical properties of piecewise expanding maps. A feasibility study,* Nonlinearity, 3 (2001), pp. 463–490.

[20] G.N. Nair and R.J. Evans, *Stabilization with data-rate-limited feedback: Tightest attainable bounds,* Systems Control Lett., 41 (2000), pp. 49–56.

[21] G.N. Nair and R.J. Evans, *Exponential stabilisability of finite-dimensional linear systems with limited data rates,* Automatica, 39 (2002), pp. 585–593.

[22] I.R. Petersen and A.V. Savkin, *Multirate stabilization of multivariable discrete-time linear systems via a limited capacity communication channel,* in Proceedings of CDC Conference, Las Vegas, 2002, pp. 304–309.

[23] B. Picasso, F. Gouaisbaut, and A. Bicchi, *Construction of invariant and attractive sets for quantized-input linear systems,* in Proceedings of CDC Conference, Las Vegas, 2002, pp. 824–829.

[24] B. Saussol, *Absolutely continuous invariant measures for multidimensional expanding maps,* Israel J. Math., 116 (2000), pp. 223–248.

[25] S. Tatikonda and S.K. Mitter, *Control under communication constraints,* IEEE Trans. Automat. Control, AC-49 (2004) pp. 1056–1068.

[26] W.S. Wong and R.W. Brockett, *Systems with finite communication bandwidth constraints II: Stabilization with limited information feedback,* IEEE Trans. Automat. Control, AC-44 (1999), pp. 1049–1053.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 3, pp. 867–884

FAITHFUL REPRESENTATIONS FOR CONVEX HAMILTON–JACOBI EQUATIONS*

FRANCO RAMPAZZO†

Abstract. When a Hamiltonian $H = H(t, x, p)$ is convex in the adjoint variable p, the corresponding Hamilton–Jacobi equation

$$u_t + H(t, x, u_x) = 0 \tag{0.1}$$

is known to be the Bellman equation of a suitable optimal control problem. Of course, the latter is not unique, so it is interesting to select a *good* optimal control problem among those representing (0.1). We call such a representation *faithful* if (i) it involves a dynamics which is locally Lipschitz continuous in the state variable—so that a unique trajectory corresponds to any given control and initial point—and (ii) *the Lagrangian displays the same regularity as H in the x variable.* The main result of the present paper establishes the existence of faithful representations for a large class of Hamiltonians, including those for which the standard comparison theorems (of viscosity solution theory) are valid. Moreover, our investigation includes t-measurable Hamiltonians as well.

If a faithful control-theoretical representation does exist (and (0.1) enjoys uniqueness properties), one can infer sharp regularity results for the solution of (0.1) just by studying the regularity of the value function of the associated optimal control problem. A further application consists of a simple interpretation of the front propagation phenomenon in terms of optimal trajectories of the underlying minimum problem.

Key words. HJ equations, representation of Hamiltonians, parameterization of set-valued maps

AMS subject classifications. 70H20, 49J24, 35E10

DOI. 10.1137/S0363012903436855

1. Introduction.

1.1. Some notation and conventions. We shall call *modulus* any increasing, continuous function $\omega : [0, +\infty[\to [0, +\infty[$ such that $\omega[0] = 0$. A *local modulus* will be a continuous map $\omega : [0, +\infty[\times [0, +\infty[\to [0, +\infty[$ that is increasing in the first variable and is a modulus in the second variable. The closed ball of $\mathbb{R}^n$ of radius $R \geq 0$ will be denoted by $\mathbf{B}_R$, and $\mathbf{B}$ will stand in place of $\mathbf{B}_1$. For each map $\varphi : \mathbb{R}^n \to \mathbb{R} \cup +\infty$, the domain of φ, i.e., the subset of those $v \in \mathbb{R}^n$ such that $\varphi(v) < +\infty$, will be denoted by $\mathrm{dom}(\varphi(\cdot))$. For any map $H : [0, T] \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, H^* will denote the conjugate map with respect to the third variable; that is, we shall set

$$H^*(t, x, v) \doteq \sup_{p \in \mathbb{R}^n} \{p \cdot v - H(t, x, p)\}$$

for all $(t, x, v) \in [0, T] \times \mathbb{R}^n \times \mathbb{R}^n$.

If $w = w(y_1, \ldots, y_q)$ is a map of many (possibly vector-valued) variables, for any $i = 1, \ldots, q$ we shall use w_{y_i} to denote the gradient with respect to the y_i variable. It will be clear by the context whether this has to be intended in the sense of viscosity solution theory.

*Received by the editors October 29, 2003; accepted for publication (in revised form) October 18, 2004; published electronically September 15, 2005. This work was partially supported by the MIUR COFIN project "Metodi di viscositá, metrici e di teoria del controllo in equazioni alle derivate parziali nonlineari."

http://www.siam.org/journals/sicon/44-3/43685.html

†Dipartimento di Matematica Pura ed Applicata, Università di Padova, Via Belzoni 7, 35131 Padova, Italy (rampazzo@math.unipd.it).

1.2. Statement of the problem. For every $(t,x) \in [0,T] \times \mathbb{R}^n$ let us consider the Bolza optimal control problem

$$(\mathcal{P}_{t,x}) \qquad \begin{array}{ll} \text{minimize} & \int_t^T l(s,y(s),a(s))ds + g(y(T)), \\ & \dot{y}(s) = f(s,y(s),a(s)), \\ & y(t) = x, \end{array}$$

where the controls $a(\cdot)$ (are measurable maps on $[t,T]$ and) take values in some subset A of a Euclidean vector space and the *Lagrangian-dynamics* pair (l,f) verify suitable hypotheses which will be made precise later. The function g will be assumed continuous, even though weaker assumptions could be considered (see Remark 2.1 below).

The Bellman–Cauchy problem corresponding to the family of optimal control problems $\{\mathcal{P}_{t,x},\ (t,x) \in [0,T] \times \mathbb{R}^n\}$ is defined as the Hamilton–Jacobi equation

$$u_t + H(t,x,u_x) = 0 \qquad \text{in }]0,T[\times \mathbb{R}^n \tag{1.1}$$

with the initial condition

$$u(0,x) = g(x) \qquad \forall x \in \mathbb{R}^n, \tag{1.2}$$

where

$$H(t,x,p) \doteq \sup_{a \in A} \{p \cdot f(t,x,a) - l(t,x,a)\}. \tag{1.3}$$

As is well known, the connection between the Bolza problems $(\mathcal{P}_{t,x})$ and the initial value problem (1.1)–(1.2) relies on the fact that if $V(t,x)$ is the value function of $(\mathcal{P}_{t,x})$, that is,

$$V(t,x) = \inf_{a(\cdot)} \int_t^T l(s,y(s),a(s))ds + g(y(T)), \tag{1.4}$$

then the map $u(t,x) = V(T-t,x)$ is a solution (e.g., viscosity [BCD] or minmax [Su] solution) of (1.1)–(1.2). Notice, in particular, that the Hamiltonian is convex in the gradient variable.

Let us consider the converse question. Suppose the Cauchy problem (1.1)–(1.2) is given, with only the information that H is convex in the gradient variable (plus other technical conditions which guarantee existence and uniqueness of *solutions* to (1.1)–(1.2)). Then it is natural to wonder whether (1.1)–(1.2) is the Bellman–Cauchy problem of a family $\{\mathcal{P}_{t,x},\ (t,x) \in [0,T] \times \mathbb{R}^n\}$ of optimal control problems. This means that one looks for a triple (A,f,l) such that (1.3) is verified. Such a triple will be called a (*control theoretical*) *representation of* H.

It is easily seen that if a representation of H exists, then infinitely many others exist.[1] So, we may consider a further question, namely, that of choosing a representation verifying some given properties.

[1] For instance, the map $H(p) = |p|$ is the Hamiltonian corresponding to the trivial optimal control problem

$$\begin{array}{ll} \text{minimize} & \int_t^T (1 - |a(s)|)ds, \\ & \dot{y}(s) = a(s), \qquad a(s) \in [-1,1], \\ & y(t) = x. \end{array}$$

On the other hand, for every pair $(h(\cdot),k(\cdot))$ of positive maps, $H(p) = |p|$ is also the Hamiltonian of

Indeed, this is our aim, which, loosely speaking, consists of finding representations that allow both *uniqueness of trajectories of* f (for any given control) and *a Lagrangian with the same* x*-regularity* as the given Hamiltonian.

In order to define the problem let us begin by stating properties (A1)–(A3) below, which are the properties we wish to be satisfied by a family of optimal control problems. They will imply certain conditions on H, which have to be considered as sort of *minimal assumptions* for our problem.

Given a family $\{\mathcal{P}_{t,x},\ (t,x) \in [0,T] \times \mathbb{R}^n\}$ of Bolza optimal control problems, we shall consider the following hypotheses on the triple (A, f, l):

(A1) *There exists a constant* Q *such that*

$$|f(t,0,a,)|, |l(t,0,a)| \le Q$$

for all $t \in [0,T]$ *and* $a \in A$.

(A2) *The maps* f *and* l *are continuous from* $[0,T] \times \mathbb{R}^n \times A$ *into* $\mathbb{R}^n$ *and* $\mathbb{R}$, *respectively, and for every* $R > 0$ *there exists a nonnegative number* E_R *such that*

$$|f(t,x,a) - f(t,y,a)| \le E_R|x-y|, \tag{1.5}$$

$$|l(t,x,a) - l(t,y,a)| \le \nu[R, |x-y|] \tag{1.6}$$

for all $(t,x,a), (t,y,a) \in [0,T] \times \mathbf{B}_R \times A$, *where* ν *is a suitable local modulus.*

(A3) *There is* $C > 0$ *such that*

$$|f(t,x,a)| \le C(1+|x|)$$

for all $(t,x,a) \in [0,T] \times \mathbb{R}^n \times A$.

From a control theoretical viewpoint these are rather standard hypotheses for the triple (A, f, l). In turn, it is straightforward to verify that they imply the following properties for the Hamiltonian H defined in (1.3):

(H1) *For any* $(t,x) \in [0,T] \times \mathbb{R}^n$, *the map* $q \mapsto H(t,x,q)$ *is convex from* $\mathbb{R}^n$ *into* $\mathbb{R}$.

(H2) *There exist local moduli* ω_1, ω_2, *and* ω_3 *such that for any* $R > 0$, *one has*

$$|H(t,x,p) - H(t,y,p)| \le \omega_1[R, |x-y|(1+|p|)] \tag{1.7}$$

and

$$|H(t,x,p) - H(s,x,p)| \le |p|\omega_2[R, |t-s|] + \omega_3[R, |t-s|] \tag{1.8}$$

for all $x, y \in \mathbf{B}_R$, $p \in \mathbb{R}^n$, *and* $t, s \in [0,T]$.

(H3) *There exists a constant* C *such that*

$$|H(t,x,p) - H(t,x,q)| \le C(1+|x|)|p-q| \tag{1.9}$$

for all $(t,x) \in [0,T] \times \mathbb{R}^n$ *and* $p, q \in \mathbb{R}^n$.

(H4) *For every* $R > 0$, *there exists a nonnegative number* N_R *such that*

$$|H^*(t,x,v)| \le N_R$$

for all $(t,x) \in [0,T] \times \mathbf{B}_R$ *and* $v \in \mathrm{dom}(H^*(t,x,\cdot))$.

---

the (much more involved) optimal control problem

$$\begin{array}{ll} \text{minimize} & \int_t^T \frac{h(y(s))}{a^2(s)}\,ds, \\ & \dot{y}(s) = \frac{a(s)k(y(s))}{1+|a(s)|k(y(s))} \cdot \frac{1-a^2(s)}{1+a^2(s)}, \qquad a(s) \in \mathbb{R}, \\ & y(t) = x. \end{array}$$

1.3. Aim. Assumptions (H1)–(H4), beyond being *necessary* for (A1)–(A3) (see Remark 2.2), are in fact verified with

$$\omega_1[R,s] = E_R(s + \nu[R,s]).$$

That is, the local modulus of continuity (in x) of H turns out to coincide—up to a sum with a linear mapping and a multiplication by a positive number, both depending on the radius R—with the local modulus of l. We say that H *inherits the same continuity (in x) from l.*

For a given Hamiltonian H verifying (H1)–(H4), we wish to find a representation (A, f, l) such that f *is locally Lipschitz continuous in x*—so that uniqueness of trajectories is guaranteed—and l *has the same kind of continuity (in x) as H.*

To be more precise, this means that we are looking for a triple (A, f, l) verifying (A1)–(A3) *with*

$$\nu[R,s] = P_R(s + \omega_1[R,s])$$

for suitable coefficients $P_R(\geq 0)$.

Remark 1.1. Up to now, the major contribution to the representation's issue for convex Hamiltonians could be found in Ishii [Is2]. As a matter of fact, in [Is2] representations were provided such that both f and l turn out to have a modulus of continuity equal to $(\omega_1)^{\frac{1}{2}}$ (while the control set turns out to be infinite-dimensional). This implies, for instance, that even in the quite regular case when $\omega_1[R,s] = L \cdot s$, f and l turn out to be just $\frac{1}{2}$-Holder continuous in x. In particular, the Lagrangian is *less regular* than the Hamiltonian, and the Cauchy problems for the control vector field f in general admit multiple solutions (for each control). On the contrary, in such a situation our result implies that both the dynamics f and the Lagrangian l are locally Lipschitz continuous in the state variable.

Remark 1.2. Problems with no convexity were investigated, e.g., by Ishii in [Is3] and by Evans and Souganidis in [ES]. Both papers aimed toward a representation of the solution in terms of an (Elliot–Kalton) upper or lower value of a suitable differential game. The dynamics of Ishii's representation involves infinite-dimensional control sets for the opponents in the game and displays a sort of Lipschitz continuity on compact sets. Instead the Lagrangian is just continuous. On the other hand, in [ES] Hamiltonians as well as initial data are restricted to Lipschitz continuous, bounded, functions. When referred to the case with convexity these results are weaker than ours. However, a comparison actually does not make sense because of the greater generality of the problems treated in [Is3] and [ES]. As a matter of fact, the lack of convexity could well be a serious drawback in the attempt to give a representation with a Lagrangian *as regular (in x)* as the Hamiltonians—apart from the Lipschitz bounded case treated in [ES].

1.4. Main results and an outline of the paper. The main contribution of this paper—see Theorems 2.1 and 2.2 below—consists of accomplishing *the twofold program of finding a locally Lipschitz continuous dynamics f and a Lagrangian l that preserves the same kind of continuity (in x) of the Hamiltonian.* Moreover, the control set A in our representation turns out to be particularly simple, namely, the unit ball of $\mathbb{R}^n$. Lastly (see section 6) we can prove extensions of these results to Hamiltonians measurably dependent on t.

As a first consequence of such results, many statements in the literature that have been proved for a Hamiltonian displaying an explicit control-theoretical form

as in (1.3) can now be updated by considering Hamiltonians H that merely verify (H1)–(H4) (and, for some specific purposes, a further technical hypothesis (H5)). Furthermore, some results concerning the solution of the Cauchy problem (1.1) can be sharpened by means of the control-theoretical representation we are providing. For instance, this is the case of the regularity of the solution to (1.1), which is addressed in section 5. Finally, already known results may be interpreted as facts concerning the optimal trajectories of the underlying optimal control problem, as it happens for the phenomenon of front propagation (see section 5).

As for the proof of the main result, let us remark that it is based essentially on the following arguments. First, in Theorem 3.2 below we establish (by means of an argument based on Kakutani's fixed point theorem) that under hypotheses (H1)–(H4) the multifunction that maps (t,x) into $F(t,x) \doteq \mathrm{dom}(H^*(t,x,\cdot))$ is *locally Lipschitz continuous* in x (and an analogous fact holds in the case of Hamiltonians measurable in t). Theorem 3.3 yields a global version of this result. Observe that the presence of the local modulus ω_1 in (H2) would suggest an (at most) ω_1-regularity for this multifunction rather than the local Lipschitz continuity actually obtained by means of our results. Secondly, we exploit a parameterization theorem for convex multifunctions proved in [O] (see also [Lo]). According to this theorem, if $F(t,x)$ is a convex multifunction satisfying suitable regularity assumptions, then there exists a map $f : [0,T] \times \mathbb{R}^n \times \mathbf{B} \to \mathbb{R}^n$ displaying an akin regularity and verifying $F(t,x) = f(t,x,\mathbf{B})$ for all (t,x). Finally, by (H4) one proves that l displays the same kind of continuity (in x) as H.

The outline of the paper is as follows. In section 2 we state the main result (Theorem 2.1) and a version of it involving global regularity. In section 3 we establish that the multivalued map $(t,x) \mapsto \mathrm{dom}(H^*(t,x,\cdot))$ is (continuous and) locally Lipschitz continuous in x. Subsequently, a global version of this result is proven as well. In section 4 we conclude the proof of the main result by exploiting the parameterization theorem for multifunctions mentioned above. Section 5 is devoted to applications to regularity questions and to a control theoretical interpretation of the front propagation phenomenon. Finally, in section 6, we extend the results of the previous sections to the case when H is just measurable in the variable t.

2. The main result. In the next theorem we shall also consider the following hypothesis on the Hamiltonian H.

(H5) *For every $R > 0$ there exists $K_R > 0$ such that for every $(t,x) \in [0,T] \times \mathbf{B}_R$ and every $v \in \mathrm{dom}(H^*(t,x,\cdot))$, one has*

$$\mathrm{argmax}_p\{p \cdot v - H(t,x,p)\} \cap B_{K_R} \neq \emptyset.$$

Here $\mathrm{argmax}_p\{p \cdot v - H(t,x,p)\}$ denotes the set of values of p where the map $p \mapsto p \cdot v - H(t,x,p)$ attains its maximum.

THEOREM 2.1. *Let us consider a Hamiltonian H verifying hypotheses* (H1)–(H4). *Then there exist a dynamics $f = f(t,x,a)$ satisfying* (A1)–(A3) *and a continuous Lagrangian $l = l(t,x,a)$, with the control set A coinciding with the unit ball $\mathbf{B}$, such that*

$$H(t,x,p) = \sup_{a \in \mathbf{B}} \{p \cdot f(t,x,a) - l(t,x,a)\} \qquad \forall (t,x,p) \in [0,T] \times \mathbb{R}^n \times \mathbb{R}^n. \tag{2.1}$$

Furthermore, if hypothesis (H5) *is in force as well, then* (A2) *turns out to be satisfied with $\nu[R,s] \doteq \omega_1[R,(1+K_R)s] + D_R s$ for suitable coefficients D_R.*

Remark 2.1. As we have already pointed out in the introduction, the main point of Theorem 2.1 consists of the fact that, on one hand, f *turns out to be locally Lipschitz continuous* (in x), even in the case when H is *not* locally Lipschitz continuous (in x), and, on the other hand, l turns out to *inherit* the regularity (in x) of H. Finally, the control set A turns out to be quite simple, namely, it coincides with the unit ball of $\mathbb{R}^n$.

The following theorem is a global version of the previous one.

THEOREM 2.2. *Let H verify* (H1)–(H5)*, where we assume that ω_1 is a modulus (i.e., it is independent of R) and there exists a constant K such that $K_R = K$ for all $R \geq 0$. Then there exist a dynamics f and a Lagrangian l such that*

$$H(t,x,p) = \sup_{a \in \mathbf{B}} \{p \cdot f(t,x,a) - l(t,x,a)\} \qquad \forall (t,x,p) \in [0,T] \times \mathbb{R}^n \times \mathbb{R}^n \tag{2.2}$$

holds true, and (A1)–(A3) *are satisfied, with the control set A coinciding with the unit ball $\mathbf{B}$, E_R independent of R, and, for all $R \geq 0$, $\nu[R,s] = \nu(s) = \omega_1[(1+K)s + Ds]$ for a suitable $D \geq 0$.*

Remark 2.2 (*on hypotheses* (H1)–(H4)). Assumptions (H1)–(H4) are *necessary* if we look for a representation (A,f,l) verifying (A1)–(A3). Moreover, they guarantee the (existence and) uniqueness of a viscosity solution to the Cauchy problem (1.1)–(1.2) (see, e.g., [CL]).

Let us observe that by assuming (H3) we are confining our investigation to Hamiltonians which are (convex and) *Lipschitz continuous in the adjoint variable* (not necessarily uniformly with respect to x). Moreover, (H4) prescribes the boundedness of the conjugate map H^*, locally with respect to (t,x). This is motivated by the fact that, on one hand, we are looking for control-theoretical representations (f,l,A) of H such that both the sets $f(t,x,A)$ and $l(t,x,A)$ are bounded, not necessarily uniformly with respect to t and x. And, on the other hand, the sets $f(t,x,A)$ and $l(t,x,A)$ finally will coincide with $\mathrm{dom}(H^*(t,x,\cdot))$ and $(H^*(t,x,\mathbb{R}^n))\backslash\{\infty\}$, respectively. As a matter of fact, we regard this paper as a first step of a wider program which shall allow for more general conditions on H. These should include superlinearity in the adjoint variable, which in turn would force one to look for representations with noncompact (possibly unbounded) control sets.

Finally, let us notice that assumption (H2) is quite standard for comparison (and hence uniqueness) results for a viscosity solution of (1.1); see, e.g., [BCD] and [Ba]. (However, let us remark that there are boundary value problems for which (H2) is no longer sufficient to guarantee uniqueness of the solution. In this case, a faithful representation of the Hamiltonian could be still exploited in order to provide a representation of *all* solutions of the boundary value problem, as, e.g., in [So], where the Hamiltonian is a control-theoretical one.)

Remark 2.3 (*on hypothesis* (H5)). Let us point out that, unlike hypotheses (H1)–(H4), hypothesis (H5) is not necessary for the existence of a representation (A,f,l) verifying (A1)–(A3), i.e., for the theses of Theorems 2.1 and 2.2 to hold true. For instance, let us consider the Hamiltonian

$$\tilde{H}(x,p) = (1+p^2)^{\frac{1}{2}} - 1 - \psi(x),$$

where ψ is just a continuous function. It is straightforward to verify that this Hamiltonian satisfies hypotheses (H1)–(H4), but *it does not satisfy hypothesis* (H5). On the other hand, it is easy to check that *the triple*

$$(\tilde{A}, \tilde{f}, \tilde{l}) = ([-1,1], a, 1 - (1-a^2)^{\frac{1}{2}} + \psi(x))$$

is a representation of H verifying (A1)–(A3), that is,

$$\tilde{H}(x,p) = \sup_{a\in[-1,1]} \{p\cdot a - \tilde{l}(x,a)\}.$$

At present we are unable to foresee how (H5) could be weakened, so we leave this question as an open problem.

Remark 2.4. Let us just mention that the representation question can also be addressed by considering only calculus of variations problems (see, e.g., [L] and also [G]) at the cost of allowing the *extended* Lagrangian H^*. Of course there is an intimate relation between the two approaches: roughly speaking, in the control-theoretical approach one is looking for a dynamics-Lagrangian pair so that, in particular, the *forbidden* velocities, that is, those mapped to $+\infty$ by H^*, are not contained in the dynamics.

3. The map $(t,x)\mapsto \mathbf{dom}(H^*(t,x,\cdot))$. In this section we prove that the (convex) multifunction $F(t,x)\doteq \mathrm{dom}(H^*(t,x,\cdot))$ is Lipschitz continuous in x. As a matter of fact, the proofs of Theorems 2.1 and 2.2, given in the next section, will be based essentially on the Lipschitz continuity of F and on the application of a parameterization theorem for convex-valued multifunctions; see Theorem 4.1 below.

Let us consider the set-valued map

$$(t,x)\mapsto F(t,x)\doteq \mathrm{dom}(H^*(t,x,\cdot)),$$

which is defined on $[0,T]\times\mathbb{R}^n$.

LEMMA 3.1. *The set-valued map F has nonempty, convex, compact values.*

Proof. Since for every $(t,x)\in[0,T]\times\mathbb{R}^n$ the map $v\mapsto H^*(t,x,v)$ is convex, proper (i.e., not everywhere equal to $+\infty$), lower semicontinuous, and bounded on its domain, $F(t,x)$ is a nonempty, convex, closed subset of $\mathbb{R}^n$(see, e.g., [RW]). Moreover, hypothesis (H3) implies that $F(t,x)\subset B_{C(1+|x|)}$ for every $(t,x)\in[0,T]\times\mathbb{R}^n$. Hence, for every $(t,x)\in[0,T]\times\mathbb{R}^n$, $F(t,x)$ is a compact convex subset of $\mathbb{R}^n$.

Throughout this paper the Hausdorff distance between two nonempty, compact subsets $A,B\subset\mathbb{R}^n$ will be denoted by $\delta(A,B)$; that is, we set

$$\delta(A,B)\doteq \max\left\{\max_{a\in A} d(a,B),\ \max_{b\in B} d(b,A)\right\}.$$

Let us recall that δ is a metric on the class $\mathcal{K}$ of nonempty, compact subsets of $\mathbb{R}^n$. In what follows, a multivalued map F with compact values from $[0,T]\times\mathbb{R}^n$ into $\mathbb{R}^n$ is said to be Lipschitz continuous (resp., continuous) if it is Lipschitz continuous (resp., continuous) when considered as a (univalued) map from $\mathbb{R}^n$ into the set $\mathcal{K}$ endowed with the metric δ. Actually, for maps with compact values, these definitions are equivalent to the usual ones (see [AC]).

F is said to be locally Lipschitz continuous if it is Lipschitz continuous on compact subsets of $[0,T]\times\mathbb{R}^n$.

THEOREM 3.2. *Let us assume hypotheses* (H1)–(H4). *Then, the set-valued map $x\mapsto F(t,x)$ is locally Lipschitz continuous in x, uniformly in t. That is, for every $R>0$, there exists a number $M_R\geq 0$ such that*

$$\delta(F(t,x),F(t,y))\leq M_R|x-y| \tag{3.1}$$

for all $x,y\in\mathbf{B}_R$ and $t\in[0,T]$.

Moreover, for every $R > 0$, there exists a number $\tilde{M}_R \geq 0$ such that

$$\delta(F(t,x), F(s,x)) \leq \tilde{M}_R \omega_2[R, |t-s|] \tag{3.2}$$

for every $x \in \mathbf{B}_R$ and $t, s \in [0,T]$. In particular, for each $x \in \mathbb{R}^n$, $t \mapsto F(t,x)$ is continuous.

In order to prove Theorem 2.2 we also need the following version of the previous result, which involves the global Lipschitz continuity of the map $x \mapsto F(t,x)$.

THEOREM 3.3. *Let us assume that hypotheses* (H1)–(H5) *are verified with both the local modulus ω_1 and the parameter K_R being in fact independent of R (that is, ω_1 is a modulus, and there exists a constant K such that $K_R = K$ for all $R \geq 0$.) Then, the set-valued map $x \mapsto F(t,x)$ is Lipschitz continuous in x, uniformly in t, that is,* (3.1) *holds true, and there exists a constant M such that $M_R = M$ for all R.*

Finally, let us state a simple property of the map H^* that will be used to prove both the regularity of the Lagrangian l and the global issue stated in Theorem 3.3.

PROPOSITION 3.4. *Assume hypotheses* (H1)–(H5)*, and fix $R > 0$. Then, for all $t \in [0,T]$, $x, y \in \mathbf{B}_R$, and $v \in F(t,x)$, $w \in F(t,y)$, one has*

$$|H^*(t,x,v) - H^*(t,y,w)| \leq \omega_1[R, (1+K_R)|x-y|] + K_R|v-w|.$$

Moreover, for all $t, s \in [0,T]$, $x \in \mathbf{B}_R$, and $v \in F(t,x)$, $w \in F(s,x)$, one has

$$|H^*(t,x,v) - H^*(s,x,w)| \leq \omega_2[R, K_R|t-s|] + \omega_3[R, |t-s|] + K_R|v-w|.$$

Proof of Theorem 3.2. Let us prove (3.1). Assume by contradiction that there exist sequences (x_n), (y_n) in $\mathbf{B}_R$ such that $x_n \neq y_n$ for every n and

$$\lim_{n\to\infty} \frac{\delta(F(t,x_n), F(t,y_n))}{|x_n - y_n|} = +\infty. \tag{3.3}$$

Up to the identification of the sequence (x_n, y_n) with a suitable subsequence, condition (3.3) yields either the existence of a selection $v_n \in F(t,x_n) \backslash F(t,y_n)$ verifying

$$\lim_{n\to\infty} \frac{d(v_n, F(t,y_n))}{|x_n - y_n|} = +\infty \tag{3.4}$$

or the existence of a selection $v_n' \in F(t,y_n) \backslash F(t,x_n)$ verifying

$$\lim_{n\to\infty} \frac{d(v_n', F(t,x_n))}{|x_n - y_n|} = +\infty. \tag{3.5}$$

Suppose that (3.4) is actually verified ((3.5) implying perfectly symmetric considerations). Then, for any selection $w_n \in F(t,y_n)$, one has

$$\lim_{n\to\infty} \frac{|v_n - w_n|}{|x_n - y_n|} = +\infty. \tag{3.6}$$

Setting

$$p_n \doteq \frac{v_n - w_n}{|v_n - w_n||x_n - y_n|},$$

one obtains

$$\begin{aligned}\omega_1[R,|x_n-y_n|(1+|p_n|)] &= \omega_1[R,|x_n-y_n|+1]\\ &\geq H(t,x_n,p_n)-H(t,y_n,p_n)\\ &\geq p_n\cdot v_n - H^*(t,x_n,v_n)\\ &\quad - \max_{w\in\mathrm{dom}(H^*(t,y_n,\cdot))}\{p_n\cdot w - H^*(t,y_n,w)\}.\end{aligned} \tag{3.7}$$

In order to achieve a contradiction, let us choose w_n to be a fixed point of the map

$$\eta_n(w)\doteq \mathrm{argmax}\left\{\frac{(v_n-w)\cdot\xi}{|x_n-y_n||v_n-w|}-H^*(t,y_n,\xi),\quad \xi\in\mathrm{dom}(H^*(t,y_n,\cdot))\right\}.$$

In view of Lemma 3.5 (where one sets $\varphi = H^*(t,y_n,\cdot)$ and $r=|x_n-y_n|$), such a point does exist. Hence one has

$$\frac{(v_n-w_n)\cdot w_n}{|x_n-y_n||v_n-w_n|}-H^*(t,y_n,w_n)\geq \frac{(v_n-w_n)\cdot\xi}{|x_n-y_n||v_n-w_n|}-H^*(t,y_n,\xi) \tag{3.8}$$

for all $\xi\in\mathrm{dom}(H^*(t,y_n,\cdot))$. By (3.7)–(3.8) and hypothesis (H4) one obtains

$$\begin{aligned}&\omega_1[R,2R+1]\geq\omega_1[R,|x_n-y_n|+1]\\ &\geq p_n\cdot(v_n-w_n)-H^*(t,x_n,v_n)+H^*(t,y_n,w_n)\geq p_n\cdot(v_n-w_n)-2N_R,\end{aligned} \tag{3.9}$$

which is a contradiction, for the right-hand side tends to $+\infty$ while the left-hand side is bounded.

In order to prove (3.2) one has to exploit the same arguments with suitable adjustments: more precisely, one has to replace the sequences x_n, y_n with sequences $t_n, s_n\in[0,T]$, and the v_n and w_n must belong to $F(t_n,x)$ and $F(s_n,x)$, respectively. Moreover, the quantities $|x_n-y_n|$ have to be replaced with $\omega_2[R,|t_n-s_n|]$. In particular, one has

$$\lim_{n\to\infty}\frac{v_n-w_n}{\omega_2[R,|t_n-s_n|]}=+\infty$$

instead of (3.6). Setting

$$p_n\doteq\frac{v_n-w_n}{|v_n-w_n|\omega_2[R,|t_n-s_n|]}$$

one can conclude by arguing as in the first part.

Proof of Theorem 3.3. If ω_1 and K do not depend on R, in order to prove global Lipschitz continuity let us argue as in the previous proof until estimate (3.9), except for the fact that now (x_n) and (y_n) lie in $\mathbb{R}^n$. In particular, the last inequality of (3.9) is no longer valid. Yet, it is not restrictive to assume that $|x_n-y_n|\leq\frac{1}{2K}$. Hence, in view of Proposition 3.4—where we take ω_1 and K independent of R—one has

$$|H^*(t,y_n,w_n)-H^*(t,x_n,v_n)|\leq\omega_1\left[\frac{1+K}{2K}\right]+K|v_n-w_n|.$$

Hence

$$\omega_1\left[\frac{1}{2K}+1\right]\geq\frac{|v_n-w_n|}{|x_n-y_n|}-\omega_1\left[\frac{1+K}{2K}\right]-K|v_n-w_n|. \tag{3.10}$$

Now, if $|v_n - w_n|$ is bounded, we contradict (3.6). If, on the contrary, (a subsequence of) $|v_n - w_n|$ tends to infinity, by the previous estimate we have

$$\omega_1\left[\frac{1}{2K}+1\right] \geq -\omega_1\left[\frac{1+K}{2K}\right] + K|v_n - w_n|, \tag{3.11}$$

which is a contradiction, in that the left-hand side is bounded while the right-hand side diverges. □

Proof of Proposition 3.4. Let p_v be an element of $\operatorname{argmax}_p\{p \cdot v - H(t,x,p)\}$. In view of hypothesis (H5) we can choose p_v such that the inequality $|p_v| \leq K_R$ is verified. Hence,

$$\begin{aligned} H^*(t,x,v) - H^*(t,y,w) &\leq p_v v - H(t,x,p_v) - p_v w + H(t,y,p_v) \\ &\leq K_R|v-w| + \omega_1[R,(1+K_R)|x-y|]. \end{aligned}$$

In an analogous way one obtains the same estimate for $H^*(t,y,w) - H^*(t,x,v)$, so the first inequality is proved. The proof of the estimate in the t-variable is akin, so we omit it. □

LEMMA 3.5. *Let $r > 0$ and let $\varphi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex, lower semicontinuous, proper map such that* $\operatorname{dom}(\varphi)$ *is compact. Let $v \in \mathbb{R}^n \backslash \operatorname{dom}(\varphi)$ and let us consider the set-valued map $\eta : \operatorname{dom}(\varphi) \to \mathcal{P}(\operatorname{dom}(\varphi))$ defined by*

$$\eta(w) \doteq \operatorname{argmax}\left\{\frac{(v-w)\cdot\xi}{r|v-w|} - \varphi(\xi), \quad \xi \in \operatorname{dom}(\varphi)\right\}.$$

Then η has a fixed point, that is, there exists $\bar{w} \in \operatorname{dom}(\varphi)$ such that $\bar{w} \in \eta(\bar{w})$.

Proof. The map η has compact convex values. Moreover, since φ is continuous on its domain, η is upper semicontinuous (and is defined on a compact convex subset of $\mathbb{R}^n$). Then the lemma follows from Kakutani's fixed point theorem (see, e.g., [AC]). □

4. Proofs of Theorems 2.1 and 2.2. To prove Theorems 2.1 and 2.2 we are going to exploit the parameterization result for convex multifunctions proved in [O] (see also [Lo]). This result involves measurability in t, which will be useful in section 6 in order to address the case with t-measurable Hamiltonians. In particular *t-measurable moduli* will be utilized. We call *t-measurable modulus* every map $w : [0,T] \times [0,*\infty[\to [0,*\infty[$ such that for every $r \in [0,*\infty[$ the map $t \mapsto w[t,r]$ is measurable and for every $t \in [0,1]$ the map $r \mapsto w[t,r]$ is a modulus. Similarly, a *local t-measurable modulus* will be a map $w : [0,*\infty[\times [0,T] \times [0,*\infty[\to [0,*\infty[$, increasing in the first variable and such that for every $R \in [0,*\infty[$ the map $(t,r) \mapsto w[R,t,r]$ is a t-measurable modulus.

Let us recall that a multifunction $\mathcal{M} : [0,T] \to \mathbb{R}^n$ is called *measurable* if for every open subset $V \subset \mathbb{R}^n$ the preimage

$$\mathcal{M}^{-1}(V) \doteq \{t \in [0,T] \,:\, \mathcal{M}(t) \cap V \neq \emptyset\}$$

is a measurable subset of $[0,T]$).

Let us consider a multivalued map $F : [0,T] \times \mathbb{R}^n \to \mathbb{R}^n$ verifying the following hypotheses.

Hypotheses (H_F):

(a) for every $(t,x) \in [0,T] \times \mathbb{R}^n$, $F(t,x)$ is a nonempty, compact, convex subset of $\mathbb{R}^n$;

(b) for every $x \in \mathbb{R}^n$ the multifunction $t \mapsto F(\cdot, x)$ is measurable;

(c) there exists a t-measurable local modulus w such that for every $R > 0$ and for almost every $t \in [0, T]$ one has

$$\delta(F(t,x), F(t,y)) \leq w[R, t, |x-y|] \tag{4.1}$$

for all $x, y \in \mathbf{B}_R$.

THEOREM 4.1 (see [O], Thm. 1). *Let F verify hypotheses* (H_F)*, and let us set*

$$M(t,x) \doteq \max\{1, |v| \ : \ v \in F(t,x)\}.$$

Then there exists a function $f : [0,T] \times \mathbb{R}^n \times \mathbf{B}$ such that

(i) *$F(t,x) = f(t,x,\mathbf{B})$ for all $x \in \mathbb{R}^n$ and for a.e. $t \in [0,T]$;*

(ii) *$f(\cdot, x, u)$ is measurable for every $(x,u) \in \mathbb{R}^n \times \mathbf{B}$;*

(iii) *there exists $N \geq 0$ such that for all $R > 0$ one has*

$$|f(t,x,u) - f(t,y,v)| \leq N(w[R,t,|x-y|] + M(t,x)|u-v|)$$

for all $x, y \in \mathbf{B}_R$ and for a.e. $t \in [0,T]$.

Moreover, if F and w are continuous, then f is continuous as well.

Remark 4.1. Actually, this theorem was proved (in [O]) under a hypothesis of uniform continuity, which means that in fact the map w is a t-measurable modulus. However, it is easy to verify (by direct inspection of the original proof) that the local statement of the present version can be proved by just replacing moduli with local moduli.

Proofs of Theorems 2.1 *and* 2.2. By Theorem 3.2 the multifunction $F(t,x) = \mathrm{dom}(H^*(t,x,\cdot))$ is continuous and agrees with the hypotheses of Theorem 4.1, with $w[R,t,r] \doteq M_R \cdot r$. Hence there exists a vector field f which verifies (A2) with $A = \mathbf{B}$ and such that $F(t,x) = f(t,x,\mathbf{B})$ for all x and a.e. $t \in [0,T]$.

Setting

$$l(t,x,a) \doteq H^*(t,x,f(t,x,a)) \qquad \forall (t,x,a) \in [0,T] \times \mathbb{R}^n \times \mathbf{B},$$

we get (2.1). Moreover, if hypothesis (H5) is in force, Proposition 3.4 and Theorem 4.1 imply the last part of the thesis. Notice that, since $A = \mathbf{B}$ is compact, (A1) is verified as well. Finally, let us prove that f satisfies the linear growth condition (A3). Indeed (H3) implies

$$|H(t,x,p)| \leq C(1+|x|)|p| + |H(t,x,0)| \tag{4.2}$$

for every $(t,x,p) \in [0,T] \times \mathbb{R}^n \times \mathbb{R}^n$. For any $(t,x,a) \in [0,T] \times \mathbb{R}^n \times \mathbf{B}$ and $\lambda > 0$ let us take $p = \lambda f(t,x,a)$, thus obtaining

$$\begin{aligned} \lambda f^2(t,x,a) - l(t,x,a) &\leq |H(t,x,\lambda f(t,x,a))| \\ &\leq \lambda C(1+|x|)|f(t,x,a)| + |H(t,x,0)|. \end{aligned} \tag{4.3}$$

If $f(t,x,a) = 0$, we are done. Otherwise, by dividing both members in (4.3) by $\lambda |f(t,x,a)|$ and letting λ go to $+\infty$ one obtains

$$|f(t,x,a)| \leq C(1+|x|).$$

In view of Theorem 3.3, Theorem 2.2 can be proved in a similar way. □

5. Some applications. Let us present two simple instances on how the representation results proved in the previous sections can be exploited both to sharpen and to interpret some facts concerning (1.1).

5.1. Regularity of the solutions of (1.1)–(1.2). The first issue concerns the regularity of the solutions to (1.1)–(1.2), which, in view of the representation results in section 2 (and of the uniqueness of the solution), is nothing but the regularity of the corresponding value function.

Let us begin by briefly recalling some well-known facts concerning the value function of an optimal control problem. Besides (A1)–(A3), let us assume the following hypothesis on the final cost g:

(A4) The map g is continuous, that is, it verifies

$$|g(x) - g(y)| \le \nu_g[R, |x-y|]$$

for all $x, y \in \mathbb{R}^n$ and a suitable local modulus ν_g.

Let us consider the value function $V = V(t,x)$ defined in (1.4) and the connected Hamiltonian

$$H(t,x,p) \doteq \sup_{a\in A} \{p \cdot f(t,x,a) - l(t,x,a)\}. \tag{5.1}$$

THEOREM 5.1. *Let us assume hypotheses* (A1)–(A4). *Then the map* $u(t,x) \doteq V(T-t,x)$ *is continuous on* $[0,T]\times\mathbb{R}^n$, *and, for any* $R>0$, *there exists a coefficient* $L_R \ge 0$ *such that*

$$\begin{aligned} |u(t,x) - u(t,y)| &\le L_R\,(|x-y| + \nu[R,|x-y|] + \nu_g[L_R R, L_R|x-y|]),\\ |u(t,x) - u(s,x)| &\le L_R\,(|t-s| + \nu[R,L_R|s-t|] + \nu_g[L_R R, L_R|t-s|]) \end{aligned}$$

for all $(t,x),(t,y),(s,x) \in [0,T]\times \mathbf{B}_R$. *Moreover,* u *is the unique viscosity solution of the Cauchy problem* (1.1)–(1.2).

We omit the proof of the regularity of V (and hence of u), which is standard and based essentially on Gronwall's lemma. For the uniqueness result see, e.g., [CL, Thm. VI.I]

As a corollary of Theorems 2.1 and 5.1 we obtain the following regularity result.

THEOREM 5.2. *Assume hypotheses* (H1)–(H5) *and let the initial datum* g *satisfy* (A4). *Then, for any* $R>0$ *there exists a coefficient* $C_R \ge 0$ *such that the solution* $u(t,x)$ *of* (1.1)–(1.2) *verifies*

$$\begin{aligned} |u(t,x) - u(t,y)| &\le C_R\big(|x-y| + \omega_1[R,|x-y|] + \nu_g[C_R R, C_R|x-y|]\big),\\ |u(t,x) - u(s,x)| &\le C_R\,(|t-s| + \omega_1[R,C_R|s-t|] + \nu_g[C_R R, C_R|t-s|]) \end{aligned}$$

for all $(t,x),(t,y) \in [0,T]\times\mathbf{B}_R$.

Example. Roughly speaking, this theorem shows that the solution *preserves* the (x-)continuity of both the Hamiltonian H and the datum g. For instance, if $\omega_1(\eta) \doteq |\eta|^\alpha$, $\alpha \le 1$, and g is β-Holder continuous, then the solution u turns out to be γ-Holder continuous, with $\gamma = \min\{\alpha,\beta\}$. As an example, consider the Cauchy problem in $[0,T]\times\mathbb{R}$:

$$u_t + \tilde H(x,u_x) = 0, \qquad u(0,x) = 0, \tag{5.2}$$

where

$$\tilde H(x,p) = |x\cdot p| - |x|^{\frac12}.$$

It is straightforward to check that the map

$$v(t,x) = 2|x|^{\frac{1}{2}}(1 - e^{-\frac{t}{2}}) \tag{5.3}$$

is a viscosity solution of (5.2), and well-known uniqueness results imply that no other solutions do exist. Since

$$|H(x,p) - H(y,p)| \leq |x-y|(1+|p|) + [|x-y|(1+|p|)]^{\frac{1}{2}}$$

and $g = 0$, Theorem 5.2 establishes that for any $R > 0$ and for all $(t,x),(t,y) \in [0,T] \times \mathbf{B}_R$ the solution of 5.2 satisfies

$$|v(t,x) - v(y,t)| \leq C_R(|x-y| + |x-y|^{\frac{1}{2}})$$

for a suitable positive number C_R. On the other hand, by the explicit expression (5.3) we know that this indeed is the case, with $C_R = 2$, for every R.

Let us note that neither the available results based on direct PDE methods nor the application of the representation provided in [Is2] would yield such sharpness in the regularity estimates. Indeed, on one hand, PDE arguments are mainly concerned with local Lipschitz continuity (see, e.g., [Ba], [CL], [Le]). On the other hand, the results in [Is2], when applied to the present example, give at most[2] $\frac{1}{4}$-Holder regularity for the solution (see Remark 1.1).

Example. An even more elementary but significative example is provided by the transport equation

$$u_t + u_x \cdot f(x) - l(x) = 0, \qquad u(0,x) = g(x),$$

where we assume that $f(x)$ and $l(x)$ verify (A2)–(A3) and g is continuous. Denoting the solution at time s of the Cauchy problem

$$\dot{y} = f(y), \qquad y(0) = x,$$

by $y(x,s)$ one can straightforwardly check that

$$u(t,x) = g(y(x,-t)) + \int_0^t l(y(x,-s))ds$$

is the unique viscosity solution of this problem. Moreover, in view of Remark 2.2, the involved Hamiltonian verifies hypotheses (H1)–(H5), with $\omega_1[R,s] = E_R s + \nu[R,s]$. So, comparing the actual regularity of u with the one which can be deduced by Theorem 5.2, we see that the latter is as sharp as possible.

Remark 5.1. As observed in the introduction, since Theorems 2.1 and 2.2 concern just the Hamiltonian H, results for different boundary value problems could be obtained as well. Similarly, the case where the datum g is no longer continuous, possibly equal to $+\infty$—which includes optimal control problems with endpoint constraints—also could be treated (by exploiting the notion of semicontinuous solution; see, e.g., [BJ91] and [Fr]).

[2]We use the expression "at most" because the fact remains that in general no uniqueness of trajectories—for a given control—would be guaranteed.

5.2. Front propagation. A second issue where a representation result can be applied concerns the phenomenon of front propagation. Let us begin with a definition. Let $\mathcal{G}$ be a class of real continuous functions on $\mathbb{R}^n$.

DEFINITION 5.3. *We say that the pair* $(H, \mathcal{G})$ *verifies the* front propagation property *if*

(i) *for every* g *belonging to* $\mathcal{G}$ *the Cauchy problem*

$$\begin{aligned} u_t + H(t, x, u_x) &= 0 \qquad \text{in }]0, T[\times \mathbb{R}^n, \\ u(0, x) &= g(x) \qquad \forall x \in \mathbb{R}^n \end{aligned}$$

has a unique (viscosity) solution, say u_g*;*

(ii) *if* $k \in \mathbb{R}$ *and* $g, \tilde{g} \in \mathcal{G}$ *are such that*

$$\begin{aligned} \Lambda_g^k(0) &\doteq \{x \in \mathbb{R}^n \,:\, g(x) < k\} = \{x \in \mathbb{R}^n \,:\, \tilde{g}(x) < k\} \doteq \Lambda_{\tilde{g}}^k(0), \\ \Gamma_g^k(0) &\doteq \{x \in \mathbb{R}^n \,:\, g(x) = k\} = \{x \in \mathbb{R}^n \,:\, \tilde{g}(x) = k\} \doteq \Gamma_{\tilde{g}}^k(0), \end{aligned}$$

then

$$\begin{aligned} \Lambda_g^k(t) &\doteq \{x \in \mathbb{R}^n \,:\, u_g(t, x) < k\} = \{x \in \mathbb{R}^n \,:\, u_{\tilde{g}}(t, x) < k\} \doteq \Lambda_{\tilde{g}}^k(t), \\ \Gamma_g^k(t) &\doteq \{x \in \mathbb{R}^n \,:\, u_g(t, x) = k\} = \{x \in \mathbb{R}^n \,:\, u_{\tilde{g}}(t, x) = k\} \doteq \Gamma_{\tilde{g}}^k(t) \end{aligned}$$

for every $t \in [0, T]$.

In other words, this condition states that the propagations of the k-level and the k-sublevel sets depend only on the k-sublevel set and the k-level set of the initial data. It is straightforward to check that property (ii) holds true for all k as soon as it is valid for one particular value of k. As is well known, a crucial role is played by the following homogeneity assumption:

(H-hom) For each $\lambda \geq 0$ one has

$$H(t, x, \lambda p) = \lambda H(t, x, p) \tag{5.4}$$

for all $(t, x, p) \in [0, T] \times \mathbb{R}^n \times \mathbb{R}^n$.

In fact, if the Hamiltonian H verifies hypotheses (H1)–(H3) *and* (H-hom) and $\mathcal{G}$ is the set of uniformly continuous functions, then the pair $(H, \mathcal{G})$ has the front propagation property (see, e.g., [BSS]). Thanks to the representation results of the previous sections—which can be applied here, for (H-hom) implies (H4) and (H5)—we can now give a simple control-theoretical explanation to this phenomenon, with $\mathcal{G}$ equal to the set of (not necessarily uniformly) continuous maps.

Remark 5.2. A control-theoretical interpretation of the front propagation phenomenon is nothing new: indeed it was originally proposed in [ES]. However, though the Hamiltonian is allowed to be nonconvex, the regularity assumptions therein assumed are much stronger than those considered here. In particular, they include the *global Lipschitz continuity of* H *in* (x, p), which in a representation like (1.3) means that f has to be *bounded*; see, for instance, assumption (1.1) in [So] in the context of front propagation along normal directions.

In section 6 we shall show that the front propagation property is still valid for Hamiltonians measurable in the variable t.

THEOREM 5.4. *Let us assume* (H1)–(H3)*, and let* $\mathcal{G} \doteq C(\mathbb{R}^n)$*. Then the following are equivalent:*

(i) H *verifies* (H-hom)*;*

(ii) *there exists a representation* (A, f, l) *of* H *satisfying* (A1)–(A3)*, with* l *equal to zero;*

(iii) *for every* $(t, x) \in [0, T] \times \mathbb{R}^n$*, the conjugate map* $v \mapsto H^*(t, x, v)$ *is constant equal to zero on its domain.*

Moreover, they imply the following:

(iv) *for all* $R > 0$ *the local modulus* $\omega_1[R, \cdot]$ *is in fact a linear mapping;*

(v) *the pair* $(H, \mathcal{G})$ *verifies the front propagation property.*

Proof. Since H is convex, the equivalence of (i) and (iii) is straightforward. Moreover, let us observe that (iii) trivially implies hypotheses (H4) and (H5), so Theorem 2.1 applies. Hence (ii) follows from (iii), since l was defined by $l(t, x, a) = H^*(t, x, f(t, x, a))$. The fact that (ii) implies (iii) and (iv) is straightforward as well.

Let us prove that (ii) implies (v). Assume by contradiction that there exist initial data g and $\tilde{g}$, both belonging to $\mathcal{G}$, and a point $(t, x) \in]0, T] \times \mathbb{R}^n$ such that $\Lambda_g^0(0) = \Lambda_{\tilde{g}}^0(0)$, $\Gamma_g^0(0) = \Gamma_{\tilde{g}}^0(0)$, while the corresponding solutions of (5.4) verify $u_g(t, x) = 0$, $u_{\tilde{g}}(t, x) \neq 0$. Let us recall that $u_g(t, x) = V_g(T - t, x)$ and $u_{\tilde{g}}(t, x) = V_{\tilde{g}}(T - t, x)$, where the value functions V_g and $V_{\tilde{g}}$ are defined as follows:

$$V_g(T - t, x) \doteq \inf g(y(T)), \quad \dot{y}(s) = f(s, y(s), a(s)), \quad y(T - t) = x,$$
$$V_{\tilde{g}}(T - t, x) \doteq \inf \tilde{g}(y(T)), \quad \dot{y}(s) = f(s, y(s), a(s)), \quad y(T - t) = x.$$

Let $\hat{a}$ be an optimal control for the datum g, which means

$$V_g(T - t, x) = g(\hat{y}(T)),$$
$$\dot{\hat{y}}(s) = f(s, \hat{y}(s), a(s)), \qquad y(T - t) = x.$$

(This control exists, for $f(s, y, \mathbf{B}) = \mathrm{dom} H^*(t, x, \cdot)$ is convex for every (s, y). However this is not crucial, for one could as well consider an ϵ-optimal control.) Now $0 = V_g(T - t, x) = g(\hat{y}(T))$, which implies $\tilde{g}(\hat{y}(T)) = 0$. Hence it cannot happen that $V_{\tilde{g}}(T - t, x) = u_{\tilde{g}}(t, x) > 0$, for one would get $\tilde{g}(\hat{y}(T)) = 0 < V_{\tilde{g}}(T - t, x)$. In a similar way, the case when $u_{\tilde{g}}(t, x) < 0$ produces a contradiction. Finally, with the same arguments one proves that it cannot happen that $u_g(t, x) > 0$ while $u_{\tilde{g}}(t, x) < 0$. □

6. *t*-measurable Hamiltonians. The results presented in the previous sections may be extended, substantially in their full strength, to the case where the Hamiltonian H is measurable in the variable t. The aim of this section is to present the corresponding statements and to point out some needed changes in the assumptions and in the proofs.

6.1. The value function and the Bellman equation. Aiming toward representations of t-measurable Hamiltonians, we have to consider optimal control problems where the data f and l are measurable in t. Accordingly, let us replace assumptions (A1)–(A3) with the following ones:

(A1′) *There exists a constant* Q *such that*

$$|f(t, 0, a)|, |l(t, 0, a)| \leq Q$$

for almost all $t \in [0, T]$ *and* $a \in A$.

(A2′) *The maps* f *and* l *are continuous in* (x, a) *from* $[0, T] \times \mathbb{R}^n \times A$ *into* $\mathbb{R}^n$ *and* $\mathbb{R}$*, respectively, and verify conditions*

$$|f(t, x, a) - f(t, y, a)| \leq E_R |x - y|, \tag{6.1}$$
$$|l(t, x, a) - l(t, y, a)| \leq \nu[R, |x - y|] \tag{6.2}$$

for all $(t,x,a),(t,y,a)\in[0,T]\times\mathbf{B}_R\times A$, *where* ν *is a suitable local modulus.*

(A3′) *There is* $C>0$ *such that*

$$f(t,x,a)\le C(1+|x|)$$

for all $(x,a)\in[0,T]\times\mathbb{R}^n\times A$ *and almost every* $t\in[0,T]$.

PROPOSITION 6.1. *The regularity results stated in Theorem* 5.1 *are still valid under the weaker hypotheses* (A1′), (A2′), (A3′), *and* (A4).

The proof of this proposition does not present substantial new difficulties with respect to the case where the data are continuous.

A *uniqueness result* analogous to the one stated in Theorem 5.1 holds true for t-measurable Hamiltonians as well, but some care is needed. To begin with, we cannot exploit the classical notion of viscosity solution, for the Hamiltonian H in (5.1) is now merely measurable in the t-variable. A suitable notion of solution for this case was introduced by Ishii in [Is2]. Successively, Lions and Perthame [LP87] provided three equivalent versions of this notion (see also [BJ87]). Recently (see [BR]) density results have been proved for this concept of solution. For the sake of self-consistency, let us recall the notion of subsolution, in one of the versions provided in [LP87].

DEFINITION 6.2. *A continuous map* $u:[0,T]\times\mathbb{R}^n$ *is a* viscosity subsolution *of* (1.1) *at* $(t_0,x_0)\in[0,T]\times\mathbb{R}^n$ *if for every* C^1 *map* ϕ *defined in a neighborhood of* (t_0,x_0) *and* $b\in L^1(0,T)$ *such that* (t_0,x_0) *is a local maximum for*

$$u(t,x)+\int_0^t b(s)ds-\phi(x)$$

one has

$$\lim_{\delta\downarrow 0^+}\operatorname*{ess\,inf}_{|t-t_0|<\delta}\inf\{H(t,x,s,p)-b(t):|x-x_0|\le\delta,|p-\nabla\phi(x_0)|\le\delta,|s-u(t_0,x_0)|\le\delta\}\le 0.$$

The definition of *viscosity supersolution* is perfectly symmetric, and a map is a *viscosity solution* if it is both a subsolution and a supersolution.

Again, it is not difficult to prove that the map $u(t,x)\doteq V(T-t,x)$ is a viscosity solution of the Cauchy problem (1.1)–(1.2).

6.2. A representation theorem for t-measurable Hamiltonians. In order to state a representation result for t-measurable Hamiltonians we shall assume suitable hypotheses. It turns out that we have to make only the obvious change due to the lack of continuity in t. Precisely we shall consider those hypotheses, which we label (H1′)–(H5′), respectively, that are obtained from (H1)–(H5) by replacing $[0,T]$ with any full-measure subset. (Of course, condition (1.8), which would imply continuity in t, is no longer assumed.)

In the new framework, the representation Theorems 2.1 and 2.2 assume the following forms, respectively.

THEOREM 6.3. *Let us consider a Hamiltonian* H *verifying hypotheses* (H1′)–(H4′). *Then there exist a dynamics* $f=f(t,x,a)$ *satisfying* (A1′)–(A3′) *and a Lagrangian* $l=l(t,x,a)$ *(continuous in* (x,a) *for almost every* $t\in[0,T]$*), with the control set* A *coinciding with the unit ball* $\mathbf{B}$, *such that*

$$H(t,x,p)=\sup_{a\in\mathbf{B}}\{p\cdot f(t,x,a)-l(t,x,a)\}\tag{6.3}$$

for almost all $t \in [0,T]$ *and for all* $(x,p) \in \mathbb{R}^n \times \mathbb{R}^n$. *Furthermore, if hypothesis* (H5′) *is in force as well, then* l *verifies* (A2′), *with* $\nu[R,s] \doteq \omega_1[R,(1+K_R)s] + D_R s$, *for suitable coefficients* D_R.

THEOREM 6.4. *Let* H *verify* (H1′)–(H5′), *with* ω_1 *being a modulus (i.e., independent of* R*) and the numbers* K_R *being equal to a constant* K *for all* R. *Then there exist a dynamics* f *and a Lagrangian* l *such that*

$$H(t,x,p) = \sup_{a \in \mathbf{B}} \{p \cdot f(t,x,a) - l(t,x,a)\} \tag{6.4}$$

holds true for almost all $t \in [0,T]$ *and for all* $(x,p) \in \mathbb{R}^n \times \mathbb{R}^n$, *conditions* (A1′)–(A3′) *are satisfied, and, moreover, the control set* A *coincides with the unit ball* $\mathbf{B}$. *Furthermore,* E_R *turns out to be independent of* R, *and* $\nu(s) = \omega_1[(1+K)s + Ds]$ *for a suitable* $D \geq 0$.

Proofs of Theorems 6.3 *and* 6.4. In view of Theorem 4.1, once we have proved that for every x the map $t \mapsto F(t,x) = \mathrm{dom} H^*(t,x,\cdot)$ is measurable (see definition in section 5) we are done. Indeed the parts of Theorems 3.2 and 3.3 concerning the variable x remain unchanged.

To prove that the map $t \mapsto F(t,x)$ is measurable we need some sharper result from set-valued analysis. Let us fix $x \in \mathbb{R}^n$. Then (see, e.g., [RW]) by the measurability of $t \mapsto H(t,x,p)$, the measurability of $t \mapsto H^*(t,x,v)$ follows, for each $v \in \mathbb{R}^n$.

Moreover, the multivalued map

$$t \mapsto \mathrm{epi}[H^*(t,x,\cdot)] \doteq \{(u,r) \in \mathbb{R}^n \times \mathbb{R} \,:\, r \geq H^*(t,x,u)\}$$

turns out to have a Castaing representation (u_n, r_n) (see, e.g., [RW]).

Hence (u_n) is a Castaing representation of the map $t \mapsto F(t,x)$, which therefore turns out to be measurable (see, e.g., [RW]). □

6.3. Regularity of solutions for *t*-measurable Hamiltonians. By the previous considerations it turns out that Theorem 5.2 on the regularity of solutions is still valid for t-measurable Hamiltonians verifying hypotheses (H1′)–(H5′) (plus some extra condition such that the uniqueness of the solution is guaranteed). Let us point out that the latter can be achieved either according to [Is1] (e.g., by imposing hypothesis (A6) therein) or by following the approach in [BR], which relies on the approximability of H by continuous Hamiltonians.

6.4. Front propagation for *t*-measurable Hamiltonians. Thanks to the representation provided by Theorem 6.3, the front propagation phenomenon can be studied for t-measurable Hamiltonians as well, as soon as the latter verify (H1′)–(H3′). For this purpose let us consider the following weakened version of assumption (H-hom):

(H′-hom) For each $\lambda \geq 0$ one has

$$H(t,x,\lambda p) = \lambda H(t,x,p) \tag{6.5}$$

for all $(t,x,p) \in [0,T] \backslash \mathcal{N} \times \mathbb{R}^n \times \mathbb{R}^n$ where $\mathcal{N}$ has measure zero.

With an unchanged proof with respect to Theorem 5.4, one obtains the following result.

THEOREM 6.5. *Let us assume* (H1′)–(H3′), *and let* $\mathcal{G} \doteq C(\mathbb{R}^n)$. *Then the following are equivalent:*

(i) H *verifies* (H′-hom);

(ii) *there exists a representation* (f, l) *of* H *satisfying* (A1′)–(A3′)*, with* l *equal to zero in* $[0,T]\backslash\mathcal{N}\times\mathbb{R}^n\times\mathbf{B}$*, for a suitable subset* $\mathcal{N}$ *of measure zero;*

(iii) *there is a zero measure subset* $\mathcal{N}$ *such that, for every* $(t,x)\in[0,T]\backslash\mathcal{N}\times\mathbb{R}^n$*, the conjugate map* $v\mapsto H^*(t,x,v)$ *is constant equal to zero on its domain.*

Moreover, each of them implies the following two conditions:

(iv) *for all* $R>0$ *the modulus* $\omega_1[R,\cdot]$ *is in fact a linear mapping;*

(v) *the pair* $(H,\mathcal{G})$ *verifies the front propagation property.*

Acknowledgment. The author is indebted to Michel Valadier, who suggested an argument related to the measurability issue in Theorems 6.3 and 6.4.

REFERENCES

[AC] J. P. Aubin and A. Cellina, *Differential Inclusions*, Springer-Verlag, Berlin, 1984.

[Ba] G. Barles, *Solutions de Viscositè des Equations de Hamilton-Jacobi*, Springer-Verlag, Berlin, 1994.

[BCD] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solution of Hamilton-Jacobi-Bellman Equations*, Birkäuser Boston, Boston, 1997.

[BJ87] E. N. Barron and R. Jensen, *Generalized viscosity solution for Hamilton-Jacobi equations with time-measurable Hamiltonians*, J. Differential Equations, 68 (1987), pp. 10–21.

[BJ91] E. N. Barron and R. Jensen, *Semicontinuous viscosity solutions of Hamilton-Jacobi with convex Hamiltonians*, Comm. Partial Differential Equations, 15 (1990), pp. 1713–1742.

[BSS] G. Barles, H. M. Soner, and P. E. Souganidis, *Front propagation and phase field theory*, SIAM J. Control Optim., 31 (1993), pp. 439–469.

[BR] A. Briani and F. Rampazzo, *A density approach to Hamilton-Jacobi equations with t-measurable Hamiltonians*, NoDEA Nonlinear Differential Equations Appl., 12 (2005), pp. 71–92.

[CL] M. G. Crandall and P. L. Lions, *Remarks on the existence and uniqueness of unbounded viscosity solutions of Hamilton-Jacobi equations*, Illinois J. Math., 31 (1987), pp. 665–688.

[Fr] H. Frankowska, *Lower semicontinuous solutions of Hamilton–Jacobi–Bellman equations*, SIAM J. Control Optim., 31 (1993), pp. 257–272.

[ES] L. C. Evans and P. E. Souganidis, *Differential games and representation formulas for solutions of Hamilton-Jacobi equations*, Indiana Univ. Math. J., 33 (1984), pp. 773–797.

[G] G. N. Galbraith, *Extended Hamilton–Jacobi characterization of value functions in optimal control*, SIAM J. Control Optim., 39 (2000), pp. 281–305.

[Is1] H. Ishii, *Hamilton-Jacobi equations with discontinuous Hamiltonians on arbitrary open sets*, Bull. Fac. Sci. Engrg. Chuo Univ., 28 (1985), pp. 33–77.

[Is2] H. Ishii, *On representations of solutions of Hamilton-Jacobi equations with convex Hamiltonians*, in Recent Topics in Nonlinear PDE, II, North–Holland Math. Stud. 128, K. Masuda and M. Mimura, eds., North–Holland, Amsterdam, 1985, pp. 15–52.

[Is3] H. Ishii, *Representation of solutions of Hamilton-Jacobi equations*, Nonlinear Anal., 12 (1988), pp. 121–146.

[Le] O. Ley, *Lower bound gradient estimates for first-order Hamilton-Jacobi equations and applications to the regularity of propagating fronts*, Adv. Differential Equations, 6 (2001), pp. 547–576.

[L] P. L. Lions, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman, Boston, 1982.

[LP87] P. L. Lions and B. Perthame, *Remarks on Hamilton-Jacobi equations with measurable time-dependent Hamiltonians*, Nonlinear Anal., 11 (1987), pp. 613–621.

[Lo] S. Lojasiewicz, jr., *Parameterization of convex sets*, in Progress in Approximation Theory, Academic Press, Boston MA, 1991, pp. 629–648.

[O] A. Ornelas, *Parameterization of Carathéodory multifunctions*, Rend. Sem. Mat. Univ. Padova, 83 (1990), pp. 33–44.

[RW] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Grundlehren Math. Wiss. 317, Springer-Verlag, Berlin, 1988.

[So] P. Soravia, *Generalized motion of a front propagating along its normal direction: A differential game approach*, Nonlinear Anal., 22 (1994), pp. 1247–1262.

[Su] A. I. Subbotin, *Minimax solutions of first-order partial differential equations*, Russian Math. Surveys, 51 (1996), pp. 283–313.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 3, pp. 885–908

LOW-ORDER CONTROLLABILITY AND KINEMATIC REDUCTIONS FOR AFFINE CONNECTION CONTROL SYSTEMS*

FRANCESCO BULLO† AND ANDREW D. LEWIS‡

Abstract. Controllability and kinematic modeling notions are investigated for a class of mechanical control systems. First, low-order controllability results are given for the class of mechanical control systems. Second, a precise connection is made between those mechanical systems which are dynamic (i.e., have forces as inputs) and those which are kinematic (i.e., have velocities as inputs). Interestingly and surprisingly, these two subjects are characterized and linked by a certain intrinsic vector-valued quadratic form that can be associated to an affine connection control system.

Key words. affine connection control systems, controllability, mechanics, driftless systems

AMS subject classifications. 70Q05, 93B03, 93B05, 93B29

DOI. 10.1137/S0363012903421182

1. Introduction.

The determination of useful necessary and sufficient conditions for local controllability of nonlinear systems remains an open problem, although significant progress has been made [2, 4, 19, 20, 34, 36]. In this paper, we investigate local controllability for a class of nonlinear systems with a rich geometric structure, namely, affine connection control systems. For these systems, we provide first-order (in the sense that the conditions involve first derivatives of the system data) local controllability conditions. The results use a certain intrinsic vector-valued quadratic form. The use of vector-valued quadratic forms in control theory has been noticed in the context of optimal control (which has, of course, a relationship with controllability) by Agrachev [3], and they have been utilized explicitly for providing conditions for local controllability by Basto-Gonçalves [6] and Hirschorn and Lewis [21]. Other uses of vector-valued quadratic forms in control are outlined in [10]. The controllability conditions we provide in section 4 bear a strong resemblance to the more general conditions of Hirschorn and Lewis [21], but we are able to provide more detail in this case because of the additional structure of the class of systems under consideration.

Affine connection control systems are a slight generalization of a class of mechanical control systems, namely, those which are Lagrangian with kinetic energy Lagrangian, and possibly with nonholonomic constraints. An initial systematic investigation of the local controllability properties of this class of systems was undertaken by Lewis and Murray [27].

The conditions for local accessibility in this work are characterized geometrically by the same authors [28] by utilizing the characterization of the so-called symmetric product provided by Lewis [24]. However, the sufficient conditions for local controllability provided by Lewis and Murray, following Sussmann [36], are not entirely

*Received by the editors January 10, 2003; accepted for publication (in revised form) January 12, 2005; published electronically September 15, 2005. This work was supported by the National Science Foundation under award CMS-0301423 and by the Natural Sciences and Engineering Research Council of Canada.

http://www.siam.org/journals/sicon/44-3/42118.html

†Department of Mechanical and Environmental Engineering, University of California at Santa Barbara, Santa Barbara, CA 93106-5070 (bullo@engineering.ucsb.edu).

‡Department of Mathematics and Statistics, Queen's University, Kingston, ON K7L 3N6, Canada (andrew@mast.queensu.ca).

satisfactory. One of the reasons for this is that these conditions are not feedback-invariant. The consequences of the lack of feedback invariance can be seen even in very simple examples, where a system can fail the sufficient condition test but still be controllable. This points out the need to better understand local controllability, and one way to do this is to obtain conditions which are not dependent on a choice of basis for the input distribution. It is this that we do in this paper, at least for systems whose controllability can be determined by brackets of low order.

A second objective of this paper is to characterize affine connection control systems in terms of equivalent lower-dimensional kinematic (or driftless) systems. The interest in low-complexity representations of affine connection control systems can be related to numerous previous efforts, including work on hybrid models for motion control systems [9], motion description languages [30], consistent control abstractions [32], hierarchical steering algorithms [31], and maneuver automata [18]. The key advantage of a low-complexity or reduced-order representation is the subsequent simplification of various control problems, including planning, stabilization, and optimal control.

In section 5, we introduce and characterize the notion of kinematic reductions as a reduced-order modeling technique adapted to affine connection control systems. This novel concept extends and unifies previous results by Lewis [25] and Bullo and Lynch [13]; see also the motivating work [5, 29, 15]. A kinematic model for an affine connection control system is one such that every controlled trajectory for the kinematic model can be realized as a trajectory, with a possible reparameterization, of the full affine connection control system with some appropriate control. We also introduce and characterize the notion of maximally reducible affine connection control systems. For such systems, every trajectory of the affine connection control system, starting from initial velocities in the input distribution, can be implemented as a controlled trajectory of a maximal kinematic reduction. Some open problems concerning inverse kinematics and sufficient conditions for controllability are presented by Cortés, Martínez, and Bullo [16].

As a third contribution of this paper, the existence of, and the controllability properties of, kinematic reductions are related to the low-order controllability properties of the corresponding affine connection control system. Interestingly, all these concepts are characterized in terms of the vector-valued quadratic form mentioned above. Insightful relationships are established and presented in Figure 5.4. We illustrate our results with some example systems. For instance, it appears that numerous (but not all) interesting mechanical devices satisfying the low-order sufficient controllability condition are also kinematically controllable. This is surprising because the concept of kinematic controllability is not a priori related to the conditions for low-order controllability. We refer to [12] for a catalog of examples.

One of the byproducts of the intrinsic formulation of the controllability and kinematic reduction results we give is that they provide a fairly complete characterization of what can be done. The incompleteness of the characterizations we give results from a possible degeneracy of the vector-valued quadratic forms. Here, one will generally have to go to higher-order conditions for controllability. Sometimes it is possible to give results using quadratic forms, even in degenerate cases, and this is being explored in a paper by Tyner and Lewis [39], currently in preparation.

Let us briefly describe the layout of the paper. We begin in section 2 with a general discussion of affine connection control systems, giving clear statements of the results of Lewis and Murray [27]. Background on vector-valued quadratic forms is presented in section 3, along with the construction of a vector-valued quadratic form that can

be associated with an affine connection control system. Our controllability results are motivated, stated, and proved in section 4. Similarly, our kinematic reductions are discussed in section 5. In this section are also presented a couple of physical examples, and a discussion of the relationships between low-order controllability and kinematic reductions.

2. Affine connection control systems. The basic differential geometric notation we use is that of [1]. When it is convenient to do so, we shall use the summation convention where summation over repeated indices is implied. For a vector bundle $\pi\colon E \to Q$, 0_q will denote the zero vector in the fiber E_q. Objects will be assumed real analytic (which we simply call "analytic") unless otherwise stated. We denote by $\Gamma(E)$ the set of analytic sections of the vector bundle $\pi\colon E \to Q$. Thus, in particular, $\Gamma(TQ)$ is the set of analytic vector fields on a manifold Q. The set of analytic functions on a manifold Q we denote by $\mathcal{F}(Q)$. We will assume the reader is familiar with affine differential geometry to the extent that it is used in [27]. An excellent reference is [22]. Affine connection control systems represent a class of mechanical control systems. We shall not devote any space to the physics involved in this representation, but refer to [27] for a few words along these lines. These issues are addressed also in the books [8, 11].

We begin with the essential definitions for affine connection control systems and provide definitions for what Lewis and Murray call "configuration controllability." Then we give the results of those authors which provide a launching point for what we do in the present paper. We provide fairly strong statements of the results of Lewis and Murray—stronger in fact than the original statements. All that we say, however, is readily implicit in the calculations of their original work.

2.1. Basic definitions. In this paper, an *affine connection control system* is a 5-tuple $\Sigma = (Q, \nabla, \mathsf{D}, \mathscr{Y}, U)$, where

1. Q is an analytic, finite-dimensional, manifold,
2. ∇ is an analytic affine connection on Q,
3. D is a constant-rank analytic distribution on Q having the property that ∇ restricts to D (i.e., $\nabla_X Y \in \Gamma(\mathsf{D})$ for all $Y \in \Gamma(\mathsf{D})$ and for all $X \in \Gamma(TQ)$),
4. $\mathscr{Y} = \{Y_1, \dots, Y_m\}$ is a collection of analytic vector fields on Q taking values in D, and
5. $U \subset \mathbb{R}^m$.

The distribution D will not concern us much here, and we allow it in order to correctly model systems with nonholonomic constraints [26]. The essential geometry of our results is captured by thinking of $\mathsf{D} = TQ$. We will frequently be interested only in 4-tuples $(Q, \nabla, \mathsf{D}, \mathscr{Y})$ satisfying the above conditions. Let us therefore agree to call this an *affine connection precontrol system*. This notion will be useful in discussions of properties of affine connection control systems that are independent of the control set U.

Associated with an affine connection control system $\Sigma = (Q, \nabla, \mathsf{D}, \mathscr{Y}, U)$ is the set of second-order control equations

$$\nabla_{\gamma'(t)}\gamma'(t) = \sum_{a=1}^{m} u^a(t) Y_a(\gamma(t)) \tag{2.1}$$

on Q. Thus a *controlled trajectory* for Σ is taken to be a pair (γ, u), where

1. $\gamma\colon I \to Q$ and $u\colon I \to U$ are both defined on the same interval $I \subset \mathbb{R}$,
2. u is locally integrable,

3. $\gamma'(t) \in \mathsf{D}_{\gamma(t)}$ for a.e. $t \in I$, and
4. (γ, u) together satisfy (2.1).

We denote by $\operatorname{conv}(U)$ and $\operatorname{aff}(U)$ the convex hull and affine hull, respectively, of $U \subset \mathbb{R}^m$. Thus $\operatorname{conv}(U)$ is the smallest convex set in $\mathbb{R}^m$ containing U, and $\operatorname{aff}(U)$ is the smallest affine subspace (i.e., shifted subspace) containing U. The control set U is *proper* (resp., *almost proper*) if $0 \in \operatorname{int}(\operatorname{conv}(U))$ (resp., if $\operatorname{aff}(U) = \mathbb{R}^m$ and $0 \in \operatorname{conv}(U)$). (One may verify that for a control-affine system the property of the control set being almost proper is exactly that which ensures that the Lie algebra rank condition is equivalent to the reachable set having nonempty interior.) We denote by Y the input distribution, so that

$$\mathsf{Y}_q = \operatorname{span}_{\mathbb{R}}\{Y_1(q), \dots, Y_m(q)\}.$$

More generally if $\mathscr{V} \subset \Gamma(TQ)$, then we denote by V the distribution generated by the vector fields $\mathscr{V}$: $\mathsf{V}_q = \operatorname{span}_{\mathbb{R}}\{X(q) |\ X \in \mathscr{V}\}$. We also denote by $\Gamma(\mathsf{V})$ the set of analytic vector fields taking values in V. We make no a priori assumptions on the constancy of the rank of any of the distributions we encounter, including the input distribution Y.

Remark 1. Our allowing a distribution to have variable rank has consequences for the choice of generators. Let us make some comments on this. Consider a family $\mathscr{Y}$ of analytic vector fields, letting Y be the distribution generated as above. Then $\Gamma(\mathsf{Y})$ is a submodule of $\Gamma(TQ)$. If Y has constant rank, then it is true that the vector fields $\mathscr{Y}$ generate this submodule. This is essentially due to a theorem of Swan [38]. However, if the rank of Y is *not* constant (more precisely, locally constant), then it can be the case that the vector fields $\mathscr{Y}$ are *not* generators for $\Gamma(\mathsf{Y})$. However, we shall always require that our families of vector fields have the property that they are generators for the submodule of sections of the induced distribution. Locally, and in the analytic setting, this can be done without loss of generality, due to the Noetherian property of the ring of analytic functions.

Let us clearly state our controllability definitions. First we provide notation for the reachable sets. For $T > 0$ and $q_0 \in Q$, let

$$\mathcal{R}^{\Sigma}_{TQ}(q_0, T) = \{\gamma'(T) |\ (\gamma, u) \text{ is a controlled trajectory on } [0, T] \text{ with } \gamma'(0) = 0_{q_0}\}$$

and let $\mathcal{R}^{\Sigma}_{TQ}(q_0, \leq T) = \bigcup_{0 \leq t \leq T} \mathcal{R}^{\Sigma}_{TQ}(q_0, t)$. These are therefore reachable *states* in TQ starting from zero initial velocity at the configuration q_0. We also consider the reachable configurations, which we denote by

$$\mathcal{R}^{\Sigma}_{Q}(q_0, T) = \tau_Q(\mathcal{R}^{\Sigma}_{TQ}(q_0, T)), \qquad \mathcal{R}^{\Sigma}_{Q}(q_0, \leq T) = \tau_Q(\mathcal{R}^{\Sigma}_{TQ}(q_0, \leq T)),$$

where $\tau_Q \colon TQ \to Q$ is the tangent bundle projection. Note that since D is invariant under ∇ and since the input vector fields are D-valued, solutions of (2.1) with initial conditions in D remain in D. In the following definition, $\operatorname{int}_{\mathsf{D}}(\cdot)$ means the interior in the relative topology on $\mathsf{D} \subset TQ$.

DEFINITION 2.1. *Let $\Sigma = (Q, \nabla, \mathsf{D}, \mathscr{Y}, U)$ be an affine connection control system and let $q_0 \in Q$.*

(i) *$(Q, \nabla, \mathsf{D}, \mathscr{Y})$ is* accessible *from q_0 if, for every almost proper control set, there exists $T > 0$ such that $\operatorname{int}_{\mathsf{D}}(\mathcal{R}^{\Sigma}_{TQ}(q_0, \leq t)) \neq \emptyset$ for $t \in (0, T]$.*

(ii) *$(Q, \nabla, \mathsf{D}, \mathscr{Y})$ is* configuration accessible *from q_0 if, for every almost proper control set, there exists $T > 0$ such that $\operatorname{int}(\mathcal{R}^{\Sigma}_{Q}(q_0, \leq t)) \neq \emptyset$ for $t \in (0, T]$.*

(iii) *Σ is* small-time locally controllable (STLC) *from q_0 if there exists $T > 0$ such that $0_{q_0} \in \operatorname{int}_{\mathsf{D}}(\mathcal{R}^{\Sigma}_{TQ}(q_0, \leq t)) \neq \emptyset$ for $t \in (0, T]$.*

(a) $(Q, \nabla, \mathsf{D}, \mathscr{Y})$ *is* properly small-time locally controllable (properly STLC) *from* q_0 *if* Σ *is STLC from* q_0 *for every proper control set* U.

(b) $(Q, \nabla, \mathsf{D}, \mathscr{Y})$ *is* small-time locally uncontrollable (STLUC) *from* q_0 *if* Σ *is not STLC from* q_0 *for any compact control set* U.

(iv) Σ *is* small-time locally configuration controllable (STLCC) *from* q_0 *if there exists* $T > 0$ *such that* $0_{q_0} \in \operatorname{int}(\mathcal{R}^{\Sigma}_{Q}(q_0, \leq t)) \neq \emptyset$ *for* $t \in (0, T]$.

(a) $(Q, \nabla, \mathsf{D}, \mathscr{Y})$ *is* properly small-time locally configuration controllable (properly STLCC) *from* q_0 *if* Σ *is STLCC from* q_0 *for every proper control set* U.

(b) $(Q, \nabla, \mathsf{D}, \mathscr{Y})$ *is* small-time locally configuration uncontrollable (STLCUC) *from* q_0 *if* Σ *is not STLCC from* q_0 *for any compact control set* U.

Remark 2.

1. Note that we are careful in these definitions to distinguish between those notions of controllability that depend only on the geometry of the affine connection precontrol system $(Q, \nabla, \mathsf{D}, \mathscr{Y})$ and those that also depend on the character of the control set U. Hirschorn and Lewis [21] illustrate various situations where the exact nature of the control set must be accounted for in the controllability analysis. For this reason we try to be careful about the exact manner in which the control set is considered.

2. A consequence of the classical theory of accessibility [37] is that for an affine connection precontrol system $(Q, \nabla, \mathsf{D}, \mathscr{Y})$, the reachable sets for $(Q, \nabla, \mathsf{D}, \mathscr{Y}, U)$ have nonempty interior for *all* almost proper control sets if and only if the reachable sets have nonempty interior for *some* almost proper control set.

3. It is clear that STLC implies STLCC and that STLCUC implies STLUC. The converse implications are generally false. What's more, even the relationships between STLCC and STLC *on the reachable set* are not completely understood at this time.

2.2. Review of existing results. Let us briefly review the results of [27]. These results rely on the *symmetric product* defined by the affine connection ∇ by $\langle X : Y \rangle = \nabla_X Y + \nabla_X Y$. First let us provide a description of the set of points accessible from the zero vector 0_q in the tangent space T_qQ. We let $\Sigma = (Q, \nabla, \mathsf{D}, \mathscr{Y}, U)$ be an affine connection control system. As above, we denote by Y the distribution generated by the vector fields $\mathscr{Y}$, and we now define a sequence $\mathsf{Sym}^{(k)}(\mathsf{Y})$, $k \in \mathbb{N}$, of distributions by

$$\begin{aligned}
&\mathsf{Sym}^{(1)}(\mathsf{Y})_q = \mathsf{Y}_q + \operatorname{span}_{\mathbb{R}}\{\langle Y_a : Y_b \rangle |\ a, b \in \{1, \dots, m\}\},\\
&\mathsf{Sym}^{(k)}(\mathsf{Y})_q = \mathsf{Sym}^{(k-1)}(\mathsf{Y})_q\\
&\quad + \operatorname{span}_{\mathbb{R}}\{\langle Y_a : Y_b \rangle |\ Y_a \in \Gamma(\mathsf{Sym}^{(k_1)}(\mathsf{Y})),\ Y_b \in \Gamma(\mathsf{Sym}^{(k_2)}(\mathsf{Y})),\ k_1 + k_2 = k - 1\}.
\end{aligned}$$

The smallest distribution containing these distributions we denote by $\mathsf{Sym}^{(\infty)}(\mathsf{Y})$, and we note that $\langle X : Y \rangle \in \Gamma(\mathsf{Sym}^{(\infty)}(\mathsf{Y}))$ for each $X, Y \in \Gamma(\mathsf{Sym}^{(\infty)}(\mathsf{Y}))$. The integrable distribution generated by $\mathsf{Sym}^{(\infty)}(\mathsf{Y})$ we denote by $\mathsf{Lie}^{(\infty)}(\mathsf{Sym}^{(\infty)}(\mathsf{Y}))$. Since this distribution is integrable, through each point $q_0 \in Q$ there is an immersed maximal integral manifold Λ_{q_0} with the property that $T_q\Lambda_{q_0} = \mathsf{Lie}^{(\infty)}(\mathsf{Sym}^{(\infty)}(\mathsf{Y}))_q$ for each $q \in \Lambda_{q_0}$. Note that since we are only thinking of local controllability, we may shrink Q so that Λ_{q_0} is an embedded submanifold, and thus $T_q\Lambda_{q_0}$ has its usual definition.

With this notation, we have the following theorem which describes the reachable set from $0_{q_0} \in TQ$. Note that the description we provide here is a little more complete

than that originally given by Lewis and Murray, but what we state here is certainly implicit in the original paper.

THEOREM 2.2. *Let $\Sigma = (Q, \nabla, \mathsf{D}, \mathscr{Y}, U)$ be an affine connection control system with U almost proper. Let Λ_{q_0} be the maximal integral manifold of $\mathsf{Lie}^{(\infty)}(\mathsf{Sym}^{(\infty)}(\mathsf{Y}))$ through $q_0 \in Q$, which we assume without loss of generality to be an embedded submanifold of Q. Let $\mathsf{S}(\mathsf{Y}, q_0)$ be the vector bundle over Λ_{q_0} whose fiber at $q \in \Lambda_{q_0}$ is $\mathsf{Sym}^{(\infty)}(\mathsf{Y})_q$. We have the following statements.*

(i) *There exists $T > 0$ such that for each $t \in (0, T]$, $\mathcal{R}^{\Sigma}_{TQ}(q_0, \leq t)$ is contained in $\mathsf{S}(\mathsf{Y}, q_0)$ and contains a nonempty open subset of $\mathsf{S}(\mathsf{Y}, q_0)$.*

(ii) *In particular, there exists $T > 0$ such that for each $t \in (0, T]$, $\mathcal{R}^{\Sigma}_{Q}(q_0, \leq t)$ is contained in Λ_{q_0} and contains a nonempty open subset of Λ_{q_0}.*

Theorem 2.2 obviously leads to the following corollary.

COROLLARY 2.3. *An affine connection precontrol system $(Q, \nabla, \mathsf{D}, \mathscr{Y})$*

(i) *is accessible from q_0 if and only if $\mathsf{Sym}^{(\infty)}(\mathsf{Y})_{q_0} = \mathsf{D}_{q_0}$, and*

(ii) *is configuration accessible from q_0 if and only if $\mathsf{Lie}^{(\infty)}(\mathsf{Sym}^{(\infty)}(\mathsf{Y}))_{q_0} = T_{q_0}Q$.*

Now we turn to local configuration controllability. Let $P(\mathscr{Y})$ denote the set of iterated symmetric products of vector fields in $\mathscr{Y}$. A product $P_0 \in P(\mathscr{Y})$ is *bad* when it is composed of an even number of each of the vector fields from $\mathscr{Y}$ and is otherwise *good*. The *degree* of $P_0 \in P(\mathscr{Y})$ is the total number of vector fields from $\mathscr{Y}$ which participate in P_0, counting multiplicities. Thus, for example, $\langle Y_a : \langle Y_b : Y_b \rangle \rangle$ is good and of degree 3, and $\langle \langle Y_a : Y_b \rangle : \langle Y_a : Y_b \rangle \rangle$ is bad and of degree 4. Let S_m be the symmetric group on m symbols. For $P_0 \in P(\mathscr{Y})$ and $\sigma \in S_m$, let $\sigma(P_0) \in P(\mathscr{Y})$ be obtained by permuting the occurrences of the vector fields from $\mathscr{Y}$ by σ. For example, if $P_0 = \langle Y_a : \langle Y_b : Y_c \rangle \rangle$ and if $\sigma = \left(\begin{smallmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{smallmatrix}\right)$, then $\sigma(P_0) = \langle Y_b : \langle Y_c : Y_a \rangle \rangle$. With this notation, we have the following definition.

DEFINITION 2.4. An affine connection precontrol system $(Q, \nabla, \mathsf{D}, \mathscr{Y})$ *satisfies the good/bad hypothesis at q_0* if, for each bad symmetric product $P_0 \in P(\mathscr{Y})$, there exist good symmetric products $P_1, \ldots, P_k \in P(\mathscr{Y})$ of degree strictly less than P_0 and such that

$$\sum_{\sigma \in S_m} \sigma(P_0)(q_0) = \sum_{j=1}^{k} c_j P_j(q_0)$$

for some $c_1, \ldots, c_k \in \mathbb{R}$.

The following result of Lewis and Murray [27] is derived from a result of Sussmann [36]. Again, we provide a somewhat more thorough statement of the result than is given in [27].

THEOREM 2.5. *Let $\Sigma = (Q, \nabla, \mathsf{D}, \mathscr{Y}, U)$ be an affine connection control system with U proper, and let $q_0 \in Q$. If $(Q, \nabla, \mathsf{D}, \mathscr{Y})$ satisfies the good/bad hypothesis at $q_0 \in Q$, then there exists $T > 0$ such that for each $t \in (0, T]$ the set $\mathcal{R}^{\Sigma}_{TQ}(q_0, \leq t)$ contains a neighborhood of 0_{q_0} in the vector bundle $\mathsf{S}(\mathscr{Y}, q_0)$ over Λ_{q_0}.*

The result essentially says that when the good/bad hypothesis is satisfied, the system is locally controllable when restricted to its reachable set. In particular, we have the following corollary.

COROLLARY 2.6. *Let $\Sigma = (Q, \nabla, \mathsf{D}, \mathscr{Y}, U)$ be an affine connection control system with U proper and such that the precontrol system $(Q, \nabla, \mathsf{D}, \mathscr{Y})$ satisfies the good/bad hypotheses at $q_0 \in Q$. Then*

(i) *Σ is locally controllable at q_0 if it is locally accessible at q_0, and*

(ii) Σ *is locally configuration controllable at* q_0 *if it is locally configuration accessible at* q_0.

The above results all follow from a detailed analysis of the Lie algebra of vector fields associated with the control system (2.1) when it is thought of as a control-affine system with state manifold TQ. The results reflect the fact that, when evaluated at zero velocity points, this Lie algebra structure simplifies enormously. We shall exploit this further when we prove our main results in section 4. We remark that the structure of the Lie algebra at points of nonzero velocity is not currently well understood.

3. Vector-valued quadratic forms. In our controllability analysis we are led to investigate symmetric bilinear maps $B\colon V \times V \to W$ from a finite-dimensional $\mathbb{R}$-vector space V into a finite-dimensional $\mathbb{R}$-vector space W. In this section we first look at such objects in general, and then we construct a specific such object associated to an affine connection control system. Some other control theoretic problems where vector-valued quadratic forms arise are given by Bullo et al. [10].

3.1. Basic definitions and properties. Let V and W be finite-dimensional $\mathbb{R}$-vector spaces and let $\Sigma_2(V;W)$ denote the set of symmetric $\mathbb{R}$-bilinear maps from $V \times V$ to W. For $B \in \Sigma_2(V;W)$, we define $Q_B\colon V \to W$ by $Q_B(v) = B(v,v)$. For $\lambda \in W^*$, we define $\lambda B\colon V \times V \to \mathbb{R}$ by $\lambda B(v_1, v_2) = \langle \lambda; B(v_1, v_2)\rangle$.

DEFINITION 3.1. *Let* $B \in \Sigma_2(V;W)$.

(i) B *is* definite *if there exists* $\lambda \in W^*$ *such that* λB *is positive-definite.*

(ii) B *is* essentially indefinite *if, for each* $\lambda \in W^*$, λB *is either*

(a) *zero or*

(b) *neither positive nor negative-semidefinite.*

The following properties of symmetric bilinear maps will be important for us. The proof follows more or less directly from the definitions.

LEMMA 3.2. *Let* V *and* W *be finite-dimensional* $\mathbb{R}$*-vector spaces with* $B \in \Sigma_2(V;W)$. *Suppose that* $V \neq \{0\}$. *The following statements hold:*

(i) *if* $W = \{0\}$, *then* B *is essentially indefinite;*

(ii) *if* $W \neq \{0\}$, *then* B *is essentially indefinite if and only if*

$$0 \in \mathrm{int}_{\mathrm{aff}(\mathrm{image}(Q_B))}(\mathrm{conv}(\mathrm{image}(Q_B)));$$

(iii) *if* $W \neq \{0\}$, *then* B *is definite if and only if there exists a hyperplane* P *through* $0 \in W$ *such that*

(a) $\mathrm{image}(Q_B)$ *lies on one side of* P *and*

(b) $\mathrm{image}(Q_B) \cap P = \{0\}$.

The matter of deciding whether a vector-valued quadratic form is essentially indefinite is known to be NP-complete, at least in the case when $\dim(W) > 1$.[1]

The following result gives some properties of $\mathbb{R}$-valued quadratic forms that will be useful in our discussion. We refer to Hirschorn and Lewis [21] for a proof.

LEMMA 3.3. *Let* V *be a finite-dimensional* $\mathbb{R}$*-vector space and let* $B \in \Sigma_2(V;\mathbb{R})$. *For a basis* $\mathscr{V} = \{v_1, \ldots, v_n\}$ *for* V, *let* $[B]_{\mathscr{V}}$ *be the* $n \times n$ *matrix representation of* B. *The following statements are equivalent:*

(i) *there exists a basis* $\mathscr{V}$ *for* V *for which the sum of the diagonal entries in the matrix* $[B]_{\mathscr{V}}$ *is zero;*

(ii) *there exists a basis* $\mathscr{V}$ *for* V *for which the diagonal entries in the matrix* $[B]_{\mathscr{V}}$ *are all zero;*

(iii) B *is essentially indefinite.*

[1] This was pointed out to the authors by a reviewer for [10].

3.2. Vector-valued quadratic forms and affine connection control systems. Let $\Sigma = (Q, \nabla, \mathsf{D}, \mathscr{Y}, U)$ be an affine connection control system and let $q \in Q$. If $S_q \subset T_qQ$ is a subspace, then we define $B_{\mathsf{Y}_q}(S_q)\colon \mathsf{Y}_q \times \mathsf{Y}_q \to T_qQ/S_q$ as the T_qQ/S_q-valued symmetric, bilinear mapping on Y_q given by

$$B_{\mathsf{Y}_q}(S_q)(v_1, v_2) = \pi_{S_q}(\langle V_1 : V_2\rangle(q)), \tag{3.1}$$

where V_1 and V_2 are vector fields extending $v_1, v_2 \in \mathsf{Y}_q$, and where $\pi_{S_q}\colon T_qQ \to T_qQ/S_q$ is the canonical projection. Note that $B_{\mathsf{Y}_q}(S_q)$ is not necessarily well-defined.

LEMMA 3.4. *If $\mathsf{Y}_q \subset S_q$, then $B_{\mathsf{Y}_q}(S_q)$ is well-defined.*

Proof. We need to show that the definition in (3.1) does not depend on the extensions V_1 and V_2 of v_1 and v_2. This will follow if $\pi_{S_q}(\langle V_1 : V_2\rangle(q))$ depends only on the values of V_1 and V_2 at q, and not on their derivatives. Let $\phi_1, \phi_2 \in \mathcal{F}(Q)$ and compute

$$\langle \phi_1 V_1 : \phi_2 V_2\rangle = \phi_1\phi_2\langle V_1 : V_2\rangle + \phi_1(\mathscr{L}_{V_1}\phi_2)V_2 + \phi_2(\mathscr{L}_{V_2}\phi_1)V_1.$$

Thus $\pi_{S_q}(\langle \phi_1 V_1 : \phi_2 V_2\rangle(q)) = \phi_1(q)\phi_2(q)\pi_{S_q}(\langle V_1 : V_2\rangle(q))$, showing that $\pi_{S_q}(\langle V_1 : V_2\rangle(q))$ does not depend on the derivatives of V_1 and V_2 at q, and so the result follows. ∎

Remark 3. Note that $(T_qQ/S_q)^* \simeq \operatorname{ann}(S_q)$. Therefore, the definition of $\lambda B_{\mathsf{Y}_q}(S_q)$, $\lambda \in (T_qQ/S_q)^*$ is concrete in that one needs to worry about objects in the quotient.

If Y has constant rank, then one can define a TQ/Y-valued quadratic form B_{Y} globally by

$$B_{\mathsf{Y}}(V_1, V_2) = \pi_{\mathsf{Y}}(\langle V_1 : V_2\rangle)$$

for $V_1, V_2 \in \Gamma(\mathsf{Y})$, where $\pi_{\mathsf{Y}}\colon TQ \to TQ/\mathsf{Y}$ is the projection.

4. Controllability results. In this section we undertake the formulation and discussion of novel controllability results. Our objective is to obtain controllability conditions that are independent of the basis for the input distribution Y. We achieve this by means of controllability tests that do not entail good/bad conditions but rather are expressed in terms of properties of a vector-valued quadratic form. Before we state the results we need some preliminary constructions.

4.1. Constructions concerning vanishing input vector fields. We let $\Sigma = (Q, \nabla, \mathsf{D}, \mathscr{Y}, U)$ be an analytic affine connection control system and we let $q_0 \in Q$. One of the generalizations we wish to allow is the case when q_0 may not be a regular point for the distribution Y generated by $\mathscr{Y}$. In this case the vector fields $\mathscr{Y}$ cannot be linearly independent at q_0. It may also happen that even if q_0 *is* a regular point for Y, the vector fields may still not be linearly independent. For example, if one wishes to globally define a control system for which the input distribution Y has constant rank but is not trivial, then one will necessarily have to choose more input vector fields than $\operatorname{rank}(\mathsf{Y})$, implying that the input vector fields will never be linearly independent. It will be convenient to organize the vector fields in $\mathscr{Y}$ in a manner consistent with these possibilities. The following result gives a useful way of doing this.

LEMMA 4.1. *Let $(Q, \nabla, \mathsf{D}, \mathscr{Y} = \{Y_1, \ldots, Y_m\})$ be an analytic affine connection precontrol system with $q_0 \in Q$. There exists $\boldsymbol{T} \in GL(m; \mathbb{R})$ with the property that if $\tilde{Y}_a = T_a^b Y_a$, $a \in \{1, \ldots, m\}$, then*

(i) *$\{\tilde{Y}_1(q_0), \ldots, \tilde{Y}_k(q_0)\}$ form a basis for Y_{q_0} and*

(ii) *the vector fields $\tilde{Y}_{k+1}, \ldots, \tilde{Y}_m$ vanish at q_0.*

Proof. We let $k = \dim(\mathsf{Y}_{q_0})$. Since $\mathscr{Y}$ generates Y, we may find $\boldsymbol{R} \in GL(m;\mathbb{R})$ with the property that if $X_a = R_a^b Y_b$, $a \in \{1,\dots,m\}$, then $\{X_1(q_0),\dots,X_k(q_0)\}$ form a basis for Y_{q_0}. Now let $L_{q_0}\colon \mathbb{R}^m \to \mathsf{Y}_{q_0}$ be defined by $L_{q_0}(\boldsymbol{u}) = \sum_{a=1}^m u^a X_a(q_0)$. Let $\boldsymbol{u}_{k+1},\dots,\boldsymbol{u}_m \in \mathbb{R}^m$ be a basis for $\ker(L_{q_0})$ and define $\boldsymbol{S} \in GL(m;\mathbb{R})$ by

$$\boldsymbol{S} = \left[\ \boldsymbol{e}_1 \ \middle|\ \cdots \ \middle|\ \boldsymbol{e}_k \ \middle|\ \boldsymbol{u}_{k+1} \ \middle|\ \cdots \ \middle|\ \boldsymbol{u}_m \ \right].$$

It is then clear that if we take $\tilde{Y}_a = S_a^b X_b$, $a \in \{1,\dots,m\}$, then $\{\tilde{Y}_1(q_0),\dots,\tilde{Y}_k(q_0)\}$ form a basis for Y_{q_0}, and that $\tilde{Y}_{k+1},\dots,\tilde{Y}_m$ vanish at q_0. Now we take $\boldsymbol{T} = \boldsymbol{R}\boldsymbol{S}$. ∎

Remark 4.

1. If the vector fields $\mathscr{Y}$ are linearly independent at q_0, then one may take $\boldsymbol{T} = \boldsymbol{I}_m$ in the lemma.

2. Suppose that we have a control set U for $(Q,\nabla,\mathsf{D},\mathscr{Y})$. If we take $\boldsymbol{T} \in GL(m;\mathbb{R})$ and $\tilde{\mathscr{Y}} = \{\tilde{Y}_1,\dots,\tilde{Y}_m\}$ as in the lemma, and if we define $\tilde{U} = \{\boldsymbol{T}^{-1}\boldsymbol{u} \mid\ \boldsymbol{u} \in U\}$, this gives an affine connection control system $\tilde{\Sigma} = (Q,\nabla,\mathsf{D},\tilde{\mathscr{Y}},\tilde{U})$. Clearly the controlled trajectories for $\Sigma = (Q,\nabla,\mathsf{D},\mathscr{Y},U)$ and $\tilde{\Sigma}$ agree, so we can without loss of generality assume that the input vector fields for an affine connection control system satisfy conditions (i) and (ii) of the lemma. Input vector fields satisfying these conditions at q_0 will be said to be *adapted at* q_0.

Let $X, Y \in \Gamma(Q)$. If $X(q_0) = 0_{q_0}$, then the expression $\langle X : Y\rangle(q_0)$ may be verified (in coordinates, for example) to depend only on the value of Y at q_0. That is to say, we may define a linear map $\mathrm{sym}_X\colon T_{q_0}Q \to T_{q_0}Q$ by $v \mapsto \langle X : V\rangle(q_0)$, where V is any extension of $v \in T_{q_0}Q$. If $\mathscr{Y}$ is adapted at q_0, then we denote by $Z_{q_0}(\mathscr{Y})$ the set of linear maps sym_{Y_a}, $a \in \{k+1,\dots,m\}$, where $k = \dim(\mathsf{Y}_{q_0})$. For an $\mathbb{R}$-vector space W, an arbitrary subset $\mathscr{L}$ of linear transformations of W, and a subspace $S \subset W$, we denote by $\langle \mathscr{L}, S\rangle$ the smallest subspace of W containing S and which is an invariant subspace for each of the linear maps from $\mathscr{L}$. One readily verifies that $\langle \mathscr{L}, S\rangle$ is generated by vectors of the form

$$L_1 \circ \cdots \circ L_{k-1}(v), \qquad L_1,\dots,L_{k-1} \in \mathscr{L},\ v \in S,\ k \in \mathbb{N}. \tag{4.1}$$

We will be interested in subspaces of the form $\langle Z_{q_0}(\mathscr{Y}), S_{q_0}\rangle$, where S_{q_0} is a subspace of $T_{q_0}Q$. In order for such constructions to make sense (in that they are independent of the choice of adapted family of vector fields) the subspace S_{q_0} should have some properties.

LEMMA 4.2. *Let $\Sigma = (Q,\nabla,\mathsf{D},\mathscr{Y},U)$ and $\tilde{\Sigma} = (Q,\nabla,\mathsf{D},\tilde{\mathscr{Y}},\tilde{U})$ be affine connection control systems satisfying*

(i) $\mathsf{Y} = \tilde{\mathsf{Y}}$ *and*

(ii) $\mathscr{Y}$ *and* $\tilde{\mathscr{Y}}$ *are adapted at* q_0.

Then $\langle Z_{q_0}(\tilde{\mathscr{Y}}), S_{q_0}\rangle = \langle Z_{q_0}(\mathscr{Y}), S_{q_0}\rangle$ for any subspace S_{q_0} containing Y_{q_0}.

Proof. We write $\mathscr{Y} = \{Y_1,\dots,Y_m\}$ and $\tilde{\mathscr{Y}} = \{\tilde{Y}_1,\dots,\tilde{Y}_{\tilde{m}}\}$. Since $\mathsf{Y} = \tilde{\mathsf{Y}}$, we must have

$$\tilde{Y}_\alpha = \sum_{a=1}^m \Lambda_\alpha^a Y_a, \qquad \alpha \in \{1,\dots,\tilde{m}\},$$

for functions $\Lambda_\alpha^a\colon Q \to \mathbb{R}$, $a \in \{1,\dots,m\}$, $\alpha \in \{1,\dots,\tilde{m}\}$. (Here we make use of the assumption stated in Remark 1.) Assume that $\dim(\mathsf{Y}_{q_0}) = k$ so that both $\{Y_1(q_0),\dots,Y_k(q_0)\}$ and $\tilde{Y}_1(q_0),\dots,\tilde{Y}_k(q_0)\}$ are bases for Y_{q_0} and so that $Y_{k+1},\dots,Y_m$

and $\tilde{Y}_{k+1}, \dots, \tilde{Y}_{\tilde{m}}$ all vanish at q_0. Note that $\langle Z_{q_0}(\mathscr{Y}), S_{q_0}\rangle$ is generated by those tangent vectors at q_0 of the form

$$\mathrm{sym}_{Y_{a_{\ell-1}}} \circ \cdots \circ \mathrm{sym}_{Y_{a_1}}(v), \qquad a_1, \dots, a_{\ell-1} \in \{k+1, \dots, m\},\ \ell \in \mathbb{N},\ v \in S_{q_0}.$$

We will show by induction on ℓ that each of these generators lies in $\langle Z_{q_0}(\tilde{\mathscr{Y}}), S_{q_0}\rangle$. This is clearly true for $\ell = 1$, so suppose it true for $\ell = j$ and let $a_j \in \{k+1, \dots, m\}$. Then for any $V \in \Gamma(TQ)$, we have

$$\langle Y_{a_j} : V\rangle = \langle \Lambda^\alpha_{a_j}(\tilde{Y}_\alpha) : V\rangle = \Lambda^\alpha_a \langle \tilde{Y}_\alpha : V\rangle + \sum_{\alpha=1}^{\tilde{m}} (\mathscr{L}_V \Lambda^\alpha_{a_j})\tilde{Y}_\alpha,$$

from which we ascertain that

$$\mathrm{sym}_{Y_{a_j}} = \sum_{\alpha=k+1}^{\tilde{m}} \Lambda^\alpha_{a_j}(q_0)\mathrm{sym}_{\tilde{Y}_\alpha} + \sum_{\alpha=1}^{k} \tilde{Y}_\alpha(q_0) \otimes \boldsymbol{d}^\alpha_{a_j}(q_0),$$

since $\Lambda^\alpha_a(q_0) = 0$ for $\alpha \in \{1, \dots, k\}$ and $a \in \{k+1, \dots, m\}$. Therefore, by the induction hypothesis, we conclude that

$$\mathrm{sym}_{Y_{a_j}} \circ \mathrm{sym}_{Y_{a_{j-1}}} \circ \cdots \circ \mathrm{sym}_{Y_{a_1}}(v) \in \langle Z_{q_0}(\tilde{\mathscr{Y}}), S_{q_0}\rangle.$$

This shows that $\langle Z_{q_0}(\mathscr{Y}), S_{q_0}\rangle \subset \langle Z_{q_0}(\tilde{\mathscr{Y}}), S_{q_0}\rangle$. The opposite inclusion follows as above, but swapping $\mathscr{Y}$ and $\tilde{\mathscr{Y}}$. □

The preceding result shows the invariance of the definition of a subspace on the choice of adapted generators for Y. The next result gives the same conclusion for a vector-valued quadratic form.

LEMMA 4.3. *Let $\Sigma = (Q, \nabla, \mathsf{D}, \mathscr{Y}, U)$ and $\tilde{\Sigma} = (Q, \nabla, \mathsf{D}, \tilde{\mathscr{Y}}, \tilde{U})$ be affine connection control systems satisfying*

(i) $\mathsf{Y} = \tilde{\mathsf{Y}}$ *and*

(ii) $\mathscr{Y}$ *and* $\tilde{\mathscr{Y}}$ *are adapted at* q_0.

If $S_{q_0} \subset T_{q_0}Q$ is a subspace containing Y_{q_0}, then $B_{\tilde{\mathsf{Y}}_{q_0}}(S_{q_0}) = B_{\mathsf{Y}_{q_0}}(S_{q_0})$.

Proof. As in the proof of Lemma 4.2 we have

$$\tilde{Y}_\alpha = \sum_{a=1}^{m} \Lambda^a_\alpha Y_a, \qquad \alpha \in \{1, \dots, \tilde{m}\},$$

for functions $\Lambda^a_\alpha \colon Q \to \mathbb{R}$, $a \in \{1, \dots, m\}$, $\alpha \in \{1, \dots, \tilde{m}\}$. We then compute

$$\langle Y_a : Y_b\rangle = \Lambda^\alpha_a \Lambda^\beta_b \langle \tilde{Y}_\alpha : \tilde{Y}_\beta\rangle + \sum_{\alpha,\beta=1}^{\tilde{m}} \Lambda^\beta_b (\mathscr{L}_{\tilde{Y}_\beta} \Lambda^\alpha_a)\tilde{Y}_\alpha + \sum_{\alpha,\beta=1}^{\tilde{m}} \Lambda^\alpha_a (\mathscr{L}_{\tilde{Y}_\alpha} \Lambda^\beta_b)\tilde{Y}_\beta + \Lambda^\alpha_a \Lambda^\beta_b S^\delta(\tilde{Y}_\alpha, \tilde{Y}_\beta)\tilde{Y}_\delta.$$

The lemma follows directly from this formula since the terms in $\Gamma(\mathsf{Y})$ will go to zero when projected by $\pi_{S_{q_0}}$ since $\mathsf{Y}_{q_0} \subset S_{q_0}$. □

4.2. Main results. Our main results may now be stated. Let us first state a sufficient condition for controllability.

THEOREM 4.4. *Let $(Q, \nabla, \mathsf{D}, \mathscr{Y})$ be an analytic affine connection precontrol system, and suppose that $\mathscr{Y}$ is adapted at $q_0 \in Q$. Suppose that*

(i) $\mathsf{Sym}^{(\infty)}(\mathsf{Y})_{q_0} = \langle Z_{q_0}(\mathscr{Y}), \mathsf{Sym}^{(2)}(\mathsf{Y})\rangle$, *and*

(ii) $B_{\mathsf{Y}_{q_0}}(\langle Z_{q_0}(\mathscr{Y}), \mathsf{Y}_{q_0}\rangle)$ *is essentially indefinite.*

Then $(Q, \nabla, \mathsf{D}, \mathscr{Y})$ is properly STLC from q_0 if it is accessible from q_0, and is properly STLCC from q_0 if it is configuration accessible from q_0.

Proof. The proof essentially follows from Theorem 2.5. However, the extension to allow singular points for the input distribution Y does not follow directly from Theorem 2.5 but requires some manipulations with the variational cone, which we will not go into here. The idea, in essence, is that if an input vector field vanishes at the reference point, then directions generated by symmetric products using these vector fields come "for free." Since these symmetric products are simply applications of a linear map, this explains the presence of the invariant subspace characterizations of the tangent space to the reachable set. We refer to [21, Lemma 7.2] for the details behind this, noting that the discussion in that paper builds on concepts presented in [36, 7]. A consequence of these discussions, once they are specialized to our setting, is the following result.

LEMMA 4.5. *Let $(Q, \nabla, \mathsf{D}, \mathscr{Y} = \{Y_1, \dots, Y_m\})$ be an analytic affine connection precontrol system for which $\mathscr{Y}$ is adapted at $q_0 \in Q$. Assume the following:*

(i) $\mathsf{Sym}^{(\infty)}(\mathsf{Y}) = \langle Z_{q_0}(\mathscr{Y}), \mathsf{Sym}^{(2)}(\mathsf{Y})_{q_0}\rangle$*;*

(ii) *there exist $\tilde{m} \geq m$ and a full-rank matrix $\boldsymbol{T} \in \mathbb{R}^{m\times\tilde{m}}$ such that if $\tilde{Y}_\alpha = T^a_\alpha Y_\alpha$, then*

$$\sum_{a=1}^{\tilde{m}} \langle \tilde{Y}_\alpha : \tilde{Y}_\alpha\rangle(q_0) \in \langle Z_{q_0}(\mathscr{Y}), \mathsf{Y}_{q_0}\rangle.$$

Then $(Q, \nabla, \mathsf{D}, \mathscr{Y})$ is properly STLC from q_0 if it is accessible from q_0 and is properly STLCC from q_0 if it is configuration accessible from q_0.

We shall show that if the hypotheses of Theorem 4.4 are satisfied at q_0, then the hypotheses of Lemma 4.5 are satisfied for some possibly different collection of input vector fields. From this the conclusion of Theorem 4.4 will follow.

For brevity let us denote $S_{q_0} = \langle Z_{q_0}(\mathscr{Y}), \mathsf{Y}_{q_0}\rangle$ and $B = B_{\mathsf{Y}_{q_0}}(S_{q_0})$. First we need to find an appropriate collection of input vector fields. Choose $v_1, \dots, v_\ell \in \mathsf{Y}_{q_0}$ such that $0_{q_0} + S_{q_0} \in \mathsf{Sym}^{(\infty)}(\mathsf{Y})_{q_0}/S_{q_0}$ lies in the interior of the convex hull of the vectors $B(v_1, v_1), \dots, B(v_\ell, v_\ell)$. That this is possible is guaranteed by the hypotheses of Theorem 4.4 and by Lemma 3.2. If necessary, add vectors $v_{\ell+1}, \dots, v_{\tilde{k}}$ such that the vectors $v_1, \dots, v_{\tilde{k}}$ span Y_{q_0}. It now follows that the vectors $B(v_1, v_1), \dots, B(v_{\tilde{k}}, v_{\tilde{k}})$ contain $0_{q_0} + S_{q_0} \in \mathsf{Sym}^{(\infty)}(\mathsf{Y})_{q_0}/S_{q_0}$ in the interior of their convex hull. Thus the vectors $v_1, \dots, v_{\tilde{k}}$ may be rescaled by strictly positive constants (for simplicity, let us denote the rescaled vectors also by $v_1, \dots, v_{\tilde{k}}$) so that

$$\sum_{\alpha=1}^{\tilde{k}} B(v_\alpha, v_\alpha) = 0_{q_0} + S_{q_0} \in \mathsf{Sym}^{(\infty)}(\mathsf{Y})_{q_0}/S_{q_0}. \tag{4.2}$$

It is now possible to define vector fields $\tilde{\mathscr{Y}} = \{\tilde{Y}_1, \dots, \tilde{Y}_{\tilde{m}}\}$ such that, if $\dim(\mathsf{Y}_{q_0}) = k$, then

1. $\tilde{Y}_{\tilde{k}+a} = Y_{k+a}$, $a \in \{1, \dots, m-k\}$, and
2. $\tilde{Y}_\alpha = \sum_{a=1}^{k} \tilde{T}^a_\alpha Y_a$, $\alpha \in \{1, \dots, \tilde{k}\}$, for a full-rank matrix $\tilde{\boldsymbol{T}} \in \mathbb{R}^{k\times\tilde{k}}$.

Clearly this then implies the existence of a full-rank matrix $\boldsymbol{T} \in \mathbb{R}^{m \times \tilde{m}}$ such that $\tilde{Y}_\alpha = T^a_\alpha Y_a$, $\alpha \in \{1, \dots, \tilde{m}\}$. From (4.2) it immediately follows that $(Q, \nabla, \mathsf{D}, \mathscr{Y})$ satisfies the hypotheses of Lemma 4.5, and so Theorem 4.4 follows. □

Remark 5. Our use of the vector fields $Z_{q_0}(\mathscr{Y})$ from $\mathscr{Y}$ that vanish at q_0 is similar in spirit to how the vanishing of the drift vector appears in the work of Sussmann [36] and Bianchini and Stefani [7]. The idea is that brackets generated by such vanishing vector fields can be achieved "for free," without invoking bad brackets.

A necessary condition for controllability is the following.

THEOREM 4.6. *Let $(Q, \nabla, \mathsf{D}, \mathscr{Y})$ be an analytic affine connection precontrol system for which $\mathscr{Y}$ is adapted at $q_0 \in Q$. Suppose that*

(i) *q_0 is a regular point for Y and*

(ii) *$B_{\mathsf{Y}_{q_0}}(\mathsf{Y}_{q_0})$ is definite.*

Then $(Q, \nabla, \mathsf{D}, \mathscr{Y})$ is STLCUC from q_0.

Proof. We work locally. Therefore, we may assume the vector fields $\{Y_1, \dots, Y_m\}$ are linearly independent in a neighborhood of q_0. First we show that the system is not STLC from q_0 using calculations of Hirschorn and Lewis [21]. We will not provide here a self-contained justification for all of our computations, since they take considerable space, but we refer to the paper [21]. The calculation uses the Chen–Fliess–Sussmann series [14, 17, 35]. For an analytic control-affine system

$$\xi'(t) = f_0(\xi(t)) + \sum_{a=1}^{m} u_a(t) f_a(\xi(t)), \qquad \xi(t) \in M,$$

on a manifold M with a compact control set, and for an analytic function ϕ, the Chen–Fliess–Sussmann series gives the following formula for the value of ϕ along a controlled trajectory (ξ, u):

$$\phi(\xi(t)) = \sum_J U_J(t) f_J \phi(\xi(0)).$$

The sum is over multi-indices $J = (a_1, \dots, a_k)$ in $\{0, 1, \dots, m\}$,

$$U_J(t) = \int_0^t u_{a_k}(t_k) \int_0^{t_k} u_{a_{k-1}}(t_{k-1}) \cdots \int_0^{t_2} u_{a_1}(t_1)\, \mathrm{d}t_1 \dots \mathrm{d}t_{k-1}\, \mathrm{d}t_k$$

and

$$f_J \phi = f_{a_1} f_{a_2} \cdots f_{a_k} \phi.$$

We adopt the convention that $u_0 = 1$. We also regard an affine connection control system as a control-affine system in the usual manner by taking f_0 to be the geodesic spray for ∇ and $f_1, \dots, f_m$ to be the vertical lifts of $Y_1, \dots, Y_m$ [27].

The function we evaluate is defined as follows. We let λ be an analytic covector field defined in a neighborhood of q_0 with the following properties:

1. λ annihilates the distribution Y;
2. $\lambda(q_0) B_{\mathsf{Y}_{q_0}} | \mathsf{Y}_{q_0}$ is negative-definite.

By a linear input transformation one can ensure that the input vector fields diagonalize $\lambda(q_0) B_{\mathsf{Y}_{q_0}}$ with the diagonal entries being -1. We assume this input transformation to have been made. We then define a function $\phi_\lambda \colon TQ \to \mathbb{R}$ by $\phi_\lambda(v_q) = \lambda(q) \cdot v_q$, and we also define

$$\Phi^+_\lambda = \{v_q \in TQ |\ \phi_\lambda(v_q) > 0\}, \quad \Phi^-_\lambda = \{v_q \in TQ |\ \phi_\lambda(v_q) < 0\}.$$

Note that in any neighborhood V of 0_{q_0} in TQ, the sets $V \cap \Phi_\lambda^-$ and $V \cap \Phi_\lambda^+$ will be nonempty, since ϕ_λ is linear on the fibers of TQ. Therefore, we can show that $(Q, \nabla, \mathsf{D}, \mathscr{Y})$ is STLUC from q_0 by showing that ϕ_λ has constant sign along any controlled trajectory. One may directly verify that ϕ_λ has the following properties:

1. $f_a\phi_\lambda$, $a \in \{1, \dots, m\}$, is zero in a neighborhood of 0_{q_0};
2. $\mathrm{ad}_{f_0}^k f_a \phi_\lambda(0_{q_0}) = 0$, $a \in \{1, \dots, m\}$, $k \in \mathbb{N}$;
3. $[f_a, [f_0, f_a]]\phi_\lambda(0_{q_0}) = -1$, $a \in \{1, \dots, m\}$ (this and the next fact use the formula $[f_a, [f_0, f_b]] = \mathrm{verlift}(\langle Y_a : Y_b \rangle)$, $a, b \in \{1, , \dots, m\}$);
4. $[f_a, [f_0, f_b]]\phi_\lambda(0_{q_0}) = 0$, $a, b \in \{1, \dots, m\}$, $a \neq b$.

For an input $u \colon [0, T] \to U$, let us define

$$\|u\|_{2,t} = \max\Big\{\Big(\int_0^t |u_a(t)|^2\Big)^{1/2} \Big| \ a \in \{1, \dots, m\}\Big\}.$$

The calculations of Hirschorn and Lewis [21] now immediately give the following inequality for $\phi_\lambda(\gamma'(t))$ along a controlled trajectory (γ, u) for an affine connection control system like that under consideration here:

$$\phi_\lambda(\gamma'(t)) \geq \frac{1}{2}(\|u\|_{2,t})^2 - |E(t)|.$$

According to the analysis in Hirschorn and Lewis, the map $t \mapsto E(t)$ satisfies the bound $|E(t)| \leq tE_0(\|u\|_{2,t})^2$ for some $E_0 > 0$. For t sufficiently small, this shows that $\phi_\lambda(\gamma'(t))$ has constant sign. This shows that $(Q, \nabla, \mathsf{D}, \mathscr{Y})$ is STLCUC from q_0.

Now let us show that our above constructions also preclude the system from being locally *configuration* controllable. Choose a coordinate chart (V, χ) for Q around q_0 with the following properties: (1) $\chi(q_0) = \mathbf{0}$ and (2) $\mathrm{d}q^n(q_0) = \lambda(q_0)$. Let us define a function ψ_λ on the coordinate domain V by $\psi_\lambda(q) = q^n$ such that the sets

$$\Psi_\lambda^+ = \{q \in Q| \ \psi_\lambda(q) > 0\}, \quad \Psi_\lambda^- = \{q \in Q| \ \psi_\lambda(q) < 0\}$$

each intersect any neighborhood of $q_0 \in Q$. Along any nontrivial trajectory $t \mapsto \gamma(t)$ we have

$$\frac{\mathrm{d}\psi_\lambda(\gamma(t))}{\mathrm{d}t}\Big|_{t=0} = \mathrm{d}\psi_\lambda(\gamma'(0)) = \phi_\lambda(\gamma'(0)) < 0.$$

Since $\psi_\lambda(q_0) = 0$, this means that, for sufficiently small t, $\psi_\lambda(\gamma(t)) < 0$, and this shows that the points in Ψ_λ^+ are not reachable in small time, and so Σ is not locally configuration controllable. □

Remark 6. The spirit of the preceding proof is that of the single-input necessary condition appearing as Proposition 6.3 in the paper by Sussmann [35]. However, the modifications to the multi-input case by Hirschorn and Lewis [21] require some care.

Let us provide an example that nicely illustrates Theorems 4.4 and 4.6. This example is a slight modification of an example in [33].

Example 1. We take $Q = \mathbb{R}^2$ with (x, y) the usual Cartesian coordinates. We choose the affine connection on $\mathbb{R}^2$ with all vanishing Christoffel symbols except for $\Gamma_{xx}^y = x$. We choose the single-input vector field $Y = \frac{\partial}{\partial x}$. We also take $\mathsf{D} = TQ$. One then readily computes

$$\langle Y : Y \rangle = 2x\frac{\partial}{\partial y}, \qquad \langle Y : \langle Y : Y \rangle\rangle = 2\frac{\partial}{\partial y}.$$

We consider two cases.

1. $q_0 = (0, y)$, $y \in \mathbb{R}$: We readily see that $B_{\mathsf{Y}_{q_0}}(\langle Z_{q_0}(\mathscr{Y}), \mathsf{Y}_{q_0}\rangle)$ is identically zero, and so essentially indefinite. We also have $\mathtt{Sym}^{(2)}(\mathsf{Y})_{q_0} = T_{q_0}Q$. Therefore, Theorem 4.4 shows that $(Q, \nabla, \mathbf{D}, \{Y\})$ is properly STLC from q_0.

2. $q_0 \neq (0, y)$, $y \in \mathbb{R}$: Here we use $\mathrm{span}_{\mathbb{R}}\{\frac{\partial}{\partial y}\}$ as a model for $T_{q_0}Q/\mathsf{Y}_{q_0}$. Thus both Y_{q_0} and $T_{q_0}Q/\mathsf{Y}_{q_0}$ are one-dimensional, and so $B_{\mathsf{Y}_{q_0}}(\mathsf{Y}_{q_0})$ is essentially a quadratic function on $\mathbb{R}$. This quadratic function is then exactly $\xi \mapsto 2x\xi^2$. This function is definite, so Theorem 4.6 implies that the system is STLCUC from q_0.

Thus this example has the rather degenerate feature of being controllable on the y-axis but being uncontrollable at every point in a neighborhood of the y-axis. Note that this example is also a counterexample to a single-input result of one of the authors [23]. There it was stated that a single-input affine connection control system is STLCC if and only if the dimension of the configuration space is one. We see here that this is false. However, what is true is that a single-input affine connection control system is STLCC at all points in an *open subset* of configuration space if and only if the configuration space has dimension one.

5. Reductions of affine connection control systems. The controllability results of section 4 turn out to apply to a great many examples. That is to say, many interesting physical examples may be shown to be controllable or uncontrollable using these results. What is not obvious is that many of these systems are describable, in some sense, by a driftless system. This effectively simplifies the system, making certain control design tasks, especially motion planning, considerably simpler. In this section we introduce the framework for discussing these simplifications.

The objective in this section is then to relate second-order systems to first-order systems. In order to do this, one must be aware that the allowable inputs for the two classes of systems cannot be the same. For example, a trajectory for a first-order system using a discontinuous input will be continuous in configuration, but not in velocity. These velocity discontinuities are not allowed for second-order systems with bounded inputs. Therefore, we need to fix a set of inputs to use in each case, and they need to differ, essentially, by one integration. To be specific, we let $\mathscr{U}_{\mathrm{kin}}$ be the collection of locally absolutely continuous controls, and we let $\mathscr{U}_{\mathrm{dyn}}$ be the collection of locally integrable controls. The former will be used for first-order systems and the latter for second-order systems. In all cases, we allow controls to be defined on an arbitrary interval $I \subset \mathbb{R}$.

5.1. Kinematic reductions. In this section, in order to emphasize the difference between the two kinds of systems we are comparing, we shall denote an affine connection control system by $\Sigma_{\mathrm{dyn}} = (Q, \nabla, \mathbf{D}, \mathscr{Y}, \mathbb{R}^m)$. A *driftless system* is a triple $\Sigma_{\mathrm{kin}} = (Q, \mathscr{X} = \{X_1, \dots, X_{\tilde{m}}\}, U \subset \mathbb{R}^{\tilde{m}})$. The associated control system is then

$$\gamma'(t) = \sum_{\alpha=1}^{\tilde{m}} \tilde{u}^\alpha(t) X_\alpha(\gamma(t)), \tag{5.1}$$

so that a *controlled trajectory* is a pair $(\gamma, \tilde{u})$, where

1. $\gamma\colon I \to Q$ and $\tilde{u}\colon I \to U$ are both defined on the same interval $I \subset \mathbb{R}$,
2. $\tilde{u} \in \mathscr{U}_{\mathrm{kin}}$, and
3. $(\gamma, \tilde{u})$ together satisfy (5.1).

A driftless system $(Q, \mathscr{X}, U)$ is *STLC from* q_0 if the set of points reachable from q_0 contains q_0 in its interior, and a pair $(Q, \mathscr{X})$ is *properly STLC from* q_0 if $(Q, \mathscr{X}, U)$ is STLC from q_0 for every proper U. With our underlying assumption of analyticity, it is well known that $(Q, \mathscr{X})$ is properly STLC from q_0 if and only if $\mathtt{Lie}^{(\infty)}(\mathsf{X})_{q_0} = T_{q_0}Q$.

First we define what we mean by a kinematic reduction.

DEFINITION 5.1. *Let* $\Sigma_{\text{dyn}} = (Q, \nabla, \mathsf{D}, \mathscr{Y} = \{Y_1, \dots, Y_m\}, \mathbb{R}^m)$ *be an affine connection control system with* Y *having constant rank. A driftless system* $\Sigma_{\text{kin}} = (Q, \mathscr{X} = \{X_1, \dots, X_{\tilde{m}}\}, \mathbb{R}^{\tilde{m}})$ *is a* kinematic reduction *of* Σ_{dyn} *if*

(i) X *is a constant-rank subbundle of* D *and*

(ii) *for every controlled trajectory* (γ, u_{kin}) *for* Σ_{kin} *with* $u_{\text{kin}} \in \mathscr{U}_{\text{kin}}$*, there exists* $u_{\text{dyn}} \in \mathscr{U}_{\text{dyn}}$ *such that* (γ, u_{dyn}) *is a controlled trajectory for* Σ_{dyn}.

The rank *of the kinematic reduction* Σ_{kin} *is the rank of* X.

Thus kinematic reductions are driftless systems whose controlled trajectories, at least for controls in $\mathscr{U}_{\text{kin}}$, can be followed by controlled trajectories of Σ_{dyn}. Let us characterize kinematic reductions. To do so, recall that with our constant-rank assumptions, given an affine connection ∇ and a family of vector fields $\mathscr{Y} = \{Y_1, \dots, Y_m\}$ on Q, we may globally define B_{Y} as at the end of section 3.2. This also allows us to define a map $Q_{B_{\mathsf{Y}}} \colon \Gamma(TQ) \to \Gamma(TQ/\mathsf{Y})$ by $Q_{B_{\mathsf{Y}}}(X)(q) = B_{\mathsf{Y}}(q)(X(q), X(q))$. With this notation, we have the following result.

THEOREM 5.2. *Let* $\Sigma_{\text{dyn}} = (Q, \nabla, \mathsf{D}, \mathscr{Y}, \mathbb{R}^m)$ *be an affine connection control system with* Y *of constant rank and let* $\Sigma_{\text{kin}} = (Q, \mathscr{X}, \mathbb{R}^{\tilde{m}})$ *be a driftless system with* X *of constant rank. The following statements are equivalent:*

(i) Σ_{kin} *is a kinematic reduction of* Σ_{dyn}*;*

(ii) $\mathsf{Sym}^{(1)}(\mathsf{X}) \subset \mathsf{Y}$*;*

(iii) $\mathsf{X} \subset \mathsf{Y}$ *and* $Q_{B_{\mathsf{Y}}}|\mathsf{X} = 0$.

Proof. (i) $\Longrightarrow$ (ii) Let $X \in \Gamma(\mathsf{X})$ such that $X = \phi^\alpha X_\alpha$ for some $\phi^1, \dots, \phi^{\tilde{m}} \in \mathcal{F}(Q)$. For $q \in Q$, define controls $\tilde{u}_1, \tilde{u}_2 \in \mathscr{U}_{\text{kin}}$ by $\tilde{u}_1 = (\phi^1(q), \dots, \phi^{\tilde{m}}(q))$ and $\tilde{u}_2 = (1+t)\tilde{u}_1$. Let $(\gamma_1, \tilde{u}_1)$ and $(\gamma_2, \tilde{u}_2)$ be the corresponding controlled trajectories of Σ_{kin} satisfying $\gamma_1(0) = \gamma_2(0) = q$. Thus $\gamma_i'(t) = \sum_{\alpha=1}^{\tilde{m}} \tilde{u}_i^\alpha(t) X_\alpha(\gamma_i(t))$, $i \in \{1, 2\}$. We compute

$$\begin{aligned}\nabla_{\gamma_1'(t)}\gamma_1'(t) &= \sum_{\alpha,\beta=1}^{\tilde{m}} \nabla_{\tilde{u}_1^\alpha(t) X_\alpha(\gamma_1(t))} \tilde{u}_1^\beta(t) X_\beta(\gamma_1(t)) \\ &= \sum_{\alpha,\beta=1}^{m} \tilde{u}_1^\alpha(t)\tilde{u}_1^\beta(t) \nabla_{X_\alpha(\gamma_1(t))} X_\beta(\gamma_1(t)) + \dot{\tilde{u}}_1^\beta(t) X_\beta(\gamma_1(t)).\end{aligned}$$

Evaluating this at $t = 0$ gives

$$\nabla_{\gamma_1'(t)}\gamma_1'(t)\big|_{t=0} = \sum_{\alpha,\beta=1}^{\tilde{m}} \tilde{u}_1^\alpha(0)\tilde{u}_1^\beta(0) \nabla_{X_\alpha} X_\beta(q) + \dot{\tilde{u}}_1^\beta(0) X_\beta(q) = \nabla_X X(q).$$

Similarly, for γ_2 we have

$$\nabla_{\gamma_2'(t)}\gamma_2'(t)\big|_{t=0} = \nabla_X X(q) + X(q).$$

Therefore, since Σ_{kin} is a kinematic reduction of Σ_{dyn}, we have $\nabla_X X(q), \nabla_X X(q) + X(q) \in \mathsf{Y}_q$, or simply $X, \nabla_X X \in \Gamma(\mathsf{Y})$ since the above constructions can be performed for all $X \in \Gamma(\mathsf{X})$ and $q \in Q$. Therefore, for $X, Y \in \Gamma(\mathsf{X})$ we have the polarization identity,

$$\langle X : Y \rangle = \frac{1}{2}\big(\langle X + Y : X + Y \rangle - \langle X : X \rangle - \langle Y : Y \rangle\big) \in \Gamma(\mathsf{Y}), \tag{5.2}$$

which gives (ii).

(ii) $\Longrightarrow$ (iii) From the definition of B_{Y} we readily see that $Q_{B_{\mathsf{Y}}}|\mathsf{X} = 0$ exactly means that $\langle X : X\rangle = 2\nabla_X X \in \Gamma(\mathsf{Y})$ for each $X \in \Gamma(\mathsf{X})$. From this observation, the current implication follows easily by employing the formula for $\langle X : Y\rangle$ in (5.2).

(iii) $\Longrightarrow$ (i) As in the preceding step, we see that the condition $Q_{B_{\mathsf{Y}}}|\mathsf{X} = 0$ is equivalent to asserting that $\nabla_X X \in \Gamma(\mathsf{Y})$ for each $X \in \Gamma(\mathsf{X})$. By (5.2) this implies that $\langle X_\alpha : X_\beta\rangle \in \Gamma(\mathsf{Y})$ for $\alpha, \beta \in \{1, \dots, \tilde{m}\}$. Let $u_{\text{kin}} \in \mathscr{U}_{\text{kin}}$ and let (γ, u_{kin}) be the corresponding controlled trajectory for Σ_{kin}. We then have

$$\nabla_{\gamma'(t)}\gamma'(t) = u^\alpha_{\text{kin}}(t)u^\beta_{\text{kin}}(t)\nabla_{X_\alpha(\gamma(t))}X_\beta(\gamma(t)) + \dot{u}^\alpha_{\text{kin}}(t)X_\alpha(\gamma(t)).$$

We note that

$$u^\alpha_{\text{kin}}(t)u^\beta_{\text{kin}}(t)\nabla_{X_\alpha(\gamma(t))}X_\beta(\gamma(t)) = \frac{1}{2}u^\alpha_{\text{kin}}(t)u^\beta_{\text{kin}}(t)\langle X_\alpha(\gamma(t)) : X_\beta(\gamma(t))\rangle.$$

Since $X_\alpha, \langle X_\alpha : X_\beta\rangle \in \Gamma(\mathsf{Y})$ it now follows that $\nabla_{\gamma'(t)}\gamma'(t) \in \mathsf{Y}_{\gamma(t)}$, implying that there exists a control $u_{\text{dyn}} \in \mathscr{U}_{\text{dyn}}$ such that (γ, u_{dyn}) is a controlled trajectory for Σ_{dyn}. □

Of particular interest are kinematic reductions of rank one: $(Q, \{X_1\}, \mathbb{R})$. In this case, any vector field of the form $X = \phi X_1$, where $\phi \in \mathcal{F}(Q)$ is nowhere vanishing, is called a *decoupling vector field*. From Theorem 5.2 we have the following description of a decoupling vector field.

COROLLARY 5.3. *A vector field X is a decoupling vector field for $\Sigma_{\text{dyn}} = (Q, \nabla, \mathsf{D}, \mathscr{Y}, \mathbb{R}^m)$ if and only if $X, \nabla_X X \in \Gamma(\mathsf{Y})$.*

It is the notion of a decoupling vector field that was initially presented by Bullo and Lynch [13], and which is generalized by our idea of a kinematic reduction.

Remark 7. While in general, even when a kinematic reduction exists, it will not be easy to find, it turns out that in practice many examples exhibit kinematic reductions in a more or less obvious way. We shall see this in the examples below. Note that condition (iii) of Theorem 5.2 provides a set of algebraic equations that can, in principle, be solved to identify decoupling vector fields. This was discussed by Bullo and Lynch [13].

Next, let us consider affine connection control systems endowed with multiple kinematic reductions. It is interesting to characterize when the concatenation of controlled trajectories of the kinematic reductions gives rise to a controlled trajectory for the affine connection control system. Given two curves γ_1 and γ_2 on Q, let $\gamma_1 * \gamma_2$ be their concatenation. The following lemma follows immediately from the definition of a kinematic reduction.

LEMMA 5.4. *Consider an affine connection control system $\Sigma_{\text{dyn}} = (Q, \nabla, \mathsf{D}, \mathscr{Y}, \mathbb{R}^m)$ with two kinematic reductions $\Sigma_{\text{kin},1} = (Q, \mathscr{X}_1, \mathbb{R}^{m_1})$ and $\Sigma_{\text{kin},2} = (Q, \mathscr{X}_2, \mathbb{R}^{m_2})$. For $i \in \{1, 2\}$, let $(\gamma_i, u_{\text{kin},i})$ be a controlled trajectory for $\Sigma_{\text{kin,i}}$ defined on the interval $[0, T_i]$ with $u_{\text{kin},i} \in \mathscr{U}_{\text{kin}}$. There exists a control $u_{\text{dyn}} \in \mathscr{U}_{\text{dyn}}$ such that $(\gamma_1 * \gamma_2, u_{\text{dyn}})$ is a controlled trajectory for Σ_{dyn} if and only if $\gamma_1'(T_1) = \gamma_2'(0)$.*

Motivated by this result we make the following definition.

DEFINITION 5.5. *An affine connection control system $\Sigma_{\text{dyn}} = (Q, \nabla, \mathsf{D}, \mathscr{Y}, \mathbb{R}^m)$ is* kinematically controllable from $q_0 \in Q$ (KC from $q_0 \in Q$) *if there exists a finite collection*

$$\Sigma_{\text{kin},1} = (Q, \mathscr{X}_1, \mathbb{R}^{m_1}), \dots, \Sigma_{\text{kin},k} = (Q, \mathscr{X}_k, \mathbb{R}^{m_k})$$

of kinematic reductions for Σ_{dyn} such that $(Q, \mathscr{X}_1 \cup \cdots \cup \mathscr{X}_k)$ is properly STLC from q_0.

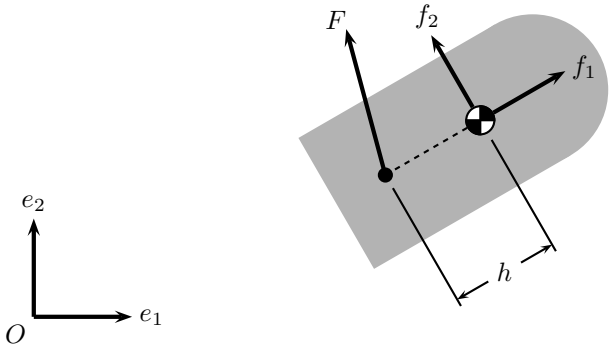


FIG. 5.1. *Planar rigid body with thruster.*

Remark 8.

1. For analytic systems, the condition that $(Q, \mathscr{X}_1 \cup \cdots \cup \mathscr{X}_k)$ be properly STLC from q_0 is equivalent to the condition that $\mathsf{Lie}^{(\infty)}(\mathsf{X}_1 + \cdots + \mathsf{X}_k)_{q_0} = T_{q_0}Q$, where $\mathsf{X}_1 + \cdots + \mathsf{X}_k$ denotes the fiberwise sum of the distributions $\mathsf{X}_1, \ldots, \mathsf{X}_k$.

2. If an affine connection control system $\Sigma_{\text{dyn}} = (Q, \nabla, \mathsf{D}, \mathscr{Y}, \mathbb{R}^m)$ is kinematically controllable from q_0, then it is STLCC from q_0. This fact is proved in Proposition 5.10 below, and we refer to section 5.3 for a discussion of the relationships between the various notions of controllability introduced in this paper.

3. Suppose the affine connection control system $\Sigma_{\text{dyn}} = (Q, \nabla, \mathsf{D}, \mathscr{Y}, \mathbb{R}^m)$ is kinematically controllable from all $q \in Q$. A standard control problem is to find a controlled trajectory connecting two given configurations $q_1, q_2 \in Q$, starting and ending with zero velocity. Lemma 5.4 says that this can be done for Σ_{dyn} by concatenating integral curves of decoupling vector fields where each segment is reparameterized to start and end at zero velocity. This is the viewpoint of Bullo and Lynch [13].

Example 2. We consider a planar rigid body with a variable-direction thruster as shown in Figure 5.1. The system has configuration manifold $SE(2)$. We use coordinates (x, y, θ) defined as follows. Let $\{e_1, e_2\}$ be an orthonormal frame in E^2 fixed at $O \in E^2$, and let $\{f_1, f_2\}$ be a body orthonormal frame attached to the center of mass and with the property that the vector f_1 points in the direction of the line connecting the center of mass with the point of application of the force (see Figure 5.1). Then (x, y) denote the position of the center of mass with respect to O, and θ is defined so that $f_1 = R(\theta)e_1$ with $R(\theta)$ the matrix giving a positive rotation by θ in E^2. With respect to these coordinates, the kinetic energy of the system is determined by the Riemannian metric

$$g = m\mathrm{d}x \otimes \mathrm{d}x + m\mathrm{d}y \otimes \mathrm{d}y + J\mathrm{d}\theta \otimes \mathrm{d}\theta,$$

where m is the mass of the body, and J is its inertia about the center of mass. Since the coefficients of this Riemannian metric are independent of the coordinates, the Christoffel symbols for the corresponding Levi-Civita affine connection are zero. As shown by Lewis and Murray [27], Newton's law with the force F as shown in Figure 5.1 is equivalent to (2.1) if the affine connection ∇ is the Levi-Civita connection associated with g and if the vector fields $\{Y_1, Y_2\}$ are chosen as follows:

$$Y_1 = \frac{\cos\theta}{m}\frac{\partial}{\partial x} + \frac{\sin\theta}{m}\frac{\partial}{\partial y}, \qquad Y_2 = -\frac{\sin\theta}{m}\frac{\partial}{\partial x} + \frac{\cos\theta}{m}\frac{\partial}{\partial y} - \frac{h}{J}\frac{\partial}{\partial\theta}.$$

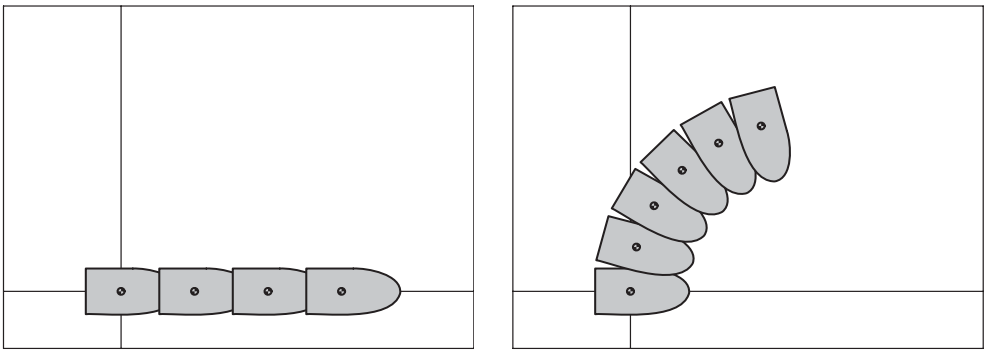


FIG. 5.2. *Decoupling motions for the planar rigid body: X_1 on the left and X_2 on the right.*

The system is unconstrained so we take $\mathsf{D} = TQ$.

We claim that the vector fields $X_1 = mY_1$ and $X_2 = mY_2$ are decoupling vector fields. Clearly, they are sections of Y. We also compute

$$\nabla_{X_1} X_1 = 0, \quad \nabla_{X_2} X_2 = \frac{mh\cos\theta}{J}\frac{\partial}{\partial x} + \frac{mh\sin\theta}{J}\frac{\partial}{\partial y}.$$

Therefore $\nabla_{X_1} X_1, \nabla_{X_2} X_2 \in \Gamma(\mathsf{Y})$, showing that X_1 and X_2 are indeed decoupling vector fields.

Let us explore the implications of the existence of these decoupling vector fields. Since X_1 and X_2 are decoupling vector fields, we may follow their integral curves. In Figure 5.2 we show motions of the body along sample integral curves of X_1 and X_2. In actuality, one can follow not only the integral curves of the decoupling vector fields but also any reparameterization of these vector fields. With this in mind, one has the following possible methodology for moving the body around in the plane:

1. Given $q_1, q_2 \in Q$, find a concatenation of the integral curves of X_1 and X_2 that connects q_1 with q_2. (This is possible since $\mathsf{Lie}^{(\infty)}(\mathsf{X}) = TQ$.)
2. Reparameterize each segment of the preceding concatenated curve so that each segment has zero initial and final velocity.
3. Because of Lemma 5.4, the resulting reparameterized curve can be followed by controlled trajectories of Σ_{dyn}.

This method for motion planning is explained in detail in [11, Chapter 13].

5.2. Maximally reducible systems. If $\Sigma_{\text{kin}} = (Q, \mathscr{X} = \{X_1, \dots, X_{\tilde{m}}\}, \mathbb{R}^{\tilde{m}})$ is a kinematic reduction of $\Sigma_{\text{dyn}} = (Q, \nabla, \mathsf{D}, \mathscr{Y} = \{Y_1, \dots, Y_m\}, \mathbb{R}^m)$, then, by definition, any controlled trajectory of Σ_{kin} may be followed by a controlled trajectory of Σ_{dyn}. In this section we wish to consider the possibility of the converse statement. The following definition, and the attendant Theorem 5.7 below, are due to Lewis [25].

DEFINITION 5.6. *An affine connection control system $\Sigma_{\text{dyn}} = (Q, \nabla, \mathsf{D}, \mathscr{Y} = \{Y_1, \dots, Y_m\}, \mathbb{R}^m)$ with Y of constant rank is* maximally reducible *to $\Sigma_{\text{kin}} = (Q, \mathscr{X} = \{X_1, \dots, X_{\tilde{m}}\}, \mathbb{R}^{\tilde{m}})$ if Σ_{kin} is a kinematic reduction of Σ_{dyn} and if for every controlled trajectory (γ, u_{dyn}) for Σ_{dyn} satisfying $\gamma'(0) \in \mathsf{X}_{\gamma(0)}$, there exists a control $u_{\text{kin}} \in \mathscr{U}_{\text{kin}}$ such that (γ, u_{kin}) is a controlled trajectory for Σ_{kin}.*

Before we proceed to characterize maximally reducible systems, let us illustrate that a system may not be maximally reducible to a given kinematic reduction.

Example 3 (Example 2 cont'd). We claim that the affine connection control system corresponding to the planar rigid body with a thruster is not maximally re-

ducible to either of the kinematic reductions $\Sigma_{\text{kin},1} = (Q, \mathscr{X}_1 = \{X_1\}, \mathbb{R})$ or $\Sigma_{\text{kin},2} = (Q, \mathscr{X}_2 = \{X_2\}, \mathbb{R})$ exhibited in Example 2. We shall exhibit this explicitly for $\Sigma_{\text{kin},1}$ and leave the other case to the reader.

Consider the control $t \mapsto u(t) = (0,1) \in \mathscr{U}_{\text{dyn}}$, along with the initial condition $\gamma'(0) = ((0,0,0),(1,0,0)) \in TQ$. We have $\gamma'(0) \in \mathsf{X}_{1,\gamma(0)}$, where X_1 is the distribution generated by the vector field X_1. If Σ_{dyn} is to be maximally reducible to $\Sigma_{\text{kin},1}$, then we should have $\gamma'(t) \in \mathsf{X}_{1,\gamma(t)}$ for each $t > 0$. To show that this is not the case, consider the governing equations for the system with the given control:

$$\ddot{x} = -\frac{\sin\theta}{m}, \quad \ddot{y} = \frac{\cos\theta}{m}, \quad \ddot{\theta} = -\frac{h}{J}.$$

Clearly the solution to this ordinary differential equation is not a reparameterization of the integral curve for X_1 through $\gamma(0)$ since the latter is given by $t \mapsto (t,0,0)$. Thus it cannot be that $\gamma'(t) \in \mathsf{X}_{1,\gamma(t)}$ for each $t > 0$.

Now let us establish when an affine connection control system is in fact maximally reducible to *some* driftless system. Note that in the statement of the following theorem, the driftless systems to which Σ_{dyn} is maximally reducible are characterized sharply.

THEOREM 5.7. *An affine connection control system* $\Sigma_{\text{dyn}} = (Q, \nabla, \mathsf{D}, \mathscr{Y} = \{Y_1, \dots, Y_m\}, \mathbb{R}^m)$, *with* Y *of constant rank, is maximally reducible to* $\Sigma_{\text{kin}} = (Q, \mathscr{X} = \{X_1, \dots, X_{\tilde{m}}\}, \mathbb{R}^{\tilde{m}})$ *if and only if the following two conditions hold:*

(i) $\mathsf{X} = \mathsf{Y}$;

(ii) $\mathsf{Sym}^{(\infty)}(\mathsf{Y}) = \mathsf{Y}$.

Proof. In the proof it is convenient to understand that the second-order system (2.1) on Q is equivalent to the first-order system on TQ given

$$\Upsilon'(t) = Z(\Upsilon(t)) + \sum_{a=1}^{m} u^a(t)\text{verlift}(Y_a)(\Upsilon(t)) \tag{5.3}$$

for a curve Υ on TQ, where Z is the geodesic spray for ∇ and $\text{verlift}(Y_a) \in \Gamma(TTQ)$ denotes the vertical lift of Y_a. This is discussed in Lewis and Murray [27]. Further, one may easily verify that a vector field X is a section of a distribution D if and only if $\text{verlift}(X)$ is tangent to $\mathsf{D} \subset TQ$. Also, Lewis [24] shows that condition (ii) is equivalent to the assertion that Y be geodesically invariant, by which we mean that geodesics $\gamma\colon I \to Q$ satisfying $\gamma'(t_0) \in \mathsf{Y}_{\gamma(t_0)}$ for some $t_0 \in I$ satisfy $\gamma'(t) \in \mathsf{Y}_{\gamma(t)}$ for all $t \in I$. Clearly, geodesic invariance of Y is equivalent to Y being an invariant submanifold for Z.

First suppose that Σ_{dyn} is maximally reducible to a driftless system Σ_{kin}. Let $\gamma\colon [0,T] \to Q$ be a geodesic so that $(\gamma', 0)$ is a controlled trajectory for Σ_{dyn}. If we ask that $\gamma'(0) \in \mathsf{X}$, then Definition 5.6 implies that there exists $u_{\text{kin}} \in \mathscr{U}_{\text{kin}}$ such that (γ, u_{kin}) is a controlled trajectory of Σ_{kin}. Indeed, u_{kin} is defined by

$$\gamma'(t) = \sum_{\alpha=1}^{\tilde{m}} u^\alpha_{\text{kin}}(t) X_\alpha(\gamma(t))$$

and so is smooth. Further, this implies that X is geodesically invariant. The remainder of this part of the proof will be directed towards showing that $\mathsf{X} = \mathsf{Y}$.

Let $\boldsymbol{e}_a$ be the ath standard basis vector for $\mathbb{R}^m$ and let $u_a\colon [0,T] \to \mathbb{R}^m$ be the control defined by $u_a(t) = \boldsymbol{e}_a$. Let $\Upsilon\colon [0,T] \to TQ$ be an integral curve for the

vector field $Z + \text{verlift}(Y_a)$, so that (Υ, u_a) satisfies (5.3). By Definition 5.6, Υ must be tangent to X. Since X is geodesically invariant, Z is tangent to X, and therefore $\text{verlift}(Y_a)$ must be tangent to X. This implies that $\mathsf{Y} \subset \mathsf{X}$.

To show that $\mathsf{X} \subset \mathsf{Y}$ we employ the following lemma.

LEMMA 5.8. *If a distribution D is geodesically invariant for an affine connection ∇, then for each $q \in Q$ and each $X \in \mathsf{D}_q$ there exist $T > 0$ and a smooth curve $\gamma\colon [0,T] \to Q$ with the following properties:*

(i) $\gamma'(t) \in \mathsf{D}_{\gamma(t)}$ *for* $t \in (0,T]$;

(ii) $\nabla_{\gamma'(0)}\gamma'(0) = X$.

Proof. Let (U, χ) be a normal coordinate chart [22, Proposition 8.4] with $\chi(q) = \mathbf{0}$. In such a chart the Christoffel symbols for ∇ satisfy $\Gamma^i_{jk}(\mathbf{0}) + \Gamma^i_{kj}(\mathbf{0}) = 0$, $i, j, k \in \{1, \dots, n\}$. Let $\tilde{T} > 0$ be small if necessary and let $\tilde{\gamma}\colon [0, \tilde{T}] \to Q$ be the geodesic satisfying $\tilde{\gamma}'(0) = X$. Let us denote the local representative of $\tilde{\gamma}$ in our normal coordinate chart by $t \mapsto (\tilde{q}^1(t), \dots, \tilde{q}^n(t))$. We must then have $\ddot{\tilde{q}}^i(0) = 0$, $i \in \{1, \dots, n\}$, since $\tilde{\gamma}$ is a geodesic and we are using normal coordinates. Since D is geodesically invariant, $\tilde{\gamma}'(t) \in \mathsf{D}_{\tilde{\gamma}(t)}$ for $t \in (0, \tilde{T}]$. Now define $\tau\colon [0, \tilde{T}] \to [0, \frac{1}{2}\tilde{T}^2]$ by $\tau(t) = \frac{1}{2}t^2$. Let $T = \frac{1}{2}\tilde{T}^2$, define $\gamma\colon [0,T] \to Q$ by $\gamma = \tilde{\gamma} \circ \tau$, and denote by $t \mapsto (q^1(t), \dots, q^n(t))$ the local representative of γ. Then we have

$$\dot{q}^i(t) = \frac{2t\dot{\tilde{q}}^i(t)}{T}, \qquad i \in \{1, \dots, n\},$$
$$\ddot{q}^i(0) = \dot{\tilde{q}}^i(0), \qquad i \in \{1, \dots, n\}.$$

Since $\tilde{\gamma}'(0) = X$ the result follows, and the proof of the lemma is complete. ∎

Now let $q \in Q$ and $X \in \mathsf{X}_q$. Choose a curve $\gamma\colon [0,T] \to Q$ as in the lemma. Define a smooth map $u_{\text{kin}}\colon [0,T] \to \mathbb{R}^{\tilde{m}}$ by asking that it satisfy

$$\gamma'(t) = \sum_{\alpha=1}^{\tilde{m}} u^\alpha_{\text{kin}}(t) X_\alpha(\gamma(t)).$$

Then (γ, u_{kin}) is a controlled trajectory for Σ_{kin}. Therefore, by Definition 5.6, there exists a map $u_{\text{dyn}}\colon [0,T] \to \mathbb{R}^m$ such that $(\gamma', u_{\text{dyn}})$ is a controlled trajectory for $(TQ, \mathscr{X}_{\Sigma_{\text{dyn}}}, \mathbb{R}^m)$. Indeed, since γ' is smooth, u_{dyn} will also be smooth. Furthermore, we have

$$X = \nabla_{\gamma'(0)}\gamma'(0) = \sum_{a=1}^{m} u^a_{\text{dyn}}(0) Y_a(\gamma(0)).$$

This shows that $\mathsf{X} \subset \mathsf{Y}$, which completes the proof of the "only if" part of the theorem.

Now suppose that parts (i) and (ii) of the theorem hold. Let us work locally, so we may as well assume that the vector fields $\{Y_1, \dots, Y_m\}$ and $\{X_1, \dots, X_{\tilde{m}}\}$ are linearly independent (and so $\tilde{m} = m$). First, part (ii) implies that Y is an invariant submanifold for the system $(TQ, \mathscr{X}_{\Sigma_{\text{dyn}}}, \mathbb{R}^m)$, since $\text{verlift}(Y_a)$, $a \in \{1, \dots, m\}$, is tangent to Y. If $(\Upsilon, u_{\text{dyn}})$ is a controlled trajectory of $(TQ, \mathscr{X}_{\Sigma_{\text{dyn}}}, \mathbb{R}^m)$, then $\Upsilon\colon [0,T] \to TQ$ is absolutely continuous, and so $\gamma \triangleq \tau_Q \circ \Upsilon$ is also absolutely continuous. In fact, $\Upsilon = \gamma'$, and so not only is γ absolutely continuous but γ' is also absolutely continuous. If we further suppose that $\gamma'(0) \in \mathsf{Y}_{\gamma(0)}$, then $\gamma'(t) \in \mathsf{Y}_{\gamma(t)}$ for $t \in [0,T]$. We may then define $u_{\text{kin}}\colon [0,T] \to \mathbb{R}^{\tilde{m}}$ by $\gamma'(t) = u^\alpha_{\text{kin}}(t) X_\alpha(\gamma(t))$, which uniquely defines u_{kin} since $(TQ, \mathscr{X}_{\Sigma_{\text{dyn}}}, \mathbb{R}^m)$ leaves Y, and hence X, invariant. It is clear that u_{kin} is absolutely continuous.

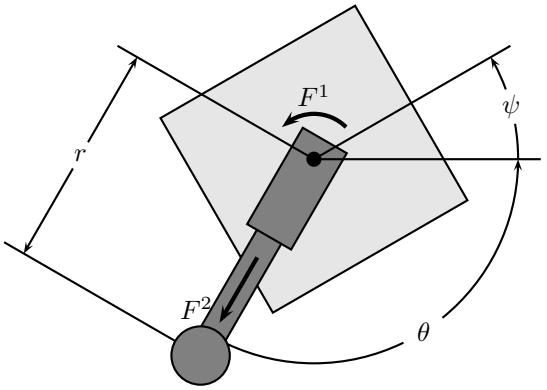


FIG. 5.3. *The robotic leg.*

Finally, let (γ, u_{kin}) be a controlled trajectory for Σ_{kin}. Thus γ' is absolutely continuous. Since Y, and therefore X, are geodesically invariant, $\nabla_{\gamma'(t)}\gamma'(t) \in \mathsf{Y}_{\gamma(t)}$ for $t \in [0,T]$. Thus we may write

$$\nabla_{\gamma'(t)}\gamma'(t) = \sum_{a=1}^{m} u^a_{\text{dyn}}(t) Y_a(\gamma(t)),$$

which defines $u_{\text{dyn}} \colon [0,T] \to \mathbb{R}^m$. It is clear that u is locally integrable, and this completes the proof. □

Remark 9. Note that all driftless systems to which a given affine connection control system $\Sigma_{\text{dyn}} = (Q, \nabla, \mathsf{D}, \mathscr{Y} = \{Y_1, \dots, Y_m\}, \mathbb{R}^m)$ is maximally reducible are essentially the same, by which we mean that for two such driftless systems, $\Sigma_{\text{kin}} = (Q, \mathscr{X} = \{X_1, \dots, X_m\}, \mathbb{R}^m)$ and $\tilde{\Sigma}_{\text{kin}} = (Q, \tilde{\mathscr{X}} = \{\tilde{X}_1, \dots, \tilde{X}_{\tilde{m}}\}, \mathbb{R}^{\tilde{m}})$, we have $\mathsf{X} = \tilde{\mathsf{X}}$. Thus, without loss of generality, we may take $(Q, \{Y_1, \dots, Y_m\}, \mathbb{R}^m)$ as the system to which Σ_{dyn} is maximally reducible. For this reason, it makes sense to simply say that Σ_{dyn} is *maximally reducible* if it is maximally reducible to *some* driftless system.

Let us give an example of a system that is maximally reducible.

Example 4. We consider the robotic leg system depicted in Figure 5.3. The configuration space for the system is $Q = \mathbb{R}_+ \times \mathbb{S}^1 \times \mathbb{S}^1$, and the coordinates we use are (r, θ, ψ), as indicated in Figure 5.3. The Riemannian metric for the system is

$$g = m(\mathrm{d}r \otimes \mathrm{d}r + r^2 \mathrm{d}\theta \otimes \mathrm{d}\theta) + J\mathrm{d}\psi \otimes \mathrm{d}\psi,$$

where m is the mass of the particle on the end of the extensible massless leg, and J is the moment of inertia of the base rigid body about the pivot point. The nonzero Christoffel symbols for the associated affine connection are $\Gamma^r_{\theta\theta} = -r$ and $\Gamma^\theta_{r\theta} = \Gamma^\theta_{\theta r} = \frac{1}{r}$. Lewis and Murray [27] show that if we define Y_1 and Y_2 by

$$Y_1 = \frac{1}{mr^2}\frac{\partial}{\partial\theta} - \frac{1}{J}\frac{\partial}{\partial\psi}, \quad Y_2 = \frac{1}{m}\frac{\partial}{\partial r},$$

then the equations of motion for the system are of the form (2.1), where ∇ is the Levi-Civita connection associated with g. There are no constraints on the system, so we take $\mathsf{D} = TQ$.

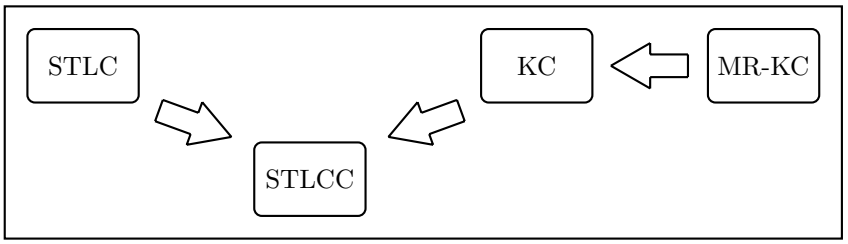


FIG. 5.4. *Relationships between various forms of controllability for affine connection control systems.*

One readily computes

$$\langle Y_1 : Y_1 \rangle = -\frac{2}{m^2 r^3} \frac{\partial}{\partial r}, \quad \langle Y_1 : Y_2 \rangle = 0, \qquad \langle Y_2 : Y_2 \rangle = 0.$$

This shows that Y is geodesically invariant. Thus the corresponding affine connection control system Σ_{dyn} is maximally reducible to $(Q, \{Y_1, Y_2\}, \mathbb{R}^2)$.

Since $\mathtt{Sym}^{(\infty)}(\mathsf{Y}) = \mathsf{Y}$ for an affine connection control system that is maximally reducible to a driftless system, by Remark 8(2) such an affine connection control system, if analytic, is STLCC from $q \in Q$ if and only if $\mathtt{Lie}^{(\infty)}(\mathsf{Y})_q = T_q Q$. Thus we make the following definition.

DEFINITION 5.9. *A maximally reducible affine connection control system* $\Sigma_{\text{dyn}} = (Q, \nabla, \mathsf{D}, \mathscr{Y}, \mathbb{R}^m)$ *is* maximally reducibly kinematically controllable from $q_0 \in Q$ (MR-KC from q_0) *if* $(Q, \mathscr{Y})$ *is properly STLC from* q_0.

5.3. Relationships to controllability. The appearance in Theorem 5.2 of the vector-valued quadratic form B_{Y} raises questions about how the notion of kinematic reductions is related to the low-order controllability results of section 4. In this section we describe the proper relationships. In [12] counterexamples are provided to show that one cannot generally improve on the relationships presented here.

Let $\Sigma_{\text{dyn}} = (Q, \nabla, \mathsf{D}, \mathscr{Y}, \mathbb{R}^m)$ be an affine connection control system. First let us list the various types of controllability we have at hand for Σ_{dyn} from a point $q_0 \in Q$:

1. small-time local controllability (STLC);
2. small-time local configuration controllability (STLCC);
3. kinematic controllability (KC);
4. maximal reducible kinematic controllability (MR-KC).

The relationships between these concepts are demonstrated in Figure 5.4. Let us show that these implications do indeed hold.

PROPOSITION 5.10. *For an analytic affine connection control system* $\Sigma_{\text{dyn}} = (Q, \nabla, \mathsf{D}, \mathscr{Y}, \mathbb{R}^m)$ *and for* $q_0 \in Q$, *the implications of Figure* 5.4 *hold.*

Proof. The implications STLC $\Longrightarrow$ STLCC and MR-KC $\Longrightarrow$ KC follow directly from the definitions of the various notions of controllability involved. Thus we need only show that KC $\Longrightarrow$ STLCC. We let

$$\Sigma_{\text{kin},1} = (Q, \mathscr{X}_1, \mathbb{R}^{m_1}), \ldots, \Sigma_{\text{kin},k} = (Q, \mathscr{X}_k, \mathbb{R}^{m_k})$$

be a collection of kinematic reductions for which $\mathtt{Lie}^{(\infty)}(\mathsf{X}_1 + \cdots + \mathsf{X}_k)_{q_0} = T_{q_0} Q$, where $\mathsf{X}_1 + \cdots + \mathsf{X}_k$ denotes the fiberwise sum of the distributions $\mathsf{X}_1, \ldots, \mathsf{X}_k$. Let $\mathscr{X} = \mathscr{X}_1 \cup \cdots \cup \mathscr{X}_k$. Note that since $\mathsf{X}_i \subset \mathsf{Y}$, Σ_{dyn} is STLCC from q_0 if $(Q, \nabla, \mathsf{D}, \mathscr{X})$

is properly STLCC from q_0. Select vector fields $X_{a_1}, \dots, X_{a_\ell}$ from the family $\mathscr{X}$ such that $\{X_{a_1}(q_0), \dots, X_{a_\ell}(q_0)\}$ is a basis for X_{q_0}. For brevity, let us denote by $B \in \Sigma_2(\mathsf{Y}_{q_0}; T_{q_0}Q/\mathsf{Y}_{q_0})$ the vector-valued quadratic form $B_{\mathsf{Y}}(q_0)$. By Theorem 5.2 we know that $Q_B|\mathsf{X}_{i,q_0} = 0$, $i \in \{1, \dots, k\}$. It therefore follows that for each $\lambda \in \operatorname{ann}(\mathsf{Y}_{q_0})$, $\lambda B(X_{a_j}(q_0), X_{a_j}(q_0)) = 0$, $j \in \{1, \dots, \ell\}$. From Lemma 3.3 this means that λB is essentially indefinite, and since this holds for every $\lambda \in \operatorname{ann}(\mathsf{Y}_{q_0})$, B is itself essentially indefinite. Therefore, by Theorem 4.4, $(Q, \nabla, \mathsf{D}, \mathscr{X})$ is properly STLCC if $\mathsf{Lie}^{(\infty)}(\mathsf{X})_{q_0} = T_{q_0}Q$. The result now follows directly. □

Remark 10. Note that all implications in Figure 5.4 are local. There are implications for global notions of controllability that follow from the local notions, but we do not consider this in a systematic way, as the understanding of global controllability of affine connection control systems is, as yet, poor.

REFERENCES

[1] R. Abraham, J. E. Marsden, and T. S. Ratiu, *Manifolds, Tensor Analysis, and Applications*, 2nd ed., Appl. Math. Sci. 75, Springer-Verlag, New York, Heidelberg, Berlin, 1988.

[2] A. A. Agrachev, *A necessary condition for second order optimality in the general nonlinear case*, Mat. Sb. (N.S.), 102 (1977), pp. 551–568 (in Russian); translation in Math. USSR-Sb.

[3] A. A. Agrachev, *Quadratic mappings in geometric control theory*, J. Soviet Math., 51 (1990), pp. 2667–2734.

[4] A. A. Agrachev and R. V. Gamkrelidze, *Local controllability and semigroups of diffeomorphisms*, Acta Appl. Math., 32 (1993), pp. 1–57.

[5] H. Arai, K. Tanie, and N. Shiroma, *Nonholonomic control of a three-DOF planar underactuated manipulator*, IEEE Trans. Robotics Automat., 14 (1998), pp. 681–695.

[6] J. Basto-Gonçalves, *Second-order conditions for local controllability*, Systems Control Lett., 35 (1998), pp. 287–290.

[7] R. M. Bianchini and G. Stefani, *Controllability along a trajectory: A variational approach*, SIAM J. Control Optim., 31 (1993), pp. 900–927.

[8] A. M. Bloch, *Nonholonomic Mechanics and Control*, Interdiscip. Appl. Math. 24, Springer-Verlag, New York, Heidelberg, Berlin, 2003.

[9] R. W. Brockett, *Hybrid models for motion control systems*, in Essays in Control: Perspectives in the Theory and Its Applications, Progr. Systems Control Theory 14, Birkhäuser Boston, Boston, 1993, pp. 29–53.

[10] F. Bullo, J. Cortés, A. D. Lewis, and S. Martínez, *Vector-valued quadratic forms in control theory*, in Sixty Open Problems in Mathematical Systems and Control Theory, V. Blondel and A. Megretski, eds., Princeton University Press, Princeton, NJ, 2004, pp. 315–320.

[11] F. Bullo and A. D. Lewis, *Geometric Control of Mechanical Systems: Modeling, Analysis, and Design for Simple Mechanical Systems*, Texts Appl. Math. 49, Springer-Verlag, New York, Heidelberg, Berlin, 2004.

[12] F. Bullo, A. D. Lewis, and K. M. Lynch, *Controllable kinematic reductions for mechanical systems: Concepts, computational tools, and examples*, in Proceedings of the Fifteenth International Symposium on Mathematical Theory of Networks and Systems, South Bend, IN, 2002.

[13] F. Bullo and K. M. Lynch, *Kinematic controllability and decoupled trajectory planning for underactuated mechanical systems*, IEEE Trans. Robotics Automat., 17 (2001), pp. 402–412.

[14] K.-T. Chen, *Integration of paths, geometric invariants and a generalized Baker-Hausdorff formula*, Ann. of Math. (2), 65 (1957), pp. 163–178.

[15] P. Choudhury and K. M. Lynch, *Trajectory planning for kinematically controllable underactuated mechanical systems*, in Workshop on Algorithmic Foundations of Robotics, Nice, France, 2002.

[16] J. Cortés, S. Martínez, and F. Bullo, *Motion planning and control problems for underactuated robots*, in Control Problems in Robotics, A. Bicchi, H. Christensen, and D. Prattichizzo, eds., Springer Tracts Adv. Robotics 4, Springer-Verlag, New York, Heidelberg, Berlin, 2002, pp. 59–74.

[17] M. Fliess, *Fonctionnelles causales non linéaires et indéterminées non commutatives*, Bull. Soc. Math. France, 109 (1981), pp. 3–40.

[18] E. Frazzoli, *Robust Hybrid Control for Autonomous Vehicle Motion Planning*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 2001.
[19] H. Hermes, *On local and global controllability*, SIAM J. Control, 12 (1974), pp. 252–261.
[20] H. Hermes, *On local controllability*, SIAM J. Control Optim., 20 (1982), pp. 211–220.
[21] R. M. Hirschorn and A. D. Lewis, *Geometric Local Controllability: Second-order Conditions*, preprint, 2002; available online at http://penelope.mast.queensu.ca/~andrew/.
[22] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry, Volume* I, Interscience Tracts Pure Appl. Math. 15, Interscience, New York, 1963.
[23] A. D. Lewis, *Local configuration controllability for a class of mechanical systems with a single input*, in Proceedings of the European Control Conference, Brussels, Belgium, 1997.
[24] A. D. Lewis, *Affine connections and distributions with applications to nonholonomic mechanics*, Rep. Math. Phys., 42 (1998), pp. 135–164.
[25] A. D. Lewis, *When is a mechanical control system kinematic?*, in Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, AZ, 1999, pp. 1162–1167.
[26] A. D. Lewis, *Simple mechanical control systems with constraints*, IEEE Trans. Automat. Control, 45 (2000), pp. 1420–1436.
[27] A. D. Lewis and R. M. Murray, *Configuration controllability of simple mechanical control systems*, SIAM J. Control Optim., 35 (1997), pp. 766–790.
[28] A. D. Lewis and R. M. Murray, *Decompositions of control systems on manifolds with an affine connection*, Systems Control Lett., 31 (1997), pp. 199–205.
[29] K. M. Lynch, N. Shiroma, H. Arai, and K. Tanie, *Collision-free trajectory planning for a 3-DOF robot with a passive joint*, Internat. J. Robotics Res., 19 (2000), pp. 1171–1184.
[30] V. Manikonda, P. S. Krishnaprasad, and J. Hendler, *Languages, behaviors, hybrid architectures, and motion control*, in Mathematical Control Theory, J. Baillieul and J. C. Willems, eds., Springer-Verlag, New York, Heidelberg, Berlin, 1998, pp. 199–226.
[31] K. McIsaac and J. P. Ostrowski, *Steering algorithms for dynamic robotic locomotion systems*, in Algorithmic and Computational Robotics: New Directions, B. R. Donald, K. M. Lynch, and D. Rus, eds., A. K. Peters, Natick, MA, 2001, pp. 221–231.
[32] G. J. Pappas, G. Lafferriere, and S. S. Sastry, *Hierarchically consistent control systems*, IEEE Trans. Automat. Control, 45 (2000), pp. 1144–1160.
[33] J. Shen, A. K. Sanyal, and N. H. McClamroch, *Controllability analysis of a two degree of freedom nonlinear attitude control system*, in Proceedings of the Fifteenth International Symposium on Mathematical Theory of Networks and Systems, South Bend, IN, 2002.
[34] H. J. Sussmann, *A sufficient condition for local controllability*, SIAM J. Control Optim., 16 (1978), pp. 790–802.
[35] H. J. Sussmann, *Lie brackets and local controllability: A sufficient condition for scalar-input systems*, SIAM J. Control Optim., 21 (1983), pp. 686–713.
[36] H. J. Sussmann, *A general theorem on local controllability*, SIAM J. Control Optim., 25 (1987), pp. 158–194.
[37] H. J. Sussmann and V. Jurdjevic, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.
[38] R. G. Swan, *Vector bundles and projective modules*, Trans. Amer. Math. Soc., 105 (1962), pp. 264–277.
[39] D. R. Tyner and A. D. Lewis, *Controllability properties of affine connection control systems*, in preparation.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 3, pp. 909–938

A SMALL-GAIN THEORY FOR LIMIT CYCLES OF SYSTEMS ON LURÉ FORM*

ULF T. JÖNSSON† AND ALEXANDRE MEGRETSKI‡

Abstract. Local exponential stability and local robustness for limit cycle solutions of ordinary differential equations can be verified using the characteristic multipliers. These well-known results are here generalized to a class of infinite-dimensional systems. Stability and robustness are now verified using certain invertibility conditions on the linear equations that are obtained when the system is linearized along the limit cycle. The new criterion reduces to the classical condition on the characteristic multipliers when we consider a finite-dimensional system which is perturbed by a bounded but possibly infinite-dimensional operator. The computation of a robustness margin, i.e., a bound on the maximally allowed perturbation, is also considered.

Key words. limit cycles, uncertain system, robustness

AMS subject classifications. 93D09, 49N20, 37C27

DOI. 10.1137/S0363012903437575

1. Introduction.

Autonomous oscillations appear frequently in physical systems [4]. Such oscillations may appear naturally as in population dynamics and the motion of the planets or through intentional engineering design such as in electronic and mechanical oscillators. The system models used to generate such periodic solutions are often based on finite-dimensional ordinary differential equations (ODE). The robustness of these mathematical models is an important topic of investigation. For example, if a population model predicts an oscillatory solution, will there remain a nearby oscillation if unmodeled species are taken into account? Similarly, will an electronic oscillator function in the presence of stray capacitances and other unmodeled dynamics?

There is rich literature treating stability and robustness of periodic solutions of autonomous ODE. Stability and perturbation results were obtained early in [11, 2] and are discussed in many books on ODE theory and periodic systems; see, e.g., [3, 6, 4, 17]. These stability and robustness criteria are stated as a condition on the characteristic multipliers corresponding to the variational system which is obtained when the system is linearized around the nominal periodic solution. By using an extension of the implicit function theorem it is possible to determine bounds on the allowed perturbation [10, 4]. An extension of the stability results to infinite-dimensional systems has been obtained in [12]; see also [4]. Here we consider the above questions for systems consisting of a feedback interconnection of a linear time invariant (LTI) transfer function with a nonlinear function. No assumption is made on the dimension of the transfer function. This class of systems appear frequently in control applications under the name of the Luré system. Our main results are stated as invertibility conditions for certain linear operators corresponding to the variational system that

*Received by the editors November 11, 2003; accepted for publication (in revised form) January 17, 2005; published electronically September 15, 2005. The work was supported by the Swedish Research Council for Engineering Sciences, the NSF, and the AFOSR.
http://www.siam.org/journals/sicon/44-3/43757.html
†Optimization and Systems Theory, Royal Institute of Technology, 10044 Stockholm, Sweden (ulfj@math.kth.se).
‡Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139 (ameg@mit.edu).

appear when the system is linearized along the periodic trajectory. This allows us to obtain robustness results that can be verified using techniques from robust control.

There are to our knowledge only a few results on robustness of oscillations in autonomous systems. A recent contribution by Georgiou and Smith considers robustness of a one-dimensional relaxation oscillator consisting of a relay hysteresis and an integrator [5]. By using the gap metric topology and suitably chosen function spaces they can prove that there remains an oscillatory solution (not necessarily periodic) as long as the perturbation of the integrator is sufficiently small in the gap topology. Varigonda extended the result to a new case in [14]. Yakubovich gave sector conditions for oscillatory solutions of a class of nonlinear systems in [15]. In [13], Stokes consider functional differential equations perturbed by an operator that vanish asymptotically as time tends to infinity.

The results obtained in this paper are different from the above in both the assumptions on the system and the obtained results. We consider a class of continuously differentiable systems and provide conditions which guarantee existence of an exponentially stable periodic solution for a general class of perturbations.

2. Problem formulation. We consider the following class of infinite-dimensional systems:

$$y(t) = \int_{-\infty}^{t} h(t-\tau,\theta)\varphi(y(\tau),\theta)\,d\tau, \tag{2.1}$$

where the impulse response function $h(t,\theta)$ and the nonlinearity $\varphi(y,\theta)$ are C^1 with respect to both arguments. We will state the exact assumptions on these functions below, but for now it is enough to think of the system as a feedback interconnection of an exponentially stable LTI plant and a memoryless nonlinearity. The system is called *nominal* when $\theta = 0$ and we assume the nominal system has a nontrivial isolated T_0-periodic solution $y_0(t) = y_0(t+T_0)\ \forall t$. Such solutions are called *limit cycles*. The case when $h(t,\theta)$ is finite-dimensional when $\theta = 0$ and infinite-dimensional for $\theta \neq 0$ is particularly interesting in applications because system design and system modeling are often done based on finite-dimensional approximations. The theory developed in this paper allows the systems analyst to rigorously verify that a modeled or designed limit cycle will also appear in the true infinite-dimensional system. The parameter θ should be viewed as a scaling of a class of infinite-dimensional perturbations. A typical case is illustrated in Figure 2.1, where S_0 denotes a finite-dimensional system with a periodic solution, Δ is a perturbation described by some norm bound, and θ scales the perturbation. In the next section we make some further connections to the standard models of robust control.

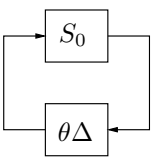


FIG. 2.1. *S_0 is a low-dimensional nominal system with limit cycle solution and $\theta\Delta$ is an infinite-dimensional but bounded perturbation. We derive conditions on S_0 that ensure existence of a stable limit cycle when θ is bounded away from zero.*

We also note that the classical model

$$\dot{x}(t) = f(x(t),\theta), \tag{2.2}$$

where f is C^1, a special case of system (2.1). Indeed, (2.2) can be written as $\dot{x} = Ax + \varphi(x,\theta)$, where A is stable and $\varphi(x,\theta) = f(x,\theta) - Ax$. Hence, the system can be represented as in (2.1) with $y = x$ and $h(t,\theta) = e^{At}\nu(t)$, where $\nu(t)$ is the unit step function. We assume that when $\theta = 0$, then (2.2) has a nontrivial T_0-periodic solution x_0. Conditions for existence and exponential stability of limit cycles are in the classical literature derived using the linearization of (2.2) along the nominal periodic solution

$$\dot{v} = f'_x(x_0(t),0)v.$$

This system cannot be asymptotically stable since $v(t) = f(x_0(t),0)$ is a nontrivial solution. This means that the monodromic matrix defined as $\Phi(T_0,0)$, where

$$\frac{d}{dt}\Phi(t,0) = f'_x(x_0(t),0)\Phi(t,0), \quad \Phi(0,0) = I,$$

has at least one eigenvalue at 1. The eigenvalues of $\Phi(T_0,0)$ are called the *characteristic multipliers* of $f'_x(x_0(t),0)$. We have thus seen that one characteristic multiplier must be 1. It can be proven that if $\lambda = 1$ is a simple characteristic multiplier of $f'_x(x_0(t),0)$, then for all sufficiently small θ there exists a periodic solution x_θ with period T_θ, which both are C^1 functions with respect to θ. Moreover, if all other characteristic multipliers are strictly inside the unit disc, then x_θ is locally exponentially stable. Proofs and exact formulations of these results can be found in [4, 3].

In this paper we extend the classical results to systems of the form (2.1). The development bears some similarities with the classical finite-dimensional theory. There will be, just as in the finite-dimensional case, a neutrally stable mode in the linearization of the dynamics (2.1) along the periodic solution. The classical results were derived using the implicit function theorem in a suitably chosen coordinate system, where the coordinate corresponding to the neutrally stable mode of the linearized dynamics can be removed. Here we use similar ideas in an operator setting. To prove exponential stability we use the concept *stability defect*, which allows us to move the neutrally stable mode into the unstable region, and then the implicit function theorem can be used. To prove structural robustness we use a version of the implicit function theorem that only require right invertibility of the linearized dynamics and thus overcome the problem with the neutrally stable mode.

2.1. Notation and assumptions. For a large part of this paper we consider a version of system (2.1) where the period time is normalized to 1. It is then natural to consider as solution space the set of continuous 1-periodic functions equipped with the norm $\|v\|_{C(1)} = \sup_{t\in[0,1]}|v(t)|$, which here is denoted by $C(1)$. For computational reasons we will state many of our results in terms of operators defined on $\mathbf{L}_2(1)$, the space of locally square integrable 1-periodic functions with the norm $\|v\|^2_{\mathbf{L}_2(1)} = \int_0^1 |v(t)|^2\,dt$. All our main results can, due to this choice of function space, be verified using methods from linear quadratic optimization.

The exponentially weighted $\mathbf{L}_2$ space ($\alpha > 0$)

$$\mathbf{L}_{2\alpha}[0,\infty) = \left\{ e(t) \in \mathbf{L}_2[0,\infty) : \int_0^\infty e^{2\alpha t}|e(t)|^2\,dt < \infty \right\} \tag{2.3}$$

will be used to define and prove exponential stability. The norm on the usual $\mathbf{L}_2[0,\infty)$ space is denoted as $\|\cdot\|$ while the norm on $\mathbf{L}_{2\alpha}[0,\infty)$ is denoted and defined as $\|v\|_\alpha = (\int_0^\infty e^{2\alpha t}|v(t)|^2\,dt)^{1/2}$. The spatial norm will always be the Euclidean norm

$|v| = (\sum_{i=1}^n v_i^2)^{1/2}$. At several places we consider the space $C(1) \times \mathbf{R}$ with the norm $\|(v,T)\|_{C(1)\times\mathbf{R}} = (\|v\|_{C(1)}^2 + |T|^2)^{1/2}$. Similarly, $\mathbf{L}_2(1) \times \mathbf{R}$ is equipped with the norm $\|(v,T)\|_{\mathbf{L}_2(1)\times\mathbf{R}} = (\|v\|_{\mathbf{L}_2(1)}^2 + |T|^2)^{1/2}$.

We also use that the characteristic multipliers (the Floquet multiplier) of a periodic matrix $A(t) = A(t+T_0)$ are the eigenvalues of the monodromy matrix $\Phi(T_0, 0)$, where

$$\frac{d}{dt}\Phi(t,0) = A(t)\Phi(t,0), \quad \Phi(0,0) = I.$$

The impulse response function in (2.1) is assumed to be a strictly proper exponentially stable system with the decay rate α.

DEFINITION 2.1 (strictly proper exponentially stable system (SPES)). *The impulse response function $h : \boldsymbol{R}^+ \to \boldsymbol{R}^p$ is exponentially stable if there exists $\epsilon > 0$ such that $e^{\epsilon t}h(t) \in \boldsymbol{L}_1[0,\infty)$. It is exponentially stable with the decay rate α if $e^{\alpha t}h(t) \in \boldsymbol{L}_1[0,\infty)$. We further say that h is a strictly proper exponentially stable system if additionally the differential of h has the form*

$$dh(t) = \dot{h}_c(t)\,dt + \sum_{k=0}^{\infty} h_k \delta(t - t_k)\,dt,$$

where $\delta(\cdot)$ denotes the Dirac impulse, $0 = t_0 < t_1 < t_2 \cdots$, and

$$e^{\epsilon t}\dot{h}_c \in \boldsymbol{L}_1[0,\infty), \quad \sum_{k=0}^{\infty} e^{\epsilon t_k}|h_k| < \infty$$

for some $\epsilon > 0$. If h is SPES, then the system output $y(t) = \int_{-\infty}^t h(t-\tau)v(\tau)\,d\tau$ belongs to $C(1)$ for $v \in C(1)$ and is differentiable with

$$\begin{aligned}\dot{y}(t) &= h(0)v(t) + \int_{-\infty}^t dh(t-\tau)v(\tau) \\ &= h(0)v(t) + \int_{-\infty}^t \dot{h}_c(t-\tau)v(\tau)\,d\tau + \sum_{k=0}^{\infty} h_k v(t-t_k),\end{aligned}$$

which also belongs to $C(1)$. As the norm on the convolution operators defined by h and dh we use

$$\begin{aligned}\|h\|_1 &= \int_0^\infty |h(t)|\,dt, \\ \|dh\|_1 &= \int_0^\infty |\dot{h}_c(t)| + \sum_{k=0}^{\infty} |h_k|.\end{aligned} \tag{2.4}$$

If h is SPES with the decay rate α, then the Laplace transforms $H(s)$ and $sH(s)$ are (i) *analytic in* $\operatorname{Re} s > -\alpha$, (ii) *continuous on* $-\alpha + i\boldsymbol{R}$, *and* (iii) *bounded such that for* $\operatorname{Re} s \geq -\alpha$ *we have* $\max(|sH(s,\theta)|, |H(s,\theta)|) \leq b$ *for some number b.*

REMARK 1. *The norms in* (2.4) *provide bounds on the induced norm of the convolution operators defined by h and dh in all applications of the paper. Sometimes we use the spaces $\boldsymbol{L}_2[0,\infty)$ and $\boldsymbol{L}_2(1)$ and then better estimates on the induced norm can be obtained for systems involving convolution with h.*

The results of the following lemma will be used in the paper.

LEMMA 2.2. *Suppose h is a strictly proper exponentially stable system. Then*
(a) $\|tdh(t)\|_1 < \infty$,
(b) $h(t)$ *is a bounded function which converges to zero with exponential rate, i.e.,* $|h(t)| \le ce^{-\epsilon t}$ *for some* $c, \epsilon > 0$,
(c) *if the decay rate is* α, *then* $e^{\alpha t}h(t) \in \boldsymbol{L}_2[0, \infty)$.

We let $\mathcal{L}(V_1, V_2)$ denote the vector space of bounded linear operators that map a normed vector space V_1 into another normed vector space V_2. The induced norm of $F \in \mathcal{L}(V_1, V_2)$ is denoted by $\|F\|_{V_1 \to V_2}$ unless $V_1 = V_2 = \mathbf{L}_2$ in which case we use the simplified notation $\|F\|$.

Next follows some terminology from nonlinear functional analysis; see, for example, [1] for further reference. Let V_1, V_2, V_3 be normed vector spaces and let $U_1 \subset V_1$, $U_2 \subset V_2$ be open subsets. A nonlinear operator $F : U_1 \to V_3$ is said to be continuously differentiable (C^1) if there exists a continuous operator $F' : U_1 \to \mathcal{L}(V_1, V_3)$ such that

$$\lim_{u \to u_0} \frac{\|F(u) - F(u_0) - F'(u_0)(u - u_0)\|_{V_3}}{\|u - u_0\|_{V_1}} = 0 \tag{2.5}$$

for each $u_0 \in U_1$. The derivative F' is called the Fréchet derivative. If F is a C^1 function of two variables, i.e., $F : U_1 \times U_2 \to V_3$, then the partial derivatives are denoted by $F'_{u_i} : U_1 \times U_2 \to \mathcal{L}(V_i, V_3)$ for $i = 1, 2$.

The kernel and the image of a linear operator $L \in \mathcal{L}(V, V)$ are defined as $\operatorname{Ker} L = \{v \in V : Lv = 0\}$ and $\operatorname{Im} L = \{Lv : v \in V\}$. The codimension of $\operatorname{Im} L$ is the dimension of the quotient space $V/\operatorname{Im} L$. An operator $L \in \mathcal{L}(V, V)$ is called a Fredholm operator if $\operatorname{Ker} L$ and the codimension of $\operatorname{Im} L$ both are finite-dimensional.

2.2. Summary of problem formulation. We consider system (2.1) under the following assumption.

Assumption 1. For system (2.1) we assume that
(i) the impulse response function $h(t, \theta)$ and the nonlinearity $\varphi(y, \theta)$ are defined for all θ on an open interval I_θ, which contains 0;
(ii) φ is continuously differentiable with respect to both arguments;
(iii) h is continuously differentiable with respect to θ and SPES with the decay rate α for every $\theta \in I_\theta$;
(iv) there exists a T_0-periodic solution y_0 of (2.1) for the case when $\theta = 0$.

In order to define exponential stability of systems of the form (2.1) we consider a system without perturbation. An absolutely continuous function $y_0(t)$ is called a T-periodic solution of the system equation if $y_0(t) = y_0(t + T)$ $\forall t$ and

$$y_0(t) = \int_{-\infty}^{t} h(t - \tau)\varphi(y_0(\tau))\, d\tau \quad \forall t. \tag{2.6}$$

To introduce the notion of local exponential stability of a given T-periodic solution $y_0(t)$, we consider the *non-steady-state* version of (2.6), defined as

$$y(t) = f(t) + \int_{0}^{t} h(t - \tau)\varphi(y(\tau))\, d\tau, \quad t \ge 0. \tag{2.7}$$

In (2.7), $f(\cdot)$ represents initial conditions and external disturbances. The choice

$$f_0(t) = \int_{-\infty}^{0} h(t - \tau)\varphi(y_0(\tau))\, d\tau \tag{2.8}$$

gives the T-periodic solution $y_0(t)$, since (2.7) has a unique solution for any locally integrable function $f(\cdot)$.

By exponential stability of the solution y_0 we will mean that for all f *close* to f_0 the solution y of (2.7) will *converge exponentially* to y_0.

DEFINITION 2.3. *The T-periodic solution y_0 is said to be* locally exponentially stable *if there exist $\alpha > 0$, $\delta > 0$, and $c > 0$ such that for any f satisfying the condition*

$$|f(t) - f_0(t)| \le \delta \quad \forall t \ge 0; \tag{2.9}$$

the corresponding solution y of (2.7) *satisfies the inequality*

$$\int_0^\infty e^{2\alpha t}|y(t) - y_0(t+d)|^2\,dt + |d|^2 \le c\int_0^\infty e^{2\alpha t}|f(t) - f_0(t)|^2\,dt \tag{2.10}$$

for some $d \in \boldsymbol{R}$.

The presence of the *phase shift* parameter d in (2.10) is necessary. It can be shown that with d fixed at $d = 0$, no nonequilibrium solution y_0 of (2.6) satisfies (2.10).

The following problem is considered in the paper.

PROBLEM 1. *Given Assumption* 1*, derive a sufficient condition for the existence of an open interval $\mathcal{I}_\theta \subset I_\theta$ around the origin with the property that for each $\theta \in \mathcal{I}_\theta$, there exists a unique exponentially stable periodic solution $y(t,\theta)$ of* (2.1) *with period $T(\theta)$ and with the property that $y(t,\theta)$ and $T(\theta)$ are C^1 in θ with $y(t,0) = y_0(t)$ and $T(0) = T_0$.*

REMARK 2. *We often use the notation $y_\theta(t) := y(t,\theta)$ and $T_\theta := T(\theta)$ for brevity.*

We will also derive a numerical procedure to verify a bound on θ for which there exists an exponentially stable periodic solution of (2.1) with orbit and period time within a given prespecified tolerance of the nominal solution. To do this we will normalize the nominal period time by the transformation $t/T_0 \to t$, which gives the nominal dynamics

$$y_0(t) = \int_{-\infty}^t T_0 h(T_0(t-\tau),0)\varphi(y_0(\tau),0)\,d\tau.$$

Hence, by redefining $T_0 h(T_0 t, 0) \to h(t,0)$ we can assume $T_0 = 1$. A general T-periodic solution of (2.1) can thus be written as

$$y(t) = \int_{-\infty}^t Th(T(t-\tau),\theta)\varphi(y(\tau),\theta)\,d\tau, \tag{2.11}$$

where y is the trajectory with period normalized to 1 and T is the period time. For the normalized dynamics in (2.11) we use an equivalent formulation of Assumption 1.

Assumption 2. For system (2.11) we assume (i)–(iii) in Assumption 1 together with

(iv′) there exists a 1-periodic solution y of (2.11) for the case when $\theta = 0$ and $T = 1$.

An advantage with the model class (2.11) is that we separate the orbit from the period time so the problem is to determine the existence of a pair (y_θ, T_θ) corresponding to each θ.

PROBLEM 2 (robustness margin). *Assume $(y_0, 1) \in C(1) \times \boldsymbol{R}$ is a nominal solution to* (2.11) *and let $\mathcal{Z} = \{(y,T) \in C(1) \times \boldsymbol{R} : \|y - y_0\|^2_{C(1)} + |T-1|^2 \le r_0^2\}$. A* robustness margin *is a bound $\overline{\theta} > 0$ such that for each $|\theta| \le \overline{\theta}$, there exists a unique exponentially stable solution $(y_\theta, T_\theta) \in \mathcal{Z}$ to* (2.11) *(here we assume $[-\overline{\theta}, \overline{\theta}] \subset I_\theta$).*

3. Main results. Our first result provides a solution to Problem 1 for the case when the nominal system is finite-dimensional.

THEOREM 3.1. *Suppose Assumption* 1 *holds and consider the system in* (2.1) *when* $h(t,0) = Ce^{At}B\nu(t)$, *where* $\nu(\cdot)$ *is the unit step function and* $\operatorname{Re}\lambda(\mathrm{A}) < -\alpha$. *If the characteristic multipliers of* $A_{cl}(t) = A + B\varphi_y'(y_0(t),0)C$ *can be sorted as*

$$1 = \lambda_1 > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n| \quad \textit{and} \quad \alpha < -\frac{\log|\lambda_2|}{T_0},$$

then for all sufficiently small θ *there exists a unique (modulo time translation) exponentially stable limit cycle solution* $y(t,\theta)$ *to* (2.1) *with period* $T(\theta)$ *and with the property that* $y(t,\theta)$ *and* $T(\theta)$ *are* C^1 *in* θ *with* $y(t,0) = y_0(t)$ *and* $T(0) = T_0$. *Moreover, the exponential decay rate in* (2.10) *can be chosen to be* α.

The proof of Theorem 3.1 and most other results in this section are collected in the appendix. The proof builds on results presented in sections 3.1–3.4.

We consider two examples. The first shows that the classical finite-dimensional result is completely recovered by Theorem 3.1 and the second treats an uncertainty model from robust control.

EXAMPLE 1. *System* (2.2) *can equivalently be written in the form* (2.1) *with* $y = x$, $\varphi(y,\theta) = f(y,\theta) - Ay$, *and* $h(t,\theta) = e^{At}\nu(t)$, *where* $\nu(\cdot)$ *is the unit step function and* A *is any Hurwitz matrix. Hence, since* $A_{cl}(t) = A + B\varphi_y'(y_0(t),0)C = f_x'(y_0(t),0)$, *it follows that the condition on the characteristic multipliers in Theorem* 3.1 *is the same as the classical criterion discussed in* [3, 4]. *Theorem* 3.1 *hence gives an extended interpretation of the well-known finite-dimensional perturbation result.*

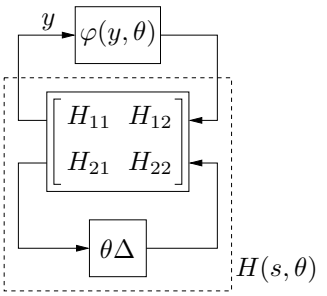


FIG. 3.1. *Block diagram corresponding to the perturbed system in* (2.1).

EXAMPLE 2. *In order to understand how our model class in* (2.1) *relates to standard uncertain system models from robust control we consider the block diagram in Figure* 3.1. *The transfer function*

$$H(s) = \begin{bmatrix} H_{11}(s) & H_{12}(s) \\ H_{21}(s) & H_{22}(s) \end{bmatrix} \in \boldsymbol{RH}_\infty$$

is assumed to be exponentially stable with the decay rate α, *i.e., the poles belong to the half space* $\operatorname{Re} s < -\alpha$. *This system corresponds to* (2.1) *with*[1] $h(t,\theta) = \mathcal{L}^{-1}(H(s,\theta))$, *where*

$$H(s,\theta) = H_{11}(s) + \theta H_{12}(s)\Delta(s)(I - \theta H_{22}(s)\Delta(s))^{-1}H_{21}(s). \tag{3.1}$$

[1]Here $\mathcal{L}$ denotes the Laplace transform.

The following result is a corollary to Theorem 3.1.

COROLLARY 3.2. *Assume*

(i) $H(s,\theta)$ *in* (3.1) *is SPES with the decay rate* α *for* $\theta \in I_\theta$,
(ii) φ *is* C^1 *with respect to both arguments,*
(iii) *for* $\theta = 0$ *the system has a nontrivial* T_0*-periodic solution* y_0,
(iv) $H_{11}(s) = C_1(sI - A_1)^{-1}B_1$, *where* $\mathrm{Re}\lambda(A_1) < -\alpha$ *and the characteristic multipliers of* $A_{cl}(t) = A_1 + B_1\varphi_y'(y_0(t),0)C_1$ *can be sorted as*

$$1 = \lambda_1 > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n| \quad \textit{and} \quad \alpha < -\frac{\log|\lambda_2|}{T_0}.$$

Then there exists a unique (modulo time translation) exponentially stable limit cycle solution for all sufficiently small θ. *Moreover, the exponential decay rate in* (2.10) *can be chosen to be* α *and the same continuity property as in Theorem* 3.1 *holds.*

We will next present a solution of Problem 2. At the same time we will touch upon some of the ideas behind the proof of Theorem 3.1. We will discuss the existence of solution and exponential stability separately and then combine them into our main small-gain theorem in section 3.3. We finally consider the case when the nominal system is finite-dimensional in section 3.4. This is where we derive the condition on the characteristic multipliers that is used to prove Theorem 3.1.

3.1. Existence of solution. To prove the existence of a periodic solution to the perturbed system we consider the system equation in the normalized time (2.11). This equation involves well-defined function spaces and the existence of solution is proven using an implicit function theorem that only requires right invertibility of the partial derivative with respect to the trajectory. To obtain a robustness bound we need to estimate the region in which the right inverse exists and for this we use Lemma 3.3. The robustness result in Lemma 3.4 is proven using Lemma 5.2 in the appendix, which is our basic result on local existence of solution.

First introduce the Banach spaces

$$X_y = C(1), \quad X_T = \mathbf{R}, \quad X_z = X_y \times X_T,$$

the open set $X_\theta = I_\theta$, and the operator $F : X_z \times X_\theta \to X_y$, which for each $z = (y,T)$ is defined as

$$F(z,\theta)(t) = y(t) - \int_{-\infty}^{t} Th(T(t-\tau),\theta)\varphi(y(\tau),\theta)\,d\tau. \tag{3.2}$$

By Assumption 2, we have $F(z_0,0) = 0$ for $z_0 = (y_0,1)$. To solve Problem 2 we need to show that for all $|\theta| \leq \overline{\theta}$ there exists a pair $z_\theta = (y_\theta, T_\theta) \in \mathcal{Z}$ such that

$$F(z_\theta,\theta) = 0. \tag{3.3}$$

This equation simply states that z_θ is a solution of (2.11). An important part of the proof of Theorem 3.1 is to use an implicit function theorem stated in the appendix. It shows that if $F_z'(z_0,0)$ has a bounded right inverse, then there exists a solution z_θ to (3.3) for all θ in some neighborhood of $\theta = 0$. To estimate the size of this neighborhood and to ensure that $z_\theta \in \mathcal{Z}$ we need to explore more structure of the operator equation (3.3). We proceed formally and differentiate the equality in (3.3), which gives

$$\frac{dz}{d\theta} = -F_z'(z(\theta),\theta)^\dagger F_\theta'(z(\theta),\theta), \tag{3.4}$$

where $F_z'(z(\theta),\theta)^\dagger$ denotes the right inverse. To evaluate the right inverse for an arbitrary solution (z_θ,θ) we use the following lemma.

LEMMA 3.3. *Let V_1, V_2 be Banach spaces and assume $H_0 \in \mathcal{L}(V_1,V_2)$ has a right inverse $G_0 \in \mathcal{L}(V_2,V_1)$. Then for each $\Delta \in \mathcal{L}(V_1,V_2)$ with $\|\Delta\|_{V_1\to V_2} < 1/\|G_0\|_{V_2\to V_1}$, there exists $G \in \mathcal{L}(V_2,V_1)$ such that $(H_0+\Delta)G = I$. One possible choice for this right inverse is*

$$G = G_0(I+\Delta G_0)^{-1}.$$

Proof. Since $H_0G_0 = I$, it follows that $(H_0+\Delta)G_0(I+\Delta G_0)^{-1} = I$. The right inverse $G = G_0(I+\Delta G_0)^{-1}$ is bounded with

$$\|G\|_{V_2\to V_1} \le \frac{\|G_0\|_{V_2\to V_1}}{1-\|\Delta\|_{V_1\to V_2}\cdot\|G_0\|_{V_2\to V_1}}. \qquad \square$$

Hence, if we let $\bar\Delta(z,\theta) = F_z'(z,\theta) - F_z'(z_0,0)$, then (3.4) becomes

$$\frac{dz}{d\theta} = -F_z'(z_0,0)^\dagger(I-\bar\Delta(z(\theta),\theta)F_z'(z_0,0)^\dagger)^{-1}F_\theta'(z(\theta),\theta) \tag{3.5}$$

which is well defined when

$$\sup_{z\in\mathcal{Z},|\theta|\le\bar\theta}\|\bar\Delta(z,\theta)\|_{X_z\to X_y}\cdot\|F_z'(z_0,0)^\dagger\|_{X_y\to X_z} < 1.$$

This small-gain condition will generally give rise to conservative estimates. It is our experience that better estimates can be obtained by extending the operators to the corresponding $\mathbf{L}_2$-spaces on which norm bounds can be computed efficiently:

$$\begin{array}{lll} \widetilde F_\theta' : \mathcal{Z}\times I_\theta \to \mathcal{L}(\tilde X_\theta,\tilde X_y), & & \tilde X_\theta = \mathbf{R}, \\ \widetilde F_z' : \mathcal{Z}\times I_\theta \to \mathcal{L}(\tilde X_z,\tilde X_y), & \text{where} & \tilde X_y = \mathbf{L}_2(1), \quad \tilde X_T = \mathbf{R}, \\ \widetilde\Delta : \mathcal{Z}\times I_\theta \to \mathcal{L}(\tilde X_z,\tilde X_y), & & \tilde X_z = \tilde X_y \times \tilde X_T. \end{array} \tag{3.6}$$

It is easy to show that we have the block representations

$$\begin{aligned} \widetilde F_z'(z,\theta) &= \begin{bmatrix}\widetilde F_y'(z,\theta) & \widetilde F_T'(z,\theta)\end{bmatrix} = \begin{bmatrix} I - L^s(z,\theta) & y_1(z,\theta)\end{bmatrix}, \\ \widetilde\Delta(z,\theta) &= \begin{bmatrix}\widetilde\Delta_1(z,\theta) & \widetilde\Delta_2(z,\theta)\end{bmatrix} = \begin{bmatrix} L^s(z_0,0) - L^s(z,\theta) & y_1(z,\theta) - y_1(z_0,0)\end{bmatrix}, \end{aligned} \tag{3.7}$$

where

$$\begin{aligned} (L^s(z,\theta)v)(t) &= \int_{-\infty}^t Th(T(t-\tau),\theta)\varphi_y'(y(\tau),\theta)v(\tau)\,d\tau, \\ (y_1(z,\theta))(t) &= \int_{-\infty}^t h(T(t-\tau),\theta)\varphi(y(\tau),\theta)\,d\tau \\ &\quad + \int_{-\infty}^t T(t-\tau)\,dh(T(t-\tau),\theta)\varphi(y(\tau),\theta). \end{aligned} \tag{3.8}$$

The operator $\bar\Delta : [\bar\Delta_1 \quad \bar\Delta_2] : \mathcal{Z}\times I_\theta \to \mathcal{L}(X_y,X_z)$ is defined in exactly the same way as $\widetilde\Delta$ but on a different function space. We have the following lemma.

LEMMA 3.4. *Suppose Assumption 2 holds and let* $\widetilde{F}'_z(z,\theta)$ *and* $\widetilde{\Delta}(z,\theta)$ *be defined as in* (3.6)–(3.8). *If*

(i) $\widetilde{F}'_z(z_0,0)$ *has a bounded right inverse denoted* $\widetilde{F}'_z(z_0,0)^\dagger$,

(ii) $\sup_{z\in\mathcal{Z},|\theta|\le\overline{\theta}} \|\widetilde{\Delta}(z,\theta)\|_{\tilde{X}_z\to\tilde{X}_y} \cdot \|\widetilde{F}'_z(z_0,0)^\dagger\|_{\tilde{X}_y\to\tilde{X}_z} < 1$,

(iii) $\sup_{z\in\mathcal{Z},|\theta|\le\overline{\theta}} \|\widetilde{F}'_z(z_0,0)^\dagger(I-\widetilde{\Delta}(z,\theta)\widetilde{F}'_z(z_0,0)^\dagger)^{-1}\widetilde{F}_\theta(z,\theta)\|_{\tilde{X}_\theta\to X_z} < \frac{r_0}{\overline{\theta}}$,

then there exists a unique (modulo time translation) solution $z_\theta = (y_\theta, T_\theta) \in \mathcal{Z}$ *to* (2.11) *for all* $|\theta| \le \overline{\theta}$.

3.2. Exponential stability. To prove exponential stability we consider a linearization of the nonsteady state dynamics in (2.7). The linearized model does not have a stable inverse due to a neutrally stable mode which is identified in Proposition 3.5. In order to prove convergence we remove the effect of the neutrally stable mode by compensating with a term corresponding to the phase delay. This leads us to define the notion stability defect in Definition 3.6. The main stability result in Theorem 3.7 is proven in the appendix.

We derive a sufficient condition for exponential stability by using the standard implicit function theorem. In order to obtain an appropriate topology on which Frechét derivatives can be defined we introduce the vector space $V = \{v : v \in \mathbf{L}_{2\alpha}[0,\infty) \cap \mathbf{L}_\infty[0,\infty)\}$ with the norm $\|v\|_V = \max(\|v\|_\alpha, \|v\|_\infty)$. The first norm is used to define the exponential decay while it is the second that allow us to compute the derivative. We let $\mathcal{X}_y = V$, $\mathcal{X}_d = \mathbf{R}$, and $\mathcal{X}_z = \mathcal{X}_y \times \mathcal{X}_d$ and consider the operator $\Psi : \mathcal{X}_z \times \mathcal{X}_y \to \mathcal{X}_y$, which for each pair $z = (\delta y, d) \in \mathcal{X}_z$ and $\delta f \in \mathcal{X}_y$ is defined as

$$(\Psi((\delta y,d),\delta f))(t) = y_0(t+d) + \delta y(t) - \int_0^t h(t-\tau)\varphi(y_0(\tau+d)+\delta y(\tau))\,d\tau - f_0(t) - \delta f(t), \tag{3.9}$$

where y_0 is a T-periodic solution of (2.7) and f_0 is defined in (2.8). Note that the equation $\Psi(z,\delta f) = 0$ is equivalent to (2.7) with $f(t) = f_0(t) + \delta f(t)$ and $y(t) = y_0(t+d)+\delta y(t)$. In particular, we know from (2.6) that $\Psi(0,0) = 0$. We need to show that for any δf in some neighborhood of 0 there exists a unique pair $z = (\delta y, d) \in \mathcal{X}_z$ such that $\Psi((\delta y,d),\delta f) = 0$ and such that the bound (2.10) holds. The first condition follows from the implicit function theorem if we can prove that the Frechét derivative with respect to the first argument, $\Psi'_z(0,0)$, has a bounded inverse. To prove that each such triplet $(\delta y, d, \delta f)$ satisfies the exponential decay condition in (2.10) we use the extension of the derivative to $\mathbf{L}_{2\alpha}$-space. With $\widetilde{\mathcal{X}}_y = \mathbf{L}_{2\alpha}[0,\infty)$, $\widetilde{\mathcal{X}}_d = \mathbf{R}$, and $\widetilde{\mathcal{X}}_z = \widetilde{\mathcal{X}}_y \times \widetilde{\mathcal{X}}_d$ the extension $\widetilde{\Psi}'_z(0,0) \in \mathcal{L}(\widetilde{\mathcal{X}}_z, \widetilde{\mathcal{X}}_y)$ is defined as

$$\widetilde{\Psi}'_z(0,0) = \begin{bmatrix}\widetilde{\Psi}'_y(0,0) & \widetilde{\Psi}'_d(0,0)\end{bmatrix} = \begin{bmatrix} I - L & e\end{bmatrix}, \tag{3.10}$$

where $L \in \mathcal{L}(\mathbf{L}_{2\alpha}, \mathbf{L}_{2\alpha})$ and $e \in \mathcal{L}(\mathbf{R}, \mathbf{L}_{2\alpha})$ are defined as

$$(Lv)(t) = \int_0^t h(t-\tau)\varphi'(y_0(\tau))v(\tau)\,d\tau \tag{3.11}$$

and

$$e(t) = \int_{-\infty}^0 h(t-\tau)\varphi'(y_0(\tau))\dot{y}_0(\tau)\,d\tau. \tag{3.12}$$

In the proof of Theorem 3.7 we shall show that (3.10) is the correct derivative by using Proposition 3.5 and that $e \in \mathbf{L}_{2\alpha}[0,\infty)$. The proof also shows that if $\widetilde{\Psi}'_z(0,0)$ has a bounded inverse, then the exponential decay condition in (2.10) holds and, moreover, that the original operator $\Psi'_z(0,0)$ has a bounded inverse (which proves the existence of $(\delta y, d)$).

The operator $I - L$ does not have a bounded inverse which follows from the next lemma.

PROPOSITION 3.5. *If $y_0 \neq$ const is a T-periodic solution of* (2.6) *and L and e are defined as in* (3.11)–(3.12)*, then $\dot{y}_0 \in (I-L)^{-1}e$.*

Proof. Let us differentiate $y_0(t)$ in (2.6). This gives

$$\begin{aligned}\dot{y}_0(t) &= \frac{d}{dt}\int_{-\infty}^{t} h(t-\tau)\varphi(y_0(\tau))\,d\tau \\ &= h(0)\varphi(y_0(t)) + \int_{-\infty}^{t} dh(t-\tau)\varphi(y_0(\tau)) \\ &= h(0)\varphi(y_0(t)) + \lim_{T\to-\infty}[-h(t-\tau)\varphi(y_0(\tau))]_T^t + \int_{-\infty}^{t} h(t-\tau)\varphi'(y_0(\tau))\dot{y}_0(\tau)\,d\tau \\ &= e(t) + \int_0^t h(t-\tau)\varphi'(y_0(\tau))\dot{y}_0(\tau)\,d\tau,\end{aligned}$$

where we used that $\lim_{T\to-\infty} h(t-T)\varphi(y_0(T)) = 0$ since h is exponentially stable and continuous, see Lemma 2.2. ∎

From this result it follows that $e \notin \operatorname{Im}(I-L)$ and one expects that (3.10) has a bounded inverse if $(I-L)$ has codimension 1. To make this precise we introduce the notion of α-defect of the operator L.

DEFINITION 3.6. *Let $L \in \mathcal{L}(\mathcal{X},\mathcal{X})$ be a bounded operator on a Banach space $\mathcal{X}$. Suppose that $I-L$ is a Fredholm operator with* $\operatorname{Ker}(I-L) = 0$*. Then the* stability defect *$def(L)$ is defined as the codimension of the subspace*

$$\mathcal{X}_L = \{(I-L)u : u \in \mathcal{X}\} \subset \mathcal{X}.$$

For $L \in \mathcal{L}(\tilde{X}_y, \tilde{X}_y)$ in (3.11) *the stability defect is called* α-defect*, denoted by $def_\alpha(L)$, due to the underlying space $\widetilde{\mathcal{X}}_y = \boldsymbol{L}_{2\alpha}[0,\infty)$.*

THEOREM 3.7. *The T-periodic solution y_0 of* (2.6) *is exponentially stable if $def_\alpha(L) = 1$ for L defined in* (3.11).

It will be convenient in computations to work with operators defined on $\mathbf{L}_2$ instead of $\mathbf{L}_{2\alpha}$. If $e^{\alpha t}h(t) \in \mathbf{L}_1[0,\infty)$, then the next lemma shows that the α-defect can be computed on $\mathbf{L}_2$ by using the operator $L_\alpha : \mathbf{L}_2[0,\infty) \to \mathbf{L}_2[0,\infty)$ defined by

$$(L_\alpha v)(t) = \int_0^t h_\alpha(t-\tau)\varphi'(y_0(\tau))v(\tau)\,d\tau, \tag{3.13}$$

where $h_\alpha(t) = e^{\alpha t}h(t)$.

LEMMA 3.8. *With $L : \boldsymbol{L}_{2\alpha} \to \boldsymbol{L}_{2\alpha}$ and $L_\alpha : \boldsymbol{L}_2 \to \boldsymbol{L}_2$ defined in* (3.11) *and* (3.13)*, respectively, we have $def_\alpha(L) = def(L_\alpha)$.*

Proof. The proof is easy and is given in [8]. ∎

3.3. A small-gain theorem. We will next combine the results in sections 3.1 and 3.2 to obtain a solution of Problem 2. First we derive a condition for exponential stability of all solutions $z_\theta = (y_\theta, T_\theta) \in \mathcal{Z}$ of (2.11) by proving that all possible

linearizations have α-defect 1. This is done in Lemma 3.9 which is a zero exclusion principle that is valid because the stability defect is robust to small perturbations.

Since we will do all computations in $\mathbf{L}_2$-space this leads us to consider the operators $L_\alpha : \mathcal{Z} \times I_\theta \to \mathcal{L}(\mathbf{L}_2, \mathbf{L}_2)$ defined by

$$(L_\alpha(z,\theta)v)(t) = \int_0^t Th_\alpha(T(t-\tau),\theta)\varphi_y'(y(\tau),\theta)v(\tau)\,d\tau, \tag{3.14}$$

where $h_\alpha(t,\theta) = e^{\alpha t}h(t,\theta)$. The nominal operator $L_\alpha^0 = L_\alpha(z_0,0)$ (where $z_0 = (y_0,1)$) is in our applications easy to work with. If $\mathrm{def}_\alpha(L) = 1$, then $\mathrm{Im}\,(I-L)$ is a closed subspace and it follows by the Banach inverse theorem that there exists $c > 0$ such that

$$\|(I - L_\alpha^0)v\| \geq c\|v\| \qquad \forall v \in \mathbf{L}_2. \tag{3.15}$$

We will use this bound in the next lemma, which shows that the α-defect is robust to perturbations of the system.

LEMMA 3.9. *Suppose Assumption* 2 *holds. Let* $L_\alpha : \mathcal{Z} \times I_\theta \to \mathcal{L}(\boldsymbol{L}_2, \boldsymbol{L}_2)$ *be defined as in* (3.14) *and let* $L_\alpha^0 = L_\alpha(z_0,0)$ *and* $\Delta L_\alpha(z,\theta) = L_\alpha(z,\theta) - L_\alpha^0$. *If*

(i) $def(L_\alpha^0) = 1$,

(ii) *there exists* $0 < \varepsilon < c$ *such that*

$$\sup_{z\in\mathcal{Z},|\theta|\leq\bar\theta} \|\Delta L_\alpha(z,\theta)\| \leq c - \varepsilon,$$

where $c > 0$ *satisfies* (3.15),

then $def(L_\alpha(z,\theta)) = 1$ *for all* $z \in \mathcal{Z}$ *and* $|\theta| \leq \bar\theta$. *In particular, every* 1*-periodic solution* $(y,T) \in \mathcal{Z}$ *of* (2.11) *is exponentially stable.*

The next theorem, which solves Problem 2, follows from Lemmas 3.4 and 3.9.

THEOREM 3.10. *Suppose Assumption* 2 *holds and let* $\widetilde{F}_z'(z,\theta)$ *and* $\widetilde{\Delta}(z,\theta)$ *be defined as in* (3.6)–(3.8). *Further let* $L_\alpha(z,\theta)$ *be defined as in* (3.14) *and define* $L_\alpha^0 = L_\alpha(z_0,0)$ *and* $\Delta L_\alpha(z,\theta) = L_\alpha(z,\theta) - L_\alpha^0$. *If*

(i) $\widetilde{F}_z'(z_0,0)$ *has a bounded right inverse denoted by* $\widetilde{F}_z'(z_0,0)^\dagger$,

(ii) $\sup_{z\in\mathcal{Z},|\theta|\leq\bar\theta} \|\widetilde{\Delta}(z,\theta)\|_{\tilde X_z\to\tilde X_y} \cdot \|\widetilde{F}_z'(z_0,0)^\dagger\|_{\tilde X_y\to\tilde X_z} < 1$,

(iii) $\sup_{z\in\mathcal{Z},|\theta|\leq\bar\theta} \|\widetilde{F}_z'(z_0,0)^\dagger(I - \widetilde{\Delta}(z,\theta)\widetilde{F}_z'(z_0,0)^\dagger)^{-1}\widetilde{F}_\theta(z,\theta)\|_{\tilde X_\theta\to X_z} < \frac{r_0}{\bar\theta}$,

(iv) $def(L_\alpha^0) = 1$,

(v) *there exists* $0 < \varepsilon < c$ *such that*

$$\sup_{z\in\mathcal{Z},|\theta|\leq\bar\theta} \|\Delta L_\alpha(z,\theta)\| \leq c - \varepsilon, \tag{3.16}$$

where $c > 0$ *satisfies* (3.15),

then there exists a unique (modulo time translation) exponentially stable solution $(y_\theta, T_\theta) \in \mathcal{Z}$ *to* (2.11) *for all* $\theta \in [-\bar\theta, \bar\theta]$.

3.4. Estimation of norms. We here consider the case when the nominal system in (2.11) is finite-dimensional. This means that $h(t,0) = Ce^{At}B\nu(t)$, where A is Hurwitz and $\nu(t)$ is the unit step function. In this case all conditions in Theorem 3.10 can be verified numerically. We show how right invertibility and the condition on the stability defect can be proven in the finite-dimensional case in Propositions 3.12 and 3.14, respectively. These two results are the foundation for the proof of Theorem 3.1. We also discuss how to estimate some of the relevant norms in Theorem 3.10.

3.4.1. Right invertibility of $\widetilde{F}'_{z}(z_0, 0)$. We first derive a state-space realization of the operator $\widetilde{F}'_z(z_0, 0) : (v, \delta T) \mapsto w$ in (3.6)–(3.8).

LEMMA 3.11. *Let $h(t, 0) = Ce^{At}B\nu(t)$. Then the Frechét derivative*

$$\widetilde{F}'_z(z_0, 0) = \begin{bmatrix} I - L^s(z_0, 0) & y_1(z_0, 0) \end{bmatrix},$$

where

$$(L^s(z_0, 0)v)(t) = \int_{-\infty}^{t} h(t - \tau, 0)\varphi'_y(y_0(\tau), 0)v(\tau)\, d\tau,$$
$$(y_1(z_0, 0))(t) = \int_{-\infty}^{t} h(t - \tau, 0)\varphi(y_0(\tau), 0)\, d\tau + \int_{-\infty}^{t} (t - \tau)\, dh(t - \tau, 0)\varphi(y_0(\tau), 0)$$

has the state space realization $\widetilde{F}'_z(z_0, 0) : (v, \delta T) \mapsto w$ defined by

$$\dot{x} = Ax + B\varphi'_y(y_0, 0)v + \dot{x}_0\delta T, \quad x(1) = x(0), \quad w = v - Cx, \tag{3.17}$$

where $\dot{x}_0(t) = Ax_0(t) + B\varphi(y_0(t), 0)$ is the nominal 1-periodic state trajectory.

Proof. It is straightforward to see that $L^s(z_0, 0) : v \mapsto w_1$ has the state space realization

$$\dot{x}_1 = Ax_1 + B\varphi'_y(y_0, 0)v, \quad x_1(1) = x_1(0), \quad w = v - Cx_1.$$

In order to obtain a state-space realization for $y_1(z_0, 0)$ we notice that

$$\begin{aligned}(y_1(z_0, 0))(t) &= C\int_{-\infty}^{t} (I + (t - \tau)A)e^{A(t-\tau)}B\varphi(y_0(\tau))\, d\tau \\ &= C\int_{-\infty}^{t} e^{A(t-\tau)}\left(\frac{d}{d\tau}\int_{-\infty}^{\tau} e^{A(\tau-s)}B\varphi(y_o(s))\, ds\right) d\tau.\end{aligned}$$

Hence $y_1(z_0, 0) : \delta T \to w_2$ has the state-space realization

$$\dot{x}_2 = Ax_2 + \dot{x}_0\delta T, \quad x_2(1) = x_2(0), \quad w_2 = Cx_2.$$

If we let $x = x_1 + x_2$, then we obtain the state-space realization in (3.17). ∎

The next proposition shows that the right invertibility condition (i) in Theorem 3.10 follows from the classical finite-dimensional condition on the characteristic multipliers of the system matrix for the linearized dynamics $A_{cl}(t) = A + B\varphi'_y(y_0(t), 0)C$. Note that one characteristic multiplier must be equal to 1 since $\frac{d}{dt}\dot{x}_0(t) = A_{cl}(t)\dot{x}_0(t)$, which implies that $\dot{x}_0(0) = \Phi_{cl}(1, 0)\dot{x}_0(0)$ due to the periodicity of $\dot{x}_0$.

PROPOSITION 3.12. *Consider the operator $\widetilde{F}'_z(z_0, 0)$ defined in (3.6)–(3.8) in the finite-dimensional case when $h(t, 0) = Ce^{At}B\nu(t)$, where A is Hurwitz. If $n-1$ of the characteristic multipliers of $A_{cl}(t) = A + B\varphi'_y(y_0(t), 0)C$ are different from 1, then $\widetilde{F}'_z(z_0, 0)$ has a bounded right inverse. One possible right inverse is $\widetilde{F}'_z(z_0, 0)^\dagger : w \mapsto (v, \delta T)$ defined by*

$$\begin{aligned}\dot{x} &= (A + B\varphi'_y(y_0, 0)C)x + B\varphi'_y(y_0, 0)w + \dot{x}_0 kx(0), \quad x(1) = x(0), \\ (v, \delta T) &= (w + Cx, kx(0)),\end{aligned} \tag{3.18}$$

where the row vector k must be chosen such that $1 \notin \operatorname{eig}(\Phi_{cl}(1,0) + \dot{x}_0(0)k)$, e.g., $k = \dot{x}_0(0)^T$. Here $\Phi_{cl}(t,0)$ is the transition matrix corresponding to A_{cl}.

Proof. The operator $\widetilde{F}'_z(z_0,0)$ can be represented by the state-space realization in (3.17). We proceed formally and construct a candidate right inverse by using $v = w + Cx$ in the first equation of (3.17) and $\delta T = kx(0)$. This gives rise to a map $w \mapsto (v, \delta T)$ defined by (3.18). In order for this to be well defined on $\mathbf{L}_2(1)$ it is necessary that the following equation have a solution for all $w \in \mathbf{L}_2(1)$:

$$x(0) = (\Phi_{cl}(1,0) + \dot{x}_0(0)k)x(0) + \int_0^1 \Phi_{cl}(1,\tau)B\varphi'_y(y_0(\tau),0)w(\tau)\,d\tau,$$

where we used that $\int_0^1 \Phi_{cl}(1,\tau)\dot{x}_0(\tau)\,d\tau kx(0) = \dot{x}_0(0)kx(0)$. Since $\operatorname{span}\{\dot{x}_0(0)\} = \operatorname{Ker}(I - \Phi_{cl}(1,0))$ it follows that there exists a vector k such that $I - \Phi_{cl}(1,0) - \dot{x}_0(0)k$ is invertible. Indeed, one possible choice is $k = \dot{x}_0(0)^T$. It is now easy to see that when the initial condition of (3.18) is the same as that of (3.17) then the composition of (3.17) with (3.18) is the identity operator. This proves the existence of a right inverse. ∎

3.4.2. Verification of condition (iii) in Theorem 3.10. To verify condition (iii) in Theorem 3.10 we exploit the structure of the operators. It follows from (3.18) in Proposition 3.12 that the nominal right inverse will have the block structure

$$\widetilde{F}'_z(z_0,0)^\dagger = \begin{bmatrix} 1 + G_1 \\ G_2 \end{bmatrix}. \tag{3.19}$$

Indeed, if we assume $k = \dot{x}_0(0)$, then the initial condition of (3.18) must be

$$x(0) = (I - \Phi_{cl}(1,0) - kk^T)^{-1} \int_0^1 \Phi_{cl}(1,\tau)B_{cl}(\tau)w(\tau)\,d\tau.$$

If we let

$$g_1(t,\tau) = \begin{cases} (\Gamma(t)\Phi_{cl}(1,t) + C)\Phi_{cl}(t,\tau)B_{cl}(\tau), & t > \tau, \\ \Gamma(t)\Phi_{cl}(1,\tau)B_{cl}(\tau), & t < \tau, \end{cases}$$
$$g_2(t,\tau) = k(I - \Phi_{cl}(1,0) - kk^T)^{-1}\Phi_{cl}(t,\tau)B_{cl}(\tau),$$

where $B_{cl}(t) = B\varphi'_y(y_0(t),0)$ and $\Gamma(t) = C(\Phi_{cl}(t,0) + \dot{x}_0(t)k)(I - \Phi_{cl}(1,0) - kk^T)^{-1}$, then we have the representation

$$(\widetilde{F}'_z(z_0,0)^\dagger w)(t) = (w(t) + (g_1 * w)(t), (g_2 * w)(1))$$

which corresponds to the block structure in (3.19). Here

$$(g_i * w)(t) = \int_0^1 g_i(t,\tau)w(\tau)\,d\tau, \quad i = 1,2. \tag{3.20}$$

We will also use the strictly proper part defined by

$$(G_{sp}w)(t) = ((g_1 * w)(t), (g_2 * w)(1)). \tag{3.21}$$

We have the following result.

PROPOSITION 3.13. *Define*[2]

$$\widetilde{\gamma}_1 = \|\widetilde{F}'_z(z_0,0)^\dagger\|_{\tilde{X}_y\to\tilde{X}_z}, \qquad \widetilde{\gamma}_2 = \|G_{sp}\|_{\tilde{X}_y\to X_z},$$
$$\widetilde{\gamma}_\Delta = \sup_{z\in\mathcal{Z},|\theta|\le\overline{\theta}} \|\widetilde{\Delta}(z,\theta)\|_{\tilde{X}_z\to\tilde{X}_y},$$
$$\gamma_{\Delta_1} = \sup_{z\in\mathcal{Z},|\theta|\le\overline{\theta}} \|\bar{\Delta}_1(z,\theta)\|_{X_y\to X_y}, \quad \gamma_{\Delta_2} = \sup_{z\in\mathcal{Z},|\theta|\le\overline{\theta}} \|\bar{\Delta}_2(z,\theta)\|_{X_T\to X_y},$$
$$\widetilde{\gamma}_\theta = \sup_{z\in\mathcal{Z},|\theta|\le\overline{\theta}} \|\widetilde{F}'_\theta(z,\theta)\|_{\tilde{X}_\theta\to\tilde{X}_y}, \quad \gamma_\theta = \sup_{z\in\mathcal{Z},|\theta|\le\overline{\theta}} \|F'_\theta(z,\theta)\|_{X_\theta\to X_y}.$$

A sufficient condition for (iii) *in Theorem* 3.10 *is*

$$\left(\frac{(\gamma_{\Delta_1}^2+\gamma_{\Delta_2}^2)^{1/2}}{1-\gamma_{\Delta_1}}+1\right)\frac{\widetilde{\gamma}_2\cdot\widetilde{\gamma}_\theta}{1-\widetilde{\gamma}_\Delta\cdot\widetilde{\gamma}_1}+\frac{\gamma_\theta}{1-\gamma_{\Delta_1}}<\frac{r_0}{\overline{\theta}}. \tag{3.22}$$

Proof. We start to derive a bound on the norm for fixed $z\in\mathcal{Z}$ and $\theta\in I_\theta$. We will use the block representations

$$\widetilde{F}'^{\,\dagger}_{z_0} := \widetilde{F}'_z(z_0,0)^\dagger = \begin{bmatrix}1+G_1\\ G_2\end{bmatrix}, \quad G_{sp}=\begin{bmatrix}G_1\\ G_2\end{bmatrix}, \quad \widetilde{\Delta}=[\widetilde{\Delta}_1 \quad \widetilde{\Delta}_2],$$

where G_1 and G_2 are defined in terms of the convolutions (3.20) and $\widetilde{\Delta}$ is defined as in (3.7). We have

$$\begin{aligned}\widetilde{F}'^{\,\dagger}_{z_0}(I+\widetilde{\Delta}\,\widetilde{F}'^{\,\dagger}_{z_0})^{-1} &= \begin{bmatrix}(I+\widetilde{\Delta}_1)^{-1}\\ 0\end{bmatrix} - (I+\widetilde{\Delta}_1)^{-1}\begin{bmatrix}\widetilde{\Delta}_1 & \widetilde{\Delta}_2\\ 0 & 0\end{bmatrix} G_{sp}(I+\widetilde{\Delta}\,\widetilde{F}'^{\,\dagger}_{z_0})^{-1}\\ &\quad + G_{sp}(I+\widetilde{\Delta}\,\widetilde{F}'^{\,\dagger}_{z_0})^{-1}\end{aligned} \tag{3.23}$$

which follows since by the identity $A^{-1}-B^{-1}=B^{-1}(B-A)A^{-1}$ we have

$$\begin{aligned}(I+\widetilde{\Delta}\,\widetilde{F}'^{\,\dagger}_{z_0})^{-1}-(I+\widetilde{\Delta}_1)^{-1} &= -(I+\widetilde{\Delta}_1)^{-1}(\widetilde{\Delta}_1G_1+\widetilde{\Delta}_2G_2)(I+\widetilde{\Delta}\,\widetilde{F}'^{\,\dagger}_{z_0})^{-1}\\ &= -(I+\widetilde{\Delta}_1)^{-1}\widetilde{\Delta}G_{sp}(I+\widetilde{\Delta}\,\widetilde{F}'^{\,\dagger}_{z_0})^{-1}.\end{aligned}$$

From (3.23) we get

$$\begin{aligned}\|\widetilde{F}'^{\,\dagger}_{z_0}(I+\widetilde{\Delta}\,\widetilde{F}'^{\,\dagger}_{z_0})^{-1}\widetilde{F}'_\theta\|_{\tilde{X}_\theta\to X_z} &\le \frac{\|F'_\theta\|_{X_\theta\to X_y}}{1-\|\bar{\Delta}_1\|_{X_y\to X_y}}\\ &\quad+\left(\frac{(\|\bar{\Delta}_1\|^2_{X_y\to X_y}+\|\bar{\Delta}_2\|^2_{X_\theta\to X_y})^{1/2}}{1-\|\bar{\Delta}_1\|_{X_y\to X_y}}+1\right)\frac{\|G_{sp}\|_{\tilde{X}_y\to X_z}\|\widetilde{F}'_\theta\|_{\tilde{X}_\theta\to\tilde{X}_y}}{1-\|\widetilde{\Delta}\|_{\tilde{X}_z\to\tilde{X}_y}\|\widetilde{F}'^{\dagger_{z_0}}\|_{\tilde{X}_y\to\tilde{X}_z}}.\end{aligned} \tag{3.24}$$

Note that some induced norms are over the original X-spaces ($C(1)$), others are over the $\tilde{X}$-spaces ($\mathbf{L}_2(1)$), and some are from $\tilde{X}$-space to X-space. If we optimize over $z\in\mathcal{Z}$ and $|\theta|\le\overline{\theta}$, we see from (3.24) that (3.22) is sufficient for (iii) in Theorem 3.10. ∎

The computation of $\widetilde{\gamma}_\Delta$-$\widetilde{\gamma}_\theta$ in Proposition 3.13 depends very much on the uncertainty structure and must be treated on a case-by-case basis. For $\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ it is possible to derive systematic algorithms for computing bounds on these parameters.

[2]Recall that the operators $\bar{\Delta}$ and $\widetilde{\Delta}$ are both defined as in (3.6)–(3.8), but on different spaces.

3.4.3. Computation of $\widetilde{\gamma}_1$. For a given choice of k we compute

$$\widetilde{\gamma}_1 = \|\widetilde{F}_z'(z_0,0)^\dagger\|_{\tilde{X}_y\to\tilde{X}_z} = \|\widetilde{F}_z'(z_0,0)^\dagger\|_{\mathbf{L}_2(1)\to\mathbf{L}_2(1)\times\mathbf{R}}$$

by solving the optimization problem

$$\widetilde{\gamma}_1^2 = \inf\gamma^2 \quad \text{subject to} \quad J(x,w,\gamma)\le 0 \quad \forall (x,w)\in\mathcal{L}, \tag{3.25}$$

where

$$J(x,w,\gamma) = |kx(0)|^2 + \int_0^1 (|w+Cx|^2 - \gamma^2|w|^2)\,dt$$
$$\mathcal{L} = \{(x,w)\in\mathbf{L}_2(1) : \dot{x}(t) = A_{cl}(t)x(t) + B_{cl}(t)w(t) + \dot{x}_0(t)kx(0),\ x(1) = x(0)\}$$

and $A_{cl}(t) = A + B\varphi_y'(y_0(t),0)C$ and $B_{cl}(t) = B\varphi_y'(y_0(t),0)$. We obtain an upper bound by using LQ optimal control techniques. If we let $Q_0 = k^Tk$, $Q = C^TC$, $S = C^T$, $R = (1-\gamma^2)I$, then γ is an upper bound on the optimization problem (3.25), i.e., $\gamma > \|\widetilde{F}_z'(z_0,0)^\dagger\|_{\tilde{X}_y\to\tilde{X}_z}$ if and only if there exists $\epsilon > 0$ such that

$$\sup_{(x,w)\in\mathcal{L}} x(0)^TQ_0x(0) + \int_0^1 (x^TQx + 2x^TSw + w^TRw)\,dt \le -\epsilon|x(0)|^2. \tag{3.26}$$

Necessary and sufficient conditions for this condition to hold can be obtained from the Pontryagin maximum principle. The full details are given in [8].

3.4.4. Computation of $\widetilde{\gamma}_2$. In the linear systems theory the $\mathbf{H}_2$-norm can be used to estimate the $\|\cdot\|_{\mathbf{L}_2\to\mathbf{L}_\infty}$ norm of a transfer function. We will here estimate the corresponding norm for the strictly proper operator G_{sp} in (3.21). By the convolution formula in (3.21) and the definition of $\tilde{X}_y$ and X_z, we get

$$\|G_{sp}w\|_{X_z} = \|G_{sp}w\|_{C(1)\times\mathbf{R}} \le (\|G_1\|^2_{\mathbf{L}_2(1)\to C(1)} + \|G_2\|^2_{\mathbf{L}_2(1)\to\mathbf{R}})^{1/2}\|w\|_{\mathbf{L}_2(1)},$$

where

$$\|G_1\|^2_{\mathbf{L}_2(1)\to C(1)} := \max_{t\in[0,1]}\int_0^1 |g_1(t,s)|^2\,ds,$$
$$\|G_2\|^2_{\mathbf{L}_2(1)\to\mathbf{R}} := \int_0^1 |g_2(1,s)|^2\,ds.$$

Hence, we get the bound

$$\widetilde{\gamma}_2 = \|G_{sp}\|_{\tilde{X}_y\to X_z} = \|G_{sp}\|_{\mathbf{L}_2(1)\to C(1)\times\mathbf{R}} \le (\|G_1\|^2_{\mathbf{L}_2(1)\to C(1)} + \|G_2\|^2_{\mathbf{L}_2(1)\to\mathbf{R}})^{1/2}. \tag{3.27}$$

3.4.5. Verification of the α-defect. We will next derive a condition for the α-defect to be 1 in the finite-dimensional case.

PROPOSITION 3.14. *Consider the operator L defined in* (3.11) *in the finite-dimensional case when $h(t) = Ce^{At}B\nu(t)$, where* $\mathrm{Re}\lambda(A) < -\alpha$. *If the characteristic multipliers of $A_{cl}(t) = A + B\varphi'(y_0(t))C$ can be sorted as*

$$1 = \lambda_1 > |\lambda_2| \ge |\lambda_3| \ge \cdots \ge |\lambda_n|,$$

then $def_\alpha(L) = 1$ *for* $\alpha \in (0, -\frac{\log|\lambda_2|}{T})$.

We next discuss how to verify conditions (iv) and (v) in Theorem 3.10. For condition (iv) we introduce $L^0 \in \mathcal{L}(\mathbf{L}_2, \mathbf{L}_2)$ defined by

$$(L^0 v)(t) = \int_0^t h(t-\tau, 0)\varphi'(y(\tau), 0)v(\tau)\, d\tau.$$

If $h(t,0) = Ce^{At}B\nu(t)$, then $\mathrm{def}_\alpha(L^0) = 1$ if the condition in Proposition 3.14 holds, which by Lemma 3.8 implies $\mathrm{def}\,(L^0_\alpha) = 1$.

To verify condition (v) in Theorem 3.10 we need to derive a bound $c > 0$ (as large as possible) such that (3.15) holds, i.e.,

$$\|(I - L^0_\alpha)v\| \geq c\|v\| \quad \forall v \in \mathbf{L}_2[0,\infty), \tag{3.28}$$

where $L^0_\alpha = L_\alpha(z_0, 0)$ is defined as in (3.14) with $h_\alpha(t,0) = Ce^{(A+\alpha I)t}B\nu(t)$. The condition (3.28) holds if there exists $\epsilon > 0$ such that

$$\int_0^\infty (x^T C^T C x - 2x^T C^T v + (1-c^2)v^T v)\, dt \geq \epsilon(\|x\|^2 + \|v\|^2) \tag{3.29}$$

for all $(x,v) \in \{(x,v) \in \mathbf{L}_2[0,\infty) : \dot{x}(t) = (A+\alpha I)x(t) + B_{cl}(t)v(t),\ x(0) = 0\}$, where $B_{cl}(t) = B\varphi'(y_0(t), 0)$. Several equivalent conditions for the last inequality to hold are given in [16].

4. Numerical example. We consider the Wien bridge oscillator in Figure 4.1. The model is adopted from [9]. If we let $x_1 = v_{C_1}$ and $x_2 = v_{C_2}$, the voltages over C_1

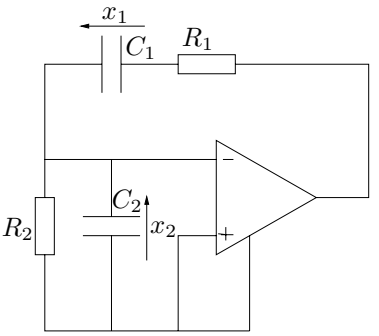


FIG. 4.1. *The Wien bridge oscillator.*

and C_2 be the state variables, then the system equation becomes

$$\dot{x}_1 = \frac{1}{C_1 R_1}(-x_1 + x_2 - \varphi(x_2)),$$
$$\dot{x}_2 = -\frac{1}{C_2 R_1}(-x_1 + x_2 - \varphi(x_2)) - \frac{1}{C_2 R_2}x_2,$$

where $\varphi(\cdot)$ is the model of the operational amplifier. With the values $C_1 = C_2 = 1$, $R_1 = R_2 = 1/6.32$, and

$$\varphi(y) = 3.234y - 2.195y^3 + 0.666y^5$$

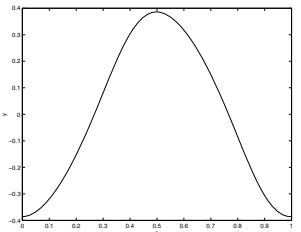


FIG. 4.2. *A period of the* 1*-periodic solution to the Wien bridge oscillator. Only the output* $y = x_2$ *is shown.*

we obtain the 1-periodic solution in Figure 4.2. The nominal system can be written as

$$y(t) = \int_{-\infty}^{t} h(t-\tau)\varphi(y(\tau))\,d\tau,$$

where $h(t) = Ce^{At}B\nu(t)$. Here $\nu(\cdot)$ is the unit step function and

$$A = \frac{1}{6.32}\begin{bmatrix} -1 & 1 \\ 1 & -2 \end{bmatrix}, \quad B = \frac{1}{6.32}\begin{bmatrix} -1 \\ 1 \end{bmatrix}, \quad C = [0 \quad 1].$$

We will investigate robustness to additive uncertainties of the form

$$H(s,\theta) = H(s) + \theta\Delta(s),$$

where $H(s) = C(sI - A)^{-1}B$ and $\Delta(s)$ is a stable transfer function. For $r_0 = 0.01$ we will show that $\bar{\theta} = 0.0007$ is a robustness margin. We let the exponential decay parameter be $\alpha = 0.7$. We assume

$$\sup_{|T-1|\le r_0} \|\Delta(s/T)\|_{C(1)\to C(1)} \le 1, \qquad \sup_{|T-1|\le r_0} \|(s/T)\Delta'(s/T)\|_{C(1)\to C(1)} \le 1,$$
$$\sup_{|T-1|\le r_0} \|\Delta((s-\alpha)/T)\|_{\mathbf{H}_\infty} \le 1, \qquad \sup_{|T-1|\le r_0} \|((s-\alpha)/T)\Delta'((s-\alpha)/T)\|_{\mathbf{H}_\infty} \le 1.$$

We start to verify conditions (i)–(iii) in Theorem 3.10. The characteristic multipliers of $A_{cl} = A + B\varphi'(y_0(t))C$ are 1 and 0.24, which by Proposition 3.12 implies that $\widetilde{F}'_z(z_0, 0)$ has a bounded right inverse. It also motivates the choice of α since $\alpha = 0.7 < -\log|\lambda_2|/T_0 = 1.43$.

We next need to compute bounds on the norms in Proposition 3.13.

- We obtain the bound $\widetilde{\gamma}_1 = \|\widetilde{F}'_z(z_0,0)^\dagger\|_{\tilde{X}_y\to\tilde{X}_z} \le 9.0$ by verifying (3.26).
- The bound $\widetilde{\gamma}_2 = \|G_{sp}\|_{\tilde{X}_y\to X_z} \le 17.0$ is obtained using (3.27).
- $\widetilde{\gamma}_\theta = \sup_{z\in\mathcal{Z},\ |\theta|\le\bar{\theta}} \|\widetilde{F}'_\theta(z,\theta)\|_{\tilde{X}_\theta\to\tilde{X}_y} \le \sup_{\|y-y_0\|_{C(1)}\le r_0} \|\varphi(y)\|_{\mathbf{L}_2(1)} \approx 0.6032.$
- $\gamma_\theta = \sup_{z\in\mathcal{Z},\ |\theta|\le\bar{\theta}} \|F'_\theta(z,\theta)\|_{X_\theta\to X_y} \le \sup_{\|y-y_0\|_{C(1)}\le r_0} \|\varphi(y)\|_{C(1)} \approx 1.072.$

The norm of the operator $\widetilde{\Delta}$ requires more work. This operator can be represented as ($z = (y, T)$)

$$\begin{aligned}\widetilde{\Delta}(z,\theta)(v,\delta T) &= \widetilde{\Delta}_1(y,T,\theta)v + \widetilde{\Delta}_2(y,T,\theta)\delta T \\ &= [H(s/T,\theta)\varphi'(y) - H(s)\varphi'(y_0)]v \\ &\quad + [(s/T)H'(s/T,\theta)\varphi(y) - sH'(s)\varphi(y_0)]\delta T.\end{aligned}$$

For each $z=(y,T)\in\mathcal{Z}$ and $|\theta|\le\bar\theta$ we use the bounds

(4.1)
$$\begin{aligned}\|\widetilde\Delta_1(z,\theta)\|_{\mathbf{L}_2(1)\to\mathbf{L}_2(1)} &\le \|H(s/T)-H(s)\|_{\mathbf{H}_\infty}\cdot\|\varphi'(y)\|_{C(1)}\\ &\quad+\|H(s)\|_{\mathbf{H}_\infty}\cdot\|\varphi'(y)-\varphi'(y_0)\|_{C(1)}+\bar\theta\|\varphi'(y)\|_{C(1)}=:N_1(z),\\ \|\widetilde\Delta_2(z,\theta)\|_{\mathbf{R}\to\mathbf{L}_2(1)} &\le \|(s/T)H'(s/T)-sH'(s)\|_{\mathbf{H}_\infty}\cdot\|\varphi(y)\|_{\mathbf{L}_2(1)}\\ &\quad+\|sH'(s)\|_{\mathbf{H}_\infty}\cdot\|\varphi(y)-\varphi(y_0)\|_{\mathbf{L}_2(1)}+\bar\theta\|\varphi(y)\|_{\mathbf{L}_2(1)}=:N_2(z).\end{aligned}$$

Hence using (4.1) we get

$$\begin{aligned}\widetilde\gamma_\Delta &= \sup_{z\in\mathcal{Z},|\theta|\le\bar\theta}\|\widetilde\Delta(z,\theta)\|_{\tilde X_z\to\tilde X_y}=\sup_{z\in\mathcal{Z},|\theta|\le\bar\theta}\|\widetilde\Delta(z,\theta)\|_{\mathbf{L}_2(1)\times\mathbf{R}\to\mathbf{L}_2(1)}\\ &\le \sup_{z\in\mathcal{Z},|\theta|\le\bar\theta}\sqrt{\|\widetilde\Delta_1(z,\theta)\|^2_{\mathbf{L}_2(1)\to\mathbf{L}_2(1)}+\|\widetilde\Delta_2(z,\theta)\|^2_{\mathbf{R}\to\mathbf{L}_2(1)}}\\ &\le \sup_{z\in\mathcal{Z}}\sqrt{N_1(z)^2+N_2(z)^2}\le 0.0194,\end{aligned}$$

where the last inequality was obtained numerically.

To obtain the remaining norms we use convolution representations of $\bar\Delta_1$ and $\bar\Delta_2$. If we let $h_{\bar\Delta(z,\theta)}(t,s)$ be the convolution kernel corresponding to the operator $H(s/T,\theta)\varphi'(y(t))-H(s,0)\varphi'(y_0(t))$ defined on $C(1)$ and $g_{\bar\Delta(z,\theta)}(t)$ correspond to the function $sH'(s,0)\varphi(y_0)-(s/T)H'(s/T,\theta)\varphi(y)\in C(1)$, then we can write ($z=(y,T)$)

$$\begin{aligned}(\bar\Delta_1(z,\theta)v)(t)&=\int_0^1 h_{\bar\Delta(z,\theta)}(t,s)v(s)\,ds,\\ (\bar\Delta_2(z,\theta)\delta T)(t)&=g_{\bar\Delta(z,\theta)}(t)\delta T.\end{aligned}$$

Note that

$$h_{\bar\Delta(z,\theta)}(t,s)=h_z(t,s)-h_{z_0}(t,s)+\theta\delta_T(t,s)\varphi'(y(s)),$$

where $\delta_T(t,s)$ is the weighting function corresponding to $\Delta(s/T)$ and

$$h_z(t,s)=\begin{cases}TC(I-e^{AT})^{-1}e^{AT(t-s)}B\varphi'(y(s)), & t>s,\\ TC(I-e^{AT})^{-1}e^{AT}e^{AT(t-s)}B\varphi'(y(s)), & t<s.\end{cases}$$

Similarly, we have

$$g_{\bar\Delta(z,\theta)}(t)=\int_0^1\left(g_z(t,s)-g_{z_0}(t,s)+\theta\hat\delta_T(t,s)\varphi(y(s))\right)ds,$$

where $\hat\delta_T(t,s)$ is the weighting function corresponding to $(s/T)\Delta(s/T)$ and

$$g_z(t)=\begin{cases}T\tilde C(I-e^{\tilde AT})^{-1}e^{\tilde AT(t-s)}\tilde B\varphi'(y(s)), & t>s,\\ T\tilde C(I-e^{\tilde AT})^{-1}e^{\tilde AT}e^{\tilde AT(t-s)}\tilde B\varphi'(y(s)), & t<s\end{cases}$$

and

$$\tilde A=\begin{bmatrix}A & I\\ 0 & A\end{bmatrix},\quad \tilde B=\begin{bmatrix}0\\ B\end{bmatrix},\quad \tilde C=-[CA\quad C].$$

We get

$$\gamma_{\Delta_1} = \sup_{z\in\mathcal{Z}} \sup_{|\theta|\le\bar\theta} \|\bar\Delta_1(z,\theta)\|_{C(1)\to C(1)} = \sup_{z\in\mathcal{Z}} \sup_{|\theta|\le\bar\theta} \|h_{\bar\Delta(z,\theta)}\|_1 \le 0.0226$$

and similarly

$$\gamma_{\Delta_2} = \sup_{z\in\mathcal{Z}} \sup_{|\theta|\le\bar\theta} \|\bar\Delta_2(z,\theta)\|_{\mathbf{R}\to C(1)} = \sup_{z\in\mathcal{Z}} \sup_{|\theta|\le\bar\theta} \|g_{\bar\Delta(z,\theta)}\|_{C(1)} \le 0.0053.$$

Hence, we have

$$\widetilde\gamma_\Delta \cdot \widetilde\gamma_1 = \sup_{z\in\mathcal{Z},|\theta|\le\bar\theta} \|\widetilde\Delta(z,\theta)\|_{\tilde X_z\to\tilde X_y} \cdot \|\widetilde F_z'(z_0,0)^\dagger\|_{\tilde X_y\to\tilde X_z} = 0.0194\cdot 9 = 0.175 < 1,$$

which verifies condition (ii). For condition (iii) we use (3.22) in Proposition 3.13. We have

$$\left(\left(\frac{(\gamma_{\Delta_1}^2+\gamma_{\Delta_2}^2)^{1/2}}{1-\gamma_{\Delta_1}}+1\right)\frac{\widetilde\gamma_2\cdot\widetilde\gamma_\theta}{1-\widetilde\gamma_\Delta\cdot\widetilde\gamma_1}+\frac{\gamma_\theta}{1-\gamma_{\Delta_1}}\right)\frac{\bar\theta}{r_0}\approx 0.9667.$$

It remains to verify the exponential stability conditions (iv)–(v) in Theorem 3.10. Condition (iv) follows from Proposition 3.14 since $\lambda_2 = 0.24$. Next we compute a bound on $c > 0$ such that (3.28) holds. A bound can be obtained by verifying (3.29) using [16], which results in the bound $c = 0.0735$. In order to verify (3.16) we use that

$$\Delta L_\alpha(z,\theta) = H((s-\alpha)/T)\circ\varphi'(y) - H(s-\alpha)\circ\varphi'(y_0) + \theta\Delta((s-\alpha)/T)\varphi'(y).$$

We get the bound

$$\begin{aligned}\sup_{z\in\mathcal{Z},\,|\theta|\le\bar\theta}\|\Delta L_\alpha(z,\theta)\| &\le \sup_{|T-1|\le r_0}\|H((s-\alpha)/T)-H(s-\alpha)\|_{\mathbf{H}_\infty}\sup_{\|y-y_0\|_{C(1)}\le r_0}\|\varphi'(y)\|_{C(1)}\\ &\quad+\|H(s-\alpha)\|_{\mathbf{H}_\infty}\sup_{\|y-y_0\|_{C(1)}\le r_0}\|\varphi'(y)-\varphi'(y_0)\|_{C(1)}+\bar\theta\sup_{\|y-y_0\|_{C(1)}\le r_0}\|\varphi'(y)\|_{C(1)}\\ &= 0.0210 \le c = 0.074,\end{aligned}$$

which proves that condition (v) in Theorem 3.10 is true. We have thus shown that $\bar\theta = 0.0007$ is a robustness margin when $r_0 = 0.01$. The question is whether this bound is conservative. We simulated the system with $\Delta(s) = \frac{10}{s+10}$. For $\hat\theta = 0.001$ we get a perturbation $\hat r = (\|y-y_0\|_{C(1)}^2 + |T-1|^2)^{1/2} \approx 0.01$. Hence, the gap between the true robustness margin and the estimated is not at larger than $\hat\theta/\bar\theta \approx 1.4$. Table 1 shows the results of several numerical experiments.

5. Concluding remarks. We have proven that a well-known condition for robustness of limit cycles of finite-dimensional systems is also valid when the system is perturbed by a sufficiently small dynamic perturbation. We also showed how bounds on a robustness margin can be estimated using a number of small-gain conditions.

The robustness conditions in this paper are to a large extent formulated as invertibility conditions on operators defined in $\mathbf{L}_2$ spaces. An existence result and a perturbation bound as in Lemma 3.4 are easier to derive in the Banach space $C(1)$. However, the resulting conditions gave much more conservative bounds for our numerical example in section 4. It is in particular the small-gain condition

TABLE 1

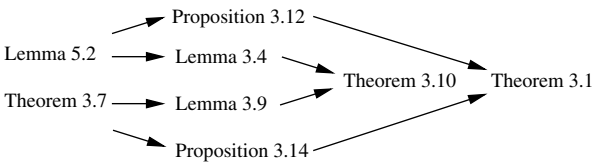
Results of numerical examples. $\overline{\theta}$ is the bound computed using our results and $\hat{\theta}$ is obtained by simulating with $\Delta(s) = 10/(s+10)$. The gap $\hat{\theta}/\overline{\theta}$ is an upper bound on how conservative our bound possibly can be.

r_0	$r_0/(\|y_0\|^2_{C(1)}+1)^{1/2}$	$\overline{\theta}$	$\frac{\overline{\theta}}{\|H\|_{\mathbf{H}_\infty}}$	$\hat{\theta}$	$\frac{\hat{\theta}}{\|H\|_{\mathbf{H}_\infty}}$	$\hat{\theta}/\overline{\theta}$
0.01	0.93%	0.0007	0.21%	0.001	0.3%	1.4
0.02	1.83%	0.0011	0.33%	0.0019	0.6%	1.7
0.03	2.8%	0.0012	0.36%	0.0026	0.8%	2.2

$\sup_{z\in\mathcal{Z},|\theta|\le\overline{\theta}} \|\widetilde{\Delta}(z,\theta)\|_{\tilde{X}_z\to\tilde{X}_y} \cdot \|\widetilde{F}'_z(z_0,0)^\dagger\|_{\tilde{X}_y\to\tilde{X}_z} < 1$ that is much less conservative and also easier to verify in $\mathbf{L}_2(1)$.

The problem of computing robust stability bounds for limit cycles is an important and challenging problem that is still in its infancy. The results provided in this paper are valid for a fairly large class of systems that appear frequently in applications. By using a minimum of structure in the problem we obtain a set of small gain conditions that can be used to estimate a robustness margin. The verification of the small gain conditions are involved and it is desirable to find conditions that can be verified with less effort.

Appendix: Proofs. We here collect several of the proofs in the paper. We start to state a version of the implicit function theorem suitable for our purposes. We then provide proofs for our main results using the following sequence of implications:



Here Lemma 5.2 gives a sufficient condition for existence of a periodic solution locally. The upper implications are results on the existence of a periodic solution and the lower implications are results on exponential stability. Next follows the implicit function theorem, which is used to prove Lemmas 5.2 and 3.4.

THEOREM 5.1. *Let V_1, V_2, V_3 be Banach spaces and suppose $F : U_1 \times U_2 \to V_3$ is C^1, where $U_1 \subset V_1$ and $U_2 \subset V_2$ are open. Assume $F(u_{10}, u_{20}) = 0$ for some $u_{10} \in U_1$ and $u_{20} \in U_2$. If $F'_{u_1}(u_{10}, u_{20})$ has a bounded right inverse, then there exists a neighborhood U_{20} of u_{20} and a C^1 function $E : U_{20} \to U_1$ such that*

$$F(E(u_2), u_2) = 0$$

for all $u_2 \in U_{20}$. To estimate the size of U_{20} assume G_0 is a bounded right inverse of $H_0 = D_{u_1}F(u_{10}, u_{20})$ and

$$\begin{aligned} B^1_{r_1}(u_{10}) &= \{u_1 : \|u_1 - u_{10}\| \le r_1\} \subset U_1, \\ B^2_{r_2}(u_{20}) &= \{u_2 : \|u_2 - u_{20}\| \le r_2\} \subset U_2, \\ B_r(u_0) &= B^1_{r_1}(u_{10}) \times B^2_{r_2}(u_{20}) \subset U_1 \times U_2 \end{aligned}$$

are such that

$$L_r = \sup_{u\in B_r(u_0)} \|F'_{u_1}(u) - F'_{u_1}(u_0)\|,$$
$$L_{r_1} = \sup_{u_1\in B^1_{r_1}(u_{10})} \|F'_{u_1}(u_1, u_{20}) - F'_{u_1}(u_{10}, u_{20})\|,$$
$$L_{r_2} = \sup_{u\in B_r(u_0)} \|F'_{u_2}(u)\|$$

satisfy

$$L_r\|G_0\| < 1, \tag{5.1}$$
$$\frac{L_{r_2}\|G_0\|}{1 - L_{r_1}\|G_0\|} r_2 \le r_1. \tag{5.2}$$

Then we can use $U_{20} = B^2_{r_2}(u_{20})$ *and* $E(u_2) \in B^1_{r_1}(u_{10})$ *for all* $u_2 \in B^2_{r_2}(u_{20})$.

Proof. We only sketch on a proof. The easiest way to prove the first part of the theorem is to use the standard implicit function theorem. Let $\hat{U}_k = U_k - u_{k0}$, $k = 1, 2$, and $\hat{F} : \hat{U}_1 \times \hat{U}_2 \to V_3$ be defined as

$$\hat{F}(\hat{u}_1, \hat{u}_2) = F(u_{10} + G_0\hat{u}_1, u_{20} + \hat{u}_2).$$

We have $\hat{F}(0, 0) = 0$ and $\hat{F}'_{\hat{u}_1}(0, 0) = I$. Hence, it follows from the standard implicit function theorem that there exists a neighborhood $\hat{U}_{20}$ of 0 and a unique C^1 function $\hat{E} : \hat{U}_{20} \to \hat{U}_1$ such that $\hat{F}(\hat{E}(\hat{u}_2), \hat{u}_2) = 0$ for all $\hat{u}_2 \in \hat{U}_{20}$; see, e.g., [1]. The first claim of the theorem follows by using $U_{20} = \hat{U}_{20} + u_{20}$ and $E(u_2) = u_{10} + G_0\hat{E}(u_2 - u_{20})$.

It remains to verify that the estimate $U_{20} = B^2_{r_2}(u_{20})$ is valid. The following sketch of this part also provides the idea behind a more constructive way of proving the theorem. We continue to use the translated variables and the ˆ notation, e.g., $\hat{B}^1_{r_1}(0) = \{\hat{u}_1 : \|\hat{u}_1\| \le r_1\} = B_{r_1}(u_{10}) - u_{10}$ but now with $\hat{F}(\hat{u}_1, \hat{u}_2) = F(u_{10} + \hat{u}_1, u_{20} + \hat{u}_2)$. The idea is to consider the fixed point iteration

$$x_{k+1} = L(x_k, \hat{u}_2) := G_0(\hat{F}'_{\hat{u}_1}(0, 0)x_k - \hat{F}(x_k, \hat{u}_2)).$$

If we can prove the existence of a fixed point $x = \hat{E}(\hat{u}_2)$, then this implies that $\hat{F}(\hat{E}(\hat{u}_2), \hat{u}_2) = 0$. We use the following inequalities.

(i) For any pairs $(x_2, \hat{u}_2) \in \hat{B}_r(0)$, $(x_1, \hat{u}_2) \in \hat{B}_r(0)$ we have

$$\|L(x_2, \hat{u}_2) - L(x_1, \hat{u}_2)\| \le L_r\|G_0\| \cdot \|x_2 - x_1\| < \|x_2 - x_1\|,$$

where the second inequality follows from (5.1).

(ii) For all $(x, \hat{u}_2) \in \hat{B}_r(0)$ we have

$$\begin{aligned}\|L(x, \hat{u}_2)\| &\le \|G_0\| \cdot \|\hat{F}'_{\hat{u}_1}(0, 0)x - \hat{F}(x, 0)\| + \|G_0\| \cdot \|\hat{F}(x, \hat{u}_2) - \hat{F}(x, 0)\| \\ &\le \|G_0\|(L_{r_1}\|x\| + L_{r_2}\|\hat{u}_2\|) \le \|G_0\|(L_{r_1}r_1 + L_{r_2}r_2) \le r_1,\end{aligned}$$

where the last inequality follows from (5.2).

These two inequalities show that $L(\cdot, \hat{u}_2)$ is a contraction on $\hat{B}^1_{r_1}(0)$ for each $\hat{u}_2 \in \hat{B}_{r_2}(0)$, which proves the existence of the function $\hat{E}(\cdot)$ on $\hat{B}^2_{r_2}(0)$. The C^1 property of $\hat{E}(\cdot)$ on $\hat{B}^2_{r_2}(0)$ can be proven fairly easily. ∎

Lemma 5.2 on the existence of a solution. Recall the notation $\tilde{X}_y = \mathbf{L}_2(1)$ and $\tilde{X}_z = \mathbf{L}_2(1) \times \mathbf{R}$. We have the following lemma.

LEMMA 5.2. *Suppose Assumption* 2 *holds and let* $z_0 = (y_0, 1)$. *If the operator* $\widetilde{F}'_z(z_0, 0) \in \mathcal{L}(\tilde{X}_z, \tilde{X}_y)$ *defined by*

$$(\widetilde{F}'_z(z_0,0)(v,\delta T))(t) = ((I - L^s)v)(t) - y_1^0(t)\delta T,$$

$$(L^s v)(t) = \int_{-\infty}^{t} h(t-\tau,0)\varphi'_y(y_0(\tau),0)v(\tau)\,d\tau,$$

$$y_1^0(t) = \int_{-\infty}^{t} h(t-\tau,0)\varphi(y(\tau),0)\,d\tau + \int_{-\infty}^{t} (t-\tau)\,dh((t-\tau),0)\varphi(y(\tau),0)$$

has a bounded right inverse, then for each sufficiently small $\theta \in \boldsymbol{R}$ *there exists* $y_\theta \in C(1)$ *and* $T_\theta > 0$ *that satisfies* (2.11). *The perturbed solution* $y_\theta = y(\theta)$, $T_\theta = T(\theta)$ *are* C^1 *functions of* θ *such that* $y(0) = y_0$ *and* $T(0) = 1$. *The perturbed solution is unique modulo phase delays (time translation).*

Proof. Let $F : X_z \times X_\theta \to X_y$ be defined as in (3.2). By assumption we have $F(z_0, 0) = 0$. The following conclusion follows from Theorem 5.1 with $V_1 = U_1 = X_z = C(1) \times \mathbf{R}$, $V_2 = X_\theta = \mathbf{R}$, $U_2 = I_\theta$, and $V_3 = X_y = C(1)$: if $F'_z(z_0, 0)$ has a bounded right inverse, then there exists a $z_\theta = (y_\theta, T_\theta) \in X_z$ satisfying $F(z_\theta, \theta) = 0$ for all $\theta \in (-\theta_0, \theta_0)$, where $\theta_0 > 0$ is some sufficiently small number. In other words, there exists a periodic solution of (2.11) for all sufficiently small θ. The C^1 condition on z_θ also follows from Theorem 5.1.

The Fréchet derivative of F with respect to its first argument at $(z_0, 0)$ is defined as

$$(F'_z(z_0,0)(v,\delta T))(t) = v(t) - \int_{-\infty}^{t} h(t-\tau,0)\varphi'_y(y_0(\tau),0)v(\tau)\,d\tau - y_1^0(t)\delta T.$$

We will show that the existence of a bounded right inverse to $\widetilde{F}'_z(z_0, 0)$ implies the existence of a bounded right inverse to $F'_z(z_0, 0)$, which proves Lemma 5.2. Note that these operators are defined in exactly the same way but on different spaces.

For notational convenience, we let

$$H^s := \widetilde{F}'_z(z_0,0) = \begin{bmatrix} I - L^s & y_1^0 \end{bmatrix} \in \mathcal{L}(\mathbf{L}_2(1) \times \mathbf{R}, \mathbf{L}_2(1)).$$

We have the topological inclusion $C(1) \hookrightarrow L_2(1)$ since $\|v\|_{\mathbf{L}_2(1)} \le \|v\|_{C(1)}$ for all $v \in C(1)$. This shows that the restriction $H^s|_{C(1)\times\mathbf{R}} = F'_z(z_0, 0)$. By assumption, there exists a right inverse of H^s, i.e., a bounded linear operator $G^s : \mathbf{L}_2(1) \to \mathbf{L}_2(1) \times \mathbf{R}$ such that

$$(H^s G^s)w = w \qquad \forall w \in \mathbf{L}_2(1).$$

We will show that $G = G^s|_{C(1)}$ is a bounded right inverse of $H^s|_{C(1)\times\mathbf{R}}$. Let $w \in C(1)$; then

$$w = (H^s \circ G)w = H^s(G_1 w, G_2 w) = (I - L^s)G_1 w + y_1^0 G_2 w,$$

which shows that $G_1 w = w + L^s G_1 w - y_1^0 G_2 w \in C(1)$. To prove boundedness we start with the second component. We have

$$\begin{aligned} \|G_2\|_{C(1)\to\mathbf{R}} &= \sup\{|G_2 w| : \|w\|_{C(1)} \le 1,\ w \in C(1)\} \\ &\le \sup\{|G_2 w| : \|w\|_{\mathbf{L}_2(1)} \le 1,\ w \in C(1)\} \\ &\le \sup\{|G_2^s w| : \|w\|_{\mathbf{L}_2(1)} \le 1,\ w \in \mathbf{L}_2(1)\} = \|G_2^s\|_{\mathbf{L}_2(1)\to\mathbf{R}}. \end{aligned}$$

For the first component we use

$$\begin{aligned}\|G_1 w\|_{C(1)} &= \|w + L^s G_1 w - y_1 G_2 w\|_{C(1)} \\ &\le \|w\|_{C(1)} + \|L^s\|_{\mathbf{L}_2(1)\to C(1)} \cdot \|G_1^s\|_{\mathbf{L}_2(1)\to\mathbf{L}_2(1)}\|w\|_{\mathbf{L}_2(1)} \\ &\quad + \|y_1^0\|_{C(1)} \cdot \|G_2\|_{C(1)\to\mathbf{R}}\|w\|_{C(1)} \\ &\le (1 + \|L^s\|_{\mathbf{L}_2(1)\to C(1)} \cdot \|G_1^s\|_{\mathbf{L}_2(1)\to\mathbf{L}_2(1)} + \|y_1^0\|_{C(1)}\cdot\|G_2^s\|_{\mathbf{L}_2(1)\to\mathbf{R}})\|w\|_{C(1)}.\end{aligned}$$

Note that $\|L^s\|_{\mathbf{L}_2(1)\to C(1)}$ is finite because we have (here $H = \mathcal{L}h$, $\mathcal{L}$ denotes the Laplace transform, H_n denotes the nth row of H, and $\hat{w}_k$ is the kth Fourier coefficient of w)

$$\begin{aligned}\Big|\int_{-\infty}^{t} h(t-\tau)w(\tau)\,d\tau\Big| &= \Big|\sum_{k=-\infty}^{\infty} H(i2\pi k)\hat{w}_k e^{-i2\pi kt}\Big| \le \sum_{k=-\infty}^{\infty}\sum_{n=1}^{p}\Big|H_n(i2\pi k)\hat{w}_k\Big| \\ &\le \sum_{n=1}^{p}\Big(\sum_{k=-\infty}^{\infty}|H_n(i2\pi k)|^2\Big)^{1/2}\Big(\sum_{k=-\infty}^{\infty}|\hat{w}_k|^2\Big)^{1/2} \le \Big(\sum_{n=1}^{p}\|H_n\|_2\Big)\|w\|_{\mathbf{L}_2(1)},\end{aligned}$$

where $\|H_n\|_2^2 = \sum_{k=-\infty}^{\infty}|H_n(i2\pi k)|^2 < \infty$, which follows since the SPES property of h can be used to show that $|H_n(i2\pi k)| \le \text{const}/k$ for all $n = 1, \dots, p$. Hence,

$$\|L^s\|_{\mathbf{L}_2(1)\to C(1)} \le \Big(\sum_{n=1}^{p}\|H_n\|_2\Big)\cdot\|\varphi_y'(y_0,0)\|_{C(1)}.$$

We have thus shown that $F_z'(z_0,0)$ has a bounded right inverse $F_z'(z_0,0)^\dagger = \widetilde{F}_z'(z_0,0)^\dagger|_{X_y}$, which, as we pointed out earlier, proves the existence.

Next we discuss the uniqueness. First notice that for each choice of right inverse the proof of Theorem 5.1 results in a unique solution that we can write as $u_1 = E_{G_0}(u_2)$. The solution is thus only unique modulo the particular choice of right inverse. To understand the implication for the perturbed solution we consider (for arbitrary θ)

$$F_z'(z_\theta,0) = [I - L^s(z_\theta,\theta) \quad y_1(z_\theta,\theta)] \in \mathcal{L}(C(1)\times\mathbf{R}, C(1)),$$

where $L^s(z_\theta,\theta)$ and $y_1(z_\theta,\theta)$ are defined as in (3.8). We will next show that

$$\operatorname{Ker} F_z'(z_\theta,\theta) = (\dot{y}_\theta, 0)$$

whenever the operator has a right inverse. This follows since, for any time translation d,

$$y_\theta(t+d) = \int_{-\infty}^{t} h(t-\tau,\theta)\varphi(y_\theta(\tau+d),\theta)\,d\tau.$$

Differentiation with respect to d at $d = 0$ gives

$$\dot{y}_\theta(t) = \int_{-\infty}^{t} h(t-\tau,\theta)\varphi_y'(y_\theta(\tau),\theta)\dot{y}_\theta(\tau)\,d\tau = (L^s(z_\theta,\theta)\dot{y}_\theta)(t).$$

This shows that $\dot{y}_\theta$ is an eigenfunction of $L^s(z_\theta,\theta)$ corresponding to the eigenvalue 1. We obtain the following necessary (and sufficient) condition for $F_z'(z_\theta,\theta)$ to have a

bounded right inverse:

(i) 1 is a simple eigenvalue of $L^s(z_\theta,\theta)$, which implies that $\operatorname{Im}(I-L^s)$ has codimension 1 since $L^s(z_\theta,\theta)$ is a compact operator;

(ii) $y_1 \notin \operatorname{Im}(I-L^s)$.

Hence, under the assumption of the theorem $\operatorname{Ker} F'_z(z_0,0) = (\dot y_0, 0)$. The implication is that the degree of freedom in the choice of the right inverse corresponds to the time translation of the nominal periodic solution. The limit cycle is a one-dimensional manifold whenever the right inverse exists and this implies that our perturbed solution $z_\theta = (y_\theta, T_\theta)$ is unique modulo time translation. A region of uniqueness can be proven using similar Lipschitz properties as was used in the proof of Theorem 5.1. □

Proof of Lemma 3.4. We start by repeating some of the discussions in section 3.1. Let $F : X_z \times X_\theta \to X_y$ be defined as in (3.2). A solution $z(\theta) = (y(\theta), T(\theta))$ to

$$F(z(\theta),\theta) = 0 \tag{5.3}$$

corresponds to a 1-periodic solution of (2.11). Differentiation of (5.3) gives

$$F'_z(z(\theta),\theta)\frac{dz}{d\theta} = -F'_\theta(z(\theta),\theta).$$

Now consider the operator $\widetilde F'_z : \mathcal{Z} \times I_\theta \to \mathcal{L}(\tilde X_z \times \tilde X_\theta, \tilde X_y)$ defined in (3.6)–(3.8). We note that the restriction of $\widetilde F'_z$ to $X_z \times X_\theta$ satisfies $\widetilde F'_z(z(\theta),\theta)|_{X_z} = F'_z(z(\theta),\theta)$. By assumption $\widetilde F'_z(z_0,0)$ has a bounded right inverse $\widetilde F'_z(z_0,0)^\dagger$, which by the proof of Lemma 5.2 implies that $\widetilde F'_z(z_0,0)|_{X_z} = F'_z(z_0,0)$ also has a bounded right inverse. In view of Lemma 3.3, the continuity of $F'_z(z,\theta)$ with respect to its arguments, and the continuity of the perturbed solution $z(\theta)$, it follows that for sufficiently small θ there exists a right inverse $F'_z(z(\theta),\theta)^\dagger$ such that

$$\frac{dz}{d\theta} = -F'_z(z(\theta),\theta)^\dagger F'_\theta(z(\theta),\theta). \tag{5.4}$$

In order to find a bound on θ for which this expression is valid we use Lemma 3.3. This leads to expression (3.5), which can be formulated equally well using the extensions of the operators to $\mathbf{L}_2(1)$ space. We can thus write

$$\frac{dz}{d\theta} = -\widetilde F'_z(z_0,0)^\dagger (I - \widetilde\Delta(z(\theta),\theta)\widetilde F'_z(z_0,0)^\dagger)^{-1}\widetilde F'_\theta(z(\theta),\theta). \tag{5.5}$$

Let Θ be the set of $\theta \in [-\bar\theta,\bar\theta]$ such that there exists a solution $z(\theta) \in \mathcal{Z}$ to (5.3). We use a homotopic argument based on the following observations to prove $\Theta = [-\bar\theta,\bar\theta]$:

(a) $0 \in \Theta$ since $z(0) = z_0 \in \mathcal{Z}$;

(b) if $[0,\theta] \in \Theta$, then (similarly with $[-\theta,0] \in \Theta$)

$$\begin{aligned}\|z(\theta) - z_0\|_{X_z} &\le \left\| \int_0^\theta \widetilde F'_z(z_0,0)^\dagger (I - \widetilde\Delta(z(\sigma),\sigma)\widetilde F'_z(z_0,0)^\dagger)^{-1}\widetilde F'_\theta(z(\sigma),\sigma)\,d\sigma \right\|_{X_z}\\ &\le \bar\theta \cdot \sup_{z\in\mathcal{Z},\ |\theta|\le\bar\theta} \|\widetilde F'_z(z_0,0)^\dagger (I - \widetilde\Delta(z,\theta)\widetilde F'_z(z_0,0)^\dagger)^{-1}\widetilde F'_\theta(z(\theta),\theta)\|_{\tilde X_\theta \to X_z} < r_0;\end{aligned}$$

(c) Θ is open as a subset of $[-\bar\theta,\bar\theta]$ since for any $\theta\in\Theta$ we have $\|\widetilde\Delta(z(\theta),\theta)\|_{\tilde X_z \to \tilde X_y} < 1/\|\widetilde F'_z(z_0,0)\|_{\tilde X_y \to \tilde X_z}$ by condition (ii). Then $\widetilde F'_z(z(\theta),\theta)$ is right invertible by Lemma 3.3, which by Lemma 5.2 implies the existence of a solution of (5.3) in

a neighborhood of θ. An explicit lower bound on this interval can be obtained since all derivatives are uniformly bounded in $[-\bar{\theta}, \bar{\theta}]$ and $\mathcal{Z}$. Indeed, we can use the second part of the implicit function theorem in Theorem 5.1 to derive a uniform bound on the neighborhood.[3]

Conclusions (a) and (c) imply that the set of θ such that there is a solution $z(\theta) \in \mathcal{Z}$ is nonempty and open as a subset of the interval $[-\bar{\theta}, \bar{\theta}]$ and (b) shows that it must be the full interval. Note that on the domain $[-\bar{\theta}, \bar{\theta}]$ the solution of (5.5) is unique and thus z_θ is unique modulo the choice of right inverse $\tilde{F}'_z(z_0, 0)^\dagger$. By the proof of Lemma 5.2 this corresponds to a degree of freedom due to the time translation. ∎

Proof of Theorem 3.7. We will use the standard implicit function theorem to prove this result. The Frechét derivative of (3.9) is

$$(\Psi'_z(0,0)(v, \delta d))(t) = \dot{y}_0(t)\delta d + v(t) - \int_0^t h(t-\tau)\varphi'(y_0(\tau))(\dot{y}_0(\tau)\delta d + v(\tau))\, d\tau.$$

To prove that $\Psi'_z(0,0)$ has a bounded inverse we need to show that there exists $c > 0$ such that for each $w \in V$ the equation

$$\dot{y}_0(t)\delta d + v(t) - \int_0^t h(t-\tau)\varphi'(y_0(\tau))(\dot{y}_0(\tau)\delta d + v(\tau))\, d\tau = w(t) \tag{5.6}$$

has a unique solution, which is bounded as $\|v\|_V^2 + |\delta d|^2 \le c\|w\|_V^2$. Let us restrict attention to solutions in $\mathbf{L}_{2\alpha}$ for a while. From the definition of L in (3.11) and from the proof of Proposition 3.5 it follows that system (5.6) (now extended to $\mathbf{L}_{2\alpha}$) can be rewritten as

$$(I - L)v + e\delta d = w, \tag{5.7}$$

where

$$e(t) = \int_{-\infty}^0 h(t-\tau)\varphi'(y_0(\tau))\dot{y}_0(\tau)\, d\tau = \dot{y}_0(t) - \int_0^t h(t-\tau)\varphi'(y_0(\tau))\dot{y}_0(\tau)\, d\tau \in \mathbf{L}_{2\alpha}[0,\infty).$$

The conclusion $e(t) \in \mathbf{L}_{2\alpha}[0,\infty)$ follows since h is exponentially stable with the decay rate α. Indeed, we have

$$e^{\alpha t}e(t) = \int_0^\infty e^{-\alpha s}e^{\alpha(t+s)}h(t+s)\varphi'(y_0(-s))\dot{y}_0(-s)\, ds.$$

This implies

$$\begin{aligned}\|e\|_\alpha^2 &\le \int_0^\infty \int_0^\infty e^{-\alpha s_1}e^{-\alpha s_2}\, ds_1\, ds_2 \|e^{\alpha t}h(t)\|^2 \cdot \|\varphi'(y_0)\dot{y}_0\|_\infty^2 \\ &\le \frac{1}{\alpha^2}\|e^{\alpha t}h(t)\|^2 \cdot \|\varphi'(y_0)\dot{y}_0\|_\infty^2,\end{aligned}$$

which proves the claim since $\|e^{\alpha t}h(t)\|$ is bounded by Lemma 2.2(c).

By assumption $\mathrm{def}_\alpha(L) = 1$, which implies that $V_1 = \mathrm{Im}\,(I - L)$ has codimension 1. We thus have a direct sum decomposition $\mathbf{L}_{2\alpha}[0,\infty) = V_1 \oplus V_1^\perp$, where $V_1^\perp$ is one-dimensional. The following properties will be used.

[3]The implicit function theorem can be directly used to estimate a robustness margin using computations in $C(1)$. However, this generally leads to conservative bounds.

(i) $(I-L):\mathbf{L}_{2\alpha}[0,\infty)\to V_1$ has a bounded inverse, i.e., there exists $\gamma>0$ such that $\|(I-L)^{-1}|_{V_1}\|\le\gamma$, which follows since when the domain of the inverse is restricted to V_1, the (closed) image of $(I-L)$, then the inverse is bounded by the Banach inverse theorem since $\operatorname{Ker}(I-L)=0$.

(ii) $(I-P_{V_1})e\neq 0$, where P_{V_1} is the orthogonal projection onto V_1. This follows from Proposition 3.5 since $(I-L)^{-1}e\notin\mathbf{L}_{2\alpha}$.

From (5.7) we obtain $(I-P_{V_1})e\delta d=(I-P_{V_1})w$, which has a unique solution by (ii) above. This gives the bound

$$|\delta d|\le c_1\|w\|_V,\quad c_1=1/\|(I-P_{V_1})e\|_V.$$

With this δd, we have $(I-L)v=w-e\delta d\in V_1$, which has a unique solution by (i) above. Moreover, we immediately get the norm bound

$$\|v\|_\alpha\le\gamma\|w-e\delta d\|_\alpha\le\gamma(\|w\|_\alpha+|\delta d|\cdot\|e\|_\alpha)\le\gamma(1+c_1\|e\|_\alpha)\|w\|_V.$$

However, we need a bound in terms of the norm $\|\cdot\|_V$. A $\|\cdot\|_\infty$ bound is obtained from the following derivation:

$$\begin{aligned}\|v\|_\infty&=\|Lv+w-e\delta d\|_\infty\le\|L\|_{\mathbf{L}_{2\alpha}\to\mathbf{L}_\infty}\cdot\|v\|_\alpha+\|w\|_\infty+c_1\|e\|_\infty\cdot\|w\|_V\\&\le(\|L\|_{\mathbf{L}_{2\alpha}\to\mathbf{L}_\infty}\gamma(1+c_1\|e\|_\alpha)+c_1\|e\|_\infty+1)\|w\|_V,\end{aligned}$$

where the norm $\|L\|_{\mathbf{L}_{2\alpha}\to\mathbf{L}_\infty}$ is bounded since h is SPES, see Lemma 3 in [7] for a related proof. Hence, we have $\|v\|_V\le c_2\|w\|_V$, where

$$c_2=\max(\gamma(1+c_1\|e\|_\alpha),\|L\|_{\mathbf{L}_{2\alpha}\to\mathbf{L}_\infty}\gamma(1+c_1\|e\|_\alpha)+c_1\|e\|_\infty+1).$$

This together with the previous bound on $|\delta d|$ shows that $\Psi_z'(0,0)$ has a bounded inverse. In fact, we have shown $\Psi_z'(0,0)^{-1}:\mathcal{X}_y\to\mathcal{X}_z$ satisfies $\|\Psi_z'(0,0)^{-1}\|_{\mathcal{X}_y\to\mathcal{X}_z}\le c$, where $c=(c_1^2+c_2^2)^{1/2}$.

Boundedness of $\Psi_z'(0,0)^{-1}$ implies by the implicit function theorem that there exists an open set $V_0\subset\mathcal{X}_y$ containing 0 and a C^1 function $E:V_0\to\mathcal{X}_z$ such that $\Psi(E(\delta f),\delta f)=0$ for all $\delta f\in V_0$. It is no restriction to assume that V_0 is convex since otherwise we can consider a ball $\{v\in V:\|v\|_V\le\eta\}\subset V_0$. We further have

$$E'(\delta f)=-\Psi_z'(E(\delta f),\delta f)^{-1}\Psi_{\delta f}'(E(\delta f),\delta f)=\Psi_z'(E(\delta f),\delta f)^{-1}.$$

We obtain the following bound in the $\widetilde{\mathcal{X}}_z=\mathbf{L}_{2\alpha}\times\mathbf{R}$-space:

$$\begin{aligned}\|\delta y\|_\alpha^2+|\delta d|^2&=\|E_1(\delta f)\|_\alpha^2+|E_2(\delta f)|^2=\|E(\delta f)-E(0)\|_{\mathbf{L}_{2\alpha}\times\mathbf{R}}^2\\&=\Big\|\int_0^1E'(s\delta f)\cdot\delta f\,ds\Big\|_{\mathbf{L}_{2\alpha}\times\mathbf{R}}^2\\&\le\sup_{s\in[0,1]}\|E'(s\delta f)\|_{\mathbf{L}_{2\alpha}\to\mathbf{L}_{2\alpha}\times\mathbf{R}}^2\|\delta f\|_\alpha^2\\&\le\sup_{v\in V_0}\|E'(v)\|_{\mathbf{L}_{2\alpha}\to\mathbf{L}_{2\alpha}\times\mathbf{R}}^2\|\delta f\|_\alpha^2\\&=\sup_{v\in V_0}\|\widetilde{\Psi}_z'(E(\delta f),\delta f)^{-1}\|_{\mathbf{L}_{2\alpha}\to\mathbf{L}_{2\alpha}\times\mathbf{R}}^2\|\delta f\|_\alpha^2\end{aligned}$$

for all $\delta f\in V_0$. Here $\widetilde{\Psi}_z'\in\mathcal{L}(\tilde{X}_z,\tilde{X}_y)$ is the extension of Ψ_z' to $\mathbf{L}_{2\alpha}$. Since Ψ is C^1 it follows that for any $\epsilon>0$ we get the bound

$$\|\delta y\|_\alpha^2+|d|^2\le c(\epsilon)^2\|\delta f\|_\alpha^2$$

by choosing V_0 small enough, where

$$c(\epsilon)^2 = \widetilde{c}_1^2 + \gamma^2(1+\widetilde{c}_1\|e\|_\alpha)^2 + \epsilon, \quad \widetilde{c}_1 = 1/\|(I-P_{V_1})e\|_\alpha.$$

This proves the exponential stability inequality in (2.10). □

Proof of Lemma 3.9. According to Theorem 3.7 we need to show that $L = L^0 + \Delta L$ has α-defect 1. We have

$$\begin{aligned}\|(I - L_\alpha^0 - \Delta L_\alpha)v\| &\geq \|(I-L_\alpha^0)v\| - \|\Delta L_\alpha v\| \\ &\geq (c - \|\Delta L_\alpha\|)\|v\| \geq \varepsilon\|v\|,\end{aligned}$$

where the last inequality follows from (3.16). Hence, we have shown that $\operatorname{Ker}(I - L_\alpha^0 - \Delta L_\alpha) = 0$ for all $\|\Delta L_\alpha\| \leq c - \epsilon$. The proof follows if we in addition prove that $\operatorname{codim}\operatorname{Im}(I - L_\alpha^0 - \Delta L_\alpha) = 1$. Since the codimension of $\operatorname{Im}(I - L_\alpha^0)$ is 1 it follows that we can find $v_1 \in \mathbf{L}_2[0,\infty)$ with $\|v_1\| = 1$ and $v_1 \perp \operatorname{Im}(I - L_\alpha^0)$. If we define $H : \mathbf{L}_2[0,\infty) \times \mathbf{R} \to \mathbf{L}_2[0,\infty)$ by

$$H(v,t_1) = (I - L_\alpha^0)v + ct_1v_1,$$

then H is a bijection and it follows from the Banach inverse theorem that it has a bounded inverse. Since

$$\|(I - L_\alpha^0)v + ct_1v_1\| \geq c(\|v\|^2 + |t_1|^2)^{1/2}$$

we have $\|H^{-1}\| \leq c^{-1}$. With $\Delta H(v,t_1) = -\Delta L_\alpha v$ we see that $H + \Delta H = H(I + H^{-1}\Delta H)$ is invertible since $\|H^{-1}\Delta H\| \leq 1 - \epsilon/c < 1$. It follows that $\operatorname{codim}\operatorname{Im}(I - L_\alpha^0 - \Delta L_\alpha) = 1$. □

Proof of Proposition 3.14. From Lemma 3.8 it follows that we can equivalently consider the operator L_α on $\mathbf{L}_2$ when deciding the stability defect. We will show

(i) $\operatorname{Ker}(I - L_\alpha) = 0$,

(ii) $\operatorname{codim}\operatorname{Im}(I - L_\alpha) = 1$.

Conditions (i) and (ii) show that L_α is a Fredholm operator with index 1. From Banach's isomorphism theorem it follows that $I - L_\alpha$ is nonsingular. This proves the theorem.

To prove (i) we assume that there exists nonzero $v \in \mathbf{L}_2$ such that $(I - L_\alpha)v = 0$. In the state-space domain this means that

$$\dot{x} = (A + \alpha I)x + B\varphi'(y_0(t))v, \quad x(0) = 0, \quad 0 = v - Cx$$

which implies that $v = Cx$ and $\dot{x} = (A + \alpha I + B\varphi'(y_0(t))C)x$, $x(0) = 0$. This contradicts the assumption that v is nonzero. Hence, $\operatorname{Ker}(I - L_\alpha) = 0$.

To prove (ii) we use $(\operatorname{Im}(I - L_\alpha))^\perp = \operatorname{Ker}(I - L_\alpha)^*$. One possible state-space representation of the adjoint system $v \mapsto w = (I - L_\alpha^*)v$ is

$$\begin{aligned}\dot{x} &= -(A + \alpha I)^T x + C^T v, \ x(\infty) = 0, \\ w &= v + \varphi'(y_0(t))^T B^T x.\end{aligned}$$

Any $v \in \operatorname{Ker}(I - L_\alpha^*)$ must satisfy $v = -\varphi'(y_0(t))^T B^T x$, where

$$\dot{x} = -(A + \alpha + B\varphi'(y_0(t))C)^T x, \quad x(\infty) = 0. \tag{5.8}$$

A result by Lyapunov shows that there exists a time-periodic coordinate transformation that turns system (5.8) into a linear system with constant coefficients; see, e.g., [4]. It is no restriction to assume the new coordinates are chosen such that

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \begin{bmatrix} -\alpha & 0 \\ 0 & A_2 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}, \qquad \begin{bmatrix} z_1(\infty) \\ z_2(\infty) \end{bmatrix} = 0,$$

where $A_2 \in \mathrm{C}^{(n-1)\times(n-1)}$ has all its characteristic multipliers outside the unit disc since $-\alpha T - \log|\lambda_2| > 0$. If the coordinates are related as

$$x(t) = [P_1(t) \quad P_2(t)], \begin{bmatrix} z_1(t) \\ z_2(t) \end{bmatrix},$$

where $P(t) = [P_1(t) \quad P_2(t)]$ is invertible and T is periodic, then we see that

$$\operatorname{Ker}(I - L_\alpha)^* = \{v(t) = -\varphi'(y_0(t))^T B^T P_1(t) e^{-\alpha t} z_1(0) : z_1(0) \in \mathbf{R}\}.$$

This is a one-dimensional space. Hence, we have shown that $\operatorname{Ker}(I - L_\alpha) = 0$ and $\operatorname{codim} \operatorname{Im}(I - L_\alpha) = 1$. This implies $\operatorname{def}(L_\alpha) = 1$, which by Lemma 3.8 implies $\operatorname{def}_\alpha(L) = 1$. □

Proof of Theorem 3.1. First note that the characteristic multipliers do not change if we normalize the nominal period time to $T_0 = 1$ in (2.1). Existence of a solution in a neighborhood of $\theta = 0$ follows from Lemma 5.2 if $\widetilde{F}'_z(z_0, 0)$ has a bounded right inverse. From Proposition 3.12, we see that this is the case since $n - 1$ of the characteristic multipliers are different from 1.

To prove exponential stability we consider the operator L in (3.11), which becomes

$$(L(\theta)v)(t) = \int_0^t T(\theta) h(T(\theta)(t-\tau), \theta) \varphi'_y(y_\theta(\tau), \theta) v(\tau)\, d\tau.$$

It follows from Proposition 3.14 that $L(0)$ has the α-defect 1. The same arguments that we used to prove Lemma 3.9 shows that the α-defect remains constant for sufficiently small θ since $L(\theta)$ depends continuously on θ. Hence, $\operatorname{def}_\alpha(L(\theta)) = 1$ for sufficiently small θ, which by Theorem 3.7 proves exponential stability. □

REFERENCES

[1] R. Abraham, J. E. Marsden, and T. Ratiu, *Manifolds, Tensor Analysis, and Applications,* Springer-Verlag, New York, 1988.

[2] A. Andronov and A. Witt, *On Lyapunov stability,* 3, J. Electr. Techn. Phys. (1933) (in Russian).

[3] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations,* McGraw-Hill, New York, 1955.

[4] M. Farkas, *Periodic Motions,* Springer-Verlag, New York, 1994.

[5] T. T. Georgiou and M. C. Smith, *Robustness of a relaxtion oscillator,* Internat. J. Robust Nonlinear Control, 10 (2000), pp. 1005–1024.

[6] P. Hartman, *Ordinary Differential Equations,* 2nd ed., Birkhäuser, Boston, 1982.

[7] U. Jönsson, C. Kao, and A. Megretski, *Robustness analysis of periodic trajectories,* IEEE Trans. Automat. Control, 47 (2002), pp. 1842–1856.

[8] U. Jönsson and A. Megretski, *A Small Gain Theory for Limit Cycles,* Technical report TRITA/MAT-03-OS07, Department of Mathematics, Royal Institute of Technology, 2003.

[9] H. K. Khalil, *Nonlinear Systems,* Macmillan, New York, 1996.

[10] D. C. Lewis, *Periodic solutions of differential equations containing a parameter,* Duke Math. J., 22 (1955), pp. 39–56.
[11] H. Poincaré, *Les Méthodes nouvelles de la mécanique céleste,* Vols. I, II, III, Gauthier-Villar, Paris, 1892, 1893, 1899.
[12] A. P. Stokes, *On the stability of an autonomous functional equation,* Contrib. Differential Equations, 3 (1963), pp. 121–139.
[13] A. P. Stokes, *Some implications of orbital stability in Banach spaces,* SIAM J. Appl. Math., 17 (1969), pp. 1317–1325.
[14] S. Varigonda, *Robustness analysis of a relay oscillator,* in 15th Triennial World Congress of IFAC, 2002.
[15] V. A. Yakubovich, *Frequency domain criteria for oscillations in nonlinear systems with one stationary nonlinear component,* Sibirsk. Mat. Zh., 14 (1973), pp. 1100–1129.
[16] V. A. Yakubovich, *A linear-quadratic optimization problem and the frequency theorem for periodic systems.* I, Sibirsk. Mat. Zh., 27 (1986), pp. 181–200.
[17] V. A. Yakubovich and V. M. Starzhinskij, *Linear Differential Equations with Periodic Coefficients,* Nauka, Moscow, 1972 (in Russian).

SIAM J. CONTROL OPTIM.
Vol. 44, No. 3, pp. 939–968

MINIMAX OPTIMAL CONTROL*

R. B. VINTER†

Abstract. This paper provides a framework for deriving necessary conditions, in the form of a maximum principle, for minimax optimal control problems. The distinguishing feature of these problems is that the data depends on a vector α of unknown parameters, and "optimality" is defined on a worst case basis, as α ranges over the parameter set $\mathcal{A}$. The centerpiece, a minimax maximum principle, is a set of optimality conditions for such problems. Here, the parameter set $\mathcal{A}$ is taken to be an arbitrary compact metric space and the hypotheses imposed on the dynamics and endpoint constraints are of an unrestrictive nature. The minimax maximum principle captures as special cases necessary conditions for optimal control problems with minimax costs, for problems involving "semi-infinite" endpoint constraints, and also a maximum principle for state constrained optimal control problems.

Key words. optimal control, minimax problems, nonsmooth analysis, robust control

AMS subject classifications. Primary, 49L20, 49N25; Secondary, 34A37, 49L99

DOI. 10.1137/S0363012902415244

1. Introduction.

The purpose of this paper is to derive, in a unified fashion, necessary conditions of optimality for optimal control problems involving an unknown vector parameter. In these problems, "optimality" is typically defined in terms of worst case performance, i.e., the cost of a particular control strategy is that associated with the strategy and a system response corresponding to the least favorable value of the unknown parameter, and constraints are required to be satisfied for all values of the unknown parameter.

Fix a compact metric space $(\mathcal{A}, \rho_{\mathcal{A}}(.,.))$. Take functions $f : [0,1] \times R^n \times R^m \times \mathcal{A} \to R^n$ and $g : R^n \times \mathcal{A} \to R$, a vector $x_0 \in R^n$, a time dependent set $\Omega(t) \subset R^m$, $0 \le t \le 1$, and a family of closed sets $\{C(\alpha) \subset R^n \mid \alpha \in \mathcal{A}\}$.

A *control function* is a measurable function $u : [0,1] \to R^m$ satisfying $u(t) \in \Omega(t)$ *a.e.* The set of control functions is written $\mathcal{U}$. A *process* $(u, \{x(.;\alpha) \mid \alpha \in \mathcal{A}\})$ comprises a control function u and a family $\{x(.;\alpha) \in W^{1,1}([0,1]; R^n) \mid \alpha \in \mathcal{A}\}$ of arcs satisfying, for each $\alpha \in \mathcal{A}$,

$$\begin{cases} \dot{x}(t;\alpha) = f(t, x(t;\alpha), u(t), \alpha) \quad a.e. \\ x(0;\alpha) = x_0. \end{cases}$$

The process is termed *feasible* if the $x(.;\alpha)$s satisfy the terminal constraints

$$x(1;\alpha) \in C(\alpha) \quad \text{for all } \alpha \in \mathcal{A}.$$

The optimization problem of interest in this paper, which will be referred to as the

*Received by the editors September 27, 2002; accepted for publication (in revised form) January 4, 2005; published electronically September 15, 2005.
http://www.siam.org/journals/sicon/44-3/41524.html
†Department of Electrical and Electronic Engineering, Imperial College of Science Technology and Medicine, Exhibition Road, London SW7 2BT, UK (r.vinter@imperial.ac.uk).

general minimax optimal control problem, is as follows:

$$(\mathrm{P})\quad \left\{\begin{array}{l} \text{Minimize } \max_{\alpha\in\mathcal{A}} g(x(1;\alpha),\alpha) \\ \qquad \text{over measurable functions } u:[0,1]\to R^m \text{ such that} \\ u(t)\in\Omega(t) \quad \textit{a.e. } t\in[0,1] \\ \qquad \text{and arcs } \{x(.;\alpha):[0,1]\to R^n \mid \alpha\in\mathcal{A}\} \text{ such that, for each } \alpha\in\mathcal{A}, \\ \dot{x}(t;\alpha)=f(t,x(t;\alpha),u(t),\alpha) \quad \textit{a.e. } t\in[0,1], \\ x(0;\alpha)=x_0 \quad \text{and} \quad x(1;\alpha)\in C(\alpha). \end{array}\right.$$

Briefly stated, the problem is to minimize $\sup_{\alpha\in\mathcal{A}} g(x(1;\alpha),\alpha)$ over feasible processes $(u,\{x(.;\alpha)\mid\alpha\in\mathcal{A}\})$.

A feasible process $(\bar{u},\{\bar{x}(.;\alpha)\mid\alpha\in\mathcal{A}\})$ is said to be a strong local minimizer when there exists $\epsilon>0$ such that

$$\sup_{\alpha\in\mathcal{A}} g(x(1;\alpha),\alpha) \;\geq\; \sup_{\alpha\in\mathcal{A}} g(\bar{x}(1;\alpha),\alpha)$$

for all feasible processes $(u,\{x(.;\alpha),\alpha\in\mathcal{A}\})$ such that

$$\|x(.;\alpha)-\bar{x}(.;\alpha)\|_C\leq\epsilon \quad \text{for all } \alpha\in\mathcal{A}\,.$$

The implications of our necessary conditions for various special cases of interest will also be investigated.

Our framework permits the set $\mathcal{A}$ of unknown parameter values to be an arbitrary compact metric space. It therefore covers minimax optimal control problems in which components of α comprise unknown gain values lying within specified bounds, magnitudes of step disturbances, etc., important cases that would be excluded by the requirement that $\mathcal{A}$ be a finite set.

The presence of, possibly, an infinite number of elements in $\mathcal{A}$ is the principal source of difficulty in the derivation of necessary conditions for minimax optimal control problems. In case $\mathcal{A}$ is a finite set $\{\alpha_1,\alpha_2,\ldots,\alpha_N\}$, the minimax optimal control problems studied here can be reformulated as standard optimal control problems, for which necessary conditions are already known. (See section 2.)

We comment on related earlier research. The most extensively studied minimax optimal control problems are zero sum differential games, in which a minimizer is chosen from a class of closed loop controls, appropriately defined, and the parameter set $\mathcal{A}$, from which a "worst case" element is selected, comprises open loop control functions of an opposing player [1], [4]. The fact that differential games are posed over closed loop controls gives them a quite different character to the problems studied here, in which the choice variables are open loop controls. Analysis of solutions to differential games is almost exclusively of a global nature, centering on the relationship between the value of the differential game and the solutions to the Hamilton–Jacobi equation; variants on the pontryagin maximum principle, such as featured in this paper, have a limited role in the analysis of optimal feedback strategies.

Versions of the open loop minimax optimal control problem were previously investigated by Warga, in the context of "relaxed and hyper-relaxed adverse controls." Warga adopts a broader framework than ours, in which the parameter set can include open control functions of an opposing player as well as finite dimensional vector parameters. Furthermore, he addresses questions of existence of solutions to minimax optimal control problems and appropriate relaxation schemes as well as local optimality conditions. Our minimax maximum principle, involving a Hamiltonian averaged

with respect to some measure, is implicit in the necessary conditions in ([11], Chapters IX and X). Warga's necessary conditions apply only in cases when the endpoint constraint sets are closed, convex sets with nonempty interiors and for smooth dynamics. The necessary conditions of this paper are proved by quite different methods and under significantly weaker hypotheses (for the minimax problems here considered). Furthermore, we give new insights into the limits of validity of the kinds of necessary conditions investigated here, by presenting some counterexamples where they no longer apply. Optimality conditions akin to those of section 2 below are featured also in [2], but only in the elementary case when the parameter set is a finite set and the endpoint constraint is specified by a functional inequality. The role of measures to estimate "gradients" of max functions is evident in the early Russian optimal control literature [5] and is widely exploited in nonsmooth analysis, for example, in applications of nonsmooth analysis to derive optimality conditions for state constrained optimal control problems [3].

Another point of contact with earlier work is semi-infinite programming. This is a branch of nonlinear programming that aims to provide efficient computational methods for optimization problems, in which constraints must be satisfied for a continuum of values of some parameter α. (See [8].) Minimax optimal control problems can be reformulated, by introduction of additional variables, as semi-infinite programming problems over function spaces with dynamic constraints.

One possible approach to the computation of solutions to a minimax optimal control problem is to approximate it by a (finite-dimensional) semi-infinite programming problem by means of time discretization and to apply semi-infinite programming algorithms. The emphasis in this paper is on structural properties of solutions to minimax optimal control problems. But the necessary conditions of optimality we provide may ultimately find application in convergence analysis of algorithms for minimax optimal control, based on semi-infinite programming or other approaches.

We allow nonsmooth data and express necessary conditions in terms of "limiting subdifferentials" and other constructs of nonsmooth analysis. We stress, however, that it is the unrestrictive nature of the conditions that we place on the parameter set $\mathcal{A}$, "$\mathcal{A}$ is an arbitrary compact metric space," which is the most significant feature of our analysis. The main optimality conditions supplied here (the maximum principle for the general minimax optimal control problem of section 3 and the implications explored in section 5) are new, even when specialized to the smooth case.

Finally, some notation. Throughout, $|\,.\,|$ denotes the Euclidean norm. We write B for the closed unit ball in Euclidean space. $B_{\mathcal{A}}(\alpha, \epsilon)$ denotes the set $\{\alpha' \in \mathcal{A} \mid \rho_{\mathcal{A}}(\alpha, \alpha') \leq \epsilon\}$.

$W^{1,1}([0,1]; R^n)$ is the space of absolutely continuous R^n-valued functions on $[0,1]$. Take a compact metric space A. $C(A)$ denotes the space of continuous real valued functions on A. We write $||.||_C$ for the supremum norm on this space. $C^*(A)$ denotes the topological dual of $C(A)$ with the norm topology. We use the fact that elements in $C^*(A)$ can be identified with the space of Radon measures on the Borel subsets of A. The dual norm of an element $\mu \in C^*(A)$ is written $||\mu||_{T.V.}$, a choice of notation that reflects the fact that the dual norm of μ coincides with the total variation of the Radon measure that represents μ.

The graph of a multifunction $D : A \rightsquigarrow R^k$ is denoted by $Gr\,D$,

$$Gr\,D \;:=\; \{(a,d) \in A \times R^k \,|\, d \in D(a)\}\,.$$

For a given set $E \subset R^d$, $d_E(.)$ denotes the Euclidean distance function

$$d_E(z) \;:=\; inf_{e \in E}|z - e|.$$

The *limiting normal cone* to a given closed set $C \subset R^k$ at $x \in R^k$ is the set

$$N_C(x) := \big\{\xi \in R^n \mid \exists\, \xi_i \to \xi, x_i \overset{C}{\to} x \text{ and } \{M_i\} \subset R^+ \\ \text{such that, for each } i,\ \xi_i \cdot (x - x_i) \le M_i |x - x_i|^2\ \forall\, x \in C\big\}.$$

Here "$x_i \overset{C}{\to} x$" means "$x_i \to x$ and $x_i \in C$ for all i." Note that $N_C(x) = \emptyset$, in the case $x \notin C$.

Take a function $f : R^n \to R \cup \{+\infty\}$ and a point $x \in dom\, f$. Here, $dom\, f$ is taken to be the set

$$dom\, f = \{y \in R^n \mid f(y) < +\infty\}.$$

The epigraph set of f is the set

$$epi\, f := \{(x, \alpha) \in R^n \times R \mid \alpha \ge f(x)\}.$$

The *limiting subdifferential* $\partial f(x)$ of $f : R^n \to R \cup \{+\infty\}$ at a point $x \in dom\, f$ is the set

$$\partial f(x) := \big\{\eta \mid (\eta, -1) \in N_{\mathrm{epi} f}(x, f(x))\big\}.$$

The partial limiting subdifferential $\partial_x f(x, y)$ of an extended valued function f of two variables x and y is the limiting subdifferential of $x \to f(x, y)$ for fixed y.

$N_C(x)$ and $\partial f(x)$ are widely used constructs from nonsmooth analysis in optimal control, that generalize classical notions of the set of outward normal vectors to a set with smooth boundary and of the gradient of a continuously differentiable function. They are also referred to as the normal cone and the subdifferential, respectively. For a review of their properties (and historical comments), see, for example, [7], [9], [10].

2. The finite parameter set case. Necessary conditions for minimax problems involving an arbitrary compact metric space parameter set $\mathcal{A}$ will be derived by approximating $\mathcal{A}$ by a finite set $\mathcal{A}_N$, by establishing properties of approximate minimizers for problems involving A_N, and passage to the limit. Necessary conditions for problems with finite parameter sets have an important intermediate role then in the proof of more general necessary conditions. This is one reason for attending to the finite parameter set case at this early stage. But studying this special case also gives insights into the necessary conditions we should expect to be valid in more general circumstances.

We shall invoke the following hypotheses on the data for the general minimax optimal control problem, in which $(\bar{u}, \{\bar{x}(.; \alpha) \mid \alpha \in \mathcal{A}\})$ is the strong local minimizer under consideration. For some $\delta > 0$,

(H1) The function $f(., x, ., \alpha)$ is $\mathcal{L} \times \mathcal{B}^m$ measurable for each $(x, \alpha) \in R^n \times \mathcal{A}$. ($\mathcal{L}$ denotes the Lebesgue subsets of $[0, 1]$ and $\mathcal{B}^m$ denotes the Borel subsets of R^m.) $t \rightsquigarrow \Omega(t)$ has a Borel measurable graph.

(H2) There exists a Borel measurable function $k_f : [0, 1] \times R^m$ such that $t \to k_f(t, \bar{u}(t))$ is integrable and, for each $\alpha \in \mathcal{A}$,

$$|f(t, x, u, \alpha) - f(t, x', u, \alpha)| \le k_f(t, u)|x - x'|$$

for all $x, x' \in \bar{x}(t; \alpha) + \delta B$, $u \in \Omega(t)$, *a.e.* $t \in [0, 1]$.

(H3) The function $g(.,\alpha)$ is Lipschitz continuous on $\bar{x}(1;\alpha)+\delta B$ for all $\alpha\in\mathcal{A}$.
Define the Hamiltonian

$$H(t,x,p,u,\alpha) := p\cdot f(t,x,u,\alpha).$$

PROPOSITION 2.1. *Let $(\bar{u},\{\bar{x}(.;\alpha)\mid\alpha\in\mathcal{A}\})$ be a strong local minimizer for the general minimax optimal control problem* (P). *Assume that $\mathcal{A}$ is a finite set and that, for some $\delta>0$, hypotheses* (H1)–(H3) *are satisfied.*

Then

$$\int H(t,\bar{x}(t;\alpha),\bar{u}(t),p(t;\alpha),\alpha)\ \Lambda(d\alpha)$$
$$= \max_{u\in\Omega(t)} \int H(t,\bar{x}(t;\alpha),u,p(t;\alpha),\alpha)\ \Lambda(d\alpha). \qquad a.e.\ t\in[0,1],$$

for some Radon probability measure $\Lambda\in C^*(\mathcal{A})$ and some family of arcs $\{p(.;\alpha)\in W^{1,1}([0,1];R^n)\mid\alpha\in\mathcal{A}\}$ such that, for Λ – a.e. $\alpha\in\mathcal{A}$,

$$\begin{aligned} -\dot{p}(t;\alpha) &\in co\,\partial_x H(t,\bar{x}(t;\alpha),\bar{u}(t),p(t;\alpha),\alpha) \quad a.e.,\\ -p(1;\alpha) &\in \bigcup_{0\le r\le 1}\{rG_0(\bar{x}(1;\alpha),\alpha)+(1-r)N(\bar{x}(1;\alpha),\alpha)\} \end{aligned} \tag{2.1}$$

and

$$supp\,\Lambda \subset \{\alpha\mid either\ G_0(\bar{x}(1;\alpha),\alpha)\neq\emptyset\ or\ N(\bar{x}(1;\alpha),\alpha)\neq\emptyset\}.$$

Here,

$$G_0(x,\alpha) := \begin{cases}\partial_x g(x,\alpha) & \text{if } g(x,\alpha)=max_{\alpha'\in\mathcal{A}}\ g(x,\alpha')\\ \emptyset & \text{otherwise}\end{cases}$$

and

$$N(x,\alpha) := \{\xi\in N_{C(\alpha)}(x)\,|\,|\xi|=1\}.$$

In condition (2.1), we allow the possibilities that (for some values of α) $G_0(x,\alpha)=\emptyset$ or $N(x,\alpha)=\emptyset$. If $G_0(x,\alpha)=\emptyset$, then $rG_0(x,\alpha)$ is defined only if $r=0$; in this case $rG_0(x,\alpha):=\{0\}$. If $N(x,\alpha)=\emptyset$, then $(1-r)N(x,\alpha)$ is defined only if $r=1$; in this case $(1-r)N(x,\alpha):=\{0\}$. Thus (2.1) implies that if $\Lambda(\{\alpha\})>0$, then the parameter α is "active" in either the endpoint constraint or in the objective, in the sense that

$$g(\bar{x}(1,\alpha),\alpha)=\max_{\alpha'\in\mathcal{A}} g(\bar{x}(1,\alpha'),\alpha') \quad\text{or}\quad \bar{x}(1:\alpha)\in bdy\,C(\alpha)\,.$$

($bdy\,C(\alpha)$ denotes the "boundary of the set $C(\alpha)$.")

Proof of Proposition 2.1. List the elements in the finite set $\mathcal{A}$ as

$$\mathcal{A}=\{\alpha_1,\alpha_2,\ldots,\alpha_N\}.$$

Denote by $\bar{x}=col\,\{\bar{x}(.;\alpha_1),\bar{x}(.;\alpha_2),\ldots,\bar{x}(.;\alpha_N)\}$ the collection of state trajectories corresponding to $\bar{u}$. Then $(\bar{u},\bar{x})$ is a strong local minimizer for the standard optimal control problem

$$(\tilde{\mathrm{P}})\begin{cases}\text{Minimize } \tilde{g}(x(1)) \text{ over } u(.) \text{ satisfying}\\ \dot{x}(t)=\tilde{f}(t,x(t),u(t)) \quad a.e.\ t\in[0,1],\\ x(0)=\tilde{x}_0,\\ x(1)\in\tilde{C},\\ u(t)\in\Omega(t) \quad a.e.\ t\in[0,1]\,,\end{cases}$$

in which the $N \times n$ dimensional state vector is partitioned as

$$\begin{aligned} x &= col\ \{x_1, x_2, \ldots, x_N\}, \\ \tilde{f}(t,x,u) &= col\ \{f(t,x_i,u,\alpha_i)\}_{i=1}^N, \\ \tilde{C} &= C(\alpha_1) \times C(\alpha_2) \times \cdots \times C(\alpha_N), \\ \tilde{x}_0 &= col\,\{x_0, x_0, \ldots, x_0\}, \\ \tilde{g}(x) &= \max_i\ g(x(.;\alpha_i),\alpha_i). \end{aligned}$$

Under the stated hypotheses, we deduce from the nonsmooth maximum principle (see, for example, [10], Theorem 6.2.1)), with the help of the max rule ([10], Theorem 5.5.2) to evaluate the limiting subdifferential of the cost function $\tilde{g}$, the following information. There exist numbers $\lambda_1, \ldots, \lambda_N \geq 0$, arcs $q(.;\alpha_i) \in W^{1,1}$, and elements $\xi_i \in N_{C(\alpha_i)}(x(1;\alpha_i))$, $i = 1, 2, \ldots, N$, such that

$$\text{(i)} \sum_{i=1}^N H(t, \bar{x}(t;\alpha_i), \bar{u}(t), q(t;\alpha_i), \alpha_i) = \max_{u \in \Omega(t)}\ \sum_{i=1}^N H(t, \bar{x}(t;\alpha_i), u, q(t;\alpha_i), \alpha_i) \quad a.e.$$

$$\text{(ii)} \sum_{i=1}^N (\lambda_i + |\xi_i|)\ =\ 1$$

and, for each i,

$$\text{(iii)} -\dot{q}(t;\alpha_i)\ \in\ \text{co}\,\partial_x H(t, \bar{x}(t;\alpha_i), \bar{u}(t), q(t;\alpha_i), \alpha_i) \quad a.e.,$$

$$\text{(iv)} -q(1;\alpha_i)\ \in\ \lambda_i \partial_x g(\bar{x}(t;\alpha_i),\alpha_i) + \xi_i,$$

$$\text{(v)} \lambda_i = 0 \quad \text{if} \quad g(\bar{x}(1;\alpha_i),\alpha_i) < \max_j\ g(\bar{x}(1;\alpha_j),\alpha_j).$$

Define Λ to be the discrete probability measure

$$\Lambda = \sum_{i=1}^N (\lambda_i + |\xi_i|)\ \delta_{\alpha_i},$$

in which δ_{α_i} denotes the unit measure concentrated at $\alpha = \alpha_i$. If $\alpha \in supp\ \{\Lambda\}$, in which case $\alpha = \alpha_i$ for some i such that $(\lambda_i + |\xi_i|) > 0$, define

$$p(t;\alpha_i) = \frac{1}{\lambda_i + |\xi_i|}\ q(t;\alpha_i).$$

If $\alpha \notin supp\ \{\Lambda\}$, choose the $W^{1,1}$ function $p(.;\alpha)$ arbitrarily.
All the assertions of the proposition can be confirmed for this choice of Λ and $\{p(.;\alpha) \mid \alpha \in \mathcal{A}\}$.

Note, in particular, that, if $\alpha_i \in supp\ \{\Lambda\}$, then

$$-p(1;\alpha_i) \in r_i \partial_x g(\bar{x}(1;\alpha_i),\alpha_i) + (1 - r_i)\{\xi \in N_{C(\alpha_i)}(\bar{x}(1;\alpha_i)) \mid |\xi| = 1\}.$$

Here, r_i, $0 \leq r_i \leq 1$, is the number

$$r_i = \frac{\lambda_i}{\lambda_i + |\xi_i|}.$$

We also observe that, for each $t \in [0,1]$ and $u \in \Omega(t)$,

$$\sum_{i=1}^{N} H(t, \bar{x}(t;\alpha_i), u, q(t;\alpha_i), \alpha_i) = \int_{\mathcal{A}} H(t, \bar{x}(t;\alpha), u, p(t;\alpha), \alpha)\ \Lambda(d\alpha),$$

i.e., the maximization of the "averaged" Hamiltonian condition is satisfied. Finally, note that for Λ – *a.e.* $\alpha \in \mathcal{A}$,

$$-\dot{p}(t;\alpha)\ \in\ \operatorname{co} \partial_x H(t, \bar{x}(t;\alpha), \bar{u}(t), p(t;\alpha), \alpha),$$

by positive homogeneity. □

3. A maximum principle for the general minimax optimal control problem. This section provides necessary conditions of optimality for the general minimax optimal control problem (P) of section 1, when the parameter set $\mathcal{A}$ is an arbitrary compact metric space.

Take $(\bar{u}, \{\bar{x}(.;\alpha) \,|\, \alpha \in \mathcal{A}\})$ to be the local minimizer for problem (P) of interest. For $\alpha \in \mathcal{A}$, define the set

$$\begin{aligned} Q_0(\alpha) := \big\{ p(.;\alpha) \in W^{1,1} \mid & -\dot{p}(t;\alpha)\ \in\ \operatorname{co} \partial_x H(t, \bar{x}(t;\alpha), \bar{u}(t), p(t;\alpha), \alpha) \quad \textit{a.e.} \\ & \text{and } -p(1;\alpha) \in \cup_{r\in[0,1]} \left(rG_0(\bar{x}(1;\alpha),\alpha) + (1-r)N(\bar{x}(1;\alpha),\alpha)\right), \end{aligned}$$

in which, for $\epsilon \in [0,1]$,

$$G_\epsilon(x,\alpha) := \begin{cases} \partial_x g(x,\alpha) & \text{if } g(x,\alpha) \geq \max_{\alpha' \in \mathcal{A}}\ g(x,\alpha') - \epsilon \\ \emptyset & \text{otherwise} \end{cases} \tag{3.1}$$

and

$$N(x,\alpha) := \big\{ \xi \in N_{C(\alpha)}(x) \mid |\xi| = 1 \big\}. \tag{3.2}$$

(Only $G_{\epsilon=0}(x,0)$ is involved in the definition of $Q_0(\alpha)$. $G_\epsilon(x,\alpha)$, with $\epsilon > 0$, is required for later analysis.)

The assertions of Proposition 2.1 can be expressed in terms of the set $Q_0(\alpha)$ as follows. If $(\bar{u}, \{\bar{x}(.;\alpha) \mid \alpha \in \mathcal{A}\})$ is a strong local minimizer and $\mathcal{A}$ is a finite set, then

$$\begin{aligned} &\int_{\mathcal{A}} H(t, \bar{x}(t;\alpha), \bar{u}(t), p(t;\alpha), \alpha)\ \Lambda(d\alpha) = \\ &\qquad \max_{u\in\Omega(t)} \int_{\mathcal{A}} H(t, \bar{x}(t;\alpha), u, p(t;\alpha), \alpha)\ \Lambda(d\alpha) \quad \textit{a.e. } t \in [0,1] \end{aligned}$$

for some Radon probability measure $\Lambda \in C^*(\mathcal{A})$ and family of arcs $\{p(.;\alpha) \mid \alpha \in \mathcal{A}\}$ such that

$$p(.;\alpha) \in Q_0(\alpha) \ \text{ for } \Lambda - \textit{a.e. } \alpha \in \mathcal{A}.$$

(Note that $Q_0(\alpha)$ may be empty unless α is "active" in the sense of our earlier remarks.) Unfortunately, the above optimality condition no longer remains valid in general, when we allow $\mathcal{A}$ to be an arbitrary compact metric space. Confirmation is provided by the counter examples of section 5. Indeed, standard variational techniques break down when $\mathcal{A}$ is an infinite set, because the multifunction

$$Q_0(.)\ :\ \mathcal{A} \to \{\text{subsets of } W^{1,1}\}$$

may lack the requisite convexity and closure properties for limit taking. To derive necessary conditions in this more general context, we need to replace $Q_0(\alpha)$ with a larger set, better matched to the limit taking operations involved.

Let $(\bar{u}, \{\bar{x}(.;\alpha) \mid \alpha \in \mathcal{A}\})$ be the process of interest. We embed $Q_0(.)$ in a family of multifunctions $\{Q_\epsilon(.) | \epsilon \geq 0\}$ defined as follows. For any $\epsilon \geq 0$ and $\alpha \in \mathcal{A}$ we define

$$Q_\epsilon(\alpha) := \big\{p(.;\alpha) \in W^{1,1} \mid \text{conditions (a) and (b) below are satisfied}\big\}$$

in which

(a)

$$-\dot{p}(t;\alpha) \in \bigcup_{x \in \bar{x}(t;\alpha)+\epsilon B} \operatorname{co} \partial_x H(t, x, \bar{u}(t), p(t;\alpha), \alpha) \quad a.e.$$

(b)

$$-p(1;\alpha) \in \bigcup_{x \in \bar{x}(1;\alpha)+\epsilon B} \bigcup_{r \in [0,1]} (rG_\epsilon(x,\alpha) + (1-r)N(x,\alpha))$$

The sets $G_\epsilon(x,\alpha)$ and $N(x,\alpha)$ appearing in these conditions were defined in (3.1) and (3.2).

The defining properties of the "costate" arcs $p(.;\alpha)$ will now include the condition

$$p(.;\alpha) \ \in \ \overline{Q}_0(\alpha),$$

where

$$\overline{Q}_0(\alpha) := \bigcap_{\epsilon>0} \overline{co} \left(\bigcup_{\alpha' \in B_{\mathcal{A}}(\alpha,\epsilon)} Q_\epsilon(\alpha') \right). \tag{3.3}$$

Here $\overline{co}$ denotes "convex closure" with respect to the strong $W^{1,1}$ topology. Note that the right side is a subset of $W^{1,1}([0,1]; R^n)$. This relationship involves a multifunction that is obtained from the multifunction $\alpha \rightsquigarrow Q_0(\alpha)$ by enlarging its graph. The enlargement is carried out in such a manner that the new multifunction has closed graph and convex values.

In certain cases, notably when the data is smooth and the right endpoint constraints are absent,

$$Q_0(\alpha) \ = \ \overline{Q}_0(\alpha).$$

But in many cases of interest, $Q_0(\alpha)$ is a strict subset of its "closed, convexified" counterpart. We discuss these points in section 5.

We now come to the main result of this paper, namely a maximum principle for the general minimax optimal control problem. Here, as usual, the Hamiltonian is

$$H(t,x,p,u,\alpha) := p \cdot f(t,x,u,\alpha).$$

The following hypotheses will be invoked, in which $(\bar{u}, \{\bar{x}(.;\alpha) \mid \alpha \in \mathcal{A}\})$ is the strong local minimizer for (P) of interest. For some $\delta > 0$,

(S1) The function $f(.,x,.,.)$ is $\mathcal{L} \times \mathcal{B}^m \times \mathcal{B}_{\mathcal{A}}$ measurable for each $x \in R^n$. ($\mathcal{B}_{\mathcal{A}}$ denotes the Borel subsets of $\mathcal{A}$.) $t \rightsquigarrow \Omega(t)$ has a Borel measurable graph.

(S2) There exists $k_f \in L^1$ and $c_f > 0$ such that

$$|f(t,x,u,\alpha) - f(t,x',u,\alpha)| \le k_f(t)|x - x'| \quad \text{and} \quad |f(t,x,u,\alpha)| \le c_f$$

for all $x, x' \in \bar{x}(t;\alpha) + \delta B$, $u \in \Omega(t)$ and $\alpha \in \mathcal{A}$, *a.e.* $t \in [0,1]$.

(S3) g is continuous and there exists $k_g > 0$ such that

$$|g(x,\alpha) - g(x',\alpha)| \le k_g|x - x'|$$

for all $x, x' \in \bar{x}(1;\alpha) + \delta B$ and $\alpha \in \mathcal{A}$.

(S4) There exists $\theta : [0,+\infty) \to [0,+\infty)$ such that $\lim_{s\downarrow 0} \theta(s) = 0$ and, for all $\alpha, \alpha' \in \mathcal{A}$,

$$\int_0^1 \sup_{x\in\bar{x}(t)+\delta B,\ u\in\Omega(t)} |f(t,x,u,\alpha) - f(t,x,u,\alpha')|\ dt \le \theta(\rho_{\mathcal{A}}(\alpha,\alpha')).$$

(S5) $\alpha \to d_{C(\alpha)}(x)$ is continuous on $\mathcal{A}$ for each $x \in R^n$.

In the following theorem, $\mathcal{A}$ is an arbitrary compact metric space.

THEOREM 3.1. *Let* $(\bar{u}, \{\bar{x}(.;\alpha) \mid \alpha \in \mathcal{A}\})$ *be a strong local minimizer for* (P). *Assume that, for some* $\delta > 0$, *Hypotheses* (S1)–(S5) *are satisfied.*

Then

$$\int H(t,\bar{x}(t;\alpha),\bar{u}(t),p(t;\alpha),\alpha)\ \Lambda(d\alpha)$$

$$= \max_{u\in\Omega(t)} \int H(t,\bar{x}(t;\alpha),u,p(t;\alpha),\alpha)\ \Lambda(d\alpha) \quad \textit{a.e. } t \in [0,1], \tag{3.4}$$

for some Radon probability measure $\Lambda \in C^*(\mathcal{A})$ *and family of arcs* $\{p(.;\alpha) \in W^{1,1} \mid \alpha \in \mathcal{A}\}$ *such that, for* Λ – *a.e.* $\alpha \in \mathcal{A}$,

$$p(.;\alpha)\ \in\ \overline{Q}_0(\alpha). \tag{3.5}$$

(Recall the definition of $\overline{Q}_0(\alpha)$ *in* (3.3).*)*

Note that the right side of (3.5) may be empty for certain values of α. The set is nonempty, however, on a set of full Λ measure.

Implicit in the optimality conditions is the assertion that the integrals in the maximization of the Hamiltonian condition (3.4) are well-defined, i.e., the function $\alpha \to H(t,\bar{x}(t;\alpha),u,p(t;\alpha),\alpha)$ is Λ-integrable for each $u \in \Omega(t)$, *a.e.* $t \in [0,1]$.

We might expect that necessary conditions of optimality are valid for a hybrid form of the minimax optimal control problem, in which the parameter set $\mathcal{A}$ separates into the union of a "discrete" and a "continuous" set, and which specializes to a version of Proposition 2.1 (valid under the stronger hypotheses of Theorem 3.1) and Theorem 3.1 in the extreme cases "$\mathcal{A}$ is purely discrete" and "$\mathcal{A}$ is purely continuous." The following theorem supplies such conditions. We explore some consequences in section 5.

THEOREM 3.2. *Let* $(\bar{u}, \{\bar{x}(.;\alpha) \mid \alpha \in \mathcal{A}\})$ *be a strong local minimizer for the general minimax optimal control problem* (P). *Assume that Hypotheses* (S1)–(S5) *are satisfied. Assume, furthermore, we can partition the compact metric space* $\mathcal{A}$ *into disjoint sets*

$$\mathcal{A} = \mathcal{A}^{(1)} \cup \mathcal{A}^{(2)},$$

in which $\mathcal{A}^{(1)}$ *is a compact metric space and* $\mathcal{A}^{(2)}$ *is a finite set.*

Then

$$\int H(t,\bar{x}(t;\alpha),\bar{u}(t),p(t;\alpha),\alpha)\ \Lambda(d\alpha)$$
$$= \max_{u\in\Omega(t)} \int H(t,\bar{x}(t;\alpha),u,p(t;\alpha),\alpha)\ \Lambda(d\alpha) \quad \textit{a.e. } t\in[0,1]$$

for some Radon probability measure $\Lambda \in C^*(\mathcal{A})$ *and family of arcs* $\{p(.;\alpha)\in W^{1,1} \mid \alpha\in\mathcal{A}\}$ *such that*

$$p(.;\alpha)\ \in\ \overline{Q}_0(\alpha) \quad \textit{for } \Lambda-\textit{a.e. } \alpha\in\mathcal{A}^{(1)}$$

and

$$p(.;\alpha)\ \in\ Q_0(\alpha) \quad \textit{for } \Lambda-\textit{a.e. } \alpha\in\mathcal{A}^{(2)}.$$

We conclude this section by stating a version of the foregoing theorems covering problems in which the endpoint constraints take the form of a finite collection of functional inequality constraints, namely problems for which each $C(\alpha)$ has the representation

$$C(\alpha)\ =\ \{x\in R^n \mid \psi(x,\alpha)\le 0\}, \tag{3.6}$$

for some function $\psi : R^n\times\mathcal{A}\to R^r$. The inequalities are interpreted in a "componentwise" sense. It will be assumed that, for some $\delta>0$, ψ satisfies the following hypothesis:

(H) ψ is continuous and there exist k_ψ such that

$$|\psi(x,\alpha)-\psi(x',\alpha)|\le k_\psi|x-x'| \text{ for all } x,x'\in\bar{x}(1;\alpha)+\delta B,\ \alpha\in\mathcal{A}.$$

Minor modifications to the proof of Theorems 3.1 and 3.2 yield the following optimality condition for problems involving endpoint functional inequality constraints:

THEOREM 3.3. *Let* $(\bar{u},\{\bar{x}(.;\alpha)\mid\alpha\in\mathcal{A}\})$ *be a strong local minimizer for* (P). *Assume that the endpoint constraint sets* $\{C(\alpha)\mid\alpha\in\mathcal{A}\}$ *take the form of a collection of functional inequality constraints* (3.6) *which satisfy Hypothesis* (H). *Then*

$$\int H(t,\bar{x}(t;\alpha),\bar{u}(t),p(t;\alpha),\alpha)\ \Lambda(d\alpha)$$
$$= \max_{u\in\Omega(t)} \int H(t,\bar{x}(t;\alpha),u,p(t;\alpha),\alpha)\ \Lambda(d\alpha) \quad \textit{a.e. } t\in[0,1],$$

for some Radon probability measure $\Lambda\in C^*(\mathcal{A})$ *and family of arcs* $\{p(.;\alpha)\in W^{1,1}\mid\alpha\in\mathcal{A}\}$ *such that,*

(a) *if* $\mathcal{A}$ *is a finite set and Hypotheses* (H1)–(H3) *are satisfied, then*

$$p(.;\alpha)\ \in\ Q_0^\psi(\alpha) \quad \textit{for } \Lambda-\textit{a.e. } \alpha\in\mathcal{A}.$$

(b) *if* $\mathcal{A}$ *is a compact metric space and Hypotheses* (S1)–(S4) *are satisfied, then*

$$p(.;\alpha)\in\bigcap_{\epsilon>0}\overline{co}\left(\bigcup_{\alpha'\in B_{\mathcal{A}}(\alpha,\epsilon)} Q_\epsilon^\psi(\alpha')\right) \quad \textit{for } \Lambda-\textit{a.e. } \alpha\in\mathcal{A}.$$

(c) *if Hypotheses* (S1)–(S4) *are satisfied and we can partition* $\mathcal{A} \subset R^k$ *into disjoint sets* $\mathcal{A} = \mathcal{A}^{(1)} \cup \mathcal{A}^{(2)}$, *in which* $\mathcal{A}^{(1)}$ *is a compact metric space and* $\mathcal{A}^{(2)}$ *is a finite set, then*

$$p(.;\alpha) \in \bigcap_{\epsilon>0} \overline{co}\left(\bigcup_{\alpha' \in B_{\mathcal{A}}(\alpha,\epsilon)} Q_\epsilon^\psi(\alpha') \right) \quad \text{for } \Lambda - \text{a.e. } \alpha \in \mathcal{A}^{(1)} \tag{3.7}$$

and

$$p(.;\alpha) \in Q_0^\psi(\alpha) \quad \text{for } \Lambda - \text{a.e. } \alpha \in \mathcal{A}^{(2)}.$$

In the above optimality conditions the set $Q_\epsilon^\psi(\alpha)$, $\epsilon \geq 0$, shares the defining relationships of $Q_\epsilon(\alpha)$ (see (3.3)) in all respects except that the set $N(x,\alpha)$ in condition (b), namely

$$N(x,\alpha) = \{\xi \in N_{C(\alpha)}(x) \mid |\xi| = 1\},$$

is replaced by the set

$$N^\psi(x,\alpha) := \left\{ \sum_j \lambda_j \nabla_x \psi_j(x,\alpha) \mid (\lambda_1,\ldots,\lambda_r) \in \mathcal{S}(r) \text{ such that } \lambda_i = 0 \text{ if } \psi_i(x,\alpha) < 0 \right\}.$$

in which

$$\mathcal{S}(r) := \left\{ \lambda \in R^r \mid \lambda_i \geq 0,\ i = 0,\ldots,r \text{ and } \sum_{i=0}^r \lambda_i = 1 \right\}.$$

It is a straightforward matter to derive variants on Theorem 3.3. We could, for example, assume that the endpoint constraint sets $C(\alpha)$ take the form $\{x \mid \psi(x,\alpha) \leq 0\}$ for $\alpha \in \mathcal{A}^{(2)}$ and are arbitrary closed sets for $\alpha \in \mathcal{A}^{(1)}$. In this case the necessary conditions will incorporate transversality conditions from both Theorems 3.2 and 3.3.

4. Discussion. Theorem 3.1 captures only a coarse version of Proposition 2.1 when specialized to the finite set case. This is because Proposition 2.1 asserts the existence of costate arcs in the sets $Q_0(\alpha)$, $\alpha \in \mathcal{A}$, with respect to which an "averaged" maximum principle is valid. On the other hand, Theorem 3.1 asserts the existence of costates with this property, chosen from the larger sets $\overline{Q}_0(\alpha)$, obtained by convexifying the values of $\alpha \to Q_0(\alpha)$ and closing its graph, in some sense. Minimax maximum principles involving $\overline{Q}_0(\alpha)$ provide significantly less information about minimizers than those involving $Q_0(\alpha)$. For further elucidation of this point, consider the case of (P) when the endpoint constraint sets are

$$C(\alpha) = \{x \mid \psi(x,\alpha) = 0\} \quad \text{for all } \alpha \in \mathcal{A}.$$

Here $\psi : R^n \times \mathcal{A} \to R$ is a given function. Assume that, for some fixed $\bar{\alpha}$, $g(.,\bar{\alpha})$, $\psi(.,\bar{\alpha})$ and $f(t,.,u,\bar{\alpha})$ are smooth functions and that $(\bar{u}, \{\bar{x}(.;\alpha) \mid \alpha \in \mathcal{A}\})$ is a feasible process for which

(A): $g_x(\bar{x}(1;\bar{\alpha}),\bar{\alpha})$ and $\psi_x(\bar{x}(1;\bar{\alpha}),\bar{\alpha})$ are linearly independent.

Let n be the vector of unit length

$$n = \frac{\psi_x(\bar{x}(1;\bar{\alpha}),\bar{\alpha})}{|\psi_x(\bar{x}(1;\bar{\alpha}),\bar{\alpha})|}.$$

Then we easily calculate that

$$Q_0(\bar\alpha) \;=\; \Big\{p(.;\bar\alpha)\in W^{1,1} \;|\; -\dot p(t;\bar\alpha)=H_x \text{ and } -p(1;\bar\alpha)\in co\{g_x,+n\}\cup co\{g_x,-n\}\Big\}$$

while $\overline{Q}_0(\bar\alpha)$ contains the subset

$$\Big\{p(.;\bar\alpha)\in W^{1,1} \;|\; -\dot p(.;\bar\alpha)=H_x \text{ and } -p(.;\bar\alpha)\in \text{co}\{g_x,+n,-n\}\Big\}. \tag{4.1}$$

Here g_x is evaluated at $\bar x(1;\bar\alpha)$.

Notice that the element $p(.;\bar\alpha)\equiv 0$ lies in the set (4.1), since $-\dot p(t;\bar\alpha)=H_x$ is a linear differential equation and $0\in\text{co}\{g_x,+n,-n\}$. This means that the optimality conditions of Theorem 3.1 are satisfied by any feasible process $(\bar u,\{\bar x(.;\alpha)\,|\,\alpha\in\mathcal{A}\})$ with the trivial choice of multipliers

$$\Lambda=\delta_{\{\bar\alpha\}} \text{ and } p(.;\alpha)\equiv 0 \quad \text{ for all } \alpha\in\mathcal{A}.$$

Theorem 3.1 therefore conveys no useful information about minimizers in this case. By contrast, we have

$$(p(.;\bar\alpha)\equiv 0) \;\notin\; Q_0(\bar\alpha)$$

since, under the hypothesis (A), $0\notin\text{co}\{g_x,+n\}\cup\text{co}\{g_x,-n\}$; thus the optimality conditions of Theorem 3.1 are not, in this case, automatically satisfied by any feasible process $(u,\{x(.;\alpha)\mid\alpha\in\mathcal{A}\})$.

On the other hand, consider a modification of the above special case, in which the former equality endpoint constraints are replaced by inequality constraints

$$C(\alpha)=\{x\mid\psi(x,\alpha)\le 0\}$$

and assume that

$$\psi(\bar x(1;\bar\alpha),\bar\alpha) \;=\; 0.$$

Then, under unrestrictive hypotheses,

$$\begin{aligned} Q_0(\bar\alpha)&=\overline{Q}_0(\bar\alpha)\\ &=\Big\{p(.;\bar\alpha)\in W^{1,1}\;|\;-\dot p(.;\bar\alpha)=H_x \text{ and } -p(1;\bar\alpha)\in\text{co}\{\nabla g,n\}\Big\}.\end{aligned}$$

Here, there is no loss of information in passing from $Q_0(\bar\alpha)$ to $\overline{Q}_0(\alpha)$.

These observations highlight the fact that the maximum principle for minimax optimal control problems with parameter set a general compact metric space, Theorem 3.1, will find primary application in situations where the endpoint constraints (if they are present) take the form of functional inequality constraints and their generalizations. Theorem 3.1 is not well-suited to problems with endpoint equality constraints.[1] It is therefore of interest to know whether Theorem 3.1 can be refined,

[1] Of course it can be argued that minimax problems of this nature are, broadly speaking, artificial: often such problems will have no minimizers because of the absence of feasible processes, i.e., control functions whose corresponding state trajectories satisfy the equality endpoint constraints for *all* values of the parameter α. Nontrivial maximum principles covering those few cases of interest involving equality endpoint constraints (e.g., cases where the equality constraints involve only those aspects of the dynamics which do not depend on α) can be developed along the lines of Theorem 3.2.

to provide necessary conditions for problems with parameter set a general compact metric space, in which $Q_0(\alpha)$ replaces $\overline{Q}_0(\alpha)$.

We now study two examples, the purpose of which is to demonstrate that this is not possible, in the absence of additional hypotheses.

EXAMPLE 4.1. *Consider the problem*

$$\begin{cases} \textit{Minimize } \sup_{\alpha\in[-1,+1]} -|x(1)-\alpha| \textit{ over } u(.) \textit{ such that} \\ \dot{x}(t) = u(t) \quad \textit{a.e. } t \in [0,1], \\ x(0) = 0, \\ u(t) \in [-1,+1] \quad \textit{a.e. } t \in [0,1]\,. \end{cases}$$

This is an example of the general minimax problem in which the parameter set is the interval $\mathcal{A} = [-1,1]$. *The cost function depends on* α, *but the dynamics do not. We denote processes* (u,x), *since all state trajectories corresponding to a given control function* u *coincide. Clearly,* $(\bar{u} \equiv 0,\ \bar{x} \equiv 0)$ *is a minimizer.*

Suppose that the assertions of Proposition 2.1 *were valid here. Then there would exist a probability measure* Λ *with support in*

$$\left\{\alpha \mid -|\bar{x}(1)-\alpha| = max_{\alpha'\in[-1,+1]}\left(-|\bar{x}(1)-\alpha'|\right)\right\} = \{0\}, \tag{4.2}$$

and a family of costate arcs $\{p(.;\alpha)\,|\,\alpha\in\mathcal{A}\}$ *such that* $p(.;\alpha)\in Q_0(\alpha)$ *for* Λ – *a.e.* $\alpha\in\mathcal{A}$ *and* (3.3) *is satisfied. But* (4.2) *implies that*

$$\Lambda = \delta_{\{0\}}.$$

Thus, supp $\{\Lambda\} = \{0\}$ *and the only relevant value of* α *is* $\alpha = 0$. *We calculate*

$$\begin{aligned} Q_0(\alpha=0) &= \left\{p\in W^{1,1} \mid -\dot{p} = 0,\ \ -p(1)\in\{-1\}\cup\{+1\}\right\} \\ &= \{p\equiv -1\}\cup\{p\equiv +1\}. \end{aligned}$$

We have then, for each $t\in[0,1]$,

$$\int_{\mathcal{A}} H(t,\bar{x}(t),u,p(t;\alpha),\alpha)\ \Lambda(d\alpha) = \begin{cases} +u & \text{if } p(.;\alpha=0)\equiv +1 \\ -u & \text{if } p(.;\alpha=0)\equiv -1, \end{cases}$$

for any $u\in[-1,+1]$ *and any family of costate arcs* $\{p(.;\alpha)\mid\alpha\in\mathcal{A}\}$ *such that* $p(.;\alpha)\in Q_0(\alpha)$ *for* Λ – *a.e.* $\alpha\in\mathcal{A}$.

We see that $u\to\int_{\mathcal{A}} H(t,\bar{x}(t),u,p(t),\alpha)\ \Lambda(d\alpha)$ *cannot be maximized at* $u=\bar{u}(t)$ *for a.e.* $t\in[0,1]$. *This shows that the assertions of Theorem* 3.1 *may fail to be true, if* $\mathcal{A}$ *is allowed to be an infinite set. On the other hand,*

$$\{p(.;\alpha=0)\equiv 0\}\ \in\ \overline{Q}_0(\alpha=0)$$

and so the maximization of the Hamiltonian condition is satisfied with Λ *taken to be the unit measure concentrated at* $\alpha=0$ *and with* $\{p(.;\alpha)|\alpha\in\mathcal{A}\}$ *an arbitrary collection of* $W^{1,1}$ *functions such that* $p(.;\alpha=0)\equiv 0$.

Example 4.1 involves nonsmooth data. The following more elaborate example illustrates that we cannot replace $p(\alpha)\in\overline{Q}_0(\alpha)$ by $p(\alpha)\in Q_0(\alpha)$, even for problems with smooth data.

EXAMPLE 4.2. *Consider the following example of the minimax optimal control problem, in which the state $x = (x_1, x_2)$ is a 2-vector and the control u is scalar:*

$$(\mathrm{P})\ \begin{cases} \textit{Minimize } \max_{\alpha\in\mathcal{A}} g(x(1;\alpha),\alpha) \\ \quad \textit{over } (u, \{x(.;\alpha) \mid \alpha\in\mathcal{A}\}) \textit{ satisfying} \\ \dot{x}(t;\alpha) = f(t, x(t;\alpha), u(t), \alpha) \quad \textit{a.e.,} \\ u(t)\in\Omega \quad \textit{a.e.,} \\ x(0;\alpha) = x_0 \quad \textit{and} \quad x(1;\alpha)\in C(\alpha). \end{cases}$$

Here, $x_0 = col\,(0;0)$, $\Omega = [-1,+1]$,

$$\mathcal{A} = \left[\frac{1}{3}, \frac{2}{3}\right] \cup \{1\},$$

and, for each $\alpha \in [\frac{1}{3}, \frac{2}{3}] \cup \{1\}$, $f = col(f_1, f_2)$ is the function

$$f_1(t,x,u,\alpha) = \begin{cases} u & \textit{if } 0\le t\le\alpha \\ 0 & \textit{if } \alpha < t \le 1 \end{cases}$$

$$f_2(t,x,u,\alpha) = \begin{cases} -u^2 & \textit{if } \frac{1}{3}\le t\le\frac{2}{3} \\ 0 & \textit{otherwise,} \end{cases}$$

$$g(x,\alpha) = \begin{cases} x_2 & \textit{if } \alpha = 1 \\ -1 & \textit{if } \alpha\in[\frac{1}{3},\frac{2}{3}] \end{cases}$$

and

$$C(\alpha) = \begin{cases} \{0\}\times R & \textit{if } \alpha\in[\frac{1}{3},\frac{2}{3}] \\ R\times R & \textit{if } \alpha = 1. \end{cases}$$

Noting the interpretation of this example provided below, we easily check that ($\bar{u}\equiv 0$, $\{\bar{x}(.;\alpha)\equiv(0,0)\,|\,\alpha\in\mathcal{A}\}$) is a minimizer. Suppose that

$$\int_{\mathcal{A}} H(t,\bar{x}(t;\alpha),\bar{u}(t),p(t;\alpha),\alpha)\ \Lambda(d\alpha)$$
$$= \max_u \int_{\mathcal{A}} H(t,\bar{x}(t;\alpha),u,p(t;\alpha),\alpha)\ \Lambda(d\alpha) \quad \textit{a.e. } t\in[0,1]$$

is satisfied for some probability measure Λ and family of arcs $\{p(.:\alpha)\}$ such that

$$p(.;\alpha)\in Q_0(\alpha) \qquad \Lambda - a.e. \quad \alpha\in\mathcal{A}.$$

Partition the adjoint arcs $p(.;\alpha) = (p_1(.;\alpha), p_2(.;\alpha))$. Then for Λ – a.e. $\alpha\in[\frac{1}{3},\frac{2}{3}]$

$$\begin{array}{ll} -\dot{p}_1(.;\alpha)\equiv 0, & -\dot{p}_2(.;\alpha)\equiv 0 \\ -p_1(1;\alpha) = m(\alpha), & -p_2(1;\alpha) = 0 \end{array}$$

in which $m(\alpha)$ is a Borel measurable function such that

$$m(\alpha) = -1 \textit{ or } +1 \quad \textit{for all } \alpha\in\left[\frac{1}{3},\frac{2}{3}\right].$$

Also

$$-\dot{p}_1(.:\alpha=1)\equiv 0,\ -p_1(1;\alpha=1) = 0$$
$$-\dot{p}_2(.:\alpha=1)\equiv 0,\ -p_2(1;\alpha=1) = +1.$$

Writing $a \vee b := max\{a, b\}$, *we deduce from the maximization of the Hamiltonian condition that*

$$u \to \int_{[\frac{1}{3} \vee t, \frac{2}{3}]} m(\alpha)\ \Lambda(d\alpha)u + u^2 \chi_{[\frac{1}{3},\ \frac{2}{3}]}(t)\Lambda(\{1\})$$

is maximized over $[-1, +1]$ *at* $u = 0$ *a.e.* $t \in [0, 1]$. *Here* χ_D *denotes the indicator function of the set* D. *It follows that, for a.e.* $t \in [0, 1]$,

$$\int_{[\frac{1}{3} \vee t, \frac{2}{3}]} m(\alpha)\ \Lambda(d\alpha)\ =\ 0 \tag{4.3}$$

and $\Lambda(\{1\}) = 0$. *But* (4.3) *implies*

$$\int_{[\frac{1}{3}, \frac{2}{3}]} m(\alpha)\ \Lambda(d\alpha) = 0$$

and

$$\int_{[t, \frac{2}{3}]} m(\alpha)\ \Lambda(d\alpha) = 0$$

for all $t \in F$, *where* F *is some countable dense subset of* $[\frac{1}{3}, \frac{2}{3}]$. *But since sets of the form* $[\frac{1}{3}, \frac{2}{3}]$ *and* $[t, \frac{2}{3}]$ *(for* $t \in F$*) generate the Borel subsets of* $[\frac{1}{3}, \frac{2}{3}]$, *we see that*

$$\int_B m(\alpha)\ \Lambda(d\alpha) = 0 \tag{4.4}$$

for all Borel sets $B \in \mathcal{A}$. *Let* $\mathcal{A}^{\pm} = \{\alpha \mid m(\alpha) = \pm 1\}$. *Since* $\mathcal{A}^- \cup \mathcal{A}^+ = \mathcal{A}$ *and* $||\Lambda||_{T.V.} = 1$, *either* $\Lambda(\mathcal{A}^+) > 0$ *or* $\Lambda(\mathcal{A}^-) > 0$. *Without loss of generality assume the former. Then*

$$\int_B m(\alpha)\ \Lambda(d\alpha) = \int_B\ \Lambda(d\alpha)\ =\ \Lambda(\mathcal{A}^+)\ > 0,$$

when we select $B = \mathcal{A}^+$. *This contradicts* (4.4). *It follows that a version of the minimax maximum principle is not valid for this problem, in which*

$$p(.; \alpha) \in Q_0(\alpha) \quad \textit{for } \Lambda - a.e.\ \ \alpha \in \mathcal{A}.$$

The preceding example originates in an optimal control problem with pathwise equality constraints

$$(\tilde{\mathrm{P}}) \begin{cases} & \text{Minimize} \quad -\int_{\frac{1}{3}}^{\frac{2}{3}} |u(t)|^2\ dt \\ \text{s.t.} & \dot{x}(t) = u(t), \quad a.e.\ t \in [0, 1], \\ & x(0) = 0, \\ & x(t) = 0 \quad \text{for } \frac{1}{3} \le t \le \frac{2}{3}, \\ & u(t) \in R, \quad a.e.\ t \in [0, 1]. \end{cases}$$

which has been reformulated as an example of the general minimax optimal control problem (P). The fact that we cannot derive a maximum principle involving the set $Q_0(\alpha)$ in the above example reflects the fact that measure multipliers can be used

in necessary conditions for problems with pathwise *equality* constraints only in very special circumstances.

Notice that the assertions of Theorem 3.1 are consistent with Example 4.2. In this example

$$Q_0(\alpha) \;=\; \big\{(p_1(.;\alpha)\equiv 0, p_2(.;\alpha)) \mid p_2(.;\alpha)\equiv m(\alpha)\big\},$$

for Λ – a.e. $\alpha \in [\frac{1}{3}, \frac{2}{3}]$, where $m(.)$ is some Borel measurable function such that

$$m(\alpha) \;\in\; \{-1\}\cup\{+1\} \quad \Lambda - a.e.$$

These sets are too small for the maximization condition on the "averaged" Hamiltonian to hold (for any choice of $m(.)$). On the other hand, for Λ – a.e. $\alpha \in [\frac{1}{3}, \frac{2}{3}]$,

$$\overline{Q}_0(\alpha) \;=\; \big\{(p_1(.;\alpha)\equiv 0, p_2(.;\alpha)) \mid p_2(.;\alpha)\equiv \tilde{m}(\alpha)\big\}$$

in which $\tilde{m}(.)$ is some Borel measurable function such that

$$\tilde{m}(\alpha) \;\in\; [-1,+1] \quad \Lambda - a.e.$$

The maximization condition does hold (in a trivial sense), with respect to $(p_1(.;\alpha), p_2(.;\alpha))$s chosen from this larger set; we can take $(p_1(.;\alpha), p_2(.;\alpha)) \equiv (0,0)$ Λ – *a.e.*

5. Special cases. In this section, we examine implications of the minimax maximum principle for a number of special cases of interest. Utmost generality is not a goal here; indeed, we often focus on smooth versions of the optimality conditions, when the nonsmooth version could easily be derived, better to reveal their essential character. Throughout, $\mathcal{A}$ is an arbitrary compact metric space.

Consider first the *minimax optimal control problem with no right endpoint constraints*,

$$\text{(P1)}\quad \begin{cases} \text{Minimize } \max_{\alpha\in\mathcal{A}} g(x(.;\alpha),\alpha) \\ \qquad \text{over measurable functions } u:[0,1]\to R^m \text{ such that} \\ u(t)\in\Omega(t), \quad a.e.\ t\in[0,1] \\ \qquad \text{and arcs } \{x(.;\alpha):[0,1]\to R^n \mid \alpha\in\mathcal{A}\} \text{ such that, for each } \alpha\in\mathcal{A}, \\ \dot{x}(t;\alpha) = f(t,x(t;\alpha),u(t),\alpha) \quad a.e.\ t\in[0,1], \\ x(0;\alpha) = x_0. \end{cases}$$

The data for (P1) comprises a compact metric space $\mathcal{A}$, functions $g: R^n\times\mathcal{A}\to R$, $f:[0,1]\times R^n\times R^m\times\mathcal{A}\to R^n$, a vector $x_0\in R^n$, and a time dependent set $\Omega(t)\subset R^m$, $0\le t\le 1$.

General necessary conditions for (P1) follow directly from Theorem 3.1. We state the conditions merely in the special case when the data are smooth.

PROPOSITION 5.1. *Let $(\bar{u}, \{\bar{x}(.;\alpha) \mid \alpha\in\mathcal{A}\})$ be a strong local minimizer for* (P1). *Assume that, for some $\delta>0$, the Hypotheses* (S1), (S2), *and* (S4) *of section* 3 *are satisfied. Assume, furthermore, that*

(a) *g is continuous, $g(.,\alpha)$ is differentiable for each $\alpha\in\mathcal{A}$ and g_x is continuous.*

(b) *$f(t,.,u,\alpha)$ is continuously differentiable on a neighborhood of $\bar{x}(t;\alpha)$ for all $u\in\Omega(t)$ and $\alpha\in\mathcal{A}$, a.e. $t\in[0,1]$, and $\alpha\to f_x(t,x,u,\alpha)$ is uniformly continuous with respect to $(t,x,u)\in\{(t',x',u')\in[0,1]\times R^n\times R^m \mid u'\in\Omega(t')\}$.*

Then

$$\int H(t,\bar{x}(t;\alpha),\bar{u}(t),p(t;\alpha),\alpha)\ \Lambda(d\alpha)$$
$$= \max_{u\in\Omega(t)} \int H(t,\bar{x}(t;\alpha),u,p(t;\alpha),\alpha)\ \Lambda(d\alpha) \quad a.e.$$

for some Radon probability measure $\Lambda \in C^*(\mathcal{A})$ *and some family of arcs* $\{p(.;\alpha) \in W^{1,1}([0,1];R^n) \mid \alpha \in \mathcal{A}\}$ *such that*

$$supp\ \{\Lambda\} \subset \left\{\alpha \in \mathcal{A} \mid g(\bar{x}(1;\alpha),\alpha) = \max_{\alpha'\in\mathcal{A}}\ g(\bar{x}(1;\alpha'),\alpha')\right\}$$

and, for Λ *– a.e.* $\alpha \in \mathcal{A}$,

(i) $-\dot{p}(t;\alpha) = f_x^T(t,\bar{x}(t;\alpha),\bar{u}(t),\alpha)\,p(t;\alpha) \quad a.e.$

(ii) $-p(1;\alpha) = g_x(\bar{x}(1;\alpha),\alpha).$

Proof. Everything follows from Thm. 3.1, when we note that, if a function $\phi : R^n \to R$ is continuously differentiable on a neighborhood of a point $\bar{x}$, then $\partial\phi(\bar{x}) = \{\phi_x(\bar{x})\}$ and, under the stated hypotheses,

$$\overline{Q}_0(\alpha) = \begin{cases} \left\{p' \in W^{1,1} \mid -\dot{p}' = f_x^T p',\ -p'(1) = g_x(\bar{x}(1;\alpha),\alpha)\right\} \\ \qquad\qquad \text{if } g(\bar{x}(1;\alpha)) = \max_{\alpha'\in\mathcal{A}} g(\bar{x}(1;\alpha'),\alpha') \\ \emptyset \qquad\qquad \text{otherwise.} \end{cases} \quad \square$$

Consider next the *optimal control problem with robust feasibility constraints*:

$$\text{(P2)} \begin{cases} \text{Minimize } g(x(1;\alpha^*)) \\ \quad \text{over measurable functions } u:[0,1]\to R^m \text{ such that} \\ u(t) \in \Omega(t), \quad a.e.\ t \in [0,1] \\ \quad \text{and arcs } \{x(.;\alpha):[0,1]\to R^n \mid \alpha\in\mathcal{A}\} \text{ such that, for each } \alpha\in\mathcal{A}, \\ \dot{x}(t;\alpha) = f(t,x(t;\alpha),u(t),\alpha) \quad a.e.\ t\in[0,1], \\ x(0;\alpha) = x_0 \quad \text{and} \quad \psi(x(1;\alpha)) \le 0. \end{cases}$$

The data for (P2) comprises a set $\mathcal{A} \subset R^k$, a point $\alpha^* \in \mathcal{A}$, functions $g : R^n \to R$ and $f : [0,1] \times R^n \times R^m \times \mathcal{A} \to R^n$ and $\psi : R^n \to R^{r'}$, a vector $x_0 \in R^n$ and a time dependent set $\Omega(t) \subset R^m$, $0 \le t \le 1$. The endpoint functional inequality terminal constraint is interpreted in the usual "componentwise" manner.

This is a formulation of optimal control problems involving an unknown parameter α, in which α is expected to take its nominal value α^*. Here, it is appropriate to choose a control to minimize the cost based on the system response for $\alpha = \alpha^*$. But our choice of control is restricted by the requirement that, even if α deviates from α^*, constraints on state variables must not be violated. Here, we regard values of α different from α^* as due to system degradation or failure, and "$\psi(x(1;\alpha)) \le 0$ for all $\alpha \in \mathcal{A}$" is the requirement that operational constraints (on displacements, velocities, pressures, etc.) are satisfied, even in the event of breakdown.

For simplicity, we assume that the data are smooth and that there is a single endpoint constraint ($r' = 1$).

PROPOSITION 5.2. *Let* $(\bar{u}, \{\bar{x}(.;\alpha) \mid \alpha \in \mathcal{A}\})$ *be a strong local minimizer for* (P2). *Assume that, for some* $\delta > 0$, *Hypotheses* (S1), (S2), *and* (S4) *are satisfied. Assume, furthermore, that* $r' = 1$ *and*

(a) g *and* ψ *are continuously differentiable on* $\bar{x}(1;\alpha^*) + \delta B$.

(b) *$f(t,.,u,\alpha)$ is continuously differentiable on a neighborhood of $\bar{x}(t;\alpha)$ for all $u \in \Omega(t)$ and $\alpha \in \mathcal{A}$, a.e. $t \in [0,1]$, and $\alpha \to f_x(t,x,u,\alpha)$ is uniformly continuous with respect to $(t,x,u) \in \{(t',x',u') \in [0,1] \times R^n \times R^m \mid u' \in \Omega(t')\}$.*

Then

$$\int H(t,\bar{x}(t;\alpha),\bar{u}(t),p(t;\alpha),\alpha)\ \Lambda(d\alpha)\ = \max_{u\in\Omega(t)} \int H(t,\bar{x}(t;\alpha),u,p(t;\alpha),\alpha)\ \Lambda(d\alpha) \quad \text{a.e. } t \in [0,1]$$

for some family of arcs $\{p(.;\alpha) \in W^{1,1}([0,1];R^n)\}$, a number $r \in [0,1]$ and a Radon probability measure $\Lambda \in C^*(\mathcal{A})$ such that

$$\text{supp } \{\Lambda\} \subset \big(\{\alpha^*\} \cup \big\{\alpha \in \mathcal{A} \mid \psi(\bar{x}(1;\alpha)) = 0\big\}\big),$$

and, for Λ – a.e. $\alpha \in \mathcal{A}$,

(i) $-\dot{p}(t;\alpha) = f_x^T(t,\bar{x}(t;\alpha),\bar{u}(t),\alpha)\, p(.;\alpha) \quad$ *a.e. $t \in [0,1]$,*

(ii) $-p(1;\alpha) = \begin{cases} \psi_x(\bar{x}(1;\alpha)) & \text{if } \alpha \neq \alpha^* \\ rg_x(\bar{x}(1;\alpha)) + (1-r)\psi_x(\bar{x}(1;\alpha)) & \text{if } \alpha = \alpha^* \end{cases}.$

Proof. It might appear that the simplest way to prove Proposition 6.2 would be to reformulate (P2) as a special case of the general minimax optimal control problem (P), in such a manner that $(\bar{u}, \{\bar{x}(.;\alpha) \mid \alpha \in \mathcal{A}\})$ remains a minimizer, by setting

$$g(x,\alpha) := \begin{cases} g(x) & \text{if } \alpha = \alpha^* \\ -K & \text{if } \alpha \neq \alpha^* \end{cases}$$

and

$$C(\alpha) := \{x \mid \psi(x) \leq 0\} \quad \text{for all } \alpha.$$

Here, K is a positive number such that, for some $\delta' > 0$,

$$\inf\big\{g(x) \mid x \in \bar{x}(1;\alpha) + \delta' B,\ \alpha \in \mathcal{A}\big\}\ >\ -K.$$

This is not helpful, however, since $\alpha \to g(x,\alpha)$ violates the continuity hypothesis (S3) for application of Theorem 3.1. Instead, we take a point $b \notin \mathcal{A}$ and associate with (P2) a general minimax problem with extended parameter set $\tilde{\mathcal{A}} := \mathcal{A} \cup \{b\}$, in which $g(x,\alpha)$ is the function

$$g(x,\alpha) := \begin{cases} g(x) & \text{if } \alpha = b \\ -K & \text{if } \alpha \in \mathcal{A} \end{cases}$$

and in which f is the extension of the function f of (P2), to allow for α's in $\mathcal{A} \cup \{b\}$,

$$f(t,x,u,\alpha = b) := f(t,x,u,\alpha^*).$$

The hypotheses are satisfied for the application of Theorem 3.2, with reference to the process $(\bar{u}, \{\bar{x}(.;\alpha) \mid \alpha \in \mathcal{A}\})$, when we partition the extended parameter set as

$$\tilde{\mathcal{A}} = \big(\mathcal{A}^1 := \mathcal{A}\big)\ \cup\ \big(\mathcal{A}^2 := \{b\}\big).$$

Straightforward calculations yield the following information: for $\alpha \in \mathcal{A}$

$$\overline{Q}_0(\alpha) \;=\; \begin{cases} \{q(.) \mid -\dot{q}(t) = f_x^T(t, \bar{x}(t;\alpha), \bar{u}(t), \alpha)q(t), \\ \qquad -q(1) = \psi_x(\bar{x}(1;\alpha))\} & \text{if } \psi(\bar{x}(1;\alpha)) = 0 \\ \emptyset & \text{if } \psi(\bar{x}(1;\alpha)) < 0 \end{cases}$$

and

$$Q_0(\alpha = b) \;=\; \{q(.) \mid -\dot{q}(t) = f_x^T(t, \bar{x}(t;\alpha^*), \bar{u}(t), \alpha^*)q(t), \;\; -q(1) = g_x(\bar{x}(1;\alpha^*))\}\,.$$

We deduce the existence of a Radon probability measure $\mu \subset C^*(\mathcal{A} \cup \{b\})$ and arcs $\{q(.;\alpha) \mid \alpha \in \mathcal{A}\} \cup \{q(.;b)\}$ such that, for μ – *a.e.* $\alpha \in \mathcal{A}$,

$$-\dot{q}(t;\alpha) \;=\; f_x^T(t, \bar{x}(t;\alpha), \bar{u}(t), \alpha)q(t;\alpha), \;\; -q(1;\alpha) \;=\; \psi_x(\bar{x}(1;\alpha)),$$

if $\alpha \in \mathcal{A}$ and

$$-\dot{q}(t;b) \;=\; f_x^T(t, \bar{x}(t;\alpha^*), \bar{u}(t), \alpha^*)q(t;b), \;\; -q(1;b) \;=\; g_x(\bar{x}(1;\alpha^*)).$$

Furthermore, $u \to \mathcal{H}(t,u)$ is maximized over $u \in \Omega(t)$ at $u = \bar{u}(t)$ for *a.e.* $t \in [0,1]$, where

$$\mathcal{H}(t,u) = \int_{\mathcal{A}\cup\{b\}} q(t;\alpha) \cdot f(t, \bar{x}(t;\alpha), u, \alpha)\; \mu(d\alpha)$$

and

$$supp\,\{\mu\} \;\subset\; \{\alpha \in \mathcal{A} \,|\, \psi(\bar{x}(1;\alpha)) = 0\} \cup \{b\}.$$

Now choose

$$r \;=\; \begin{cases} \frac{\mu(\{b\})}{\mu(\{b\})+\mu(\{\alpha^*\})} & \text{if } \mu(b) > 0 \\ 0 & \text{otherwise,} \end{cases}$$

$$p(.;\alpha) := \begin{cases} q(.;\alpha) & \text{for } \;\alpha \neq \alpha^* \\ rq(.;b) + (1-r)q(.;\alpha^*) & \text{for } \;\alpha = \alpha^*. \end{cases}$$

Choose also the Radon measure $\Lambda \in C^*(\mathcal{A})$,

$$\Lambda(E) := \begin{cases} \mu(\{b\}) + \mu(E) & \text{if } \;\alpha^* \in E \\ \mu(E) & \text{if } \;\alpha^* \notin E \end{cases}$$

for any Borel subset E of $\mathcal{A}$. Notice that $||\Lambda||_{T.V.} = ||\mu||_{T.V.} = 1$, so Λ is a probability measure. Clearly

$$supp\,\{\Lambda\} \;\subset\; \{\alpha^*\} \cup \{\alpha \in \mathcal{A} \,|\, \psi(\bar{x}(1;\alpha)) = 0\}\,.$$

We have

$$\begin{aligned}
\mathcal{H}(t,u) &= \left(\int_{\{b\}} + \int_{\{\alpha^*\}} + \int_{\mathcal{A}\setminus\{\alpha^*\}}\right) q(t;\alpha)\cdot f(t,\bar{x}(t;\alpha),u,\alpha)\ \mu(d\alpha)\\
&= \big(\mu(\{b\})q(t;b) + \mu(\{\alpha^*\})q(t;\alpha)\big)\cdot f(t,\bar{x}(t;\alpha^*),u,\alpha^*)\\
&\quad + \int_{\mathcal{A}\setminus\{\alpha^*\}} q(t;\alpha)\cdot f(t,\bar{x}(t;\alpha),u,\alpha)\ \mu(d\alpha)\\
&= \big(\mu(\{b\}) + \mu(\{\alpha^*\})\big)\big(rq(t;b) + (1-r)q(t;\alpha^*)\big)\cdot f(t,\bar{x}(t;\alpha^*),u,\alpha^*)\\
&\quad + \int_{\mathcal{A}\setminus\{\alpha^*\}} q(t;\alpha)\cdot f(t,x(t;\alpha),u,\alpha)\ \mu(d\alpha)\\
&= \Lambda(\{\alpha^*\})p(t;\alpha^*)\cdot f(t,\bar{x}(t;\alpha^*),u,\alpha^*) + \int_{\mathcal{A}\setminus\{\alpha^*\}} p(t;\alpha)\cdot f(t,\bar{x}(t;\alpha),u,\alpha)\ \Lambda(d\alpha)\\
&= \int p(t;\alpha)\cdot f(t,\bar{x}(t;\alpha),u,\alpha)\ \Lambda(d\alpha).
\end{aligned}$$

It follows that

$$\begin{aligned}
-\dot{p}(t;\alpha) &= f_x^T p(t;\alpha)\\
-p(1;\alpha) &= \psi_x(\bar{x}(1;\alpha))
\end{aligned}$$

for $\alpha \neq \alpha^*$. Also, by homogeneity,

$$\begin{aligned}
-\dot{p}(t;\alpha) &= f_x^T p(t;\alpha)\\
-p(1;\alpha) &= rg_x(\bar{x}(1;\alpha)) + (1-r)\psi_x(\bar{x}(1;\alpha))
\end{aligned}$$

for $\alpha = \alpha^*$. The proof is complete. □

Consider finally the *state constrained optimal control problem*,

$$\text{(P3)}\quad \begin{cases} \text{Minimize } g(x(1)) \text{ over measurable functions } u:[0,1]\to R^n \text{ such that}\\ \dot{x}(t) = f(t,x(t),u(t)) \quad \textit{a.e. } t\in[0,1],\\ x(0) = x_0 \quad \text{and} \quad x(1)\in C,\\ u(t)\in\Omega(t) \quad \textit{a.e. } t\in[0,1],\\ h(t,x(t)) \le 0 \qquad t\in[0,1]. \end{cases}$$

This is a "parameter-free" version of (P) (the function f no longer depends on α), to which has been appended an endpoint constraint and a pathwise state constraint

$$h(t,x(t)) \ \le\ 0 \quad \text{for all } t\in[0,1].$$

Here, $C\subset R^n$ is a given set and $h:[0,1]\times R^n\to R$ is a given function. This standard optimal control problem with state constraints would appear to have little relevance to minimax optimal control. The connection is this; (P3) can be interpreted as a minimax type optimal control problem to which the analytical tools of this paper are applicable. This is demonstrated below.

Thus, studying the state constrained optimal control problem in the present context establishes links between minimax optimal control and other well-established areas of optimal control. It also makes clear that the task of deriving necessary conditions of optimality for minimax problems is a challenging one, since it is at least as difficult as deriving necessary conditions for state constrained optimal control problems.

PROPOSITION 5.3. *Let $(\bar{u},\bar{x})$ be a strong local minimizer for* (P3). *Assume that for some $\delta>0$, the following hypotheses are satisfied.*

(a) *$f(.,x,.)$ is $\mathcal{L}\times\mathcal{B}$ measurable for each $x\in R^n$. $t\rightsquigarrow\Omega(t)$ has Borel measurable graph.*

(b) *There exist $k_f(t)\in L^1$ and $c_f>0$ such that*

$$|f(t,x,u)-f(t,x',u)|\le k_f(t)|x-x'| \quad \textit{and} \quad |f(t,x,u)|\le c_f$$

for all $x,x'\in\bar{x}(t)+\delta B$ and $u\in U(t)$, a.e. $t\in[0,1]$. Furthermore, $f(t,.,u)$ is continuously differentiable on a neighborhood of $\bar{x}(t)$ for all $u\in\Omega(t)$ a.e. $t\in[0,1]$.

(c) *g is continuously differentiable on $\bar{x}(1)+\delta B$.*

(d) *h is continuously differentiable.*

Then there exists an arc $p\in W^{1,1}([0,1];R^n)$, $\lambda\ge 0$, and a Radon measure $\mu\in C^*([0,1])$ such that

(i) $\lambda+\|\mu\|_{T.V.}+|p(1)|\neq 0$

(ii) $-\dot{p}=f_x^T(t,\bar{x}(t),\bar{u}(t))\ldots\Big(p(t)+\int_{[0,t)}h_x(s,\bar{x}(s))\mu(ds)\Big)$ *a.e.,*

(iii) $-(p(1)+\int_{[0,1]}h_x(s,\bar{x}(s))\mu(ds))\in\lambda g_x(\bar{x}(1))+N_C(\bar{x}(1))$

(iv) $\text{supp }\{\mu\}\subset\{t\mid h(t,\bar{x}(t))=0\}$

and

$$u\to\Big(p(t)+\int_{[0,t)}h_x(s,\bar{x}(s))\ \mu(ds)\Big)\cdot f(t,\bar{x}(t),u)$$

is maximized over $u\in\Omega(t)$ at $u=\bar{u}(t)$, a.e. $t\in[0,1]$.

We see that the minimax maximum principle can be used to obtain the maximum principle for state constrained problems with a general right endpoint constraint (cf. [10]).

Proof. We reformulate (P3) as a general minimax problem with parameter set $\mathcal{A}=[0,1]\cup\{2\}$. For all $\alpha\in[0,1]$ set

$$f(t,x,u,\alpha):=\begin{cases} f(t,x,u) & \text{for } 0\le t\le\alpha,\\ 0 & \text{for } t>\alpha\end{cases}$$

$$g(x,\alpha):=-K,$$

$$C(\alpha)=\{x\mid h(\alpha,x)\le 0\}.$$

Here, $-K$ is a number strictly less than $g(\bar{x}(1))$. Also set

$$f(t,x,u,\alpha=2):=f(t,x,u)$$

$$g(x,\alpha=2):=g(x),$$

$$C(\alpha=2)=C.$$

Clearly $(\bar{u},\{\bar{x}(.;\alpha)\mid\alpha\in\mathcal{A}\})$ is a strong local minimizer for the general minimax optimal control problem, with these identifications of the data, when

$$\bar{x}(t;\alpha)=\begin{cases}\bar{x}(t) & \text{for } 0\le t\le\alpha,\\ \bar{x}(\alpha) & \text{for } t>\alpha\end{cases}$$

for $\alpha\in[0,1]$ and

$$\bar{x}(.;\alpha=2)\equiv\bar{x}(.).$$

Now apply Theorem 3.3. (See also succeeding comments regarding the nature of the endpoint constraints.) Let $\{p(.;\alpha)\,|\,\alpha\in[0,1]\}$ and $p(.;\alpha=2)$ be the "costate arcs" for this problem and let $\Lambda\subset C^*([0,1]\cup\{2\})$ be the Radon probability measure whose existence is asserted in the theorem. Define $\mu'\subset C^*([0,1])$ to be the restriction of Λ to $[0,1]$. Then

$$||\mu'||_{T.V.}\ \leq\ 1.$$

We have, for $0\leq\alpha\leq 1$,

$$\begin{aligned}-\dot{p}(t;\alpha)\ &=\ \begin{cases} f_x^T(t,\bar{x}(t),\bar{u}(t))p(t;\alpha) & \text{for}\ \ 0\leq t\leq\alpha,\\ 0 & \text{for}\ \ t>\alpha,\end{cases}\\ -p(\alpha;\alpha)\ &=\ h_x(\alpha,\bar{x}(\alpha))\end{aligned}$$

and

$$\begin{aligned}-\dot{p}(t;\alpha=2)\ &=\ f_x^T(t,\bar{x}(t),\bar{u}(t))p(t;\alpha=2),\\ -p(1;\alpha=2)\ &\in\ \lambda g_x(\bar{x}(1))+(1-\lambda)\{\xi\in N_C(\bar{x}(1))\,|\,|\xi|=1\}.\end{aligned}$$

Furthermore, $\bar{u}(t)$ maximizes

$$u\to\left((1-\|\mu'\|_{T.V.})p(t;\alpha=2)+\int_{[t,1]}p(t;\alpha)\ \mu'(d\alpha)\right)\cdot f(t,\bar{x}(t),u)$$

over $u\in\Omega(t)$, *a.e.* $t\in[0,1]$, and

$$supp\ \{\mu'\}\subset\{\alpha\in[0,1]\mid h(\alpha,\bar{x}(\alpha))=0\}.$$

Let $\Phi(t,s)$ be the fundamental matrix for the linear equation $\dot{z}(t)=-f_x^T(t,\bar{x}(t),\bar{u}(t))z(t)$, i.e., for any $s\in[0,1]$, $\Phi(.,s)$ solves $\frac{d}{dt}\Phi(t,s)=-f_x^T(t,\bar{x}(t),\bar{u}(t))\Phi(t,s)$ for $0\leq t\leq 1$ and $\Phi(s,s)=I$.

Suppose $\|\mu'\|_{T.V.}<1$. Define

$$\mu:=\frac{1}{1-\|\mu'\|_{T.V.}}\mu'.$$

Then,

$$u\to\left(p(t)+\int_{[0,t)}h_x(s,\bar{x}(s))\ \mu(ds)\right)\cdot f(t,\bar{x}(t),u)$$

is maximized over $u\in\Omega(t)$ at $u=\bar{u}(t)$, *a.e.* $t\in[0,1]$, where

$$p(t):=p(t;\alpha=2)+\int_{[t,1]}p(t;\alpha)\ \mu(d\alpha)-\int_{[0,t)}h_x(\alpha,\bar{x}(\alpha))\ \mu(d\alpha).$$

We deduce from the differential equations for $p(.;\alpha=2)$ and $p(.;\alpha)$, $\alpha\in[0,1]$, that p satisfies

$$\begin{aligned}p(t)=&-\Phi(t,1)[\lambda g_x(\bar{x}(1))+(1-\lambda)\xi]\\ &-\int_{[t,1]}\Phi(t,\alpha)h_x(\alpha,\bar{x}(\alpha))\ \mu(d\alpha)-\int_{[0,t)}h_x(\alpha,\bar{x}(\alpha))\ \mu(d\alpha)\quad\text{for all } t\in[0,1]\,,\end{aligned}$$

for some $\xi \in \{\xi' \in N_C(\bar{x}(1)) \,|\, |\xi'| = 1\}$. It can be deduced from this relationship that $p(.)$ is an absolutely continuous function which satisfies

$$-\dot{p}(t) = f_x^T(t, \bar{x}(t), \bar{u}(t)) \left(p(t) + \int_{[0,t)} h_x(\alpha, \bar{x}(\alpha))\ \mu(d\alpha) \right) \quad a.e.\ t \in [0,1]$$

$$- \left(p(1) + \int_{[0,1]} h_x(\alpha, \bar{x}(\alpha))\ \mu(d\alpha) \right) = \lambda g_x(\bar{x}(1)) + (1-\lambda)\xi$$

$$\in \lambda g_x(\bar{x}(1)) + N_C(\bar{x}(1)).$$

Notice that if $\|\mu'\|_{T.V.} = 0$ and $\lambda = 0$, then $|p(1)| = |\xi| = 1$. Thus, the multiplier nondegeneracy condition is satisfied. We have confirmed the assertions of the proposition in the case $\|\mu'\|_{T.V.} < 1$.

It remains then to consider the case when $\|\mu'\|_{T.V.} = 1$. Set $\mu = \mu'$. Now condition (iv) in the theorem statement is valid with

$$p(t) = \int_{[t,1]} p(t;\alpha)\ \mu(d\alpha) - \int_{[0,t)} h_x(\alpha, \bar{x}(\alpha))\ \mu(d\alpha).$$

It can be deduced that p satisfies

$$-\dot{p}(t) = f_x^T(t, \bar{x}(t)\bar{u}(t)) \left(p(t) + \int_{[0,t)} h_x(\alpha, \bar{x}(\alpha))\ \mu(d\alpha) \right) \ a.e.\ t \in [0,1]$$

$$- \left(p(1) + \int_{[0,1]} h_x(\alpha, \bar{x}(\alpha))\ \mu(d\alpha) \right) = 0.$$

But

$$0 \in \lambda g_x(\bar{x}(1)) + N_C(\bar{x}(1)),$$

when $\lambda = 0$. The assertions of the proposition have been confirmed in this case too, and the proof is complete. □

6. Proofs of Theorems 3.1–3.3. Our analysis will require some properties of measures, summarized in the following proposition.

PROPOSITION 6.1. *Take a compact metric space $\mathcal{A}$, a sequence $\{\mu_i\}$ of nonnegative Radon measures in $C^*(\mathcal{A})$, a sequence $\{D_i : \mathcal{A} \to R^n\}$ of multifunctions and a sequence of Borel measurable functions $\{\gamma_i : \mathcal{A} \to R^n\}$. Take also a measure $\mu \in C^*(\mathcal{A})$ and a multifunction $D : \mathcal{A} \to R^n$. Assume that Gr D is compact,*

(6.1) $$D(\alpha) \text{ is convex for each } \alpha \in \mathcal{A},$$

$$\limsup_{i \to \infty} \ Gr\ D_i \subset Gr\ D,$$

$$\gamma_i(\alpha) \in D_i(\alpha) \qquad \mu_i \text{ – a.e. } \alpha \in \mathcal{A} \quad \text{for } \ i = 1, 2, \ldots$$

and

$$\mu_i \to \mu \qquad \text{weakly}^*.$$

Define $\eta_i \in C^*(\mathcal{A}; R^n)$ according to

$$\eta_i(d\alpha) = \gamma_i(\alpha)\mu_i(d\alpha) \qquad i = 1, 2, \ldots.$$

Then,

(i) *Along a subsequence,*

$$\eta_i \to \eta \qquad \textit{weakly}^*$$

for some $\eta \in C^*(\mathcal{A}; R^k)$ *and some Borel measurable function* γ *such that*

$$\eta(d\alpha) = \gamma(\alpha)\mu(d\alpha),$$

and

$$\gamma(\alpha) \in D(\alpha) \qquad \mu - a.e.$$

(ii) *Suppose* $\mathcal{A}$ *is expressible as a union of disjoint sets*

$$\mathcal{A} = \mathcal{A}^{(1)} \cup \mathcal{A}^{(2)}$$

in which $\mathcal{A}^{(1)}$ *is compact metric space and* $\mathcal{A}^{(2)}$ *is finite. Then the assertions of part* (i) *remain valid when the hypothesis* (6.1) *is replaced by*

$$D(\alpha) \textit{ is convex for each } \alpha \in \mathcal{A}^{(1)}.$$

Proof. The proof, which is similar to that of ([10], Proposition 9.2.1), is omitted. □

6.1. Proof of Theorem 3.1. We observe at the outset that we can, without loss of generality, replace (S2) and (S3) by stronger (global) hypotheses in which $\delta = +\infty$, that is, we can require the stated conditions in (S2) to hold for all $x, x' \in R^n$, not merely in $x, x' \in \bar{x}(t) + \delta B$; likewise for (S3). This can always be arranged by replacing f and g by their "localizations" $(t, x, u, \alpha) \to f(t, tr_{\bar{x}(t),\delta}(x), u, \alpha)$ and $(t, x, u, \alpha) \to g(t, tr_{\bar{x}(t),\delta}(x), \alpha)$, in which $tr_{y,\delta}(x)$ is the truncation function

$$tr_{y,\delta}(x) = \begin{cases} x & \text{if } |x - y| < \delta \\ y + \delta(x - y)/|x - y| & \text{if } |x - y| \geq \delta \end{cases}.$$

The property that $\bar{x}$ is a strong local minimizer is preserved under this modification of the data. It is a consequence of the hypotheses, strengthened in this way that to each $u \in \mathcal{U}$ and $\alpha \in \mathcal{A}$, there corresponds a unique state trajectory (on $[0, 1]$ with initial state x_0). This we write $x(.; \alpha, u)$.

The following lemma brings together some useful facts, regarding the dependence of the state trajectories on controls and parameters.

Let $\Delta : R^n \times R^n \to R$ denote the Ekeland metric on $\mathcal{U}$,

$$\Delta(u_1, u_2) := meas\{t \mid u_1(t) \neq u_2(t)\}.$$

LEMMA 6.1. *For any* $\delta > 0$, *a finite subset* $\widetilde{\mathcal{A}} \subset \mathcal{A}$ *and* $\rho > 0$ *can be chosen such that*

(i)

$$\sup_{u \in \mathcal{U}} \sup_{\alpha \in \mathcal{A}} \inf_{\alpha' \in \widetilde{A}} \|x(.; \alpha, u) - x(.; \alpha', u)\|_C < \delta$$

(ii)

$$\sup_{\alpha \in \mathcal{A}} \left\{ \|x(.; \alpha, u) - x(.; \alpha, u')\|_C \mid u, u' \in \mathcal{U},\ \Delta(u, u') < \rho \right\} < \delta.$$

These assertions are straightforward consequences of Filippov's existence theorem. (See, e.g., [10], Theorem 2.4.3.)

Take a sequence $\epsilon_i \downarrow 0$. For each i define $J_i : \mathcal{U} \to R$

$$J_i(u) := \max_{\alpha \in \mathcal{A}} \Big\{ \Big(g(x(1;\alpha,u),\alpha) - \max_{\alpha' \in \mathcal{A}} \; g(\bar{x}(1;\alpha'),\alpha') + \epsilon_i^2 \Big) \vee d_{C(\alpha)}(x(1;\alpha,u)) \Big\}.$$

Notice that $J_i(u) \geq 0$ for all $u \in \mathcal{U}$ and $J_i(\bar{u}) = \epsilon_i^2$. It follows that $\bar{u}$ is an ϵ_i^2-minimizer for the functional J_i on $\mathcal{U}$.

For each i, J_i is continuous with respect to the Δ-metric topology. We deduce from Ekeland's theorem the existence of a control function v_i, for each i, such that

$$\Delta(v_i, \bar{u}) \leq \epsilon_i$$

and

$$J_i(v_i) + \epsilon_i \Delta(v_i, v_i) \; = \; \min_{u \in \mathcal{U}} \Big\{ J_i(u) + \epsilon_i \Delta(v_i, u) \Big\}.$$

We have

$$J_i(v_i) > 0 \quad \text{for all } i \text{ sufficiently large,}$$

since $(\bar{u}, \{\bar{x}(.;\alpha) \,|\, \alpha \in \mathcal{A}\})$ is a strong local minimizer for (P) and by Lemma 6.1 (ii). Fix i. For any finite subset $\widetilde{\mathcal{A}} \subset \mathcal{A}$, which will be chosen presently, consider the functional

(6.2)

$$J_i^{\widetilde{\mathcal{A}}}(u) := \max_{\alpha \in \widetilde{\mathcal{A}}} \Big\{ \Big(g(x(1;\alpha,u),\alpha) - \max_{\alpha' \in \mathcal{A}} \; g(\bar{x}(1;\alpha'),\alpha') + \epsilon_i^2 \Big) \vee d_{C(\alpha)}(x(1;\alpha,u)) \Big\}.$$

Take $\rho > 0$. According to Lemma 6.1, the finite subset $\tilde{\mathcal{A}}$ can be chosen such that

$$J_i^{\widetilde{\mathcal{A}}}(u) \; \geq \; J_i(u) - \rho^2 \quad \text{for all } u \in \mathcal{U}.$$

Since v_i is a minimizer for $u \to J_i(u) + \epsilon_i \Delta(v_i, u)$ over $\mathcal{U}$, it follows that v_i is a ρ^2-minimizer for $u \to J_i^{\widetilde{A}}(u) + \epsilon_i \Delta(v_i, u)$ over $\mathcal{U}$. A second application of Ekeland's theorem then yields a control function $u_i \in \mathcal{U}$ such that

$$\Delta(v_i, u_i) \leq \rho$$

and

$$J_i^{\widetilde{\mathcal{A}}}(u_i) + \epsilon_i \Delta(v_i, u_i) + \rho \Delta(u_i, u_i) \; = \; \min_{u \in \mathcal{U}} \Big\{ J_i^{\widetilde{\mathcal{A}}}(u) + \epsilon_i \Delta(v_i, u) + \rho \Delta(u_i, u) \Big\}.$$

By adding extra elements to the finite subset $\widetilde{A}$ and reducing ρ if necessary, we can make the number $|J_i^{\widetilde{A}}(u_i) - J_i(v_i)|$ arbitrary small. (See Lemma 6.1.) Since $J_i(v_i) > 0$, we can arrange that

$$J_i^{\widetilde{\mathcal{A}}}(u_i) > 0.$$

Write $\mathcal{A}^i$ in place of $\widetilde{\mathcal{A}}$ and ρ_i in place of ρ, to emphasize their dependence on i.

We can carry out the above analysis for $i = 1, 2, \ldots$. By adding extra elements to each $\mathcal{A}^i$ and reducing each ρ_i, if necessary, we can arrange, also, that $\{\mathcal{A}^i\}$ is an increasing sequence and $\rho_i \downarrow 0$.

For clarity, we summarize relevant properties of the above constructs: for some sequences $\epsilon_i \downarrow 0$ and $\rho_i \downarrow 0$, sequences $\{u_i\}$ and $\{v_i\}$ in $\mathcal{U}$ and an increasing sequence of finite subsets $\{\mathcal{A}^i\}$ of $\mathcal{A}$, we have

(i) $J_i^{\mathcal{A}^i}(u_i) + \epsilon_i\Delta(v_i,u_i) + \rho_i\Delta(u_i,u_i)$
$= \min_{u\in\mathcal{U}}\left\{J_i^{\mathcal{A}^i}(u) + \epsilon_i\Delta(v_i,u) + \rho_i\Delta(u_i,u)\right\}$ for all i,

(ii) $J_i^{\mathcal{A}^i}(u_i) > 0$ for all i,

(iii) $\Delta(v_i,\bar{u}) \to 0$ and $\Delta(u_i,\bar{u}) \to 0$ as $i \to \infty$.

For each i, list the elements in $\mathcal{A}^i$,

$$\mathcal{A} = \{\alpha_1, \ldots, \alpha_{K_i}\}$$

and write $\{x_i(.;\alpha) \,|\, \alpha \in \mathcal{A}\}$ for the state trajectories corresponding to u_i. Define

$$m_i(t,u) := \begin{cases} 0 & \text{if } u = v_i(t), \\ 1 & \text{otherwise,} \end{cases} \quad \text{and} \quad n_i(t,u) := \begin{cases} 0 & \text{if } u = u_i(t), \\ 1 & \text{otherwise.} \end{cases}$$

With the help of these functions, we can express the minimizing property (i) of the u_i's in control theoretic terms, as follows. For each i, $(u_i, \{x_i(.;\alpha_k) \mid k = 1, \ldots, K_i\})$ is a minimizer for the optimal control problem

$$(\mathrm{P}_i)\begin{cases} \text{Minimize } \max_{k=1,\ldots,K_i}\Big\{\big(g(x(1;\alpha_k),\alpha_k) \\ \qquad\qquad - \max_{\alpha\in\mathcal{A}}\ g(\bar{x}(1;\alpha),\alpha) + \epsilon_i^2\big) \vee d_{C(\alpha_k)}(x(1;\alpha_k))\Big\} \\ \qquad\qquad + \epsilon_i \int_0^1 m_i(t,u(t))\,dt + \rho_i \int_0^1 n_i(t,u(t))\,dt \\ \text{over measurable functions } u \text{ and arcs } \{x(.;\alpha_1), \ldots, x(.;\alpha_{K_i})\} \text{ such that} \\ u(t) \in \Omega(t), \quad \textit{a.e. } t \in [0,1] \\ \text{and, for } k = 1, \ldots, K_i, \\ \dot{x}(t;\alpha_k) = f(t, x(t;\alpha_k), u(t), \alpha_k), \quad \textit{a.e. } t \in [0,1], \\ x(0;\alpha_k)(0) = x_0. \end{cases}$$

Since $u_i \to \bar{u}$ and $v_i \to \bar{u}$ with respect to the Δ-metric, we know that

$$\sup_{\alpha\in\mathcal{A}} \|\bar{x}(.;\alpha) - x(.;\alpha,u_i)\|_C \ \to\ 0 \qquad \text{as} \quad i \to \infty.$$

Take an infinite sequence of control functions $\{\hat{u}_j\} \in \mathcal{U}$ whose first element is $\bar{u}$. Using similar reasoning to that employed in the proof of Proposition 2.1 (note the crucial role of property (ii) above, to ensure multiplier nondegeneracy), we can deduce the following information from the nonsmooth maximum principle (see, e.g., [10], Theorem 6.2.1). For each i sufficiently large, there exist nonnegative numbers $\lambda_1^i, \ldots, \lambda_{K_i}^i$ such that

$$\sum_{k=1}^{K_i} \lambda_k^i = 1,$$

and a sequence $\epsilon_i' \downarrow 0$ with the following properties. Define the discrete probability measure

$$\Lambda_i = \sum_{k=1}^{K_i} \lambda_k^i \delta_{\alpha_k^i}.$$

Then, for each i sufficiently large and Λ_i – *a.e.* $\alpha \in \mathcal{A}$, there exists a costate arc $p_i(.;\alpha)$

satisfying

$$\text{(i)} -\dot{p}_i(t;\alpha) \in \operatorname{co} \partial_x H(t, \bar{x}(t) + \epsilon_i' B, \bar{u}(t), p_i(t;\alpha), \alpha)$$

$$\text{(ii)} -p_i(1;\alpha) \in \overline{co} \bigcup_{x \in \bar{x}(1;\alpha) + \epsilon_i' B} \bigcup_{r \in [0,1]} \left(r G_{\epsilon_i'}(x,\alpha) + (1-r) N(x,\alpha) \right)$$

where $G_{\epsilon'}(\alpha, x) =$

$$\begin{cases} \partial_x g(x,\alpha) & \text{if } g(x,\alpha) \geq \max_{\alpha' \in \mathcal{A}} g(x,\alpha') - \epsilon' \\ \emptyset & \text{otherwise} \end{cases}$$

and $N(x,\alpha) = \left\{ \xi \in N_{C(\alpha)}(x) \,|\, |\xi| = 1 \right\}$

$$\text{(iii)} \int_{\mathcal{A}} \int_0^1 p_i(t;\alpha) \cdot [f(t, \bar{x}(t;\alpha), \hat{u}_j(t), \alpha) - f(t, \bar{x}(t;\alpha), \bar{u}(t), \alpha)] \, dt \, \Lambda_i(d\alpha) \leq \epsilon_i'$$

for $j = 1, 2, \ldots$.

By extracting a subsequence, we can arrange that

$$\Lambda_i \to \Lambda \qquad \text{weakly}^* \qquad \text{as} \quad i \to \infty$$

for some Radon probability measure Λ on the Borel sets of $\mathcal{A}$.

Fix an integer N. We now apply the first part of Proposition 6.1, in which we identify μ with Λ, μ_i with the Λ_i, and take

$$D_i(\alpha) := \{(\xi_1, \ldots, \xi_N) \in R^N \,|\, \exists\, p(.;\alpha) \in Q_{\epsilon_i'}(\alpha) \text{ s.t. } \xi_j = w_j(p(.;\alpha), \alpha) \text{ for } j = 1, 2, \ldots, N\},$$

$i = 1, 2, \ldots$, and

(6.3)
$$D(\alpha) := \{(\xi_1, \ldots, \xi_N) \mid \exists\, p(.;\alpha) \in \overline{Q}_0(\alpha) \text{ s.t. } \xi_j = w_j(p(.;\alpha), \alpha),\ j = 1, 2, \ldots, N\}.$$

Here,

$$w_j(p(.), \alpha) := \int_0^1 p(t) \cdot [f(t, \bar{x}(t,\alpha), \hat{u}_j(t), \alpha) - f(t, \bar{x}(t,\alpha), \bar{u}(t), \alpha)] \, dt.$$

We deduce that

$$\int_{\mathcal{A}} \int_0^1 q_N(t,\alpha) \cdot [f(t, \bar{x}(t,\alpha), \hat{u}_j(t), \alpha) - f(t, \bar{x}(t,\alpha), \bar{u}(t), \alpha)] \, dt \, \Lambda(d\alpha) \leq 0 \tag{6.4}$$

for $j = 1, 2, \ldots, N$, in which $\{q_N(.;\alpha) \in W^{1,1} \,|\, \alpha \in \mathcal{A}\}$ is some family of arcs such that, for Λ – *a.e.* $\alpha \in \mathcal{A}$,

$$q_N(.;\alpha) \in \overline{Q}_0(\alpha).$$

For each N, we can regard $\alpha \to q_N(.;\alpha)$ as a representative of an equivalence class of Λ – *a.e.* equal elements in the Hilbert space

$$\mathcal{X} := L^2_\Lambda(\mathcal{A}; L^2([0,1]; R^n))$$

with the inner product

$$(p,q)_\Lambda = \int_{\mathcal{A}} \int_0^1 p(t;\alpha) \cdot q(t;\alpha) \, dt \, \Lambda(d\alpha).$$

The sequence $\{\alpha \to q_N(.;\alpha)\}_{N=1}^{\infty}$ is norm bounded and therefore has a weak limit, which we write $\{\alpha \to p(.;\alpha)\}$. But

$$\{d \in \mathcal{X} \mid d(\alpha) \in \overline{Q}_0(\alpha),\ \Lambda - a.e. \quad \alpha \in \mathcal{A}\}$$

is a strongly closed subset of $\mathcal{X}$. Since it is convex, it is also weakly closed. It follows that

$$p(.;\alpha) \in \overline{Q}_0(\alpha) \quad \Lambda - a.e.\ \alpha \in \mathcal{A}.$$

By weak convergence, we deduce from (6.4) that

$$\int \int_0^1 p(t;\alpha) \cdot [f(t, \bar{x}(t;\alpha), \hat{u}_j(t), \alpha) - f(t, \bar{x}(t;\alpha), \bar{u}(t), \alpha)]\ dt\ \Lambda(d\alpha) \leq 0 \tag{6.5}$$

for $j = 1, 2, \ldots$.

In view of the Castaing representation theorem (see, e.g., [10], Theorem 2.2.7), we can choose a subset $T \subset (0,1)$ of full measure and also the sequence of controls functions above, $\{\hat{u}_j\}$, to satisfy

$$\overline{\bigcup_j \int_{\mathcal{A}} p(t;\alpha) \cdot f(t, \bar{x}(t;\alpha), \hat{u}_j(t), \alpha)\Lambda(d\alpha)} \ \supset\ \int_{\mathcal{A}} p(t;\alpha) \cdot f(t, \bar{x}(t;\alpha), \Omega(t), \alpha)\Lambda(d\alpha) \tag{6.6}$$

for all $t \in T$. We can arrange that (6.5) remains valid when the countable set $\hat{u}_j(.)$ is replaced by another countable set comprising all concatenations of a finite number of segments of the original $\hat{u}_j$'s, with junction points belonging to a countable dense subset S of $[0,1]$.

Define T' to be the set of full measure, comprising points in T which are also Lebesgue points for

$$s \to \int p(s;\alpha)[f(s, \bar{x}(s,\alpha), \hat{u}_j(s), \alpha) - f(s, \bar{x}(s,\alpha), \bar{u}(s), \alpha)]\ \Lambda(d\alpha) \tag{6.7}$$

for all j. Take any $t \in T'$, $w \in \Omega(t)$ and $\beta > 0$, Then, in view of (6.6), there exists j such that

$$\int_{\mathcal{A}} p(t;\alpha) \cdot f(t, \bar{x}(t), \hat{u}_j(t), \alpha)\Lambda(d\alpha)\ \geq\ \int_{\mathcal{A}} p(t;\alpha) \cdot f(t, \bar{x}(t;\alpha), w, \alpha)\Lambda(d\alpha) - \beta. \tag{6.8}$$

Choose a sequence of intervals $\{[s_i, t_i]\}$, containing t and with endpoints in the set S and such that $s_i \to t$ and $t_i \to t$. Now let $v_i \in \{\hat{u}_j\}_{j=1}^{\infty}$ for $i = 1, 2, \ldots$, where

$$v_i := \begin{cases} \hat{u}_j(t) & \text{if } \ t \in [s_i, t_i], \\ \bar{u}(t) & \text{otherwise.} \end{cases}$$

Changing the order of integration, inserting $\hat{u}_j = v_i$ in (6.5) and dividing across by $|t_i - s_i|$ gives

$$\frac{1}{|t_i - s_i|} \int_{s_i}^{t_i} \int p(s;\alpha) \cdot [f(s, \bar{x}(s;\alpha), \hat{u}_j(s), \alpha) - f(s, \bar{x}(s;\alpha), \bar{u}(s), \alpha)]\ \Lambda(d\alpha)\ dt \leq 0$$

for each i. Since t is a Lebesgue point of the mapping (6.7), it follows that

$$\int p(t;\alpha)\cdot[f(t,\bar{x}(t;\alpha),\hat{u}_j(t),\alpha)-f(t,\bar{x}(t;\alpha),\bar{u}(t),\alpha)]\ \Lambda(d\alpha)\leq 0.$$

We conclude from (6.8) that

$$\int p(t;\alpha)\cdot[f(s,\bar{x}(t;\alpha),w,\alpha)-f(t,\bar{x}(t;\alpha),\bar{u}(t),\alpha)]\ \Lambda(d\alpha)\leq \beta.$$

But $\beta>0$ is arbitrary. So

$$\int p(t;\alpha)\cdot[f(t,\bar{x}(t;\alpha),w,\alpha)-f(t,\bar{x}(t;\alpha),\bar{u}(t),\alpha)]\ \Lambda(d\alpha)\leq 0.$$

Since the above inequality holds for any $t'\in T'$, a set of full measure, and any $w\in\Omega(t)$, the maximization of the Hamiltonian condition is confirmed. The proof is complete.

6.2. Proof of Theorem 3.2. The assertions of Theorem 3.1 are expressed in terms of selectors $p(.;\alpha)$ of the multifunction

$$\alpha\rightarrow\overline{Q}_0(\alpha)$$

in order to guarantee that $D(.)$, given by (6.3), has closed graph and convex values, and thereby to justify application of part (i) of Proposition 6.1.

In the case when $\mathcal{A}$ can be decomposed into disjoint sets $\mathcal{A}=A^{(1)}\cup A^{(2)}$ in which $A^{(2)}$ is finite, essentially the same analysis leads to optimality conditions involving a selector $p(.;\alpha)$ of the multifunction

$$\alpha\rightarrow\begin{cases}\overline{Q}_0(\alpha) & \text{if } \alpha\in A^{(1)},\\ Q_0(\alpha) & \text{if } \alpha\in A^{(2)}.\end{cases}\tag{6.9}$$

We do, however, now have to use part (ii) of Proposition 6.1 to justify (6.4), for some selector $p_N(.;\alpha)$ of the multifunction (6.9).

Also, to justify (6.5) (for some selector $p_N(.;\alpha)$ of (6.9)), we must use the facts that, if $A^{(2)}=\{b_1,\ldots,b_m\}$, then an element in

$$\mathcal{X}=L^2_\Lambda(\mathcal{A};L^2([0,1];R^n))$$

can be represented by an element in

$$\mathcal{X}'=L^2_\Lambda(\mathcal{A}^{(1)};L^2([0,1];R^n))\times L^2([0,1];R^n)^m,$$

and the weak topology on $\mathcal{X}$ is compatible with the weak product topology on $\mathcal{X}'$. It follows that, for the sequence $\{\alpha\rightarrow p_N(.;\alpha)\}_{N=1}^{\infty}$ constructed at the end of the proof of Theorem 3.1, we can arrange by subsequence extraction, that the limiting $p(.;\alpha)$ satisfies $p(.;\alpha)\in Q_0(\alpha)$ for $\Lambda-$ *a.e.* $\alpha\in A^{(2)}$.

6.3. Proof of Theorem 3.3. The proof the minimax maximum principle for problems with functional inequality endpoint constraints is along similar, but simpler, lines to that of Example 4.1. The main difference is that, for each i, we replace the cost function $J_i^{\tilde{A}}(u)$ (see (6.2)) of the earlier analysis by

$$\begin{aligned}\tilde{J}_i^{\tilde{A}}(u)\ :=\ \max_{\alpha\in\tilde{A}}\Big\{&\Big(g(x(1;\alpha,u),\alpha)-\max_{\alpha'\in\mathcal{A}}\ g(\bar{x}(1;\alpha'),\alpha')\Big)\\ &\vee\psi^1(\bar{x}(1;\alpha,u),\alpha,)\vee\cdots\vee\psi^r(\bar{x}(1;\alpha),\alpha)\Big\}.\end{aligned}\tag{6.10}$$

The proof is completed by examining properties of minimizers of perturbations of these cost functions and passage to the limit as before.

Acknowledgment. The assistance of the reviewers in improving earlier versions of this paper is gratefully acknowledged.

REFERENCES

[1] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi Equations*, Birkhaüser, Boston, 1997.
[2] V. G. Boltyansky and A. S. Poznyak, *Robust maximum principle in minimax control*, Internat. J. Control, 72 (1999), pp. 305–314.
[3] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York; reprinted as Classics in Applied Mathematics 5, SIAM, Philadelphia, 1990.
[4] N. N. Krasovskii and A. I. Subbotin, *Game Theoretic Control Problems*, Springer-Verlag, Berlin, 1988.
[5] A. Ya. Dubovitskii and A. A. Milyutin, *Problems for extremum under constraints*, Zh. Vychislit. Math. i Math. Piz., 5 (1965), pp. 395–453; English translation, U.S.S.R. Comput. Math. and Math. Physics, 5 (1965).
[6] M. Morari and E. Zafiriou, *Robust Process Control*, Prentice Hall, Englewood Cliffs, NJ, 1989.
[7] B. S. Mordukhovich, *Generalized differential calculus for nonsmooth and set-valued mappings*, J. Math. Anal. Appl., 183 (1994), pp. 250–288.
[8] E. J. Polak, *Optimization: Algorithms and Consistent Approximations*, Springer-Verlag, New York, 1997.
[9] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Grundlehren der Mathematischen Wissenschaft 317, Springer-Verlag, Berlin, 1998.
[10] R. B. Vinter, *Optimal Control*, Birkhaüser, Boston, 2000.
[11] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 3, pp. 969–990

A CANONICAL FORM FOR THE INCLUSION PRINCIPLE OF DYNAMIC SYSTEMS*

DELIN CHU† AND DRAGOSLAV D. ŠILJAK‡

Abstract. The inclusion principle provides a mathematical framework for comparing behavior of dynamic systems having different dimensions. Our main objective is to derive a canonical form for larger systems (expansions) that are obtained by expanding smaller systems (contractions). The form offers full freedom in selecting appropriate matrices for the expansion-contraction process. We will broaden the form to include feedback and propose an explicit characterization of contractible control laws subject to overlapping information structure constraints.

Key words. inclusion principle, expansion, contraction, canonical form, decentralized control

AMS subject classifications. 93B05, 93B40, 93B52, 65F35

DOI. 10.1137/040609616

1. Introduction. The inclusion principle for dynamic systems [1], which was developed in the 1980s, is now a well-established mathematical framework for comparing systems having different dimensions (for a self-contained presentation of the early results, see [2]). In particular, the principle has been established as a useful tool in formulation of control laws for systems with overlapping information structure constraints [3, 4, 5, 6, 7, 8, 9, 10]. In the past decade, the research on the inclusion principle has been focused on providing a wide variety of conditions for expansion and contraction of continuous, discrete-time, and stochastic dynamic systems [11, 12, 13, 14, 15, 16, 17, 18, 19, 20], which helped resolve both theoretical aspects and practical benefits of the principle in control designs.

Expansion, being an intersection of aggregation [21] and restriction [1, 2, 3], raises a question: What system properties are retained after the expansion-contraction process has been completed [22]? Much progress has been made in identifying the conditions that ensure the invariancy of controllability, observability, and stabilizability in the expanded systems [23, 24, 25].

A central issue in the framework of overlapping decentralized control has been the problem of contractibility of feedback control laws [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]. When a system contains overlapping subsystems, it is natural to add the locally available overlapping states to decentralized control in order to improve the performance of the overall system. This fact gives rise to the control design under overlapping information structure constraints, which is handled by expanding the systems into a larger space where the overlapping subsystems appear as disjoint. As a result of the expansion, overlapping decentralized control in the expanded space can be chosen by standard methods which are available for disjoint subsystems. After the selection is made, the expanded control law is contracted to the original space for implementation. While flexibility of the inclusion principle has been

*Received by the editors June 7, 2004; accepted for publication (in revised form) March 7, 2005; published electronically September 20, 2005.
http://www.siam.org/journals/sicon/44-3/60961.html
†Department of Mathematics, National University of Singapore, 2 Science Drive 2, Singapore 117543 (matchudl@math.nus.edu.sg).
‡Department of Electrical Engineering, Santa Clara University, Santa Clara, CA 95053-0569 (dsiljak@scu.edu).

greatly improved by the new conditions guiding the expansion-contraction process [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20], the contractibility problem has not been satisfactorily resolved. A failure of a condition to provide the required contractibility of a control law is often hard to interpret; one is not sure if the choice of condition or selection of control law is inappropriate, or if contractibility is not possible due to an inherent structure of the system.

Our objective in this paper is to derive a canonical form for the inclusion principle in the spirit of canonical forms for linear dynamical systems [26, 27, 28, 29, 30]. By providing an explicit characterization of expanded systems, the form, as expected, simplifies the study of invariant properties in the expansion-contraction process. The proposed form involves expansion of inputs, outputs, and feedback control laws, thus broadening in an essential way the scope of the canonical form derived previously for state expansion only [1, 2]. A by-product of this fact is a complete resolution of the contractibility problem of expanded control laws for both static and dynamic controllers, which has a special significance in formulations of decentralized control for complex systems under overlapping information structure constraints.

The present paper is organized as follows: In section 2 the inclusion and contractibility of dynamic systems are formulated. Canonical forms for the inclusion principle are established in section 3. In section 4, a problem related to overlapping decentralized control is solved. Next the contractibility of dynamic controllers is discussed in section 5. Finally, in section 6, we offer a few concluding remarks.

2. Inclusion and contractibility. Consider a pair of linear time-invariant systems

$$\mathbf{S}: \quad \begin{cases} \dot{x} = Ax + Bu, \\ y = Cx \end{cases} \tag{2.1}$$

and

$$\tilde{\mathbf{S}}: \quad \begin{cases} \dot{\tilde{x}} = \tilde{A}\tilde{x} + \tilde{B}\tilde{u}, \\ \tilde{y} = \tilde{C}\tilde{x}, \end{cases} \tag{2.2}$$

where $x(t) \in \mathbf{R}^n$, $u(t) \in \mathbf{R}^m$, $y(t) \in \mathbf{R}^l$ are the state, input, and output of system $\mathbf{S}$ at time $t \geq 0$, and $\tilde{x}(t) \in \mathbf{R}^{\tilde{n}}$, $\tilde{u}(t) \in \mathbf{R}^{\tilde{m}}$, $\tilde{y}(t) \in \mathbf{R}^{\tilde{l}}$ are those of $\tilde{\mathbf{S}}$, and $A \in \mathbf{R}^{n\times n}$, $B \in \mathbf{R}^{n\times m}$, $C \in \mathbf{R}^{l\times n}$, $\tilde{A} \in \mathbf{R}^{\tilde{n}\times\tilde{n}}$, $\tilde{B} \in \mathbf{R}^{\tilde{n}\times\tilde{m}}$, $\tilde{C} \in \mathbf{R}^{\tilde{l}\times\tilde{n}}$ are constant matrices. Suppose

$$n \leq \tilde{n}, \quad m \leq \tilde{m}, \quad l \leq \tilde{l},$$

that is, $\mathbf{S}$ is smaller than $\tilde{\mathbf{S}}$. Denote by $x(t; x_0, u)$ and $y[x(t)]$ the state behavior and the corresponding output of system $\mathbf{S}$ for a fixed input $u(t)$ and for an initial state $x(0) = x_0$, respectively. Similar notation $\tilde{x}(t; \tilde{x}_0, \tilde{u})$ and $\tilde{y}[\tilde{x}(t)]$ are used for the state behavior and output of system $\tilde{\mathbf{S}}$.

Let us link systems $\mathbf{S}$ and $\tilde{\mathbf{S}}$ through the following transformations:

$$V: \mathbf{R}^n \longrightarrow \mathbf{R}^{\tilde{n}}, \quad L: \mathbf{R}^m \longrightarrow \mathbf{R}^{\tilde{m}}, \quad T: \mathbf{R}^l \longrightarrow \mathbf{R}^{\tilde{l}}, \tag{2.3}$$

where

$$\operatorname{rank}(V) = n, \quad \operatorname{rank}(L) = m, \quad \operatorname{rank}(T) = l. \tag{2.4}$$

Denote the unique pseudoinverses of V, L, and T by V^+, L^+, and T^+, respectively, and recall the definition of the inclusion principle [1, 2].

DEFINITION 2.1. *The system* $\tilde{\mathbf{S}}$ *includes the system* $\mathbf{S}$*, that is,* $\mathbf{S}$ *is included by* $\tilde{\mathbf{S}}$*, if there exists a triplet* (V, L, T) *satisfying* (2.3) *and* (2.4) *such that, for any initial state* x_0 *and any fixed* $u(t)$ *of system S, the choice*

$$\tilde{x}_0 = Vx_0, \quad \tilde{u}(t) = Lu(t) \quad \forall t \geq 0 \tag{2.5}$$

of the initial state $\tilde{x}_0$ *and input* $\tilde{u}(t)$ *of the system* $\tilde{\mathrm{S}}$ *implies*

$$x(t; x_0, u) = V^+\tilde{x}(t; \tilde{x}_0, \tilde{u}), \quad y[x(t)] = T^+\tilde{y}[\tilde{x}(t)] \quad \forall t \geq 0. \tag{2.6}$$

If the system $\tilde{\mathbf{S}}$ *includes the system* $\mathbf{S}$*, then system* $\tilde{\mathbf{S}}$ *is said to be an expansion of the system* $\mathbf{S}$ *and system* $\mathbf{S}$ *is a contraction of system* $\tilde{\mathbf{S}}$*.*

The inclusion principle has been used to expand overlapping decentralized control laws into a larger space, where they appear disjoint, design disjoint laws by known methods, and *contract* them to the original space for implementation (see, e.g., [2]). The central issue in the expansion-contraction process is the problem of contractibility defined as follows [1, 4].

DEFINITION 2.2. *The control law*

$$\tilde{u} = -\tilde{K}\tilde{x} + \tilde{v}$$

given for system $\tilde{\mathbf{S}}$ *is contractible to the control law*

$$u = -Kx + v$$

for implementation in system $\mathbf{S}$ *if one of the following two statements holds:*

(a) *The choice*

$$\tilde{x}_0 = Vx_0, \quad \tilde{u}(t) = Lu(t)$$

implies

$$x(t; x_0, u) = V^+\tilde{x}(t; \tilde{x}_0, \tilde{u}), \quad LKx(t; x_0, u) = \tilde{K}\tilde{x}(t; \tilde{x}_0, \tilde{u}) \tag{2.7}$$

for all $t \geq 0$*, any initial state* x_0*, and any fixed input* $u(t)$ *of system* $\mathbf{S}$*.*

(b) *The choice*

$$\tilde{x}_0 = Vx_0, \quad u = L^+\tilde{u}$$

implies

$$x(t; x_0, u) = V^+\tilde{x}(t; \tilde{x}_0, \tilde{u}), \quad Kx(t; x_0, u) = L^+\tilde{K}\tilde{x}(t; \tilde{x}_0, \tilde{u}) \tag{2.8}$$

for all $t \geq 0$*, any initial state* x_0 *of system* $\mathbf{S}$*, and any fixed input* $\tilde{u}$ *of system* $\tilde{\mathbf{S}}$*.*

It should be pointed out that both conditions in (a) and (b) above ensure that the closed-loop system

$$\dot{\tilde{x}} = (\tilde{A} + \tilde{B}\tilde{K})\tilde{x} + \tilde{B}\tilde{v}$$

includes the closed-loop system

$$\dot{x} = (A + BK)x + Bv.$$

This property plays an important role in the application of the inclusion principle to overlapping decentralized control.

For the expansion-contraction and contractibility between systems $\mathbf{S}$ and $\tilde{\mathbf{S}}$, the conditions are provided in the following theorem [4].

THEOREM 2.3. *Given systems* $\mathbf{S}$ *and* $\tilde{\mathbf{S}}$*, and transformations* V*,* L*, and* T *satisfying* (2.3) *and* (2.4)*.*

(i) *System* $\tilde{\mathbf{S}}$ *is an expansion of system* $\mathbf{S}$ *if and only if for all* $i = 1, 2, \ldots, \tilde{n}$

$$\begin{cases} V^+(\tilde{A} - VAV^+)^i V = 0, \\ V^+(\tilde{A} - VAV^+)^{i-1}(\tilde{B}L - VB) = 0, \\ (T^+\tilde{C} - CV^+)(\tilde{A} - VAV^+)^{i-1}V = 0, \\ (T^+\tilde{C} - CV^+)(\tilde{A} - VAV^+)^{i-1}(\tilde{B}L - VB) = 0. \end{cases} \tag{2.9}$$

(ii) *The control law* $-\tilde{K}\tilde{x}$ *is contractible to the control law* $-Kx$ *if and only if either*

$$\begin{cases} V^+(\tilde{A} - VAV^+)^i V = 0, \\ V^+(\tilde{A} - VAV^+)^{i-1}(\tilde{B}L - VB) = 0, \quad i = 1, 2, \ldots, \tilde{n}, \\ (LKV^+ - \tilde{K})\tilde{A}^{i-1}\begin{bmatrix} V & \tilde{B}L \end{bmatrix} = 0 \end{cases} \tag{2.10}$$

or

$$\begin{cases} V^+(\tilde{A} - VAV^+)^i V = 0, \\ V^+(\tilde{A} - VAV^+)^{i-1}(\tilde{B} - VBL^+) = 0, \quad i = 1, 2, \ldots, \tilde{n}, \\ (KV^+ - L^+\tilde{K})\tilde{A}^{i-1}\begin{bmatrix} V & \tilde{B} \end{bmatrix} = 0. \end{cases} \tag{2.11}$$

In applications, the inclusion principle relies heavily on the proper choice of expanded matrices $\tilde{A}$, $\tilde{B}$, $\tilde{C}$, and $\tilde{K}$ which are restricted by the expandability and contractibility conditions of Theorem 2.3. In a variety of situations, the conditions have been hard to use since there are no simple rules for their interpretation, nor systematic procedures for utilizing the conditions in the computation of expanded matrices. For this reason, there are a few standard choices [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20] that have been repeatedly used in applications, while the full freedom offered by the conditions has remained unexplored. Recently, to broaden the scope of applications of the inclusion principle, new expansion-contraction conditions have been proposed, which involve additional flexibility provided by the choice of complementary matrices [14, 15, 16, 17, 18]. Even in this case, the conditions involve an intricate relationship between powers of matrices that obscures the full flexibility of the proposed choice.

In the next section we will establish a canonical form for the inclusion principle of dynamic system $\mathbf{S}$. The canonical form parameterizes explicitly all expansion-contraction matrices in the general setting of transformations V, L, T. Therefore, full freedom of the inclusion principle is readily available for control design.

3. Canonical form. Motivated by the difficulties in characterizing expansion matrices, we propose to derive a canonical form for the inclusion principle. The form resolves the difficulties by providing an explicit parameterization of the expanded system within the framework of expansion-contraction process. To show this, we need the following two lemmas [31, 32].

LEMMA 3.1. *Given* $\mathcal{A} \in \mathbf{R}^{n\times n}, \mathcal{B} \in \mathbf{R}^{n\times m}, \mathcal{C} \in \mathbf{R}^{l\times n}$*, and* $\mathcal{D} \in \mathbf{R}^{l\times m}$*.*

(i)

$$\max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} sI - \mathcal{A} & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{bmatrix} = n$$

if and only if

$$\mathcal{D} = 0$$

and

$$\max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} sI - \mathcal{A} & \mathcal{B} \\ \mathcal{C} & 0 \end{bmatrix} = n.$$

(ii) *Assume that* $(\mathcal{A}, \mathcal{B})$ *is controllable, i.e.,*

$$\operatorname{rank} \begin{bmatrix} sI - \mathcal{A} & \mathcal{B} \end{bmatrix} = n \quad \forall s \in \mathbf{C}.$$

Then

$$\max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} sI - \mathcal{A} & \mathcal{B} \\ \mathcal{C} & 0 \end{bmatrix} = n$$

if and only if

$$\mathcal{C} = 0.$$

LEMMA 3.2. *Given* $\mathcal{A} \in \mathbf{R}^{n\times n}, \mathcal{B} \in \mathbf{R}^{n\times m}, \mathcal{C} \in \mathbf{R}^{l\times n}$. *Then*

$$\mathcal{C}\mathcal{A}^i\mathcal{B} = 0 \quad \text{for } i = 0, 1, \ldots, n-1$$

if and only if

$$\max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} sI - \mathcal{A} & \mathcal{B} \\ \mathcal{C} & 0 \end{bmatrix} = n.$$

Proof. From the Kalman decomposition of a linear time-invariant system [26], there exists nonsingular matrix $\mathcal{X} \in \mathbf{R}^{n\times n}$ such that

$$\mathcal{X}\mathcal{A}\mathcal{X}^{-1} = \begin{bmatrix} \mathcal{A}_{11} & \mathcal{A}_{12} \\ 0 & \mathcal{A}_{22} \end{bmatrix} \begin{matrix} \}\tau_1 \\ \}n-\tau_1 \end{matrix}, \quad \mathcal{X}\mathcal{B} = \begin{bmatrix} \mathcal{B}_1 \\ 0 \end{bmatrix} \begin{matrix} \}\tau_1 \\ \}n-\tau_1 \end{matrix}, \quad \mathcal{C}\mathcal{X}^{-1} = \begin{bmatrix} \overset{\tau_1}{\mathcal{C}_1} & \overset{n-\tau_1}{\mathcal{C}_2} \end{bmatrix},$$

where $(\mathcal{A}_{11}, \mathcal{B}_1)$ is controllable, which implies

$$\operatorname{rank} \begin{bmatrix} \mathcal{B}_1 & \mathcal{A}_{11}\mathcal{B}_1 & \cdots & \mathcal{A}_{11}^{\tau_1-1}\mathcal{B}_1 \end{bmatrix} = \tau_1. \tag{3.1}$$

Since

$$\begin{aligned} \operatorname{rank} \begin{bmatrix} \mathcal{C}\mathcal{B} & \mathcal{C}\mathcal{A}\mathcal{B} & \cdots & \mathcal{C}\mathcal{A}^{n-1}\mathcal{B} \end{bmatrix} &= \operatorname{rank}(\mathcal{C}_1 \begin{bmatrix} \mathcal{B}_1 & \mathcal{A}_{11}\mathcal{B}_1 & \cdots & \mathcal{A}_{11}^{n-1}\mathcal{B}_1 \end{bmatrix}) \\ &= \operatorname{rank}(\mathcal{C}_1 \begin{bmatrix} \mathcal{B}_1 & \mathcal{A}_{11}\mathcal{B}_1 & \cdots & \mathcal{A}_{11}^{\tau_1-1}\mathcal{B}_1 \end{bmatrix}), \end{aligned}$$

so, the property (3.1) gives that $\mathcal{C}\mathcal{A}^i\mathcal{B} = 0$ for all $i = 0, 1, \ldots, n-1$ if and only if $\mathcal{C}_1 = 0$.

On the other hand,

$$\begin{aligned} \max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} sI - \mathcal{A} & \mathcal{B} \\ \mathcal{C} & 0 \end{bmatrix} &= (n - \tau_1) + \max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} sI - \mathcal{A}_{11} & \mathcal{B}_1 \\ \mathcal{C}_1 & 0 \end{bmatrix} \\ &= n + \left(\max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} sI - \mathcal{A}_{11} & \mathcal{B}_1 \\ \mathcal{C}_1 & 0 \end{bmatrix} - \tau_1 \right); \end{aligned}$$

thus, we have by using Lemma 3.1 that $\max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} sI-\mathcal{C}^{-}\mathcal{A} & \mathcal{B} \\ 0 & \end{bmatrix} = n$ if and only if $\mathcal{C}_1 = 0$. Hence, Lemma 3.2 follows. □

Now we are ready to present a canonical form for the expansion-contraction triplet $(\tilde{A}, \tilde{B}, \tilde{C})$ under the inclusion principle as follows.

THEOREM 3.3. *Given systems* $\mathbf{S}$ *and* $\tilde{\mathbf{S}}$*, and transformations* V*,* L*,* T *satisfying* (2.3) *and* (2.4)*, let the QR factorizations of* V*,* L*, and* T *be given by*

$$
\begin{cases}
\begin{bmatrix} \mathcal{U} & U \end{bmatrix}^T V = \begin{bmatrix} V_{11} \\ 0 \end{bmatrix} \begin{matrix} \}n \\ \}\tilde{n}-n \end{matrix}, \quad \mathcal{U} \in \mathbf{R}^{\tilde{n}\times n}, \quad U \in \mathbf{R}^{\tilde{n}\times(\tilde{n}-n)}, \\
\begin{bmatrix} \mathcal{P} & P \end{bmatrix}^T L = \begin{bmatrix} L_{11} \\ 0 \end{bmatrix} \begin{matrix} \}m \\ \}\tilde{m}-m \end{matrix}, \quad \mathcal{P} \in \mathbf{R}^{\tilde{m}\times m}, \quad P \in \mathbf{R}^{\tilde{m}\times(\tilde{m}-m)}, \\
\begin{bmatrix} \mathcal{S} & S \end{bmatrix}^T T = \begin{bmatrix} T_{11} \\ 0 \end{bmatrix} \begin{matrix} \}l \\ \}\tilde{l}-l \end{matrix}, \quad \mathcal{S} \in \mathbf{R}^{\tilde{l}\times l}, \quad S \in \mathbf{R}^{\tilde{l}\times(\tilde{l}-l)},
\end{cases} \tag{3.2}
$$

where $\begin{bmatrix} \mathcal{U} & U \end{bmatrix}$, $\begin{bmatrix} \mathcal{P} & P \end{bmatrix}$, *and* $\begin{bmatrix} \mathcal{S} & S \end{bmatrix}$ *are orthogonal, and* V_{11}*,* L_{11}*, and* T_{11} *are nonsingular. Then, system* $\tilde{\mathbf{S}}$ *is an expansion of the system* $\mathbf{S}$ *if and only if*

$$
\begin{cases}
\tilde{A} = \begin{bmatrix} V & UW \end{bmatrix} \begin{bmatrix} A & 0 & \tilde{A}_{13} \\ \tilde{A}_{21} & \tilde{A}_{22} & \tilde{A}_{23} \\ 0 & 0 & \tilde{A}_{33} \end{bmatrix} \begin{bmatrix} V^+ \\ (UW)^T \end{bmatrix}, \\
\tilde{B} = \begin{bmatrix} V & UW \end{bmatrix} \begin{bmatrix} B & \tilde{B}_{12} \\ \tilde{B}_{21} & \tilde{B}_{22} \\ 0 & \tilde{B}_{32} \end{bmatrix} \begin{bmatrix} L^+ \\ P^T \end{bmatrix}, \\
\tilde{C} = \begin{bmatrix} T & S \end{bmatrix} \begin{bmatrix} C & 0 & \tilde{C}_{13} \\ \tilde{C}_{21} & \tilde{C}_{22} & \tilde{C}_{23} \end{bmatrix} \begin{bmatrix} V^+ \\ (UW)^T \end{bmatrix},
\end{cases} \tag{3.3}
$$

where $W \in \mathbf{R}^{(\tilde{n}-n)\times(\tilde{n}-n)}$ *is an arbitrary orthogonal matrix,* μ *is an arbitrary integer between* 0 *and* $\tilde{n}-n$*, and* $\tilde{A}_{13} \in \mathbf{R}^{n\times(\tilde{n}-n-\mu)}$*,* $\tilde{A}_{21} \in \mathbf{R}^{\mu\times n}$*,* $\tilde{A}_{22} \in \mathbf{R}^{\mu\times\mu}$*,* $\tilde{A}_{23} \in \mathbf{R}^{\mu\times(\tilde{n}-n-\mu)}$*,* $\tilde{A}_{33} \in \mathbf{R}^{(\tilde{n}-n-\mu)\times(\tilde{n}-n-\mu)}$*,* $\tilde{B}_{12} \in \mathbf{R}^{n\times(\tilde{m}-m)}$*,* $\tilde{B}_{21} \in \mathbf{R}^{\mu\times m}$*,* $\tilde{B}_{22} \in \mathbf{R}^{\mu\times(\tilde{m}-m)}$*,* $\tilde{B}_{32} \in \mathbf{R}^{(\tilde{n}-n-\mu)\times(\tilde{m}-m)}$*,* $\tilde{C}_{13} \in \mathbf{R}^{l\times(\tilde{n}-n-\mu)}$*,* $\tilde{C}_{21} \in \mathbf{R}^{(\tilde{l}-l)\times n}$*,* $\tilde{C}_{22} \in \mathbf{R}^{(\tilde{l}-l)\times\mu}$*, and* $\tilde{C}_{23} \in \mathbf{R}^{(\tilde{l}-l)\times(\tilde{n}-n-\mu)}$ *are constant matrices with arbitrary elements.*

Proof. It is easy to see that

$$
V^+ = \begin{bmatrix} V_{11}^{-1} & 0 \end{bmatrix} \begin{bmatrix} \mathcal{U} & U \end{bmatrix}^T, \quad L^+ = \begin{bmatrix} L_{11}^{-1} & 0 \end{bmatrix} \begin{bmatrix} \mathcal{P} & P \end{bmatrix}^T, \quad T^+ = \begin{bmatrix} T_{11}^{-1} & 0 \end{bmatrix} \begin{bmatrix} \mathcal{S} & S \end{bmatrix}^T.
$$

In the following we prove the necessity first and then sufficiency.

Necessity. For any $\tilde{A} \in \mathbf{R}^{\tilde{n}\times\tilde{n}}$, $\tilde{B} \in \mathbf{R}^{\tilde{n}\times\tilde{m}}$ and $\tilde{C} \in \mathbf{R}^{\tilde{l}\times\tilde{n}}$, define

$$
\begin{cases}
\begin{bmatrix} \mathcal{U} & U \end{bmatrix}^T \tilde{A} \begin{bmatrix} \mathcal{U} & U \end{bmatrix} = \begin{matrix} \quad n & \tilde{n}-n \\ \end{matrix} \\ \quad \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} \\ \hat{A}_{21} & \hat{A}_{22} \end{bmatrix} \begin{matrix} \}n \\ \}\tilde{n}-n \end{matrix}, \\
\begin{bmatrix} \mathcal{U} & U \end{bmatrix}^T \tilde{B} \begin{bmatrix} \mathcal{P} & P \end{bmatrix} = \begin{matrix} \quad m & \tilde{m}-m \\ \end{matrix} \\ \quad \begin{bmatrix} \hat{B}_{11} & \hat{B}_{12} \\ \hat{B}_{21} & \hat{B}_{22} \end{bmatrix} \begin{matrix} \}n \\ \}\tilde{n}-n \end{matrix}, \\
\begin{bmatrix} \mathcal{S} & S \end{bmatrix}^T \tilde{C} \begin{bmatrix} \mathcal{U} & U \end{bmatrix} = \begin{matrix} \quad n & \tilde{n}-n \\ \end{matrix} \\ \quad \begin{bmatrix} \hat{C}_{11} & \hat{C}_{12} \\ \hat{C}_{21} & \hat{C}_{22} \end{bmatrix} \begin{matrix} \}l \\ \}\tilde{l}-l \end{matrix}.
\end{cases} \tag{3.4}
$$

Let the system $\tilde{\mathbf{S}}$ be an expansion of the system $\mathbf{S}$. By Theorem 2.3 and Lemma 3.2, we have that

$$\max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} sI-\tilde{A}+VAV^{+} & (\tilde{A}-VAV^{+})V & \tilde{B}L-VB \\ V^{+} & 0 & 0 \end{bmatrix} = \tilde{n},$$

$$\max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} sI-\tilde{A}+VAV^{+} & V & \tilde{B}L-VB \\ T^{+}\tilde{C}-CV^{+} & 0 & 0 \end{bmatrix} = \tilde{n},$$

which gives

$$\max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} sI-\tilde{A} & \tilde{A}V-VA & \tilde{B}L-VB \\ V^{+} & 0 & 0 \end{bmatrix} = \max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} sI-\tilde{A} & V & \tilde{B}L \\ T^{+}\tilde{C}-CV^{+} & 0 & 0 \end{bmatrix} = \tilde{n}.$$

Hence, by using (3.4) we get

$$\max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} sI-\hat{A}_{11} & -\hat{A}_{12} & \hat{A}_{11}V_{11}-V_{11}A & \hat{B}_{11}L_{11}-V_{11}B \\ -\hat{A}_{21} & sI-\hat{A}_{22} & \hat{A}_{21}V_{11} & \hat{B}_{21}L_{11} \\ V_{11}^{-1} & 0 & 0 & 0 \end{bmatrix}$$

$$= \max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} sI-\hat{A}_{11} & -\hat{A}_{12} & V_{11} & \hat{B}_{11}L_{11} \\ -\hat{A}_{21} & sI-\hat{A}_{22} & 0 & \hat{B}_{21}L_{11} \\ T_{11}^{-1}\hat{C}_{11}-CV_{11}^{-1} & T_{11}^{-1}\hat{C}_{12} & 0 & 0 \end{bmatrix} = \tilde{n},$$

that is,

$$\max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} sI-\hat{A}_{22} & \hat{A}_{21}V_{11} & \hat{B}_{21}L_{11} \\ -\hat{A}_{12} & \hat{A}_{11}V_{11}-V_{11}A & \hat{B}_{11}L_{11}-V_{11}B \end{bmatrix}$$

$$= \max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} sI-\hat{A}_{22} & -\hat{A}_{21} & \hat{B}_{21}L_{11} \\ T_{11}^{-1}\hat{C}_{12} & T_{11}^{-1}\hat{C}_{11}-CV_{11}^{-1} & 0 \end{bmatrix} = \tilde{n}-n,$$

which, by means of Lemma 3.1, is equivalent to

$$\begin{bmatrix} \hat{A}_{11}V_{11}-V_{11}A & \hat{B}_{11}L_{11}-V_{11}B \end{bmatrix} = 0, \quad T_{11}^{-1}\hat{C}_{11}-CV_{11}^{-1} = 0$$

and

$$\max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} sI-\hat{A}_{22} & \hat{A}_{21} & \hat{B}_{21} \\ \hat{A}_{12} & 0 & 0 \end{bmatrix} = \max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} sI-\hat{A}_{22} & \hat{A}_{21} & \hat{B}_{21} \\ \hat{C}_{12} & 0 & 0 \end{bmatrix} = \tilde{n}-n$$

or, equivalently,

$$\hat{A}_{11} = V_{11}AV_{11}^{-1}, \quad \hat{B}_{11} = V_{11}BR_{11}^{-1}, \quad \hat{C}_{11} = T_{11}CV_{11}^{-1}, \tag{3.5}$$

and

$$\max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} sI-\hat{A}_{22} & \hat{A}_{21} & \hat{B}_{21} \\ \hat{A}_{12} & 0 & 0 \\ \hat{C}_{12} & 0 & 0 \end{bmatrix} = \tilde{n}-n. \tag{3.6}$$

Now we only need to characterize $\hat{A}_{22}$, $\hat{A}_{21}$, $\hat{A}_{12}$, $\hat{B}_{21}$, and $\hat{C}_{12}$ in (3.6). It is well known [33] that there is an orthogonal matrix $W \in \mathbf{R}^{(\tilde{n}-n)\times(\tilde{n}-n)}$ and an integer μ between 0 and $\tilde{n}-n$ such that

$$\text{(3.7)}\qquad \begin{cases} \begin{matrix} & \mu & \tilde{n}-n-\mu & \\ W^T\hat{A}_{22}W = & \left[\begin{matrix} \tilde{A}_{22} & \tilde{A}_{23} \\ 0 & \tilde{A}_{33} \end{matrix}\right] & \begin{matrix} \}\mu \\ \}\tilde{n}-n-\mu \end{matrix} \end{matrix}, \\ \begin{matrix} & n & m & \\ W^T\left[\hat{A}_{21}V_{11} \mid \hat{B}_{21}L_{11}\right] = & \left[\begin{matrix} \tilde{A}_{21} & \tilde{B}_{21} \\ 0 & 0 \end{matrix}\right] & \begin{matrix} \}\mu \\ \}\tilde{n}-n-\mu \end{matrix} \end{matrix}, \\ (\tilde{A}_{22}, \left[\tilde{A}_{21} \quad \tilde{B}_{21}\right]) \text{ is controllable.} \end{cases}$$

Set

$$\left[\frac{\hat{A}_{12}}{\hat{C}_{12}}\right] W = \begin{matrix} & \mu & \tilde{n}-n-\mu & \\ & \left[\begin{matrix} \tilde{A}_{12} & V_{11}\tilde{A}_{13} \\ \tilde{C}_{12} & T_{11}\tilde{C}_{13} \end{matrix}\right] & \begin{matrix} \}n \\ \}l \end{matrix} \end{matrix}.$$

Then, (3.6) and Lemma 3.1 imply

$$\tilde{A}_{12} = 0, \quad \tilde{C}_{12} = 0,$$

that is,

$$\text{(3.8)}\qquad \left[\frac{\hat{A}_{12}}{\hat{C}_{12}}\right] W = \left[\frac{0 \quad V_{11}\tilde{A}_{13}}{0 \quad T_{11}\tilde{C}_{13}}\right].$$

Hence, (3.3) follows directly from a simple calculation using (3.4), (3.5), (3.7), and (3.8).

Sufficiency. Let (3.3) hold for an arbitrary orthogonal matrix $W \in \mathbf{R}^{(\tilde{n}-n)\times(\tilde{n}-n)}$, an arbitrary integer μ between 0 and $\tilde{n}-n$, and arbitrary matrices $\tilde{A}_{13} \in \mathbf{R}^{n\times(\tilde{n}-n-\mu)}$, $\tilde{A}_{21} \in \mathbf{R}^{\mu\times n}$, $\tilde{A}_{22} \in \mathbf{R}^{\mu\times\mu}$, $\tilde{A}_{23} \in \mathbf{R}^{\mu\times(\tilde{n}-n-\mu)}$, $\tilde{A}_{33} \in \mathbf{R}^{(\tilde{n}-n-\mu)\times(\tilde{n}-n-\mu)}$, $\tilde{B}_{12} \in \mathbf{R}^{n\times(\tilde{m}-m)}$, $\tilde{B}_{21} \in \mathbf{R}^{\mu\times m}$, $\tilde{B}_{22} \in \mathbf{R}^{\mu\times(\tilde{m}-m)}$, $\tilde{B}_{32} \in \mathbf{R}^{(\tilde{n}-n-\mu)\times(\tilde{m}-m)}$, $\tilde{C}_{13} \in \mathbf{R}^{l\times(\tilde{n}-n-\mu)}$, $\tilde{C}_{21} \in \mathbf{R}^{(\tilde{l}-l)\times n}$, $\tilde{C}_{22} \in \mathbf{R}^{(\tilde{l}-l)\times\mu}$, and $\tilde{C}_{23} \in \mathbf{R}^{(\tilde{l}-l)\times(\tilde{n}-n-\mu)}$. A direct calculation yields that

$$\begin{aligned} &\max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} sI-\tilde{A} & \tilde{A}V-VA & \tilde{B}L-VB \\ V^+ & 0 & 0 \end{bmatrix} \\ &\quad = \max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} sI-A & 0 & -\tilde{A}_{13} & 0 & 0 \\ -\tilde{A}_{21} & sI-\tilde{A}_{22} & -\tilde{A}_{23} & \tilde{A}_{21} & \tilde{B}_{21} \\ 0 & 0 & sI-\tilde{A}_{33} & 0 & 0 \\ I & 0 & 0 & 0 & 0 \end{bmatrix} \\ &\quad = n+\mu+(\tilde{n}-n-\mu) = \tilde{n} \end{aligned}$$

and

$$\begin{aligned} &\max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} sI-\tilde{A} & V & \tilde{B}L \\ T^+\tilde{C}-CV^+ & 0 & 0 \end{bmatrix} \\ &\quad = \max_{s\in\mathbf{C}} \operatorname{rank} \begin{bmatrix} sI-A & 0 & -\tilde{A}_{13} & I & 0 \\ -\tilde{A}_{21} & sI-\tilde{A}_{22} & -\tilde{A}_{23} & 0 & \tilde{B}_{21} \\ 0 & 0 & sI-\tilde{A}_{33} & 0 & 0 \\ 0 & 0 & \tilde{C}_{13} & 0 & 0 \end{bmatrix} \\ &\quad = n+\mu+(\tilde{n}-n-\mu) = \tilde{n}. \end{aligned}$$

Consequently, we obtain that

$$\max_{s\in\mathbf{C}} \operatorname{rank}\begin{bmatrix} sI-\tilde{A}+VAV^{+} & (\tilde{A}-VAV^{+})V & \tilde{B}L-VB \\ V^{+} & 0 & 0 \end{bmatrix}$$
$$= \max_{s\in\mathbf{C}} \operatorname{rank}\begin{bmatrix} sI-\tilde{A}+VAV^{+} & V & \tilde{B}L-VB \\ T^{+}\tilde{C}-CV^{+} & 0 & 0 \end{bmatrix} = \tilde{n}.$$

Therefore, by Theorem 2.3 and Lemma 3.2 the system $\tilde{\mathbf{S}}$ is an expansion of the system $\mathbf{S}$. □

Since the expansion process underlying the above canonical form (3.3) involves the inputs and outputs, it includes the canonical form obtained in [1] (see also [2]).

Remark 1. Let

$$M=\tilde{A}-VAV^{+},\quad N=\tilde{B}-VBL^{+},\quad G=\tilde{C}-TCV^{+}.$$

Matrices M, N, G defined above are complementary matrices [1, 15]. Obviously, using the same notation as in Theorem 3.3, we conclude that system $\tilde{\mathbf{S}}$ is an expansion of $\mathbf{S}$ if and only if

$$\begin{cases} M=\begin{bmatrix} V & UW \end{bmatrix}\begin{bmatrix} 0 & 0 & \tilde{A}_{13} \\ \tilde{A}_{21} & \tilde{A}_{22} & \tilde{A}_{23} \\ 0 & 0 & \tilde{A}_{33} \end{bmatrix}\begin{bmatrix} V^{+} \\ (UW)^{T} \end{bmatrix}, \\ N=\begin{bmatrix} V & UW \end{bmatrix}\begin{bmatrix} 0 & \tilde{B}_{12} \\ \tilde{B}_{21} & \tilde{B}_{22} \\ 0 & \tilde{B}_{32} \end{bmatrix}\begin{bmatrix} L^{+} \\ P^{T} \end{bmatrix}, \\ G=\begin{bmatrix} T & S \end{bmatrix}\begin{bmatrix} 0 & 0 & \tilde{C}_{13} \\ \tilde{C}_{21} & \tilde{C}_{22} & \tilde{C}_{23} \end{bmatrix}\begin{bmatrix} V^{+} \\ (UW)^{T} \end{bmatrix}, \end{cases}$$

that is, Theorem 3.3 established a canonical form for complementary matrices as well.

Remark 2. In the case that matrices V, L, and T are defined as

$$V=\begin{bmatrix} I_{n_1} & 0 & 0 \\ 0 & I_{n_2} & 0 \\ 0 & I_{n_2} & 0 \\ 0 & 0 & I_{n_3} \end{bmatrix},\quad L=\begin{bmatrix} I_{m_1} & 0 & 0 \\ 0 & I_{m_2} & 0 \\ 0 & I_{m_2} & 0 \\ 0 & 0 & I_{m_3} \end{bmatrix},\quad T=\begin{bmatrix} I_{l_1} & 0 & 0 \\ 0 & I_{l_2} & 0 \\ 0 & I_{l_2} & 0 \\ 0 & 0 & I_{l_3} \end{bmatrix} \tag{3.9}$$

with

$$n_1+n_2+n_3=n,\quad m_1+m_2+m_3=m,\quad l_1+l_2+l_3=l,$$

$$n_1+2n_2+n_3=\tilde{n},\quad m_1+2m_2+m_3=\tilde{m},\quad l_1+2l_2+l_3=\tilde{l},$$

two classes of complementary matrices have been identified in [14, 15] such that system $\tilde{\mathbf{S}}$ includes system $\mathbf{S}$; see (3.30) and (3.31) in [15]. These classes can be obtained by choosing $\tilde{A}$, $\tilde{B}$, $\tilde{C}$ in (3.3) as follows:

$$\begin{cases} \tilde{A}=\begin{bmatrix} V & U \end{bmatrix}\begin{bmatrix} A & 0 \\ X_{21} & X_{22} \end{bmatrix}\begin{bmatrix} V^{+} \\ U^{T} \end{bmatrix}, \\ \tilde{B}=\begin{bmatrix} V & U \end{bmatrix}\begin{bmatrix} B & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix}\begin{bmatrix} L^{+} \\ P^{T} \end{bmatrix}, \\ \tilde{C}=\begin{bmatrix} T & S \end{bmatrix}\begin{bmatrix} C & 0 \\ Z_{21} & Z_{22} \end{bmatrix}\begin{bmatrix} V^{+} \\ U^{T} \end{bmatrix}, \end{cases} \quad \text{or} \quad \begin{cases} \tilde{A}=\begin{bmatrix} V & U \end{bmatrix}\begin{bmatrix} A & X_{12} \\ 0 & X_{22} \end{bmatrix}\begin{bmatrix} V^{+} \\ U^{T} \end{bmatrix}, \\ \tilde{B}=\begin{bmatrix} V & U \end{bmatrix}\begin{bmatrix} B & Y_{12} \\ 0 & Y_{22} \end{bmatrix}\begin{bmatrix} L^{+} \\ P^{T} \end{bmatrix}, \\ \tilde{C}=\begin{bmatrix} T & S \end{bmatrix}\begin{bmatrix} C & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix}\begin{bmatrix} V^{+} \\ U^{T} \end{bmatrix}. \end{cases}$$

Remark 3. The two special cases of aggregation and restriction, which have been used extensively in the existing literature, can now by easily characterized by the canonical form of Theorem 3.3.

- System $\mathbf{S}$ is an aggregation of system $\tilde{\mathbf{S}}$ if and only if

$$\begin{cases} \tilde{A} = \begin{bmatrix} V & U \end{bmatrix} \begin{bmatrix} A & X_{12} \\ 0 & X_{22} \end{bmatrix} \begin{bmatrix} V^+ \\ U^T \end{bmatrix}, \\ \tilde{B} = \begin{bmatrix} V & U \end{bmatrix} \begin{bmatrix} B & Y_{12} \\ 0 & Y_{22} \end{bmatrix} \begin{bmatrix} L^+ \\ P^T \end{bmatrix}, \\ \tilde{C} = \begin{bmatrix} T & S \end{bmatrix} \begin{bmatrix} C & Z_{12} \\ 0 & Z_{22} \end{bmatrix} \begin{bmatrix} V^+ \\ U^T \end{bmatrix}. \end{cases}$$

- System $\mathbf{S}$ is a restriction of system $\tilde{\mathbf{S}}$ if and only if

$$\begin{cases} \tilde{A} = \begin{bmatrix} V & U \end{bmatrix} \begin{bmatrix} A & 0 \\ X_{21} & X_{22} \end{bmatrix} \begin{bmatrix} V^+ \\ U^T \end{bmatrix}, \\ \tilde{B} = \begin{bmatrix} V & U \end{bmatrix} \begin{bmatrix} B & 0 \\ Y_{21} & Y_{22} \end{bmatrix} \begin{bmatrix} L^+ \\ P^T \end{bmatrix}, \\ \tilde{C} = \begin{bmatrix} T & S \end{bmatrix} \begin{bmatrix} C & 0 \\ Z_{21} & Z_{22} \end{bmatrix} \begin{bmatrix} V^+ \\ U^T \end{bmatrix}. \end{cases}$$

When the underlying space of an expansion is used to design control with information structure constraints, then problems arise with control laws when they have to be contracted for implementation in the original space. The explicit contractibility conditions are provided by the following control law canonical form.

THEOREM 3.4. *Given systems* $\mathbf{S}$ *and* $\tilde{\mathbf{S}}$*, and transformations* V*,* L*,* T *satisfying* (2.3) *and* (2.4)*, the control law*

$$\tilde{u} = -\tilde{K}\tilde{x}$$

for system $\tilde{\mathbf{S}}$ *is contractible to the control law*

$$u = -Kx$$

for system $\mathbf{S}$ *if and only if one of the following two statements holds:*

(a) *Matrices* $\tilde{A}$ *and* $\tilde{B}$ *of system* $\tilde{\mathbf{S}}$ *are given by* (3.3) *and*

$$\tilde{K} = \begin{bmatrix} L & P \end{bmatrix} \begin{bmatrix} K & 0 & \tilde{K}_{13} \\ 0 & 0 & \tilde{K}_{23} \end{bmatrix} \begin{bmatrix} V^+ \\ (UW)^T \end{bmatrix}, \tag{3.10}$$

where W *is orthogonal and is the same as that in* (3.3)*, and matrices* $\tilde{K}_{13} \in \mathbf{R}^{m\times(\tilde{n}-n-\mu)}$ *and* $\tilde{K}_{23} \in \mathbf{R}^{(\tilde{m}-m)\times(\tilde{n}-n-\mu)}$ *have arbitrary elements.*

(b) *Matrices* $\tilde{A}$ *and* $\tilde{B}$ *of system* $\tilde{\mathbf{S}}$ *are given by* (3.3) *with*

$$\tilde{B}_{12} = 0, \quad \tilde{B}_{32} = 0$$

and

$$\tilde{K} = \begin{bmatrix} L & P \end{bmatrix} \begin{bmatrix} K & 0 & \tilde{K}_{13} \\ \tilde{K}_{21} & \tilde{K}_{22} & \tilde{K}_{23} \end{bmatrix} \begin{bmatrix} V^+ \\ (UW)^T \end{bmatrix}, \tag{3.11}$$

where W is orthogonal and is the same as that in (3.3)*, and matrices* $\tilde{K}_{13} \in \mathbf{R}^{m\times(\tilde{n}-n-\mu)}$ *and* $\begin{bmatrix}\tilde{K}_{21} & \tilde{K}_{22} & \tilde{K}_{23}\end{bmatrix} \in \mathbf{R}^{(\tilde{m}-m)\times\tilde{n}}$ *have arbitrary elements.*

Proof. The proof is similar to that of Theorem 3.3 and hence is omitted. □

A corollary to Theorems 3.3 and 3.4, which delineates an important class of contractible control laws [17], is now automatic.

COROLLARY 3.5. *Given a system* $\mathbf{S}$ *and transformations* V, L, T *satisfying* (2.3) *and* (2.4)*, if matrices* $\tilde{A}$ *and* $\tilde{B}$ *are given by* (3.3) *with* $\mu = 0$*, then any control law* $\tilde{u} = -\tilde{K}\tilde{x}$ *for system* $\tilde{\mathbf{S}}$ *is contractible to the control law* $u = -Kx$ *with* $K = L^{+}\tilde{K}V$ *for system* $\mathbf{S}$*.*

Remark 4. The definition in [15, 16] for the contractibility is different from that given in [4, 17, 19]. In [15, 16] it is defined that the control law $\tilde{u} = -\tilde{K}\tilde{x}$ for the expanded system $\tilde{\mathbf{S}}$ is contractible to the control law $u = -Kx$ for system $\mathbf{S}$ if the choice $\tilde{x}_0 = Vx_0$ and $\tilde{u} = Lu$ implies

$$Kx(t; x_0, u) = L^{+}\tilde{K}\tilde{x}(t; \tilde{x}_0, \tilde{u})$$

for any $t \geq 0$, any initial state x_0, and any fixed input u of system $\mathbf{S}$. If such a definition is used, then we can show that the control law $\tilde{u} = -\tilde{K}\tilde{x}$ for the expanded system $\tilde{\mathbf{S}}$ is contractible to the control law $u = -Kx$ for system $\mathbf{S}$ if and only if matrices $\tilde{A}$ and $\tilde{B}$ of system $\tilde{\mathbf{S}}$ are given by (3.3) and $\tilde{K}$ is given by (3.11).

It was observed in [15] that our ability to use generalized (system) decompositions depends crucially not only on the choice of the transformation matrices V, R, and T, but also on the selection of the expansion-contraction matrices $\tilde{A}$, $\tilde{B}$, $\tilde{C}$, and $\tilde{K}$ of expanded system $\tilde{\mathbf{S}}$. All previous results enable such selection only partially because of the usage of the forms of matrices $\tilde{A}$, $\tilde{B}$, $\tilde{C}$, and $\tilde{K}$ in system $\tilde{\mathbf{S}}$ corresponding only with some particular cases. Theorems 3.3 and 3.4 have established a canonical form for the inclusion principle of dynamic system $\mathbf{S}$, which explicitly parameterizes all admissible expansion-contraction matrices $\tilde{A}$, $\tilde{B}$, $\tilde{C}$, and $\tilde{K}$ in system $\tilde{\mathbf{S}}$ and thus provides full freedom under the inclusion principle. Therefore, the significance of Theorems 3.3 and 3.4 is obvious. We hasten to add, however, that in choosing suitable expansions in applications of the inclusion principle, the role of complementary matrices [16] is indispensable.

An important issue in the expansion-contraction process has been the conditions under which structural properties of expansions and contractions, such as controllability, observability, and stabilizability, remain invariant in the process. This issue has been raised in [22, 23, 24] regarding controllability and observability, and general conditions for their invariance have been formulated in [25]. To provide a comprehensive relationship between expansions and contractions using the present canonical forms, let us state the following definitions [34].

DEFINITION 3.6. *Given a system* $\mathbf{S}$*. The sets of the uncontrollable modes, the unobservable modes, and the invariant zeros of system* $\mathbf{S}$ *are defined, respectively, by*

$$\Sigma_c(A, B) := \{\lambda \in \mathbf{C} : \ \operatorname{rank}[\lambda I - A \quad B] < n\},$$

$$\Sigma_o(C, A) := \left\{\lambda \in \mathbf{C} : \ \operatorname{rank}\begin{bmatrix}\lambda I - A \\ C\end{bmatrix} < n\right\},$$

and

$$\Sigma_z(C, A, B) := \left\{\lambda \in \mathbf{C} : \ \operatorname{rank}\begin{bmatrix}\lambda I - A & B \\ C & 0\end{bmatrix} < \max_{s\in\mathbf{C}} \operatorname{rank}\begin{bmatrix}sI - A & B \\ C & 0\end{bmatrix}\right\}.$$

DEFINITION 3.7. *Given $A \in \mathbf{R}^{n\times n}, B \in \mathbf{R}^{n\times m}$, let X be a nonsingular matrix such that $(X^{-1}AX, X^{-1}B)$ is in its controllability canonical form, i.e.,*

$$\begin{cases} & \begin{matrix} \mu & n-\mu \end{matrix} \\ X^{-1}AX = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} & \begin{matrix} \}\mu \\ \}n-\mu \end{matrix}, \quad X^{-1}B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix} \begin{matrix} \}\mu \\ \}n-\mu \end{matrix}, \\ (A_{11}, B_1) \text{ is controllable.} \end{cases}$$

Then the controllability subspace $\mathcal{C}(A,B)$ of (A,B) is defined as

$$\mathcal{C}(A,B) = \text{Range}\left(X\begin{bmatrix} I_\mu \\ 0 \end{bmatrix}\right).$$

The desired result relating stability, controllability, observability, detectability, and stability of the invariant zeros is provided by the following.

THEOREM 3.8. *Given a system $\mathbf{S}$ and transformations V, L, and T satisfying (2.3) and (2.4), assume $n < \tilde{n}$, $m < \tilde{m}$, and $l < \tilde{l}$. Let $\bar{\mathbf{C}}^+$ denote the closed right half complex plane. Then, there exist matrices $\tilde{A}$, $\tilde{B}$, and $\tilde{C}$ such that the following properties hold simultaneously:*

(3.12) *System $\tilde{\mathbf{S}}$ is an expansion of system $\mathbf{S}$,*

(3.13) $\sigma(A) \subset \sigma(\tilde{A}), \quad \sigma(\tilde{A}) \cap \bar{\mathbf{C}}^+ = \sigma(A) \cap \bar{\mathbf{C}}^+,$

(3.14) $\Sigma_c(\tilde{A}, \tilde{B}) = \Sigma_c(A,B),$

(3.15) $\Sigma_o(\tilde{C}, \tilde{A}) = \Sigma_0(C,A),$

(3.16) $\Sigma_z(C,A,B) \subset \Sigma_z(\tilde{C}, \tilde{A}, \tilde{B}), \quad \Sigma_z(\tilde{C}, \tilde{A}, \tilde{B}) \cap \bar{\mathbf{C}}^+ = \Sigma_z(C,A,B) \cap \bar{\mathbf{C}}^+.$

Hence, stability, controllability, stabilizability, observability, detectability, and the stability of the invariant zeros can be transmitted simultaneously from system $\mathbf{S}$ to system $\tilde{\mathbf{S}}$ under the inclusion principle.

Proof. Let U, P, and Q be the same as those in Theorem 3.3. Take $\mu = 0$ in (3.3) and define

$$\begin{cases} \tilde{A} = \begin{bmatrix} V & U \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & \mathcal{A} \end{bmatrix} \begin{bmatrix} V^+ \\ U^T \end{bmatrix}, \\ \tilde{B} = \begin{bmatrix} V & U \end{bmatrix} \begin{bmatrix} B & 0 \\ 0 & \mathcal{B} \end{bmatrix} \begin{bmatrix} L^+ \\ P^T \end{bmatrix}, \\ \tilde{C} = \begin{bmatrix} T & S \end{bmatrix} \begin{bmatrix} C & 0 \\ 0 & \mathcal{C} \end{bmatrix} \begin{bmatrix} V^+ \\ U^T \end{bmatrix}, \end{cases}$$

where

$$\mathcal{A} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_{\tilde{n}-n} \end{bmatrix}, \quad \mathcal{B} = \begin{bmatrix} b_1 & 0 \\ b_2 & 0 \\ \vdots & 0 \\ b_{\tilde{n}-n} & 0 \end{bmatrix}, \quad \mathcal{C} = \begin{bmatrix} c_1 & c_2 & \cdots & c_{\tilde{n}-n} \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

and

$$\lambda_1 < \lambda_2 < \cdots < \lambda_{\tilde{n}-n} < 0, \quad b_1 c_1 > 0, \quad b_2 c_2 > 0, \ldots, b_{\tilde{n}-n} c_{\tilde{n}-n} > 0.$$

It is easy to see that

$$\sigma(\mathcal{A}) \subset \mathbf{C}/\bar{\mathbf{C}}^{+}, \quad \Sigma_c(\mathcal{A},\mathcal{B}) = \Sigma_o(\mathcal{C},\mathcal{A}) = \emptyset, \quad \Sigma_z(\mathcal{C},\mathcal{A},\mathcal{B}) \subset \mathbf{C}/\bar{\mathbf{C}}^{+}. \tag{3.17}$$

For $\tilde{A}$, $\tilde{B}$, and $\tilde{C}$ above, Theorem 3.3 implies that system $\tilde{\mathbf{S}}$ is an expansion of system $\mathbf{S}$, the property (3.13) is obvious, and properties (3.14), (3.15), and (3.16) follow directly from (3.17) and the following facts:

$$\begin{cases} \sigma(\tilde{A}) = \sigma(A) \cup \sigma(\mathcal{A}), & \Sigma_c(\tilde{A},\tilde{B}) = \Sigma_c(A,B) \cup \Sigma_c(\mathcal{A},\mathcal{B}), \\ \Sigma_o(\tilde{C},\tilde{A}) = \Sigma(C,A) \cup \Sigma(\mathcal{C},\mathcal{A}), & \Sigma_z(\tilde{C},\tilde{A},\tilde{B}) = \Sigma_z(C,A,B) \cup \Sigma_z(\mathcal{C},\mathcal{A},\mathcal{B}). \end{cases} \quad \square$$

Remark 5. The result in [22] states that when using well-known particular forms of aggregations and restrictions, controllability or observability of the original system carries over to the expanded system, but not both. This result has been shown to be false in [24], which is confirmed by Theorem 3.8. However, it is obvious from Theorem 3.3 that the result of [22] is true when $\tilde{m} = m$ and $\tilde{l} = l$.

4. Overlapping decentralized control. A wide variety of applications of the expansion-contraction concept relies on decentralized control with overlapping information structure constraints. When a plant is composed of interconnected subsystems that share common parts, decentralized control laws, which utilize the state variables of the overlapping parts, are superior to disjoint decentralized control laws. This has been the case in the platooning of vehicles on highways and in the air where state variables are shared between adjacent vehicles [4, 9, 10, 35, 36]. Similarly, in electric power systems tie-line information is used to control each individual power area by decentralized control [2, 3, 7]. Another example is a plant which is overlapped by two controllers for reliability enhancement. The controllers either simultaneously stabilize the plant or individually, whenever one of them has failed [2, 37].

Assume that the system $\mathbf{S}$ is composed of two overlapping subsystems and is represented by the matrices

$$\begin{cases} A = \begin{array}{c} \begin{matrix} n_1 & & n_2 & & n_3 \end{matrix} \\ \left[\begin{array}{c|c|c} A_{11} & A_{12} & A_{13} \\ \hline A_{21} & A_{22} & A_{23} \\ \hline A_{31} & A_{32} & A_{33} \end{array}\right] \end{array} \begin{matrix} \}n_1 \\ \\ \}n_2 \\ \\ \}n_3 \end{matrix}, \\ B = \begin{array}{c} \begin{matrix} m_1 & & m_2 & & m_3 \end{matrix} \\ \left[\begin{array}{c|c|c} B_{11} & B_{12} & B_{13} \\ \hline B_{21} & B_{22} & B_{23} \\ \hline B_{31} & B_{32} & B_{33} \end{array}\right] \end{array} \begin{matrix} \}n_1 \\ \\ \}n_2 \\ \\ \}n_3 \end{matrix}, \\ C = \begin{array}{c} \begin{matrix} n_1 & & n_2 & & n_3 \end{matrix} \\ \left[\begin{array}{c|c|c} C_{11} & C_{12} & C_{13} \\ \hline C_{21} & C_{22} & C_{23} \\ \hline C_{31} & C_{32} & C_{33} \end{array}\right] \end{array} \begin{matrix} \}l_1 \\ \\ \}l_2 \\ \\ \}l_3 \end{matrix}, \end{cases} \tag{4.1}$$

where the lines delineate the subsystems. Using standard linear transformations

defined by matrices (3.9), we obtain the expanded matrices as

$$
(4.2)\qquad
\begin{cases}
\tilde{A} = \begin{array}{c} \begin{matrix} n_1 & n_2 & | & n_2 & n_3 \end{matrix} \\ \left[\begin{array}{cc|cc} \tilde{A}_{11} & \tilde{A}_{12} & \tilde{A}_{13} & \tilde{A}_{14} \\ \tilde{A}_{21} & \tilde{A}_{22} & \tilde{A}_{23} & \tilde{A}_{24} \\ \hline \tilde{A}_{31} & \tilde{A}_{32} & \tilde{A}_{33} & \tilde{A}_{34} \\ \tilde{A}_{41} & \tilde{A}_{42} & \tilde{A}_{43} & \tilde{A}_{44} \end{array}\right] \end{array} \begin{matrix} \}n_1 \\ \}n_2 \\ \\ \}n_2 \\ \}n_3 \end{matrix}, \\
\tilde{B} = \begin{array}{c} \begin{matrix} m_1 & m_2 & | & m_2 & m_3 \end{matrix} \\ \left[\begin{array}{cc|cc} \tilde{B}_{11} & \tilde{B}_{12} & \tilde{B}_{13} & \tilde{B}_{14} \\ \tilde{B}_{21} & \tilde{B}_{22} & \tilde{B}_{23} & \tilde{B}_{24} \\ \hline \tilde{B}_{31} & \tilde{B}_{32} & \tilde{B}_{33} & \tilde{B}_{34} \\ \tilde{B}_{41} & \tilde{B}_{42} & \tilde{B}_{43} & \tilde{B}_{44} \end{array}\right] \end{array} \begin{matrix} \}n_1 \\ \}n_2 \\ \} \\ n_2 \\ \}n_3 \end{matrix}, \\
\tilde{C} = \begin{array}{c} \begin{matrix} n_1 & n_2 & | & n_2 & n_3 \end{matrix} \\ \left[\begin{array}{cc|cc} \tilde{C}_{11} & \tilde{C}_{12} & \tilde{C}_{13} & \tilde{C}_{14} \\ \tilde{C}_{21} & \tilde{C}_{22} & \tilde{C}_{23} & \tilde{C}_{24} \\ \hline \tilde{C}_{31} & \tilde{C}_{32} & \tilde{C}_{33} & \tilde{C}_{34} \\ \tilde{C}_{41} & \tilde{C}_{42} & \tilde{C}_{43} & \tilde{C}_{44} \end{array}\right] \end{array} \begin{matrix} \}l_1 \\ \}l_2 \\ \\ \}l_2 \\ \}l_3 \end{matrix},
\end{cases}
$$

where the overlapping subsystems appear as disjoint.

An interesting idea was recently proposed in [15, 16] to use complementary matrices in order to make the interconnection (off-diagonal) block matrices as sparse as possible, thus enhancing decentralized control strategies for stabilization of the overall system. Note that V, L, and T are given by (3.9), so, the matrices V^+, L^+, T^+, U, P, and S in Theorem 3.3 are given by

$$
V^+ = \begin{bmatrix} I_{n_1} & 0 & 0 & 0 \\ 0 & I_{n_2}/2 & I_{n_2}/2 & 0 \\ 0 & 0 & 0 & I_{n_3} \end{bmatrix}, \quad L^+ = \begin{bmatrix} I_{m_1} & 0 & 0 & 0 \\ 0 & I_{m_2}/2 & I_{m_2}/2 & 0 \\ 0 & 0 & 0 & I_{m_3} \end{bmatrix},
$$

$$
T^+ = \begin{bmatrix} I_{l_1} & 0 & 0 & 0 \\ 0 & I_{l_2}/2 & I_{l_2}/2 & 0 \\ 0 & 0 & 0 & I_{l_3} \end{bmatrix}
$$

and

$$
U = \begin{bmatrix} 0_{n_1 \times n_2} \\ I_{n_2}/\sqrt{2} \\ -I_{n_2}/\sqrt{2} \\ 0_{n_3 \times n_2} \end{bmatrix}, \quad P = \begin{bmatrix} 0_{m_1 \times m_2} \\ I_{m_2}/\sqrt{2} \\ -I_{m_2}/\sqrt{2} \\ 0_{m_3 \times m_2} \end{bmatrix}, \quad S = \begin{bmatrix} 0_{l_1 \times l_2} \\ I_{l_2}/\sqrt{2} \\ -I_{l_2}/\sqrt{2} \\ 0_{l_3 \times l_2} \end{bmatrix}.
$$

From Theorem 3.3 we have that all expansion matrices $\tilde{A}$, $\tilde{B}$, and $\tilde{C}$ of system $\mathbf{S}$ are of the forms

(4.3)

$$
\tilde{A} = \begin{bmatrix} I_{n_1} & 0 & 0 & 0 \\ 0 & I_{n_2} & 0 & W/\sqrt{2} \\ 0 & I_{n_2} & 0 & -W/\sqrt{2} \\ 0 & 0 & I_{n_3} & 0 \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} & A_{13} & 0 & X_{15} \\ A_{21} & A_{22} & A_{23} & 0 & X_{25} \\ A_{31} & A_{32} & A_{33} & 0 & X_{35} \\ X_{41} & X_{42} & X_{43} & X_{44} & X_{45} \\ 0 & 0 & 0 & 0 & X_{55} \end{bmatrix} \begin{bmatrix} I_{n_1} & 0 & 0 & 0 \\ 0 & I_{n_2}/2 & I_{n_2}/2 & 0 \\ 0 & 0 & 0 & I_{n_3} \\ 0 & W^T/\sqrt{2} & -W^T/\sqrt{2} & 0 \end{bmatrix},
$$

$$\tilde{B} = \begin{bmatrix} I_{n_1} & 0 & 0 & 0 \\ 0 & I_{n_2} & 0 & W/\sqrt{2} \\ 0 & I_{n_2} & 0 & -W/\sqrt{2} \\ 0 & 0 & I_{n_3} & 0 \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} & B_{13} & Y_{14} \\ B_{21} & B_{22} & B_{23} & Y_{24} \\ B_{31} & B_{32} & B_{33} & Y_{34} \\ Y_{41} & Y_{42} & Y_{43} & Y_{44} \\ 0 & 0 & 0 & Y_{54} \end{bmatrix} \begin{bmatrix} I_{m_1} & 0 & 0 & 0 \\ 0 & I_{m_2}/2 & I_{m_2}/2 & 0 \\ 0 & 0 & 0 & I_{m_3} \\ 0 & I_{m_2}/\sqrt{2} & -I_{m_2}/\sqrt{2} & 0 \end{bmatrix},$$

$$\tilde{C} = \begin{bmatrix} I_{l_1} & 0 & 0 & 0 \\ 0 & I_{l_2} & 0 & I_{l_2}/\sqrt{2} \\ 0 & I_{l_2} & 0 & -I_{l_2}/\sqrt{2} \\ 0 & 0 & I_{l_3} & 0 \end{bmatrix} \begin{bmatrix} C_{11} & C_{12} & C_{13} & 0 & Z_{15} \\ C_{21} & C_{22} & C_{23} & 0 & Z_{25} \\ C_{31} & C_{32} & C_{33} & 0 & Z_{35} \\ Z_{41} & Z_{42} & Z_{43} & Z_{44} & Z_{45} \end{bmatrix} \begin{bmatrix} I_{n_1} & 0 & 0 & 0 \\ 0 & I_{n_2}/2 & I_{n_2}/2 & 0 \\ 0 & 0 & 0 & I_{n_3} \\ 0 & W^T/\sqrt{2} & -W^T/\sqrt{2} & 0 \end{bmatrix},$$

where $W \in \mathbf{R}^{n_2 \times n_2}$ is orthogonal, X_{44}, X_{55}, X_{i5} $(i = 1, \ldots, 4)$, Y_{j4} and Z_{4j} $(j = 1, \ldots, 5)$, X_{4k}, and Y_{4k} and Z_{k5} $(k = 1, 2, 3)$ are arbitrary matrices with appropriate dimensions, and in particular $X_{44} \in \mathbf{R}^{\mu \times \mu}$, $X_{55} \in \mathbf{R}^{(n_2-\mu)\times(n_2-\mu)}$, μ is an integer between 0 and n_2. Thus, by a direct computation using (4.3) we obtain

$$\tilde{A}_{14} = A_{13}, \quad \tilde{A}_{41} = A_{31}, \quad \tilde{B}_{14} = B_{13}, \quad \tilde{B}_{41} = B_{31}, \quad \tilde{C}_{14} = C_{13}, \quad \tilde{C}_{41} = C_{13}.$$

Consequently, system $\tilde{\mathbf{S}}$ is maximally sparsified if and only if

$$\begin{cases} \tilde{A}_{31} = 0,\ \tilde{A}_{32} = 0,\ \tilde{A}_{42} = 0,\ \tilde{A}_{23} = 0,\ \tilde{A}_{24} = 0,\ \tilde{A}_{13} = 0, \\ \tilde{B}_{31} = 0,\ \tilde{B}_{32} = 0,\ \tilde{B}_{42} = 0,\ \tilde{B}_{23} = 0,\ \tilde{B}_{24} = 0,\ \tilde{B}_{13} = 0, \\ \tilde{C}_{31} = 0,\ \tilde{C}_{32} = 0,\ \tilde{C}_{42} = 0,\ \tilde{C}_{23} = 0,\ \tilde{C}_{24} = 0,\ \tilde{C}_{13} = 0. \end{cases} \tag{4.4}$$

Now, the following problem is of interest.

Problem 1. Under what conditions does there exist an expansion $\tilde{\mathbf{S}}$ of system $\mathbf{S}$ having matrices (4.4)?

It has been mentioned in [15] that in some situation Problem 1 is solvable, but no solvability conditions have been stated; Problem 1 cannot be solved simply by setting $\tilde{A} := VAV^+$, $\tilde{B} := VBL^+$, and $\tilde{C} := TCV^+$, because

$$\begin{cases} VAV^+ = \left[\begin{array}{cc|c|cc} A_{11} & A_{12}/2 & & A_{12}/2 & A_{13} \\ A_{21} & A_{22}/2 & & A_{22}/2 & A_{23} \\ \hline\hline A_{21} & A_{22}/2 & & A_{22}/2 & A_{23} \\ A_{31} & A_{32}/2 & & A_{32}/2 & A_{33} \end{array}\right], \\ VBL^+ = \left[\begin{array}{cc|c|cc} B_{11} & B_{12}/2 & & B_{12}/2 & B_{13} \\ B_{21} & B_{22}/2 & & B_{22}/2 & B_{23} \\ \hline\hline B_{21} & B_{22}/2 & & B_{22}/2 & B_{23} \\ B_{31} & B_{32}/2 & & B_{32}/2 & B_{33} \end{array}\right], \\ TCV^+ = \left[\begin{array}{cc|c|cc} C_{11} & C_{12}/2 & & C_{12}/2 & C_{13} \\ C_{21} & C_{22}/2 & & C_{22}/2 & C_{23} \\ \hline\hline C_{21} & C_{22}/2 & & C_{22}/2 & C_{23} \\ C_{31} & C_{32}/2 & & C_{32}/2 & C_{33} \end{array}\right]; \end{cases}$$

in fact, there are no *general* algorithms for producing such systems. We provide these conditions by the following.

THEOREM 4.1. *Let the triplet* (A, B, C) *of system* $\mathbf{S}$ *be as in* (4.1) *and let matrices* V, L, *and* T *be those of* (3.9). *Then, there exists an expansion* $\tilde{\mathbf{S}}$ *of system* $\mathbf{S}$ *such*

that (4.4) *holds if and only if*

$$\mathcal{C}(A_{22}, \begin{bmatrix} A_{21} & -A_{23} & B_{21} & -B_{23} \end{bmatrix}) \subset \ker \left(\begin{bmatrix} A_{12} \\ -A_{32} \\ C_{12} \\ -C_{32} \end{bmatrix} \right). \tag{4.5}$$

Furthermore, in the case that condition (4.5) *is true, triplet* $(\tilde{A}, \tilde{B}, \tilde{C})$ *of the expanded system* $\tilde{\mathbf{S}}$ *is given by*

$$\begin{cases} \tilde{A} = \left[\begin{array}{ccc|cc} A_{11} & A_{12} & | & 0 & A_{13} \\ 2A_{21} & A_{22} & | & 0 & 0 \\ \hline 0 & 0 & | & A_{22} & 2A_{23} \\ A_{31} & 0 & | & A_{32} & A_{33} \end{array}\right], \\ \tilde{B} = \left[\begin{array}{ccc|cc} B_{11} & B_{12} & | & 0 & B_{13} \\ 2B_{21} & B_{22} & | & 0 & 0 \\ \hline 0 & 0 & | & B_{22} & 2B_{23} \\ B_{31} & 0 & | & B_{32} & B_{33} \end{array}\right], \\ \tilde{C} = \left[\begin{array}{ccc|cc} C_{11} & C_{12} & | & 0 & C_{13} \\ 2C_{21} & C_{22} & | & 0 & 0 \\ \hline 0 & 0 & | & C_{22} & 2C_{23} \\ C_{31} & 0 & | & C_{32} & C_{33} \end{array}\right]. \end{cases} \tag{4.6}$$

Proof. Since (4.3) holds, hence $\tilde{A}$, $\tilde{B}$, and $\tilde{C}$ satisfy (4.4) if and only if

$$\begin{cases} \begin{bmatrix} 0 & I_{n_2} & 0 & -W/\sqrt{2} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} & A_{13} & 0 & X_{15} \\ A_{21} & A_{22} & A_{23} & 0 & X_{25} \\ A_{31} & A_{32} & A_{33} & 0 & X_{35} \\ X_{41} & X_{42} & X_{43} & X_{44} & X_{45} \\ 0 & 0 & 0 & 0 & X_{55} \end{bmatrix} \begin{bmatrix} I_{n_1} & 0 \\ 0 & I_{n_2}/2 \\ 0 & 0 \\ 0 & W^T/\sqrt{2} \end{bmatrix} = 0, \\ \begin{bmatrix} 0 & 0 & I_{n_3} & 0 \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} & A_{13} & 0 & X_{15} \\ A_{21} & A_{22} & A_{23} & 0 & X_{25} \\ A_{31} & A_{32} & A_{33} & 0 & X_{35} \\ X_{41} & X_{42} & X_{43} & X_{44} & X_{45} \\ 0 & 0 & 0 & 0 & X_{55} \end{bmatrix} \begin{bmatrix} 0 \\ I_{n_2}/2 \\ 0 \\ W^T/\sqrt{2} \end{bmatrix} = 0, \\ \begin{bmatrix} 0 & I_{n_2} & 0 & W/\sqrt{2} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} & A_{13} & 0 & X_{15} \\ A_{21} & A_{22} & A_{23} & 0 & X_{25} \\ A_{31} & A_{32} & A_{33} & 0 & X_{35} \\ X_{41} & X_{42} & X_{43} & X_{44} & X_{45} \\ 0 & 0 & 0 & 0 & X_{55} \end{bmatrix} \begin{bmatrix} 0 & 0 \\ I_{n_2}/2 & 0 \\ 0 & I_{n_3} \\ -W^T/\sqrt{2} & 0 \end{bmatrix} = 0, \\ \begin{bmatrix} I_{n_1} & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} & A_{13} & 0 & X_{15} \\ A_{21} & A_{22} & A_{23} & 0 & X_{25} \\ A_{31} & A_{32} & A_{33} & 0 & X_{35} \\ X_{41} & X_{42} & X_{43} & X_{44} & X_{45} \\ 0 & 0 & 0 & 0 & X_{55} \end{bmatrix} \begin{bmatrix} 0 \\ I_{n_2}/2 \\ 0 \\ -W^T/\sqrt{2} \end{bmatrix} = 0, \end{cases}$$

$$\begin{cases} \begin{bmatrix} 0 & I & 0 & -W/\sqrt{2} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} & B_{13} & Y_{14} \\ B_{21} & B_{22} & B_{23} & Y_{24} \\ B_{31} & B_{32} & B_{33} & Y_{34} \\ Y_{41} & Y_{42} & Y_{43} & Y_{44} \\ 0 & 0 & 0 & Y_{54} \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & I/2 \\ 0 & 0 \\ 0 & I/\sqrt{2} \end{bmatrix} = 0, \\ \begin{bmatrix} 0 & 0 & I & 0 \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} & B_{13} & Y_{14} \\ B_{21} & B_{22} & B_{23} & Y_{24} \\ B_{31} & B_{32} & B_{33} & Y_{34} \\ Y_{41} & Y_{42} & Y_{43} & Y_{44} \\ 0 & 0 & 0 & Y_{54} \end{bmatrix} \begin{bmatrix} 0 \\ I/2 \\ 0 \\ I/\sqrt{2} \end{bmatrix} = 0, \\ \begin{bmatrix} 0 & I & 0 & W/\sqrt{2} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} & B_{13} & Y_{14} \\ B_{21} & B_{22} & B_{23} & Y_{24} \\ B_{31} & B_{32} & B_{33} & Y_{34} \\ Y_{41} & Y_{42} & Y_{43} & Y_{44} \\ 0 & 0 & 0 & Y_{54} \end{bmatrix} \begin{bmatrix} 0 & 0 \\ I/2 & 0 \\ 0 & I \\ -I/\sqrt{2} & 0 \end{bmatrix} = 0, \\ \begin{bmatrix} I & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} & B_{13} & Y_{14} \\ B_{21} & B_{22} & B_{23} & Y_{24} \\ B_{31} & B_{32} & B_{33} & Y_{34} \\ Y_{41} & Y_{42} & Y_{43} & Y_{44} \\ 0 & 0 & 0 & Y_{54} \end{bmatrix} \begin{bmatrix} 0 \\ I/2 \\ 0 \\ -I/\sqrt{2} \end{bmatrix} = 0, \end{cases}$$

and

$$\begin{cases} \begin{bmatrix} 0 & I & 0 & -I/\sqrt{2} \end{bmatrix} \begin{bmatrix} C_{11} & C_{12} & C_{13} & 0 & Z_{15} \\ C_{21} & C_{22} & C_{23} & 0 & Z_{25} \\ C_{31} & C_{32} & C_{33} & 0 & Z_{35} \\ Z_{41} & Z_{42} & Z_{43} & Z_{44} & Z_{45} \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & I/2 \\ 0 & 0 \\ 0 & W^T/\sqrt{2} \end{bmatrix} = 0, \\ \begin{bmatrix} 0 & 0 & I & 0 \end{bmatrix} \begin{bmatrix} C_{11} & C_{12} & C_{13} & 0 & Z_{15} \\ C_{21} & C_{22} & C_{23} & 0 & Z_{25} \\ C_{31} & C_{32} & C_{33} & 0 & Z_{35} \\ Z_{41} & Z_{42} & Z_{43} & Z_{44} & Z_{45} \end{bmatrix} \begin{bmatrix} 0 \\ I/2 \\ 0 \\ W^T/\sqrt{2} \end{bmatrix} = 0, \\ \begin{bmatrix} 0 & I & 0 & I/\sqrt{2} \end{bmatrix} \begin{bmatrix} C_{11} & C_{12} & C_{13} & 0 & Z_{15} \\ C_{21} & C_{22} & C_{23} & 0 & Z_{25} \\ C_{31} & C_{32} & C_{33} & 0 & Z_{35} \\ Z_{41} & Z_{42} & Z_{43} & Z_{44} & Z_{45} \end{bmatrix} \begin{bmatrix} 0 & 0 \\ I/2 & 0 \\ 0 & I \\ -W^T/\sqrt{2} & 0 \end{bmatrix} = 0, \\ \begin{bmatrix} I & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} C_{11} & C_{12} & C_{13} & 0 & Z_{15} \\ C_{21} & C_{22} & C_{23} & 0 & Z_{25} \\ C_{31} & C_{32} & C_{33} & 0 & Z_{35} \\ Z_{41} & Z_{42} & Z_{43} & Z_{44} & Z_{45} \end{bmatrix} \begin{bmatrix} 0 \\ I/2 \\ 0 \\ -W^T/\sqrt{2} \end{bmatrix} = 0. \end{cases}$$

Thus, a simple calculation yields that there exists a triplet $(\tilde{A}, \tilde{B}, \tilde{C})$ of the form (4.3) such that (4.4) holds if and only if

$$\text{(4.7)} \qquad \begin{cases} A_{22} = W \begin{bmatrix} X_{44} & X_{45} \\ 0 & X_{55} \end{bmatrix} W^T, \\ \begin{bmatrix} A_{21} & -A_{23} & B_{21} & -B_{23} \end{bmatrix} = W \begin{bmatrix} X_{41} & X_{43} & Y_{41} & Y_{43} \\ 0 & 0 & 0 & 0 \end{bmatrix} / \sqrt{2} \end{cases}$$

and

$$\begin{bmatrix} A_{12} \\ -A_{32} \\ C_{12} \\ -C_{32} \end{bmatrix} = \sqrt{2} \begin{bmatrix} 0 & X_{15} \\ 0 & X_{35} \\ 0 & Z_{15} \\ 0 & Z_{35} \end{bmatrix} W^T, \tag{4.8}$$

which is equivalent to condition (4.5).

Conversely, if condition (4.5) holds, then in (4.3) we can choose an orthogonal matrix W such that

$$(W^T A_{22} W, W^T \sqrt{2} \begin{bmatrix} A_{21} & -A_{23} & B_{21} & -B_{23} \end{bmatrix})$$

is in the controllability staircase form (4.7) [33] of $(A_{22}, \sqrt{2} \begin{bmatrix} A_{21} & -A_{23} & B_{21} & -B_{23} \end{bmatrix})$, let μ be the dimension of its controllability subspace, and define

$$\begin{bmatrix} X_{44} & X_{45} \\ 0 & X_{55} \end{bmatrix}, \quad \begin{bmatrix} X_{41} & X_{43} & Y_{41} & Y_{43} \end{bmatrix}, \quad \text{and} \quad \begin{bmatrix} X_{15} \\ X_{35} \\ Z_{15} \\ Z_{35} \end{bmatrix}$$

by equations (4.7) and (4.8) with $X_{44} \in \mathbf{R}^{\mu \times \mu}$ and $X_{55} \in \mathbf{R}^{(n_2-\mu)\times(n_2-\mu)}$. Now $(X_{44}, [X_{41} \quad X_{43} \quad Y_{41} \quad Y_{43}])$ is controllable. In addition, define

$$\begin{cases} X_{25} = 0, \quad X_{42} = 0, \quad Y_{24} = 0, \quad Y_{42} = 0, \quad Z_{25} = 0, \quad Z_{42} = 0, \\ Y_{14} = B_{12}/\sqrt{2}, \quad Y_{34} = -B_{32}/\sqrt{2}, \quad Z_{41} = \sqrt{2}C_{21}, \quad Z_{43} = -\sqrt{2}C_{23}, \\ \begin{bmatrix} Y_{44} \\ Y_{54} \end{bmatrix} = W^T B_{22}, \quad \begin{bmatrix} Z_{44} & Z_{45} \end{bmatrix} = C_{22} W. \end{cases} \tag{4.9}$$

Then (4.6) follows. □

Condition (4.5) can be verified easily using the well-known controllability staircase form of linear systems (see, e.g., [33]). Theorem 4.1 defines a numerically stable method for solving Problem 1.

5. Contractibility of dynamic controllers. Now, by capitalizing on the canonical form for state feedback laws, we want to present explicit solvability conditions for contractibility of dynamic controllers. They are exhaustive and include the sufficient conditions obtained in [17, 18].

Let us consider a dynamic controller for system **S**:

$$\mathbf{C}: \quad \begin{cases} \dot{w} = Fw + Gu + Jy, \ w(0) = w_0, \\ u = Kw + Hy + v, \end{cases} \tag{5.1}$$

where $w \in \mathbf{R}^\tau$, $u \in \mathbf{R}^m$, and $y \in \mathbf{R}^l$ are the state, input, and output of **C**. An expansion $\tilde{\mathbf{C}}$ of controller **C** is defined as

$$\tilde{\mathbf{C}}: \quad \begin{cases} \dot{\tilde{w}} = \tilde{F}\tilde{w} + \tilde{G}\tilde{u} + \tilde{J}\tilde{y}, \ \tilde{w}(0) = \tilde{w}_0, \\ \tilde{u} = \tilde{K}w + \tilde{H}y + \tilde{v}, \end{cases} \tag{5.2}$$

where $\tilde{w} \in \mathbf{R}^{\tilde{\tau}}$, $\tilde{u} \in \mathbf{R}^{\tilde{m}}$, and $\tilde{y} \in \mathbf{R}^{\tilde{l}}$. We recall the following [18].

DEFINITION 5.1. *The controller* $\tilde{\mathbf{C}}$ *for system* $\tilde{\mathbf{S}}$ *is contractible to the controller* **C** *for system* **S** *if there exist matrices* V, L, T, D, *and* E *satisfying* (2.3) *and* (2.4) *and*

$$\operatorname{rank}(E) = \tau, \quad \operatorname{rank}(D) = m \tag{5.3}$$

such that one of the following two statements holds:

(a) *For any initial states x_0 and w_0 and any input u, the choice*

$$\tilde{x}_0 = Vx_0, \quad \tilde{w}_0 = Ew_0, \quad \tilde{u} = Lu$$

implies that

$$\begin{cases} x(t;x_0,u) = V^+\tilde{x}(t;\tilde{x}_0,\tilde{u}), \quad y[x(t)] = T^+\tilde{y}[\tilde{x}(t)], \\ w(t;w_0,u) = E^+\tilde{w}(t;\tilde{w}_0,\tilde{u}), \quad D(Kw+Hy) = \tilde{K}\tilde{w}+\tilde{H}\tilde{y} \quad \forall t \geq 0. \end{cases}$$

(b) *For any initial states x_0 and w_0 and any input u, the choice*

$$\tilde{x}_0 = Vx_0, \quad \tilde{w}_0 = Ew_0, \quad u = L^+\tilde{u}$$

implies that

$$\begin{cases} x(t;x_0,u) = V^+\tilde{x}(t;\tilde{x}_0,\tilde{u}), \quad y[x(t)] = T^+\tilde{y}[\tilde{x}(t)], \\ w(t;w_0,u) = E^+\tilde{w}(t;\tilde{w}_0,\tilde{u}), \quad Kw+Hy = D^+(\tilde{K}\tilde{w}+\tilde{H}\tilde{y}) \quad \forall t \geq 0. \end{cases}$$

We shall now give an explicit characterization of contractibility of controller $\tilde{\mathbf{C}}$ by the following.

THEOREM 5.2. *Given system $\mathbf{S}$ and transformation matrices V, L, T, D, and E satisfying* (2.3), (2.4), *and* (5.3), *let the QR factorizations of V, L, and T be given by* (3.2). *Furthermore, let the QR factorizations of matrices D and E be given by*

$$(5.4) \quad \begin{cases} \begin{bmatrix} \mathcal{X} & X \end{bmatrix}^T D = \begin{bmatrix} D_{11} \\ 0 \end{bmatrix} \begin{matrix} \}m \\ \}\tilde{m}-m \end{matrix}, \quad \mathcal{X} \in \mathbf{R}^{\tilde{m}\times m}, \quad X \in \mathbf{R}^{\tilde{m}\times(\tilde{m}-m)}, \\ \begin{bmatrix} \mathcal{Y} & Y \end{bmatrix}^T E = \begin{bmatrix} E_{11} \\ 0 \end{bmatrix} \begin{matrix} \}\tau \\ \}\tilde{\tau}-\tau \end{matrix}, \quad \mathcal{Y} \in \mathbf{R}^{\tilde{\tau}\times\tau}, \quad Y \in \mathbf{R}^{\tilde{\tau}\times(\tilde{\tau}-\tau)}, \end{cases}$$

where $\begin{bmatrix} \mathcal{X} & X \end{bmatrix}$ and $\begin{bmatrix} \mathcal{Y} & Y \end{bmatrix}$ are orthogonal, and D_{11} and E_{11} are nonsingular. Then, the controller $\tilde{\mathbf{C}}$ for system $\tilde{\mathbf{S}}$ is contractible to the controller $\mathbf{C}$ for system $\mathbf{S}$ if one of the following four statements holds:

(a) *Matrices $\tilde{A}$, $\tilde{B}$, and $\tilde{C}$ of system $\tilde{\mathbf{S}}$ are given by* (3.3) *and furthermore, $\tilde{F}$, $\tilde{G}$, $\tilde{J}$, $\tilde{K}$, and $\tilde{H}$ are given by*

$$(5.5) \quad \begin{cases} \tilde{F} = \begin{bmatrix} E & YZ \end{bmatrix} \begin{bmatrix} F & 0 & \tilde{F}_{13} \\ \tilde{F}_{21} & \tilde{F}_{22} & \tilde{F}_{23} \\ 0 & 0 & \tilde{F}_{33} \end{bmatrix} \begin{bmatrix} E^+ \\ (YZ)^T \end{bmatrix}, \\ \tilde{G} = \begin{bmatrix} E & YZ \end{bmatrix} \begin{bmatrix} G & \tilde{G}_{12} \\ \tilde{G}_{21} & \tilde{G}_{22} \\ 0 & \tilde{G}_{32} \end{bmatrix} \begin{bmatrix} L^+ \\ P^T \end{bmatrix}, \\ \tilde{J} = \begin{bmatrix} E & YZ \end{bmatrix} \begin{bmatrix} J & 0 \\ \tilde{J}_{21} & \tilde{J}_{22} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} T^+ \\ S^T \end{bmatrix}, \\ \tilde{K} = \begin{bmatrix} D & X \end{bmatrix} \begin{bmatrix} K & 0 & \tilde{K}_{13} \\ 0 & 0 & \tilde{K}_{23} \end{bmatrix} \begin{bmatrix} E^+ \\ (YZ)^T \end{bmatrix}, \\ \tilde{H} = \begin{bmatrix} D & X \end{bmatrix} \begin{bmatrix} H & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} T^+ \\ S^T \end{bmatrix}. \end{cases}$$

(b) *Matrices* $\tilde{A}$, $\tilde{B}$, *and* $\tilde{C}$ *of system* $\tilde{\mathbf{S}}$ *are given by* (3.3) *with* $\tilde{C}_{21} = 0$ *and* $\tilde{C}_{22} = 0$. *Furthermore,* $\tilde{F}$, $\tilde{G}$, $\tilde{J}$, $\tilde{K}$, *and* $\tilde{H}$ *are given by*

$$\begin{cases} \tilde{F} = \begin{bmatrix} E & YZ \end{bmatrix} \begin{bmatrix} F & 0 & \tilde{F}_{13} \\ \tilde{F}_{21} & \tilde{F}_{22} & \tilde{F}_{23} \\ 0 & 0 & \tilde{F}_{33} \end{bmatrix} \begin{bmatrix} E^+ \\ (YZ)^T \end{bmatrix}, \\ \tilde{G} = \begin{bmatrix} E & YZ \end{bmatrix} \begin{bmatrix} G & \tilde{G}_{12} \\ \tilde{G}_{21} & \tilde{G}_{22} \\ 0 & \tilde{G}_{32} \end{bmatrix} \begin{bmatrix} L^+ \\ P^T \end{bmatrix}, \\ \tilde{J} = \begin{bmatrix} E & YZ \end{bmatrix} \begin{bmatrix} J & \tilde{J}_{12} \\ \tilde{J}_{21} & \tilde{J}_{22} \\ 0 & \tilde{J}_{32} \end{bmatrix} \begin{bmatrix} T^+ \\ S^T \end{bmatrix}, \\ \tilde{K} = \begin{bmatrix} D & X \end{bmatrix} \begin{bmatrix} K & 0 & \tilde{K}_{13} \\ 0 & 0 & \tilde{K}_{23} \end{bmatrix} \begin{bmatrix} E^+ \\ (YZ)^T \end{bmatrix}, \\ \tilde{H} = \begin{bmatrix} D & X \end{bmatrix} \begin{bmatrix} H & \tilde{H}_{12} \\ 0 & \tilde{H}_{22} \end{bmatrix} \begin{bmatrix} T^+ \\ S^T \end{bmatrix}. \end{cases} \tag{5.6}$$

(c) *Matrices* $\tilde{A}$, $\tilde{B}$, *and* $\tilde{C}$ *of system* $\tilde{\mathbf{S}}$ *are given by* (3.3) *with* $\tilde{B}_{12} = 0$ *and* $\tilde{B}_{32} = 0$. *Furthermore, matrices* $\tilde{F}$, $\tilde{G}$, $\tilde{J}$, $\tilde{K}$, *and* $\tilde{H}$ *are given by*

$$\begin{cases} \tilde{F} = \begin{bmatrix} E & YZ \end{bmatrix} \begin{bmatrix} F & 0 & \tilde{F}_{13} \\ \tilde{F}_{21} & \tilde{F}_{22} & \tilde{F}_{23} \\ 0 & 0 & \tilde{F}_{33} \end{bmatrix} \begin{bmatrix} E^+ \\ (YZ)^T \end{bmatrix}, \\ \tilde{G} = \begin{bmatrix} E & YZ \end{bmatrix} \begin{bmatrix} G & 0 \\ \tilde{G}_{21} & \tilde{G}_{22} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} L^+ \\ P^T \end{bmatrix}, \\ \tilde{J} = \begin{bmatrix} E & YZ \end{bmatrix} \begin{bmatrix} J & 0 \\ \tilde{J}_{21} & \tilde{J}_{22} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} T^+ \\ S^T \end{bmatrix}, \\ \tilde{K} = \begin{bmatrix} D & X \end{bmatrix} \begin{bmatrix} K & 0 & \tilde{K}_{13} \\ \tilde{K}_{21} & \tilde{K}_{22} & \tilde{K}_{23} \end{bmatrix} \begin{bmatrix} E^+ \\ (YZ)^T \end{bmatrix}, \\ \tilde{H} = \begin{bmatrix} D & X \end{bmatrix} \begin{bmatrix} H & 0 \\ \tilde{H}_{21} & \tilde{H}_{22} \end{bmatrix} \begin{bmatrix} T^+ \\ S^T \end{bmatrix}. \end{cases} \tag{5.7}$$

(d) *Matrices* $\tilde{A}$, $\tilde{B}$, *and* $\tilde{C}$ *of system* $\tilde{\mathbf{S}}$ *are given by* (3.3) *with* $\tilde{B}_{12} = 0$, $\tilde{B}_{32} = 0$, $\tilde{C}_{21} = 0$, *and* $\tilde{C}_{22} = 0$. *Furthermore, matrices* $\tilde{F}$, $\tilde{G}$, $\tilde{J}$, $\tilde{K}$, *and* $\tilde{H}$ *are given by*

$$\begin{cases} \tilde{F} = \begin{bmatrix} E & YZ \end{bmatrix} \begin{bmatrix} F & 0 & \tilde{F}_{13} \\ \tilde{F}_{21} & \tilde{F}_{22} & \tilde{F}_{23} \\ 0 & 0 & \tilde{F}_{33} \end{bmatrix} \begin{bmatrix} E^+ \\ (YZ)^T \end{bmatrix}, \\ \tilde{G} = \begin{bmatrix} E & YZ \end{bmatrix} \begin{bmatrix} G & 0 \\ \tilde{G}_{21} & \tilde{G}_{22} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} L^+ \\ P^T \end{bmatrix}, \\ \tilde{J} = \begin{bmatrix} E & YZ \end{bmatrix} \begin{bmatrix} J & \tilde{J}_{12} \\ \tilde{J}_{21} & \tilde{J}_{22} \\ 0 & \tilde{J}_{32} \end{bmatrix} \begin{bmatrix} T^+ \\ S^T \end{bmatrix}, \\ \tilde{K} = \begin{bmatrix} D & X \end{bmatrix} \begin{bmatrix} K & 0 & \tilde{K}_{13} \\ \tilde{K}_{21} & \tilde{K}_{22} & \tilde{K}_{23} \end{bmatrix} \begin{bmatrix} E^+ \\ (YZ)^T \end{bmatrix}, \\ \tilde{H} = \begin{bmatrix} D & X \end{bmatrix} \begin{bmatrix} H & \tilde{H}_{12} \\ \tilde{H}_{21} & \tilde{H}_{22} \end{bmatrix} \begin{bmatrix} T^+ \\ S^T \end{bmatrix}. \end{cases} \tag{5.8}$$

In (a), (b), (c), *and* (d) *above,* $Z \in \mathbf{R}^{(\tilde{\tau}-\tau)\times(\tilde{\tau}-\tau)}$ *is an arbitrary orthogonal matrix,* ν *is an arbitrary integer between* 0 *and* $\tilde{\tau}-\tau$, *and* $\tilde{F}_{13} \in \mathbf{R}^{\tau\times(\tilde{\tau}-\tau-\nu)}$, $\tilde{F}_{21} \in \mathbf{R}^{\nu\times\tau}$, $\tilde{F}_{22} \in \mathbf{R}^{\nu\times\nu}$, $\tilde{F}_{23} \in \mathbf{R}^{\nu\times(\tilde{\tau}-\tau-\nu)}$, $\tilde{F}_{33} \in \mathbf{R}^{(\tilde{\tau}-\tau-\nu)\times(\tilde{\tau}-\tau-\nu)}$, $\tilde{G}_{21} \in \mathbf{R}^{\nu\times m}$, $\tilde{G}_{22} \in \mathbf{R}^{\nu\times(\tilde{m}-m)}$, $\tilde{J}_{12} \in \mathbf{R}^{\tau\times(\tilde{l}-l)}$, $\tilde{J}_{21} \in \mathbf{R}^{\nu\times l}$, $\tilde{J}_{22} \in \mathbf{R}^{\nu\times(\tilde{l}-l)}$, $\tilde{J}_{32} \in \mathbf{R}^{(\tilde{\tau}-\tau-\nu)\times(\tilde{l}-l)}$, $\tilde{K}_{13} \in \mathbf{R}^{m\times(\tilde{\tau}-\nu-\tau)}$, $\tilde{K}_{21} \in \mathbf{R}^{(\tilde{m}-m)\times\tau}$, $\tilde{K}_{22} \in \mathbf{R}^{(\tilde{m}-m)\times\nu}$, $\tilde{K}_{23} \in \mathbf{R}^{(\tilde{m}-m)\times(\tilde{\tau}-\nu-\tau)}$, $\tilde{H}_{12} \in \mathbf{R}^{m\times(\tilde{l}-l)}$, $\tilde{H}_{21} \in \mathbf{R}^{(\tilde{m}-m)\times l}$, *and* $\tilde{H}_{22} \in \mathbf{R}^{(\tilde{m}-m)\times(\tilde{l}-l)}$ *are matrices with arbitrary elements.*

Proof. Theorem 5.2 can be proved using Definitions 5.1(a) and (b) directly, hence its proof is omitted. □

Similarly, as in Remarks 2 and 3, if in Theorem 5.2, we take $\nu = 0$ or $\nu = \tilde{\tau} - \tau$, then we can obtain some particular solvability conditions for contractibility of dynamic controllers, which contain the results of [17, 18] as special cases.

6. Conclusions. A canonical form for expanded systems is proposed in the inclusion principle for dynamic systems. The main benefits of the form are as follows:

1. In Theorems 3.3 and 3.4 we have established canonical forms for expansion-contraction matrices $\tilde{A}$, $\tilde{B}$, $\tilde{C}$, and $\tilde{K}$, which provide an explicit parameterization of all expansion-contraction matrices. As a result, the full freedom in selecting the expansion-contraction matrices can be exploited in system analysis and design.
2. Theorem 3.8 provides a simple way to determine if stability, stabilizability, controllability, detectability, observability, and the stability of the invariant zeros carry over from a system $\mathbf{S}$ to its expansion $\tilde{\mathbf{S}}$.
3. In Theorem 4.1, we solved Problem 1, which is central to overlapping decentralized control and which has not been solved in full generality by existing methods.
4. By Theorem 5.2 we broaden the class of dynamic controllers which are contractible for implementation in the original system.

It is hoped that the proposed canonical form will simplify not only design of overlapping decentralized control, but also design of reduced-order controllers [6, 23], where the laws can be generated in the smaller space and then expanded for implementation in the original system.

REFERENCES

[1] M. Ikeda, D.D. Šiljak, and D.E. White, *An inclusion principle for dynamic systems,* IEEE Trans. Automat. Control, 43 (1984), pp. 1040–1055.

[2] D.D. Šiljak, *Decentralized Control of Complex Systems,* Academic Press, Boston, 1991.

[3] M. Ikeda, D.D. Šiljak, and D.E. White, *Decentralized control with overlapping information sets,* J. Optim. Theory Appl., 34 (1981), pp. 279–310.

[4] M. Ikeda and D.D. Šiljak, *Overlapping decentralized control with input, state and output inclusion,* Control Theory Adv. Technol., 2 (1986), pp. 155–172.

[5] Y. Ohta, D.D. Šiljak, and T. Matsumoto, *Decentralized control using quasi-block diagonal dominance of transfer function matrices,* IEEE Trans. Automat. Control, 31 (1986), pp. 420–430.

[6] M.E. Sezer and D.D. Šiljak, *Validation of reduced order models for control systems design,* J. Guidance Control Dynam., 5 (1982), pp. 430–437.

[7] S.S. Stankovic, X.-B Chen, M.R. Matausek, and D.D. Šiljak, *Stochastic inclusion principle applied to decentralized automatic generalization control,* Internat. J. Control, 72 (1999), pp. 276–288.

[8] K. Li, E.B. Kosmatopoulos, P.A. Yoannou, and H. Ryciotaki-Boussalis, *Large segmented telescopes,* IEEE Control Syst. Mag., 20 (2000), pp. 59–72.

[9] S.S. Stankovic, M.J. Stanojevic, and D.D. Šiljak, *Decentralized overlapping control of a platoon of vehicles,* IEEE Trans. Control Syst. Technol., 8 (2000), pp. 816–832.

[10] D.M. Stipanović, G. İnhalan, R. Teo, and C.J. Tomlin, *Decentralized overlapping control of a formation of unmanned aerial vehicles,* Automatica J. IFAC, 40 (2004), pp. 1285–1296.
[11] A. Iftar and U. Özgüner, *Contractible controller design and optimal control with state and input inclusion,* Automatica J. IFAC, 26 (1990), pp. 593–597.
[12] A. Iftar, *Decentralized estimation and control with overlapping input, state, and output decomposition,* Automatica J. IFAC, 29 (1993), pp. 511–516.
[13] A. Iftar, *Overlapping decentralized dynamic optimal control,* Internat. J. Control, 58 (1993), pp. 187–209.
[14] J.M. Rossell, *Contribution to Decentralized Control of Large-Scale Systems via Overlapping Models,* Ph.D. thesis, University of Catalunya, Barcelona, Spain, 1998.
[15] L. Bakule, J. Rodellar, and J. Rossell, *Structure of expansion–contraction matrices in the inclusion principle for dynamic systems,* SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1136–1155.
[16] L. Bakule, J. Rodellar, and J. Rossell, *Generalized selection of complementary matrices in the inclusion principle,* IEEE Trans. Automat. Control, 45 (2000), pp. 1237–1243.
[17] S.S. Stankovic and D.D. Šiljak, *Contractibility of overlapping decentralized control,* Systems Control Lett., 44 (2001), pp. 189–199.
[18] L. Bakule, J. Rodellar, and J. Rossell, *Contractibility of dynamic LTI controllers using complementary matrices,* IEEE Trans. Automat. Control, 48 (2003), pp. 1269–1274.
[19] S.S. Stankovic and D. Šiljak, *Inclusion principle for linear time-varying systems,* SIAM J. Control Optim., 42 (2003), pp. 321–341.
[20] S.S. Stankovic, *Inclusion principle for discrete-time time-varying systems,* Dyn. Contin. Discrete Impuls. Syst. Ser. A Math. Anal., 11 (2004), pp. 321–338.
[21] M. Aoki, *Control of large scale dynamic systems by aggregation,* IEEE Trans. Automat. Control, 13 (1968), pp. 246–253.
[22] K. Malinowski and M. Singh, *Controllability and observability of expanded systems with overlapping decompositions,* Automatica J. IFAC, 21 (1985), pp. 203–208.
[23] G.J. Pappas, G. Laferriere, and S. Sastry, *Hierarchically consistent control systems*, IEEE Trans. Automat. Control, 45 (2000), pp. 1144–1160.
[24] L. Bakule, J. Rodellar, J. Rossell, and P. Rubio, *Preservation of controllability–observability in expanded systems,* IEEE Trans. Automat. Control, 46 (2001), pp. 1155–1162.
[25] S.S. Stankovic and D. Šiljak, *Model abstraction and inclusion principle: A comparison,* IEEE Trans. Automat. Control, 47 (2002), pp. 529–532.
[26] R.E. Kalman, *Mathematical description of linear dynamic systems,* SIAM J. Control, 1 (1963), pp. 152–192.
[27] P. Brunovsky, *On stabilization of linear systems under a certain class of persistent perturbations,* Differential Equations, 2 (1966), pp. 401–405.
[28] D.G. Luenberger, *Canonical forms for linear multivariable systems,* IEEE Trans. Automat. Control, 12 (1967), pp. 290–293.
[29] V.M. Popov, *Invariant description of linear, time-invariant controllable systems,* SIAM J. Control, 10 (1972), pp. 252–264.
[30] T. Kailath, *Linear Systems,* Prentice-Hall, Upper Saddle River, NJ, 1980.
[31] D. Chu and V. Mehrmann, *Disturbance decoupling for linear time-invariant systems: A matrix pencil approach,* IEEE Trans. Automat. Control, 46 (2001), pp. 802–808.
[32] D. Chu and Y.S. Hung, *A numerical solution for the simultaneous disturbance rejection and row by row decoupling problem,* Linear Algebra Appl., 320 (2000), pp. 37–49.
[33] P. van Dooren, *The generalized eigenstructure problem in linear system theory,* IEEE Trans. Automat. Control, 6 (1981), pp. 111–129.
[34] W.M. Wonham, *Linear Multivariable Control: The Geometric Approach,* 2nd ed., Springer-Verlag, New York, 1985.
[35] M. Athans and W.S. Levine, *On the optimal error regulation of a string of moving vehicles,* IEEE Trans. Automat. Control, 11 (1966), pp. 355–361.
[36] A.H. Levis and M. Athans, *On the optimal sampled data control of string of moving vehicles,* Transportation Sci., 2 (1968), pp. 362–382.
[37] D.D. Šiljak, *Reliable control using multiple control systems,* Internat. J. Control, 31 (1980), pp. 303–339.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 3, pp. 991–1018

THE SUBOPTIMAL NEHARI PROBLEM FOR WELL-POSED LINEAR SYSTEMS*

RUTH F. CURTAIN† AND MARK R. OPMEER†

Abstract. We solve the suboptimal Nehari problem for a transfer function that has a state-space realization as a system-stable (input, output and input-output stable) well-posed linear system. We obtain an explicit solution in terms of the state-space parameters.

Key words. infinite-dimensional linear systems, J-spectral factorizations, well-posed linear systems, stability, Hankel operators, Lyapunov equations, Nehari problem

AMS subject classifications. 41A30, 47B35, 47N70, 93B28

DOI. 10.1137/S036301290444318X

1. Introduction.

The solution to the optimal Nehari problem is well known. The vector-valued case was solved by Page [22] (see also Nikol'skiĭ [20] and Peller [23]):

$$\inf_{\mathbf{K}(-s)\in\mathbf{H}_\infty(\mathcal{L}(U,Y))} \| \mathbf{G}+\mathbf{K} \|_\infty = \| H_{\mathbf{G}} \|,$$

where $\mathbf{G} \in \mathbf{L}_\infty(\mathcal{L}(U,Y))$, U, Y are separable Hilbert spaces, and $H_{\mathbf{G}}$ is the Hankel operator associated with the symbol $\mathbf{G}$. The suboptimal problem is to find for any $\sigma > \| H_{\mathbf{G}} \|$ all solutions $\mathbf{K}(-s) \in \mathbf{H}_\infty(\mathcal{L}(U,Y))$ to

$$\| \mathbf{G}+\mathbf{K} \|_\infty \le \sigma.$$

The suboptimal Nehari problem for functions on the disc has been solved in Kheifets [18] (see also Peller [23]), but for control applications we require explicit solutions in terms of state-space parameters of the continuous-time system as we explain below.

A crucial step in many control problems is solving the suboptimal Nehari problem for the stable case: $\mathbf{G} \in \mathbf{H}_\infty(\mathcal{L}(U,Y))$. In Salamon [26] it was shown that any $\mathbf{G} \in \mathbf{H}_\infty(\mathcal{L}(U,Y))$ has a system-stable well-posed realization; i.e., there exists a state space Z and (in general unbounded) operators A, B, C, where A is the infinitesimal generator of a strongly continuous semigroup on the separable Hilbert space Z and the following stability assumptions are satisfied:

$$C(sI-A)^{-1}z \in \mathbf{H}_2(Y), \quad B^*(sI-A^*)^{-1}z \in \mathbf{H}_2(U) \quad \forall z \in Z. \tag{1.1}$$

$\mathbf{G}$ is the transfer function of the well-posed linear system in the sense that

$$\mathbf{G}(s)-\mathbf{G}(\alpha) = (\alpha-s)C(sI-A)^{-1}(\alpha I-A)^{-1}B$$

for all α and s in some open right half-plane. Conversely, every system-stable well-posed linear system has a transfer function in $\mathbf{H}_\infty(\mathcal{L}(U,Y))$. Usually in control applications not $\mathbf{G}$, but A, B, and C are given, and one wants a solution $\mathbf{K}$ in terms of these state-space parameters.

*Received by the editors April 22, 2004; accepted for publication (in revised form) January 18, 2005; published electronically September 20, 2005.
http://www.siam.org/journals/sicon/44-3/44318.html
†Mathematics Institute, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands (curtain@math.rug.nl, opmeer@math.rug.nl).

Recently, it has been shown in Curtain and Sasane [9] that if $\rho(A) \cap i\mathbb{R} \neq \emptyset$, the Nehari problem for a system-stable well-posed linear system can be reduced to solving the Nehari problem for its reciprocal system which has bounded B and C operators. (The reciprocal approach to well-posed linear systems was introduced in Curtain [10] and [11].) Since problems with bounded B and C operators are technically simpler, we first consider the special case of bounded input and output operators. The next step is to use this special case to solve the general case.

Our approach to solving the suboptimal Nehari problem is to obtain solutions $\mathbf{K}$ via the J-spectral factorization problem of finding $\mathbf{X}$ such that

$$\mathbf{P}^*(i\omega)J_\sigma\mathbf{P}(i\omega) = \mathbf{X}(i\omega)J_1\mathbf{X}(i\omega)^* \quad \text{for almost all} \quad \omega \in \mathbb{R}, \tag{1.2}$$

where

$$\mathbf{P}(s) := \begin{bmatrix} I_Y & \mathbf{G}(s) \\ 0 & I_U \end{bmatrix} \quad \text{and} \quad J_\sigma := \begin{bmatrix} I_Y & 0 \\ 0 & -\sigma^2 I_U \end{bmatrix}.$$

The candidate solution for $\mathbf{X}$ involves the solutions of the Lyapunov equations

$$AL_1 + L_1A^* = -BB^*, \quad A^*L_2 + L_2A = -C^*C. \tag{1.3}$$

The smallest bounded nonnegative solutions are L_B, L_C, the controllability and observability gramians of the system $\Sigma(A, B, C, 0)$, respectively. These are not necessarily the only bounded nonnegative solutions. Once it is shown that $\mathbf{X}$ is indeed a solution to (1.2), the rest of the proof is relatively straightforward and one obtains a solution in terms of the known system parameters A, B, C, L_1, L_2, and σ.

There have been several versions of this approach in the literature; all but one (Curtain and Oostveen [6]) assume that A is the generator of an exponentially stable C_0-semigroup. We mention Curtain and Zwart [4], Glover, Curtain, and Partington [14], Ran [24], Curtain and Ran [2], Foias and Tannenbaum [13], Curtain and Zwart [3], and Curtain and Ichikawa [5], who all treat the problem under the assumption that A is the generator of an exponentially stable C_0-semigroup and varying additional assumptions.

For exponentially stable systems one can, since $i\mathbb{R} \subset \rho(A)$, verify directly that (1.2) holds for all $\omega \in \mathbb{R}$. However, there exist many systems with a stable transfer function for which A does not generate an *exponentially* stable C_0-semigroup. This motivated Curtain and Oostveen [6] to consider the class of system-stable systems satisfying (1.1) with bounded B and C and finite-dimensional U and Y. Now assumptions (1.1) provide no information about the spectrum of A and so it is not possible to verify (1.2) by a direct calculation. Unfortunately, this point was overlooked in [6]. We give an example of a system-stable system for which the candidate solution $\mathbf{X}$ does not satisfy (1.2) for a certain pair of solutions L_1, L_2 to the Lyapunov equations (1.3). This does not show that the claim in [6] is incorrect, since the claim in [6] is made only for the smallest solutions L_B and L_C. However, this counterexample does show that there is a gap in the proof in [6]. An easy remedy is to make an additional assumption on the spectrum of A, e.g., assume that $\sigma(A) \cap i\mathbb{R}$ has measure zero or that $\mathbb{C}_0^+ \subset \rho(A)$. Our major contribution is to show that if U and Y are finite-dimensional, then these assumptions are unnecessary. This new result has consequences for the recent paper by Ball, Mikkola, and Sasane [1] on the Nehari–Takagi problem, which is a generalization of the suboptimal Nehari problem. Using a J-spectral factorization approach, they solve the suboptimal Nehari–Takagi problem

for finite-dimensional U and Y under our assumptions (1.1) plus an assumption on the spectrum of A. Our result shows that the latter assumption is redundant.

Summarizing, under the assumptions (1.1), for any $\sigma > r^{1/2}(L_1L_2)$ (here $r(T)$ is the spectral radius of the operator T) we give an explicit formula for a spectral factor $\mathbf{X}$ satisfying (1.2) in terms of the system parameters A, B, C, L_1, and L_2 under either of the following additional assumptions:

A1. $\sigma(A) \cap i\mathbb{R}$ has measure zero.

A2. U and Y are finite-dimensional, and L_1 and L_2 are chosen to be the controllability and observability gramians L_B and L_C, respectively.

This leads to our second main result: a solution of the suboptimal Nehari problem in terms of the system parameters A, B, C, L_1, L_2, and σ under either of the above assumptions A1 or A2.

Our last main result is the extension of this result to the class of system-stable well-posed linear systems satisfying the assumption $\rho(A) \cap i\mathbb{R} \neq \emptyset$ and either assumption A1 or A2. We remark that in the well-posed case the standard formulas for the solution need not be well-defined, but we obtain alternative explicit formulas in terms of the reciprocal system as in Curtain and Sasane [9].

The paper is written to be as self-contained as possible. In section 2, we summarize relevant known results on state linear systems and in section 3 we prove some interesting new ones. Section 4 contains results on Riccati equations in terms of the concepts of input and output stability and stabilizability. In addition, we study two interesting Riccati equations connected to the Nehari problem. In section 5 we give an example of a system-stable system for which the candidate solution $\mathbf{X}$ does not satisfy (1.2) for a certain pair of solutions L_1, L_2 of the Lyapunov equations. However, we show that in the case that $L_1 = L_B$ and $L_2 = L_C$, we can always construct a spectral factor of (1.2). We collect several of its properties that enable us to obtain a solution of the suboptimal Nehari problem in section 6. In section 7, we obtain a parametrization of a family of solutions to the suboptimal Nehari problem for a system-stable state linear system. Finally, in section 8 we recall the concepts of system-stable well-posed linear systems and their reciprocals from [11]. Using the reciprocal approach from [9] we extend our results to obtain an explicit solution of the suboptimal Nehari problem for the class of system-stable well-posed linear systems under the assumption that $\rho(A) \cap i\mathbb{R} \neq \emptyset$ and either of the assumptions A1 or A2.

An interesting open question is whether our conclusions also hold if in assumption A2 we allow U and Y to be infinite-dimensional. The existence of frequency domain solutions is also known for this case (see Kheifets [18] or Peller [23]).

2. State linear systems: Known results. First we recall several known results for systems with bounded input and output operators. A is the generator of a strongly continuous semigroup $T(\cdot)$ on a separable Hilbert space Z, $B \in \mathcal{L}(U, Z), C \in \mathcal{L}(Z, Y), D \in \mathcal{L}(U, Y)$ with U, Y separable Hilbert spaces. Following the terminology in Curtain and Zwart [4] we call $\Sigma(A, B, C, D)$ a *state linear system*. We now define the transfer function and the characteristic function of a state linear system.

DEFINITION 2.1. *The transfer function* $\mathbf{G}$ *is defined as follows:* $\mathbf{G} - D$ *equals the Laplace transform of* $CT(t)B$ *on some right half-plane. We define the characteristic function* $\mathfrak{G}$ *for all* $s \in \rho(A)$ *by* $\mathfrak{G} = D + C(sI - A)^{-1}B$.

Remark 2.2. For s in some right half-plane we have $\mathbf{G} = \mathfrak{G}$, but they may differ outside this region. For a counterexample see Curtain and Zwart [4, Example 4.3.8]. A more detailed discussion is given in Zwart [33].

We introduce a stability concept that is weaker than exponential stability but

stronger than input-output stability. We will show that this is the right stability concept for the Nehari problem.

DEFINITION 2.3. *The state linear system* $\Sigma(A,B,C,D)$ *is* system-stable *if*

- *it is* input stable*: there exists a constant* $\beta > 0$ *such that for all* $u \in \mathbf{L}_2(0,\infty;U)$, $\| \int_0^\infty T(t)Bu(t)\,\mathrm{d}t\|^2 \le \beta \int_0^\infty \|u(t)\|^2\,\mathrm{d}t$*;*
- *it is* output stable*: there exists a constant* $\gamma > 0$ *such that for all* $z \in Z$, $\int_0^\infty \|CT(t)z\|^2\,\mathrm{d}t \le \gamma\|z\|^2$*;*
- *it is* input-output stable*: the transfer function* $\mathbf{G} \in \mathbf{H}_\infty(\mathcal{L}(U,Y))$.

Remark 2.4. We note that other authors may require the additional assumption that the semigroup be uniformly bounded for $t > 0$ for system stability (see Staffans [30] or Mikkola [19, Definition 6.1.1]). The essential difference in our definition is that we have made no stability assumptions on A and so it can have spectrum in the right half-plane $\mathbb{C}_0^+$.

DEFINITION 2.5. *The output map* $\mathcal{C} : Z \to \mathbf{L}_2(0,\infty;Y)$ *of an output stable state linear system* $\Sigma(A,B,C,D)$ *is defined by* $(\mathcal{C}z)(t) := CT(t)z$. *The input map* $\mathcal{B} : \mathbf{L}_2(0,\infty;U) \to Z$ *of an input stable state linear system is defined by*

$$\mathcal{B}u := \int_0^\infty T(s)Bu(s)\,ds.$$

The input and output stability properties are related to the existence of solutions to Lyapunov equations (see Grabowski [15] and Hansen and Weiss [16]).

LEMMA 2.6. *The state linear system* $\Sigma(A,B,C,D)$ *is* input stable *if and only if the following controllability Lyapunov equation has a bounded nonnegative solution* $L \in \mathcal{L}(Z)$:

$$ALz + LA^*z = -BB^*z \quad \forall\ z \in D(A^*). \tag{2.1}$$

In this case, the controllability gramian $L_B := \mathcal{B}\mathcal{B}^*$ *is the smallest bounded nonnegative solution of* (2.1) *and* $L_B^{1/2}T(t)^*z \to 0$ *as* $t \to \infty$ *for all* $z \in Z$.

The state linear system $\Sigma(A,B,C,D)$ *is* output stable *if and only if the following observability Lyapunov equation has a bounded nonnegative solution* $L \in \mathcal{L}(Z)$:

$$A^*Lz + LAz = -C^*Cz \quad \forall\ z \in D(A). \tag{2.2}$$

In this case, the observability gramian $L_C := \mathcal{C}^*\mathcal{C}$ *is the smallest bounded nonnegative solution of* (2.2) *and* $L_C^{1/2}T(t)z \to 0$ *as* $t \to \infty$ *for all* $z \in Z$.

The Hankel operator of a system is a fundamental concept.

DEFINITION 2.7. *For* $\mathbf{G} \in \mathbf{L}_\infty((-i\infty,i\infty);\mathcal{L}(U,Y))$ *we define the* Hankel operator with symbol $\mathbf{G}$ *as the operator* $H_{\mathbf{G}}$*:* $\mathbf{H}_2(U) \to \mathbf{H}_2(Y)$ *given by*

$$H_{\mathbf{G}}f = \Pi(\Lambda_{\mathbf{G}}f_-) \quad \textit{for} \quad f \in \mathbf{H}_2(U), \tag{2.3}$$

where $\Lambda_{\mathbf{G}}$ *is the multiplication map on* $\mathbf{L}_2((-i\infty,i\infty);U)$ *induced by* $\mathbf{G}$, Π *is the orthogonal projection from* $\mathbf{L}_2((-i\infty,i\infty);U)$ *onto* $\mathbf{H}_2(U)$, *and* $f_-(s) := f(-s)$.

Given $h \in \mathbf{L}_1^{\mathrm{loc}}([0,\infty);\mathcal{L}(U,Y))$, *we define the (time-domain)* Hankel operator Γ_h associated with h *for* $u \in \mathbf{L}_2^{\mathrm{loc}}([0,\infty);U)$ *with compact support by*

$$(\Gamma_h)(t) := \int_0^\infty h(t+\tau)u(\tau)d\tau. \tag{2.4}$$

There is a nice relationship between the time-domain and frequency-domain Hankel operators.

LEMMA 2.8. *Suppose that $\Sigma(A,B,C,0)$ is a system-stable system with impulse response $h(t)=CT(t)B$ and transfer function $\mathbf{G}$.*

1. *$\Gamma_h=\mathcal{CB}$ and it is isomorphic to $H_{\mathbf{G}}$ via*

$$\widehat{(\Gamma_h u)}(i\omega)=(H_{\mathbf{G}}\hat{u})(i\omega) \quad \text{for} \quad u\in \mathbf{L}_2([0,\infty);U). \tag{2.5}$$

Moreover,

$$\| H_{\mathbf{G}} \|=\| \Gamma_h \|= r^{\frac{1}{2}}(L_BL_C), \tag{2.6}$$

where r denotes the spectral radius and L_B, L_C are the controllability and observability gramians, respectively, of $\Sigma(A,B,C,0)$.

2. *If $\sigma > r^{\frac{1}{2}}(L_1L_2)$, where L_1, L_2 are arbitrary bounded nonnegative solutions of the Lyapunov equations (2.1), (2.2), respectively, then $N_\sigma := (I-\frac{1}{\sigma^2}L_1L_2)^{-1}\in \mathcal{L}(Z)$. Moreover, $W=N_\sigma L_1$ is nonnegative.*

Proof. 1. See Oostveen [21, Lemma 7.1.5].

2. Now $\sigma^2 > r(L_1L_2)$ implies that the spectral radius of $\frac{1}{\sigma^2}L_1L_2$ is less than 1 and so $I-\frac{1}{\sigma^2}L_1L_2$ is boundedly invertible. Noting that $L_1\geq 0$, the following shows that $W\geq 0$:

$$W=\left(I-\frac{1}{\sigma^2}L_1L_2\right)^{-1}L_1 = L_1^{1/2}\left(I-\frac{1}{\sigma^2}L_1^{1/2}L_2L_1^{1/2}\right)^{-1}L_1^{1/2}. \qquad \square$$

3. State linear systems: Some new results. In this section we develop several new results for state linear systems that we use in what follows, many of which are interesting in their own right.

First we examine the properties of the various concepts of stability from Definition 2.3 more closely.

LEMMA 3.1. *If $\Sigma(A,B,C,D)$ is output stable with observability gramian L_C, then for all $u\in \mathbf{L}_2^{\mathrm{loc}}(\mathbb{R};U)$ with compact support*

$$L_C^{1/2}\int_{-\infty}^{t} T(t-s)Bu(s)\,ds \to 0 \quad \text{as} \quad t\to\infty.$$

Proof. Let $\tau>0$ be such that $u(t)=0$ for $t>\tau$. Then

$$\begin{aligned} L_C^{1/2}\int_{-\infty}^{t} T(t-s)Bu(s)\,ds &= L_C^{1/2}\int_{-\infty}^{\tau} T(t-\tau)T(\tau-s)Bu(s)\,ds \\ &= L_C^{1/2}T(t-\tau)z(\tau), \end{aligned}$$

where $z(\tau)=\int_{-\infty}^{\tau}T(\tau-s)Bu(s)\,ds$ is independent of t. Lemma 2.6 now gives the result. $\square$

We recall that the output of a state linear system $\Sigma(A,B,C,D)$ with locally square integrable inputs u with support bounded to the left is defined by

$$y(t)=\int_{-\infty}^{t} CT(t-s)Bu(s)\,ds + Du(t). \tag{3.1}$$

Output stability does not imply input-output stability, but we do have the following partial result.

LEMMA 3.2. *If $\Sigma(A,B,C,D)$ is output stable, then for inputs $u\in \mathbf{L}_2^{\mathrm{loc}}(\mathbb{R};U)$ with compact support, the output given by (3.1) is in $\mathbf{L}_2(\mathbb{R};Y)$.*

Proof. Since u is square integrable we can assume without loss of generality that $D = 0$. Since u has compact support, there exists a $\tau > 0$ such that $u(t) = 0$ for $t > \tau$. We calculate the output $y(t)$ for $t > \tau$ as

$$\begin{aligned} y(t) &= \int_{-\infty}^{\tau} CT(t-\tau)T(\tau-s)Bu(s)\,ds \\ &= CT(t-\tau)\int_{-\infty}^{\tau} T(\tau-s)Bu(s)\,ds = CT(t-\tau)z(\tau), \end{aligned}$$

where

$$z(\tau) := \int_{-\infty}^{\tau} T(\tau-s)Bu(s)\,ds.$$

Since $\Sigma(A,B,C,D)$ is output stable we have

$$\begin{aligned} \int_{\tau}^{\infty} \| y(t) \|^2 \, dt &= \int_{0}^{\infty} \| CT(t)z(\tau) \|^2 \, dt \\ &\leq \text{const.} \| z(\tau) \|^2 < \infty. \end{aligned}$$

Since the output of a state linear system is always locally square integrable and the output has support bounded to the left by causality we have

$$\int_{-\infty}^{\infty} \| y(t) \|^2 \, dt = \int_{-\infty}^{\tau} \| y(t) \|^2 \, dt + \int_{\tau}^{\infty} \| y(t) \|^2 \, dt < \infty. \qquad \square$$

We next examine the connection between the transfer function and the characteristic function of a state linear system.

DEFINITION 3.3. *For an output stable state linear system we define* $\widehat{\mathcal{C}} : Z \to \mathbf{H}_2(Y)$ *by* $\widehat{\mathcal{C}}z := \widehat{\mathcal{C}z}$.

For an input stable state linear system we define $\widehat{\mathcal{B}}$ *for* $u \in U, z \in Z, s \in \mathbb{C}_0^+$ *by* $\langle \widehat{\mathcal{B}}(s)u, z\rangle := \langle u, \widehat{\mathcal{B}^*z}(\bar{s})\rangle$.

The following lemma shows that input or output stability ensures that the characteristic function and the transfer function are equal on the set where they are both defined. Parts of this lemma were shown for well-posed linear systems in [11, Lemma 2.3].

LEMMA 3.4.

1. *If the state linear system* $\Sigma(A,B,C,D)$ *is output stable, then*

$$\mathbf{G}(s) = D + \widehat{\mathcal{C}}(s)B \quad \forall s \in \mathbb{C}_0^+, \tag{3.2}$$

$$\mathbf{G}(s) = D + C(sI-A)^{-1}B = \mathfrak{G}(s) \quad \forall s \in \mathbb{C}_0^+ \cap \rho(A). \tag{3.3}$$

   Moreover, for all $u \in U$ *we have that* $(\mathbf{G} - D)u \in \mathbf{H}_2(Y)$.

2. *If the state linear system* $\Sigma(A,B,C,D)$ *is input stable, then* (3.3) *holds and*

$$\mathbf{G}(s) = D + C\widehat{\mathcal{B}}(s) \quad \forall s \in \mathbb{C}_0^+. \tag{3.4}$$

   Moreover, for all $u \in U, y \in Y$ *we have that* $\langle (\mathbf{G} - D)u, y\rangle \in \mathbf{H}_2$.

Proof. 1. Taking Laplace transforms of $CT(\cdot)z$, we obtain $C(sI-A)^{-1}z = \widehat{\mathcal{C}}(s)z$ and $\mathbf{G}(s) = \widehat{\mathcal{C}}(s)B + D$ for s in some right half-plane (see [4, Lemma 2.1.11]). Now since Σ is output stable, $\hat{\mathcal{C}}z \in \mathbf{H}_2(Y)$ for all $z \in Z$ and so $\widehat{\mathcal{C}}$ is holomorphic in $\mathbb{C}_0^+$ and

(3.2) holds. Again using that $\widehat{\mathcal{C}}$ is holomorphic on $\mathbb{C}_0^+$, the equality $\widehat{\mathcal{C}}(s)(sI-A)=C$ for s in some right half-plane extends to $\mathbb{C}_0^+$. Thus for $s\in\mathbb{C}_0^+\cap\rho(A)$ we have

$$\hat{\mathcal{C}}(s)(sI-A)(sI-A)^{-1}B=C(sI-A)^{-1}B,$$

which proves (3.3).

2. Similarly, input stability implies that $V(s)z\in\mathbf{H}_2(U)$ for $z\in Z$, where $V(s)z:=B^*(sI-A^*)^{-1}z$ on some right half-plane. Then $\langle u,V(\bar{s})z\rangle=\langle\widehat{\mathcal{B}}(s)u,z\rangle$ for all $z\in Z,u\in U$. So for all s in some right half-plane we have

$$\langle(sI-A)^{-1}Bu,z\rangle=\langle\widehat{\mathcal{B}}(s)u,z\rangle, \tag{3.5}$$

and letting $z=C^*x$, we obtain $\mathbf{G}=C\widehat{\mathcal{B}}+D$ on some right half-plane. Using that $\widehat{\mathcal{B}}$ is holomorphic, we obtain (3.4). To show (3.3) choose $z\in D(A^*)$, let $x=(\bar{s}I-A^*)z$, and substitute in (3.5) to obtain for s in some right half-plane

$$\langle Bu,x\rangle=\langle(sI-A)\widehat{\mathcal{B}}(s)u,x\rangle.$$

This extends to $s\in\mathbb{C}_0^+$ since $\widehat{\mathcal{B}}$ is holomorphic and to all $x\in Z$ by continuity. Thus

$$(sI-A)\widehat{\mathcal{B}}(s)=B\quad\forall\ s\in\mathbb{C}_0^+, \tag{3.6}$$

which proves (3.3). □

It turns out that the existence of boundary functions of $\mathbf{H}_2$ functions is crucial in our later proofs. We recall some basic results from [25]. Let $\omega\in\mathbb{R}$ and consider for $\alpha>0$ the cone

$$\Gamma_\alpha=\{s\in\mathbb{C}_0^+:|\mathrm{Im}(s)-\omega|<\alpha\ \mathrm{Re}(s)\}.$$

If $f\in\mathbf{H}_2(H)$, where H is a separable Hilbert space, then for almost all $\omega\in\mathbb{R}$ and all $\alpha>0$ the limit

$$\lim_{s\to i\omega,s\in\Gamma_\alpha}f(s)$$

exists. Such a limit is called a *nontangential limit*, and it associates with an $\mathbf{H}_2(H)$ function a function in $\mathbf{L}_2(i\mathbb{R},H)$ (see [25, Theorems 4.6.B and 4.8.B]).

It is well known that if U and Y are finite-dimensional and for all $u\in U$ we have $\mathbf{G}u\in\mathbf{H}_2(Y)$, then $\mathbf{G}\in\mathbf{H}_2(\mathcal{L}(U,Y))$. Hence we obtain an almost everywhere defined *boundary function* $\mathbf{G}:i\mathbb{R}\to\mathcal{L}(U,Y)$.

In the case where U and Y are infinite-dimensional we do have that for all $u\in U$ the function $\mathbf{G}(\cdot)u$ has a boundary function in $\mathbf{L}_2(i\mathbb{R};Y)$. However, in general there does not exist an almost everywhere defined function $F:i\mathbb{R}\to\mathcal{L}(U,Y)$ such that $F(i\omega)u$ equals this boundary function (see Mikkola [19, Example 3.3.6] for a counterexample).

If $\Sigma(A,B,C,D)$ is input or output stable and $\sigma(A)\cap i\mathbb{R}$ has measure zero, then $\mathbf{G}(s)=\mathfrak{G}(s)$ on $\mathbb{C}_0^+\cap\rho(A)$ by Lemma 3.4 and since $\mathfrak{G}(s)\to\mathfrak{G}(i\omega)$ as $s\to i\omega$ by continuity of the map $s\mapsto(sI-A)^{-1}$, we have $\mathbf{G}(s)\to\mathfrak{G}(i\omega)$. So if $\Sigma(A,B,C,D)$ is input or output stable and $\sigma(A)\cap i\mathbb{R}$ has measure zero, then $\mathbf{G}$ has an almost everywhere defined operator-valued boundary function. From the above we obtain the following.

LEMMA 3.5. *Let $\Sigma(A,B,C,D)$ be input or output stable and assume that either $\sigma(A)\cap i\mathbb{R}$ has measure zero or U and Y are finite-dimensional. Then there exists an*

almost everywhere defined function $\mathbf{G}_0 : i\mathbb{R} \to \mathcal{L}(U,Y)$ *such that for almost all* $\omega \in \mathbb{R}$ *and all nontangential paths we have*

$$\mathbf{G}_0(i\omega) = \lim_{s \to i\omega} \mathbf{G}(s).$$

Moreover, if $i\omega \in \rho(A)$, *then* $\mathbf{G}_0(i\omega) = \mathfrak{G}(i\omega)$.

Proof. This follows from the paragraphs preceding the lemma. ∎

We prove the following lemma that will be useful later.

LEMMA 3.6. *Let* $f : \mathbb{C}_0^+ \to \mathcal{L}(U,Y)$ *be such that for every* $u \in U$ *we have* $f(\cdot)u \in \mathbf{H}_2(Y)$. *Assume there exists a function* $f_0 \in \mathbf{L}_\infty(i\mathbb{R}; \mathcal{L}(U,Y))$ *such that for all* $u \in U$ *there exists a set* $\mathcal{N}_u$ *of measure zero such that for all* $\omega \in \mathbb{R} - \mathcal{N}_u$ *and all nontangential paths we have*

$$f_0(i\omega)u = \lim_{s \to i\omega} f(s)u.$$

Then $f \in \mathbf{H}_\infty(\mathcal{L}(U,Y))$.

Proof. Since $f(\cdot)u \in \mathbf{H}_2(Y)$ we have the Poisson representation [25, Theorem 4.8.A]

$$f(a+ib)u = \frac{1}{\pi} \int_{\mathbb{R}} \frac{b f_0(i\omega)u}{(t-a)^2 + b^2}\, dt,$$

so we have (using that the Poisson kernel has integral one)

$$\|f(a+ib)u\| \le \frac{1}{\pi} \int_{\mathbb{R}} \|f_0(i\omega)u\| \frac{b}{(t-a)^2+b^2}\, dt \le \operatorname*{ess\,sup}_{t \in \mathbb{R}} \|f_0(it)\|\ \|u\|.$$

This shows that

$$\sup_{s \in \mathbb{C}_0^+} \|f(s)\| \le \operatorname*{ess\,sup}_{t \in \mathbb{R}} \|f_0(it)\|$$

and since f is holomorphic we have $f \in \mathbf{H}_\infty(\mathcal{L}(U,Y))$. ∎

LEMMA 3.7. *Let* $\Sigma(A,B,C,D)$ *be output stable and assume that either* $\sigma(A) \cap i\mathbb{R}$ *has measure zero or* U *and* Y *are finite-dimensional. Then for inputs with compact support we have for almost all* $\omega \in \mathbb{R}$

$$\hat{y}(i\omega) = \mathbf{G}(i\omega)\hat{u}(i\omega). \tag{3.7}$$

Proof. We first prove the statement for the case that u is zero for negative time. Now on some right half-plane we have

$$\hat{y}(s) = \mathbf{G}(s)\hat{u}(s). \tag{3.8}$$

Since $\hat{y} \in \mathbf{H}_2(Y)$ by Lemma 3.2 (and causality) and $\mathbf{G}$ is holomorphic on $\mathbb{C}_0^+$ by Lemma 3.4 this extends to $\mathbb{C}_0^+$. By Lemma 3.5 we have $\mathbf{G}(s) \to \mathbf{G}(i\omega)$ in the operator norm as $s \to i\omega$. Since $\hat{u} \in \mathbf{H}_2(U)$ and $\hat{y} \in \mathbf{H}_2(Y)$, they converge to their boundary functions as $s \to i\omega$ so we obtain (3.7). The general case follows by applying the above to the function $\underline{u}(t) := u(t-\tau)$ with output $\underline{y}(t) := y(t-\tau)$, where y is the output corresponding to u and τ is chosen such that $\underline{u}$ is zero for negative time. ∎

In the proof of Lemma 3.8 we need to study systems defined on the positive time axis only and with a given initial state. We summarize some known results for this

type of system. For an input $u \in \mathbf{L}_2^{\mathrm{loc}}(0,\infty;U)$ and initial state $x_0 \in X$ the state $x(t) \in X$ at time $t \geq 0$ is defined by

$$x(t) = T(t)x_0 + \int_0^t T(t-s)Bu(s)\,ds.$$

If u is continuously differentiable and $x_0 \in D(A)$, then x as defined above is differentiable and satisfies

$$\dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = x_0.$$

The output of the state linear system is defined by

$$y(t) = Cx(t) + Du(t).$$

A state linear system is well-posed in the sense that for all $t > 0$ there exists a $K > 0$ such that for all $u \in \mathbf{L}_2^{\mathrm{loc}}(0,\infty;U)$ and all $x_0 \in X$

$$\|x(t)\|^2 + \int_0^t \|y(s)\|^2\,ds \leq K\left(\|x_0\|^2 + \int_0^t \|u(s)\|^2\,ds\right);$$

i.e., the map from the initial state and the input restricted to $(0,t)$ to the state at time t and the output restricted to $(0,t)$ is continuous from $X \times L^2(0,t;U)$ to $X \times L^2(0,t;Y)$.

LEMMA 3.8. *Let $\Sigma(A,B,C,0)$ be output stable with observability gramian L_C and let $u_i \in \mathbf{L}_2^{\mathrm{loc}}(\mathbb{R};U)$ be an input with compact support. Denote by y_i the output of $\Sigma(A,B,C,0)$ with input u_i given by* (3.1) *and by y_i^{LC} the output of $\Sigma(A,B,B^*L_C,0)$ with input u_i given by the corresponding* (3.1). *Then we have the following*:

$$\int_{-\infty}^{\infty} \langle y_1(t), y_2(t)\rangle\,dt = \int_{-\infty}^{\infty} \langle u_1(t), y_2^{LC}(t)\rangle\,dt + \int_{-\infty}^{\infty} \langle y_1^{LC}(t), u_2(t)\rangle\,dt. \tag{3.9}$$

Proof. We first note that the integrals in (3.9) are well defined since $y_i \in \mathbf{L}_2(\mathbb{R};Y)$ by Lemma 3.2 and the u_i have compact support. Set $z_i(t) = \int_{-\infty}^t T(t-s)Bu_i(s)\,ds$ for $i = 1,2$. If u_i is continuously differentiable, then $z_i(t)$ is differentiable and

$$\begin{aligned}
&\frac{d}{dt}\langle z_1(t), L_C z_2(t)\rangle \\
&= \langle Az_1(t) + Bu_1(t), L_C z_2(t)\rangle + \langle L_C z_1(t), Az_2(t) + Bu_2(t)\rangle \\
&= \langle Bu_1(t), L_C z_2(t)\rangle + \langle L_C z_1(t), Bu_2(t)\rangle - \langle Cz_1(t), Cz_2(t)\rangle,
\end{aligned}$$

where we have used (2.2). On integrating the above we obtain

$$\langle z_1(t), L_C z_2(t)\rangle \tag{3.10}$$

$$= \int_{-\infty}^t \langle u_1(s), B^*L_C z_2(s)\rangle\,ds + \int_{-\infty}^t \langle B^*L_C z_1(s), u_2(s)\rangle\,ds - \int_{-\infty}^t \langle Cz_1(s), Cz_2(s)\rangle\,ds. \tag{3.11}$$

$$= \int_{-\infty}^t \langle u_1(s), y_2^{LC}(s)\rangle\,ds + \int_{-\infty}^t \langle y_1^{LC}(s), u_2(s)\rangle\,ds - \int_{-\infty}^t \langle y_1(s), y_2(s)\rangle\,ds. \tag{3.12}$$

From Lemma 3.1 we conclude that the left-hand side of (3.12) converges to zero as $t \to \infty$. This proves (3.9) for the case of continuously differentiable inputs, and the general case follows by the following approximation argument. Let $u_i \in \mathbf{L}_2^{\mathrm{loc}}(\mathbb{R};U)$ be inputs with compact support and let $u_i^n \in \mathbf{L}_2^{\mathrm{loc}}(\mathbb{R};U)$ be continuously differentiable inputs with compact support that converge to u_i in $\mathbf{L}_2(\mathbb{R};U)$. Let τ be such that u_i and u_i^n are equal to zero on (τ,∞), and assume that u_i and u_i^n are zero on $(-\infty,0)$. By the well-posedness there exists a $K(\tau)$ such that

$$\int_{-\infty}^{\tau} \|y_i(s)-y_i^n(s)\|^2\,ds \le K(\tau)\int_{-\infty}^{\tau} \|u_i(s)-u_i^n(s)\|^2\,ds$$

and hence

$$\int_{-\infty}^{\tau} \|y_i(s)-y_i^n(s)\|^2\,ds \to 0 \ \text{ as } \ n\to\infty.$$

For $z_i^n(\tau) := \int_{-\infty}^{\tau} T(\tau-s)Bu_i(s)\,ds$ we have

$$\int_{\tau}^{\infty} \|y_i(s)-y_i^n(s)\|^2\,ds = \int_0^{\infty} \|CT(s)\left(z_i(\tau)-z_i^n(\tau)\right)\|^2\,ds.$$

Since $\Sigma(A,B,C,0)$ is output stable, there exists a $\gamma>0$ such that

$$\int_{\tau}^{\infty} \|y_i(s)-y_i^n(s)\|^2\,ds \le \gamma\ \|z_i(\tau)-z_i^n(\tau)\|^2,$$

and by the well-posedness there exists a $K(\tau)$ such that

$$\|z_i(\tau)-z_i^n(\tau)\|^2 \le K(\tau)\int_{-\infty}^{\tau} \|u_i(s)-u_i^n(s)\|^2\,ds.$$

Hence

$$\int_{\tau}^{\infty} \|y_i(s)-y_i^n(s)\|^2\,ds \to 0 \ \text{ as } \ n\to\infty.$$

So we have $y_i^n \to y_i$ in $\mathbf{L}_2(\mathbb{R};Y)$. By the compact support of the inputs u_i and u_i^n we need only $y_i^{n,LC} \to y_i^{LC}$ in $\mathbf{L}_2(-\infty,\tau;Y)$ to establish (3.9). This convergence follows from the well-posedness as above. Using this (3.9) follows. The case that u_i is not zero on $(-\infty,0)$ can be reduced to the case that this is the case by a time-shift as in the proof of Lemma 3.7. ∎

We also need to study anticausal outputs of state linear systems. The anticausal output of the state linear system $\Sigma(A,B,C,D)$ for an input $u \in \mathbf{L}_2^{\mathrm{loc}}(\mathbb{R};U)$ with support bounded to the right is defined as

$$y^a(t) := \int_t^{\infty} CT(s-t)Bu(s)\,ds + Du(t). \tag{3.13}$$

We have the following analogue of Lemma 3.2.

LEMMA 3.9. *If $\Sigma(A,B,C,D)$ is output stable, then for inputs $u \in \mathbf{L}_2^{\mathrm{loc}}(\mathbb{R};U)$ with compact support, the anticausal output given by* (3.13) *is in $\mathbf{L}_2(\mathbb{R};Y)$.*

Proof. The proof is as in the proof of Lemma 3.2. ∎

We have the following analogue of Lemma 3.7.

LEMMA 3.10. *Let $\Sigma(A,B,C,D)$ be output stable and assume that either $\sigma(A)\cap i\mathbb{R}$ has measure zero or U and Y are finite-dimensional. Then for inputs with compact support we have for almost all $\omega\in\mathbb{R}$*

$$\hat{y}^a(i\omega)=\mathbf{G}(-i\omega)\hat{u}(i\omega), \tag{3.14}$$

where y^a is the anticausal output of $\Sigma(A,B,C,D)$ defined by (3.13).

Proof. This follows as in the proof of Lemma 3.7, but now by first assuming u to be zero for positive time and using the Hardy space $\mathbf{H}_2$ over the left half-plane. Details are as follows. We first prove the statement for the case that u is zero for positive time. Now on some left half-plane we have

$$\hat{y}^a(s)=\mathbf{G}(-s)\hat{u}(s). \tag{3.15}$$

Since $\hat{y}^a\in\mathbf{H}_2(\mathbb{C}_0^-;Y)$ by Lemma 3.9 (and anticausality) and $\mathbf{G}$ is holomorphic on $\mathbb{C}_0^+$ by Lemma 3.4 this extends to $\mathbb{C}_0^-$. By Lemma 3.5 we have $\mathbf{G}(s)\to\mathbf{G}(i\omega)$ in the operator norm as $s\to i\omega$. Since $\hat{u}\in\mathbf{H}_2(\mathbb{C}_0^-;U)$ and $\hat{y}^a\in\mathbf{H}_2(\mathbb{C}_0^-;Y)$, they converge to their boundary functions as $s\to i\omega$ so we obtain (3.14). The general case follows by applying the above to the function $\underline{u}(t):=u(t+\tau)$ with output $\underline{y}^a(t):=y^a(t+\tau)$, where y is the output corresponding to u and τ is chosen such that $\underline{u}$ is zero for positive time. ∎

The next result is a consequence of Lemma 3.8.

LEMMA 3.11. *Let $\Sigma(A,B,C,0)$ be input and output stable with observability gramian L_C and let $u_i\in\mathbf{L}_2^{\mathrm{loc}}(\mathbb{R};U)$ be an input with compact support. Denote by y_i the output of $\Sigma(A,B,C,0)$ with input u_i given by* (3.1) *and by y_i^a the anticausal output of $\Sigma(A^*,L_CB,B^*,0)$ with input u_i given by the corresponding* (3.13). *Then we have $y_i^a\in\mathbf{L}_2(\mathbb{R};U)$ and the following:*

$$\int_{-\infty}^{\infty}\langle y_1(t),y_2(t)\rangle\,dt=\int_{-\infty}^{\infty}\langle y_1^a(t),u_2(t)\rangle\,dt+\int_{-\infty}^{\infty}\langle u_1(t),y_2^a(t)\rangle\,dt. \tag{3.16}$$

Proof. Since $\Sigma(A,B,C,0)$ is input stable the system $\Sigma(A^*,L_CB,B^*,0)$ is output stable. Lemma 3.9 then implies that $y_i^a\in\mathbf{L}_2(\mathbb{R};U)$. Equation (3.16) follows from (3.9) by an application of Fubini's theorem and a change of variables. ∎

LEMMA 3.12. *If $\Sigma(A,B,C,0)$ is input and output stable with the observability gramian L_C, then*

$$\begin{aligned}&\langle L_CBu,\widehat{\mathcal{B}}(s)u\rangle+\langle\widehat{\mathcal{B}}(s)u,L_CBu\rangle\\&=\|\,\mathbf{G}(s)u\,\|^2+2\mathrm{Re}\;s\;\|L_C^{1/2}\widehat{\mathcal{B}}(s)u\|^2\quad\forall\;s\in\mathbb{C}_0^+,u\in U.\end{aligned} \tag{3.17}$$

Proof. We obtain a straightforward frequency domain identity from the Lyapunov equation (2.2):

$$(\bar{s}I-A^*)L_C+L_C(sI-A)=C^*C+2\mathrm{Re}\;s\;L_C.$$

This leads to the following identity on some right half-plane:

$$\begin{aligned}&B^*L_C(sI-A)^{-1}B+B^*(\bar{s}I-A^*)^{-1}L_CB\\&=B^*(\bar{s}I-A^*)^{-1}C^*C(sI-A)^{-1}B+2\mathrm{Re}\;sB^*(\bar{s}I-A)^{-1}L_C(sI-A)^{-1}B.\end{aligned}$$

From this we obtain (3.17) for s in some right half-plane using Lemma 3.4. From the input stability of $\Sigma(A,B,C,D)$ we obtain that $\widehat{\mathcal{B}}$ and $\mathbf{G}$ are holomorphic on $\mathbb{C}_0^+$.

From the appendix (Corollaries 9.1 and 9.4) it follows that all terms in (3.17) are real-analytic on $\mathbb{C}_0^+$. By the identity theorem for real-analytic functions we obtain (3.17) for $s \in \mathbb{C}_0^+$. ∎

The next lemma is the main result of this section.

LEMMA 3.13. *Let $\Sigma(A,B,C,0)$ be input and output stable with the observability gramian L_C and assume that either $\sigma(A) \cap i\mathbb{R}$ has measure zero or U and Y are finite-dimensional. Then for almost all $\omega \in \mathbb{R}$, all $u \in U$, and all nontangential paths*

$$\lim_{s \to i\omega} \text{Re } s \parallel L_C^{1/2} \widehat{\mathcal{B}}(s)u \parallel^2 = 0. \tag{3.18}$$

Proof. Let u_i be locally square integrable with compact support. Let y_i denote the output of $\Sigma(A,B,C,0)$ for input u_i from (3.1) and let y_i^a denote the anticausal output of $\Sigma(A^*, L_C B, B^*, 0)$ for input u_i from the corresponding (3.13). By Fourier transforming (3.16) we obtain

$$\int_{-\infty}^{\infty} \langle \hat{y}_1(i\omega), \hat{y}_2(i\omega) \rangle \, d\omega = \int_{-\infty}^{\infty} \langle \hat{y}_1^a(i\omega), \hat{u}_2(i\omega) \rangle \, d\omega + \int_{-\infty}^{\infty} \langle \hat{u}_1(i\omega), \hat{y}_2^a(i\omega) \rangle \, d\omega.$$

From Lemmas 3.7 and 3.10 we obtain

$$\int_{-\infty}^{\infty} \langle \mathbf{G}(i\omega)\hat{u}_1(i\omega), \mathbf{G}(i\omega)\hat{u}_2(i\omega) \rangle \, d\omega$$
$$= \int_{-\infty}^{\infty} \langle \mathbf{G}^{LC}(-i\omega)\hat{u}_1(i\omega), \hat{u}_2(i\omega) \rangle \, d\omega + \int_{-\infty}^{\infty} \langle \hat{u}_1(i\omega), \mathbf{G}^{LC}(-i\omega)\hat{u}_2(i\omega) \rangle \, d\omega,$$

where $\mathbf{G}^{LC}$ is the transfer function of the system $\Sigma(A^*, L_C B, B^*, 0)$. Letting $u_i(t) = f(t)v$, where f is a scalar function with compact support and $v \in U$, we obtain

$$\int_{-\infty}^{\infty} |\hat{f}(i\omega)|^2 \left(\langle \mathbf{G}^{LC}(-i\omega)v, v \rangle + \langle v, \mathbf{G}^{LC}(-i\omega)v \rangle - \|\mathbf{G}(i\omega)v\|^2 \right) d\omega = 0. \tag{3.19}$$

From (3.17) we obtain using $\mathbf{G}^{LC}(s) = \widehat{\mathcal{B}}(\bar{s})^* L_C B$ that for all $s \in \mathbb{C}_0^+$

$$\langle \mathbf{G}^{LC}(\bar{s})v, v \rangle + \langle v, \mathbf{G}^{LC}(\bar{s})v \rangle - \|\mathbf{G}(s)v\|^2 = 2\text{Re } s \, \|L_C^{1/2} \widehat{\mathcal{B}}(s)v\|^2 \geq 0.$$

Taking nontangential limits (which exist for all three terms on the left-hand side) we obtain for almost all $\omega \in \mathbb{R}$

$$\langle \mathbf{G}^{LC}(-i\omega)v, v \rangle + \langle v, \mathbf{G}^{LC}(-i\omega)v \rangle - \|\mathbf{G}(i\omega)v\|^2 \geq 0. \tag{3.20}$$

Combining (3.19) and (3.20) we obtain for almost all $\omega \in \mathbb{R}$

$$|\hat{f}(i\omega)|^2 \left(\langle \mathbf{G}^{LC}(-i\omega)v, v \rangle + \langle v, \mathbf{G}^{LC}(-i\omega)v \rangle - \|\mathbf{G}(i\omega)v\|^2 \right) = 0.$$

Now let $f : \mathbb{R} \to \mathbb{C}$ be a function that has compact support and such that $\hat{f}(i\omega) \neq 0$ for almost all $\omega \in \mathbb{R}$ (for example, the function equal to 1 on [0,1] and zero elsewhere). Then we obtain for all $v \in U$ and almost all $\omega \in \mathbb{R}$

$$\langle \mathbf{G}^{LC}(-i\omega)v, v \rangle + \langle v, \mathbf{G}^{LC}(-i\omega)v \rangle - \|\mathbf{G}(i\omega)v\|^2 = 0,$$

and comparing the above with (3.17) proves (3.18). ∎

It is an interesting open question whether Lemma 3.13 is true for infinite-dimensional U and Y without the spectrum assumption.

We need the following property of the system $\Sigma(A, B, B^*L_C, 0)$ shown in Weiss and Weiss [31, Theorem 11.1] (see also Oostveen [21, Lemma 4.2.6]).

LEMMA 3.14. *If $\Sigma(A, B, C, D)$ is output stable and input-output stable with observability gramian L_C, then $\Sigma(A, -, B^*L_C, -)$ is output stable.*

By duality we obtain the following.

COROLLARY 3.15. *If $\Sigma(A, B, C, D)$ is input stable and input-output stable with controllability gramian L_B, then $\Sigma(A, L_BC^*, -, -)$ is input stable.*

Lemma 3.13 has an easy corollary.

COROLLARY 3.16. *Let $\Sigma(A, B, C, 0)$ be a system-stable state linear system and assume that either $\sigma(A) \cap i\mathbb{R}$ has measure zero or U and Y are finite-dimensional. Denote by $\widehat{\mathcal{B}}_{LB}(\bar{s})^*$ and $\widehat{\mathcal{C}}_{LC}(s)$ the holomorphic extensions to $\mathbb{C}_0^+$ of $CL_B(sI - A^*)^{-1}$ and $B^*L_C(sI - A)^{-1}$, respectively. Then for almost all $\omega \in \mathbb{R}$, all $u \in U$, and all nontangential paths*

$$\lim_{s \to i\omega} \text{Re } s \parallel L_C^{1/2} \widehat{\mathcal{B}}_{LB}(s)u \parallel^2 = 0, \tag{3.21}$$

$$\lim_{s \to i\omega} \text{Re } s \parallel L_B^{1/2} \widehat{\mathcal{C}}(\bar{s})^* y \parallel^2 = 0, \tag{3.22}$$

$$\lim_{s \to i\omega} \text{Re } s \parallel L_B^{1/2} \widehat{\mathcal{C}}_{LC}(\bar{s})^* y \parallel^2 = 0. \tag{3.23}$$

Proof. Equation (3.21) follows from Lemma 3.13, since by Lemma 3.15, $\Sigma(A, L_BC^*, C, 0)$ is input stable and output stable and has observability gramian L_C. Equation (3.22) is the dual of (3.18), and (3.23) is the dual of (3.21). □

4. Riccati equations. In this section we obtain some new results on stabilizability and Riccati equations for state linear systems that we will need in what follows. First we introduce concepts of stabilizability from Curtain [10] that are refinements of the definitions introduced in Curtain and Oostveen [7].

DEFINITION 4.1. $\Sigma(A, B, C, D)$ *is* output stabilizable *if there exists an $F \in \mathcal{L}(Z, U)$ such that $\Sigma(A + BF, B, [C; F], 0)$ is output stable.* $\Sigma(A, B, C, D)$ *is* input stabilizable *if there exists an $L \in \mathcal{L}(Y, Z)$ such that $\Sigma(A + LC, [B, L], C, 0)$ is input stable.*

The following are extensions of the results in Curtain and Oostveen [7]. In fact, they are special cases of analogous results for the very large class of well-posed linear systems in Mikkola [19]. Since the proofs there are not so accessible, we give simple proofs here.

THEOREM 4.2. *If the state linear system $\Sigma(A, B, C, 0)$ is output stabilizable, then there exists a smallest bounded nonnegative solution of the control Riccati equation for $z \in D(A)$:*

$$A^*Qz + QAz + C^*Cz - QBB^*Qz = 0. \tag{4.1}$$

Moreover, for any bounded nonnegative solution, $\Sigma(A_Q, B, [C; -B^*Q], 0)$ is output stable, where $A_Q = A - BB^*Q$. If, in addition, $\Sigma(A, B, C, 0)$ is input stabilizable, then it is system-stable. If $\Sigma(A, B, C, 0)$ is input stabilizable, then there exists a smallest bounded nonnegative solution to the filter Riccati equation for $z \in D(A^*)$:

$$APz + PA^*z - PC^*CPz + BB^*z = 0. \tag{4.2}$$

Moreover, for any bounded nonnegative solution, $\Sigma(A_P, [B, -PC^*], C, 0)$ *is input stable, where* $A_P = A - PC^*C$. *If, in addition,* $\Sigma(A, B, C, 0)$ *is output stabilizable, then it is system-stable.*

Proof. The existence of a smallest bounded nonnegative solution to the Riccati equation has been shown in Curtain and Oostveen [7]. In fact, it follows from [4, Theorem 6.2.4], since output stabilizability implies optimizability. Next we note that the output stability of $\Sigma(A_Q, B, [C; -B^*Q], 0)$ follows from the following equivalent formulation of the Riccati equation:

$$A_Q^*Qz + QA_Qz + QBB^*Qz + C^*Cz = 0 \quad \text{for} \quad z \in D(A). \tag{4.3}$$

This is the observability Lyapunov equation for $\Sigma(A_Q, B, [C; -B^*Q], 0)$, and Lemma 2.6 shows that it is output stable.

Next we observe that the input stabilizability guarantees the existence of a solution P to the dual filter Riccati equation (4.2). This in turn shows that the solutions to the Lyapunov equations of the system $\Sigma(A_Q, B, [C; -B^*Q], 0)$ are Q and $P(I+QP)^{-1}$ (we use the dual version of Lemma 9.4.10 in [4]). So this system is input stable (see Lemma 2.6). So, from Theorem 3.4, we can write the transfer function $[\mathbf{N}; \mathbf{M}]$ of the closed-loop system on $\mathbb{C}_0^+$ in two ways:

$$[\mathbf{N}; \mathbf{M}] - [0; I] = \widehat{\mathcal{C}}_Q B = [C; -B^*Q]\widehat{\mathcal{B}}_Q,$$

where $\widehat{\mathcal{C}}_Q z$ is the Laplace transform of $[C; -B^*Q]T_Q(t)z$, $\langle \widehat{\mathcal{B}}_Q u, z\rangle$ is the Laplace transform of $\langle T_Q(t)Bu, z\rangle$ for all $z \in Z$ and $u \in U$, and T_Q is the semigroup generated by A_Q. We use this latter version of $[\mathbf{N}; \mathbf{M}]$ to compute for $s \in \mathbb{C}_0^+$

$$\begin{aligned}&[\mathbf{N}(s); \mathbf{M}(s)]^*[\mathbf{N}(s); \mathbf{M}(s)] \\ &= \widehat{\mathcal{B}}_Q^*(s)[C^*C + QBB^*Q]\widehat{\mathcal{B}}_Q(s) + I - \widehat{\mathcal{B}}_Q^*(s)QB - B^*Q\widehat{\mathcal{B}}_Q(s).\end{aligned} \tag{4.4}$$

We then use the formulation (4.3) of the Riccati equation to obtain

$$C^*C + QBB^*Q = A_Q^*Q + QA_Q = (sI - A_Q)^*Q + Q(sI - A_Q) - 2\mathrm{Re}\ s\ Q.$$

We substitute this into (4.4) and use the equality (3.6) applied to the closed-loop system $\Sigma(A_Q, B, [C; -B^*Q], 0)$,

$$(s - A_Q)\widehat{\mathcal{B}}_Q(s) = B \quad \text{for} \quad s \in \mathbb{C}_0^+,$$

to obtain

$$[\mathbf{N}(s); \mathbf{M}(s)]^*[\mathbf{N}(s); \mathbf{M}(s)] = I - 2\mathrm{Re}\ s\ \widehat{\mathcal{B}}_Q^*(s)Q\widehat{\mathcal{B}}_Q(s). \tag{4.5}$$

This shows that $[\mathbf{N}(s); \mathbf{M}(s)]^*[\mathbf{N}(s); \mathbf{M}(s)] \leq I$ for all $s \in \mathbb{C}_0^+$. Thus $[\mathbf{N}; \mathbf{M}] \in \mathbf{H}_\infty(\mathcal{L}(U, Y \oplus U))$. ∎

We proceed to deduce some interesting properties of the spectrum of the closed-loop generators A_Q and A_P on the right half-plane.

LEMMA 4.3. *Suppose that the state linear system* $\Sigma(A, B, C, 0)$ *is input and output stabilizable. Then for any bounded nonnegative solutions* Q, P *to the Riccati equations* (4.1), (4.2)*, respectively, the closed-loop generators have the following properties:*

1. *The closed-loop operators* $A_Q = A - BB^*Q$ *and* $A_P = A - PC^*C$ *have the same spectrum and*

$$(I + PQ)A_Qz = A_P(I + PQ)z \quad \textit{for} \quad z \in D(A). \tag{4.6}$$

2. *Let Q_1, Q_2 be two bounded nonnegative solutions of* (4.1). *Then* $\sigma(A_{Q_1}) = \sigma(A_{Q_2})$.
3. *The spectra of the closed-loop generators A_Q and A_P in the closed right half-plane are contained in the spectrum of A.*

Proof. 1. First we prove (4.6). As in Curtain and Zwart [4, Lemma 4.1.24] we have

$$Q : D(A) \to D(A^*), \qquad P : D(A^*) \to D(A).$$

So using (4.3) we obtain for $z \in D(A)$

$$\begin{aligned}
&(I + PQ)A_Q z \\
&= A_Q z - P(A_Q^* Qz + QB^*BQ + CC)z \\
&= (A - PC^*C)z - BB^*Qz - PA^*Qz \\
&= A_P z - P(A_P^* + C^*CP)Qz - BB^*Qz \\
&= A_P z + (A_P P + BB^*)Qz - BBQz \ \text{ from (4.2)} \\
&= A_P(I + PQ)z.
\end{aligned}$$

Since P, Q are bounded nonnegative operators, $(I + PQ)$ is boundedly invertible and $\sigma(A_Q) = \sigma(A_P)$.

2. From part 1 it follows that $\sigma(A_{Q_1}) = \sigma(A_P) = \sigma(A_{Q_2})$.

3. Suppose that $\lambda \in \overline{\mathbb{C}_0^+}$ is in the point spectrum of A_Q; i.e., $A_Q x = \lambda x$ for some nonzero $x \in D(A)$. Then from (4.3) we obtain

$$\begin{aligned}
2\ \mathrm{Re}\lambda\langle Qx, x\rangle &= \langle A_Q x, Qx\rangle + \langle Qx, A_Q x\rangle \\
&= -\|B^*Qx\|^2 - \|Cx\|^2.
\end{aligned}$$

Since $Q \geq 0$ and $\mathrm{Re}\ \lambda \geq 0$ we must have $B^*Qx = 0 = Cx$, which implies that $\lambda x = A_Q x = Ax$. So λ is in the point spectrum of A. Suppose now that $\mu \in \overline{\mathbb{C}_0^+}$ is in the residual spectrum of A_Q. Then by (4.6) $\bar{\mu}$ is in $P\sigma(A_Q^*) = P\sigma(A_P^*)$ and so there exists a $y \in D(A^*)$ such that $A_P^* y = \overline{\mu} y$. Now (4.2) can be reformulated as

$$A_P Pz + PA_P^* z = -PC^*CPz - BB^*z, \tag{4.7}$$

and substituting $z = y$ and taking the inner product with y gives

$$\begin{aligned}
2\ \mathrm{Re}\mu\langle Py, y\rangle &= \langle A_P^* y, Py\rangle + \langle Py, A_P^* y\rangle \\
&= -\|B^*y\|^2 - \|CPy\|^2.
\end{aligned}$$

Since $\mathrm{Re}\ \mu \geq 0$ and $P \geq 0$, we must have $CPy = 0 = B^*y$, which implies that $\overline{\mu} y = A_P^* y = A^* y$ and so $\mu \in \sigma(A)$. Suppose now that $\lambda \in \overline{\mathbb{C}_0^+}$ is in the continuous spectrum of A_Q. Then there exists a sequence $x_n \in D(A)$ with $\|x_n\| = 1$ and $\|A_Q x_n - \lambda x_n\| \to 0$ as $n \to \infty$. Substituting in (4.3) we obtain

$$\begin{aligned}
&\langle A_Q x_n - \lambda x_n, Qx_n\rangle + \langle Qx_n, A_Q x_n - \lambda x_n\rangle \\
&= -\|B^*Qx_n\|^2 - \|Cx_n\|^2 - 2\ \mathrm{Re}\lambda\langle Qx_n, x_n\rangle.
\end{aligned}$$

Since $Q \geq 0$ and $\mathrm{Re}\ \lambda \geq 0$, we deduce that $\|B^*Qx_n\|^2 \leq 2\|Q\|\ \|x_n\|\ \|A_Q x_n - \lambda x_n\| \to 0$ as $n \to \infty$. Thus

$$\|Ax_n - \lambda x_n\| \leq \|A_Q x_n - \lambda x_n\| + \|BB^*Qx_n\| \to 0$$

as $n \to \infty$. So λ is in the approximate point spectrum of A. The above shows that $\rho(A_Q) \cap \overline{\mathbb{C}_0^+} \subset \rho(A) \cap \overline{\mathbb{C}_0^+}$. Since by part 1 we have $\rho(A_Q) = \rho(A_P)$ this proves the assertion. ∎

In [6] it was discovered that two interesting Riccati equations play a role in the solution to the Nehari problem.

THEOREM 4.4. *Let $\Sigma(A, B, C, 0)$ be input and output stable and let L_1, L_2 be arbitrary bounded nonnegative solutions to the Lyapunov equations* (2.1), (2.2)*, respectively. Let $\sigma > r^{\frac{1}{2}}(L_1L_2)$ and define $N_\sigma := (I - \sigma^{-2}L_1L_2)^{-1}$. Then*

1. *$W := N_\sigma L_1$ is a bounded nonnegative solution of the following Riccati equation for $z \in D(A^*)$:*

$$WA^*z + AWz - \sigma^{-2}WC^*CWz + N_\sigma BB^*N_\sigma^*z = 0; \tag{4.8}$$

2. *$X := L_2N_\sigma$ is a bounded nonnegative solution of the following Riccati equation for $z \in D(A)$:*

$$A^*Xz + XAz - \sigma^{-2}XBB^*Xz + N_\sigma^*C^*CN_\sigma z = 0; \tag{4.9}$$

3. *the closed-loop systems $\Sigma(A_W, [N_\sigma B; WC^*], C)$ and $\Sigma(A_X, B, [CN_\sigma; B^*X])$ are system-stable, where $A_W = A - \sigma^{-2}WC^*C$ and $A_X = A - \sigma^{-2}BB^*X$;*
4. *$\sigma(A_X) \cap \mathbb{C}_0^+ \subset \sigma(A) \cap \mathbb{C}_0^+$ and $\sigma(A_X) \cap i\mathbb{R} \subset \sigma(A) \cap i\mathbb{R}$ and the closed-loop generators are related by $A_X = N_\sigma^{-1}A_WN_\sigma$.*

Proof. 1. and 2. The proofs of Lemmas 4.1.24 and 8.3.2 in [4] show that $L_1D(A^*) \subset D(A)$, $L_2D(A) \subset D(A^*)$, $N_\sigma D(A) \subset D(A)$, and $N_\sigma^*D(A^*) \subset D(A^*)$. Thus $WD(A^*) \subset D(A)$ and $XD(A) \subset D(A^*)$. That W satisfies (4.8) and X satisfies (4.9) can be readily verified algebraically.

3. The conclusions about the stability of the closed-loop systems then follow from Theorem 4.2, noting that $\Sigma(A, 1/\sigma B, CN_\sigma)$ is input stable and output stabilizable (with $F = -1/\sigma B^*X$) and that $\Sigma(A, N_\sigma B, 1/\sigma C)$ is output stable and input stabilizable (with $L = -1/\sigma WC^*$).

4. Theorem 4.3 shows that $\sigma(A_X) \cap \mathbb{C}_0^+ \subset \sigma(A) \cap \mathbb{C}_0^+$ and $\sigma(A_X) \cap i\mathbb{R} \subset \sigma(A) \cap i\mathbb{R}$. To relate A_X and A_W consider

$$\begin{aligned} A_XN_\sigma^{-1} &= A(I - \sigma^{-2}L_1L_2) - \sigma^{-2}BB^*L_2 \\ &= A - \sigma^{-2}(AL_1 + BB^*)L_2 \\ &= A + \sigma^{-2}L_1A^*L_2 \\ &= A - \sigma^{-2}L_1(L_2A + C^*C) \\ &= (I - \sigma^{-2}L_1L_2)A - \sigma^{-2}L_1C^*C \\ &= N_\sigma^{-1}A_W. \end{aligned}$$ ∎

5. The spectral factor. As in [4] and [6] we shall approach the solution of the Nehari problem for the input and output stable state linear system $\Sigma(A, B, C, 0)$ with transfer function $\mathbf{G}$ by solving the following J-spectral factorization problem: find $\mathbf{X}$ such that

$$\mathbf{P}(i\omega)^*J_\sigma\mathbf{P}(i\omega) = \mathbf{X}(i\omega)J_1\mathbf{X}(i\omega)^* \quad \text{for almost all} \quad \omega \in \mathbb{R}, \tag{5.1}$$

where

$$\mathbf{P}(s) = \begin{bmatrix} I_Y & \mathbf{G}(s) \\ 0 & I_U \end{bmatrix} \quad \text{and} \quad J_\sigma = \begin{bmatrix} I_Y & 0 \\ 0 & -\sigma^2 I_U \end{bmatrix}.$$

Here we introduce the candidate solution and give some properties. Let L_1 and L_2 be arbitrary bounded nonnegative solutions of the controllability and observability Lyapunov equations, respectively. For $\sigma > r^{\frac{1}{2}}(L_1L_2)$ we define $N_\sigma = (I-\sigma^{-2}L_1L_2)^{-1}$ and we introduce the state linear system

$$\Sigma\left(A, \sigma^{-2}N_\sigma \begin{pmatrix} L_1C^* & -\sigma B \end{pmatrix}, \begin{bmatrix} C \\ B^*L_2 \end{bmatrix}, \begin{bmatrix} I_Y & 0 \\ 0 & \sigma I_U \end{bmatrix}\right). \tag{5.2}$$

We denote the characteristic function of the state linear system (5.2) by $\mathfrak{X}$ and its transfer function by $\mathbf{X}$ and we prove the following lemma.

LEMMA 5.1. *Let*

$$\mathfrak{P}(s) = \begin{bmatrix} I_Y & \mathfrak{G}(s) \\ 0 & I_U \end{bmatrix}.$$

Then $\mathfrak{R}(s) = \mathfrak{P}(s)^* J_\sigma \mathfrak{P}(s) - \mathfrak{X}(s) J_1 \mathfrak{X}(s)^*$ *satisfies the following for* $s \in \rho(A)$*:*

$$\begin{aligned}
\mathfrak{R}(s)_{11} &= -2\sigma^{-2}\mathrm{Re}\ sC(sI-A)^{-1}N_\sigma L_1(sI-A)^{-*}C^*, \\
\mathfrak{R}(s)_{12} &= -2\sigma^{-2}\mathrm{Re}\ sC(sI-A)^{-1}N_\sigma L_1(sI-A)^{-*}L_2B, \\
\mathfrak{R}(s)_{21} &= \mathfrak{R}(s)_{12}^*, \\
\mathfrak{R}(s)_{22} &= -2\mathrm{Re}\ sB^*(sI-A)^{-*}L_2(sI-A)^{-1}B \\
&\qquad -2\sigma^{-2}\mathrm{Re}\ sB^*L_2(sI-A)^{-1}N_\sigma L_1(sI-A)^{-*}L_2B.
\end{aligned}$$

Proof. We only prove the formula for $\mathfrak{R}(s)_{11}$; the proof for the other components is similar. We have on some right half-plane

$$\begin{aligned}
\mathfrak{R}(s)_{11} &= I - \mathfrak{X}_{11}(s)\mathfrak{X}_{11}(s)^* + \mathfrak{X}_{12}(s)\mathfrak{X}_{12}(s)^* \\
&= -\sigma^{-4}C(sI-A)^{-1}WC^*CW(sI-A)^{-*}C^* \\
&\quad +\sigma^{-2}C(sI-A)^{-1}N_\sigma BB^*N_\sigma^*(sI-A)^{-*}C^* \\
&\quad -\sigma^{-2}C(sI-A)^{-1}WC^* - \sigma^{-2}CW(sI-A)^{-*}C \\
&= \sigma^{-2}C(sI-A)^{-1}\left(-\sigma^{-2}WC^*CW + N_\sigma BB^*N_\sigma^*\right)(sI-A)^{-*}C^* \\
&\quad -\sigma^{-2}C(sI-A)^{-1}WC^* - \sigma^{-2}CW(sI-A)^{-*}C.
\end{aligned}$$

Using (4.8) we obtain

$$\begin{aligned}
\mathfrak{R}(s)_{11} &= \sigma^{-2}C(sI-A)^{-1}\left(-WA^* - AW\right)(sI-A)^{-*}C^* \\
&\quad -\sigma^{-2}C(sI-A)^{-1}WC^* - \sigma^{-2}CW(sI-A)^{-*}C \\
&= \sigma^{-2}C(sI-A)^{-1}\left(W(sI-A)^* + (sI-A)W - 2\mathrm{Re}\ s\ W\right)(sI-A)^{-*}C^* \\
&\quad -\sigma^{-2}C(sI-A)^{-1}WC^* - \sigma^{-2}CW(sI-A)^{-*}C \\
&= -2\sigma^{-2}\mathrm{Re}\ s\ C(sI-A)^{-1}N_\sigma L_1(sI-A)^{-*}C^*. \qquad \square
\end{aligned}$$

It is clear from the above that if $\sigma(A) \cap i\mathbb{R}$ has measure zero and L_1, L_2 are an arbitrary pair of solutions to the Lyapunov equations, then $\mathfrak{X}$ is a solution to (5.1). If, in addition, (5.2) is input or output stable, then it follows from Lemma 3.5 that $\mathbf{X}$ is a solution to (5.1). The following example shows that in general $\mathbf{X}$ need not provide a spectral factor.

Example 5.2. We consider an example from Curtain and Sasane [8] (see also Sasane [28]). The transfer function $\mathbf{G}_0(s) = \frac{1}{\sqrt{s^2+1}}$ was shown to have a realization

$\Sigma(A, B, B^*, 0)$ on the state space $\ell_2(\mathbb{Z})$, where $A \in \mathcal{L}(\ell_2(\mathbb{Z}))$ and $B \in \ell_2(\mathbb{Z})$ are given by

$$A_{i,i+1} = -1/2,\ A_{i+1,i} = 1/2,\ A_{i,j} = 0 \ \text{ otherwise},$$
$$B_0 = 1,\ B_i = 0 \ \text{ otherwise}.$$

The spectrum of A is purely continuous and equals $[-i, i]$. The closed-loop system $\Sigma(A - BB^*, B, B^*, 0)$ is system-stable with the transfer function $\mathbf{G}(s) = \frac{1}{1+\sqrt{s^2+1}}$. It is continuous on the imaginary axis and it satisfies

$$\mathbf{G}(i\omega) + \mathbf{G}(i\omega)^* = 2\mathbf{G}(i\omega)^*\mathbf{G}(i\omega) \ \text{ for } \ |\omega| > 1, \tag{5.3}$$

$$\mathbf{G}(i\omega) = \mathbf{G}(i\omega)^* = \frac{1}{1+\sqrt{1-\omega^2}} \ \text{ for } \ |\omega| < 1. \tag{5.4}$$

The Lyapunov equations have solutions $L_1 = L_2 = 1/2\ I$, but note that it is known from [28] that these are not the observability or controllability gramians. The advantage of using these solutions is that the calculations are simple. In this specific case the state linear system (5.2) is

$$\Sigma\left(A - BB^*, \sigma^{-2}N_\sigma \begin{bmatrix} B/2 & -\sigma B \end{bmatrix}, \begin{bmatrix} B^* \\ B^*/2 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & \sigma \end{bmatrix}\right).$$

An easy calculation shows that for $s \in \rho(A - BB^*) = \rho(A) = \mathbb{C} - [-1, 1]$ we have

$$\mathfrak{X}(s) = \begin{bmatrix} 1 & 0 \\ 0 & \sigma \end{bmatrix} + \alpha \begin{bmatrix} 2 & -4\sigma \\ 1 & -2\sigma \end{bmatrix} \mathfrak{G}(s),$$

where $\alpha = \frac{1}{4\sigma^2-1}$. This, together with the stability properties of the state linear systems and Lemma 3.5, shows that we have for almost all $\omega \in \mathbb{R}$

$$\mathbf{X}(i\omega) = \begin{bmatrix} 1 & 0 \\ 0 & \sigma \end{bmatrix} + \alpha \begin{bmatrix} 2 & -4\sigma \\ 1 & -2\sigma \end{bmatrix} \mathbf{G}(i\omega).$$

It is now easily shown using (5.3) that (5.1) holds for $|\omega| > 1$ and using (5.4) that (5.1) does not hold for $|\omega| < 1$.

If we choose the smallest bounded nonnegative solutions to the Lyapunov equations we obtain stronger properties of the candidate spectral factor.

LEMMA 5.3. *If $\Sigma(A, B, C, 0)$ is output and input-output stable and $L_2 = L_C$, the observability gramian of the system $\Sigma(A, B, C, 0)$, then (5.2) is output stable.*

Proof. This follows from Lemma 3.14. ∎

LEMMA 5.4. *If $\Sigma(A, B, C, 0)$ is system-stable and either $\sigma(A) \cap i\mathbb{R}$ has measure zero or U and Y are finite-dimensional, and $L_1 = L_B$ and $L_2 = L_C$, where L_B, L_C are the controllability and observability gramians of the system $\Sigma(A, B, C, 0)$, respectively, then $\mathbf{X}$ satisfies (5.1).*

Proof. It follows from Lemma 5.1 that on some right half-plane

$$\left\langle \mathfrak{R}(s) \begin{bmatrix} y \\ u \end{bmatrix}, \begin{bmatrix} y \\ u \end{bmatrix} \right\rangle = -\frac{2}{\sigma^2} \mathrm{Re}\ s \left(\|\alpha(s)y + \beta(s)u\|^2 + \sigma^2 \|\gamma(s)u\|^2 \right), \tag{5.5}$$

where $\mathfrak{R}$ is as in Lemma 5.1 and

$$\alpha(s) = M_\sigma L_B^{1/2}(sI - A)^{-*}C^*, \quad \beta(s) = M_\sigma L_B^{1/2}(sI - A)^{-*}L_C B,$$
$$\gamma(s) = L_C^{1/2}(sI - A)^{-1}B, \qquad M_\sigma = (I - \frac{1}{\sigma^2} L_B^{1/2} L_C L_B^{1/2})^{-1/2}.$$

From the stability properties we can replace α, β, and γ in (5.5) by their holomorphic extensions (Lemmas 3.4 and 3.14). Then as in Lemma 3.12, using the real-analyticity property, it follows that the resulting equalities hold on $\mathbb{C}_0^+$.

Using Lemma 3.13 and Corollary 3.16 we see that the right-hand side of (5.5) converges to zero as $\operatorname{Re} s \to 0$ (here α, β, and γ are replaced by their holomorphic extensions). From this we obtain the J-spectral factorization (5.1). □

In the remainder of this section we collect some properties of the spectral factor described by (5.2) and of its inverse system

$$\Sigma\left(A, \sigma^{-2}\begin{bmatrix} -L_1C^* & B \end{bmatrix}, \begin{bmatrix} C \\ \sigma^{-1}B^*L_2 \end{bmatrix} N_\sigma, \begin{bmatrix} I_Y & 0 \\ 0 & \sigma^{-1}I_U \end{bmatrix}\right). \tag{5.6}$$

We denote the characteristic function of the state linear system (5.6) by $\mathfrak{V}$ and its transfer function by $\mathbf{V}$. It is the inverse of $\mathfrak{X}$ in the following sense.

LEMMA 5.5. *Assume that $\Sigma(A, B, C, 0)$ is input and output stable. Then for $s \in \rho(A)$ we have $\mathfrak{V}(s)\mathfrak{X}(s) = I = \mathfrak{X}(s)\mathfrak{V}(s)$.*

Proof. This follows from a straightforward calculation. □

LEMMA 5.6. *Assume that $\Sigma(A, B, C, 0)$ is system-stable and that $L_1 = L_B$, the controllability gramian of the system $\Sigma(A, B, C, 0)$. Then (5.6) is input stable.*

Proof. This follows from Corollary 3.15. □

LEMMA 5.7. *Assume that $\Sigma(A, B, C, 0)$ is system-stable and that $L_1 = L_B$ and $L_2 = L_C$, the controllability and observability gramians of the system $\Sigma(A, B, C, 0)$, respectively. Then $\mathbf{V}(s)\mathbf{X}(s) = I = \mathbf{X}(s)\mathbf{V}(s)$ for all $s \in \mathbb{C}_0^+$.*

Proof. From Lemmas 3.4, 5.3, 5.5, and 5.6 we have $\mathbf{V}(s)\mathbf{X}(s) = I = \mathbf{X}(s)\mathbf{V}(s)$ for all $s \in \mathbb{C}_0^+ \cap \rho(A)$. From Lemmas 5.3 and 5.6 both $\mathbf{X}$ and $\mathbf{V}$ are holomorphic on $\mathbb{C}_0^+$ and so the equality extends to this domain. □

LEMMA 5.8. *Assume that $\Sigma(A, B, C, 0)$ is system-stable and that $L_1 = L_B$ and $L_2 = L_C$, the controllability and observability gramians of the system $\Sigma(A, B, C, 0)$, respectively. If either $\sigma(A) \cap i\mathbb{R}$ has measure zero or U and Y are finite-dimensional, then $\mathbf{V}(i\omega)\mathbf{X}(i\omega) = I = \mathbf{X}(i\omega)\mathbf{V}(i\omega)$ for almost all $\omega \in \mathbb{R}$.*

Proof. This follows from Lemmas 5.3, 5.6, 5.7, and 3.5. □

The (2,2) component of $\mathfrak{V}$ plays a special role in what follows, and we need the following extra properties.

LEMMA 5.9. *Assume that $\Sigma(A, B, C, 0)$ is input and output stable. Then the system $\Sigma(A, \sigma^{-2}B, \sigma^{-1}B^*X, \sigma^{-1}I)$ is input stable, and its characteristic function $\mathfrak{V}_{22}(s)$ is invertible for $s \in \rho(A) \cap \rho(A_X)$. Its inverse is the characteristic function of the system-stable state linear system*

$$\Sigma(A_X, B, -\sigma^{-1}B^*X, \sigma I), \quad \text{with} \quad X = L_2N_\sigma. \tag{5.7}$$

Moreover, the transfer function $\mathbf{V}_{22}(s)$ is invertible for $s \in \mathbb{C}_0^+$, and its inverse is the transfer function of (5.7).

Proof. The input stability follows from that of $\Sigma(A, B, C, 0)$. The invertibility of the characteristic function $\mathfrak{V}_{22}$ is a simple calculation, and the stability property of (5.7) follows from Theorem 4.4. The invertibility of the transfer function follows as in Lemma 5.7. □

LEMMA 5.10. *Assume that $\Sigma(A, B, C, 0)$ is input and output stable and that either $\sigma(A) \cap i\mathbb{R}$ has measure zero or U and Y are finite-dimensional. Then the boundary function of $\mathbf{V}_{22}$ is almost everywhere invertible, and its inverse is the boundary function of the transfer function of (5.7).*

Proof. This follows from Lemmas 5.9 and 3.5. □

Dual results for $\mathfrak{X}_{11}(s)$ can be proved similarly.

COROLLARY 5.11. *Assume that $\Sigma(A,B,C,0)$ is input and output stable. Then the system $\Sigma(A,\sigma^{-2}WC^*,C,I)$ is output stable, and its characteristic function $\mathfrak{X}_{11}(s)$ is invertible for $s\in\rho(A)\cap\rho(A_W)$. Its inverse is the characteristic function of the system-stable state linear system*

$$\Sigma(A_W,-\sigma^{-2}WC^*,C,I),\quad \textit{with}\ \ W=N_\sigma L_1. \tag{5.8}$$

Moreover, the transfer function $\mathbf{X}_{11}(s)$ is invertible for $s\in\mathbb{C}_0^+$, and its inverse is the transfer function of (5.8).

COROLLARY 5.12. *Assume that $\Sigma(A,B,C,0)$ is input and output stable and that either $\sigma(A)\cap i\mathbb{R}$ has measure zero or U and Y are finite-dimensional. Then the boundary function of $\mathbf{X}_{11}$ is almost everywhere invertible, and its inverse is the boundary function of the transfer function of* (5.8).

6. The central solution. We introduce the following state linear system

$$\Sigma(A_W^*,L_2B,-\sigma^{-2}CW,0), \tag{6.1}$$

where W is as in Theorem 4.4. We denote its characteristic function by $\mathfrak{Z}$ and its transfer function by $\mathbf{Z}$. The candidate solution to the Nehari problem is given by $\mathbf{K}_c(-s)=\mathbf{Z}(s)$. The state linear system (6.1) has the following properties.

LEMMA 6.1. *If $\Sigma(A,B,C,0)$ is input and output stable, then the following hold:*

1. *The state linear system* (6.1) *is output stable.*
2. *The characteristic functions of the state linear system* (6.1) *and of the state linear system $\Sigma(A_X^*,-\sigma^{-2}XB,CL_1,0)$ coincide.*
3. *The state linear system $\Sigma(A_X^*,-\sigma^{-2}XB,CL_1,0)$ is input stable.*
4. *For $s\in\rho(A^*)\cap\rho(A_W^*)$ we have $\mathfrak{Z}(s)=\mathfrak{V}_{21}(\bar{s})^*\mathfrak{V}_{22}^{-1}(\bar{s})^*$.*

Proof. 1. This follows from Theorem 4.4, part 3.

2. This is an easy calculation using Theorem 4.4, part 4.
3. This follows from Theorem 4.4, part 3.
4. This is a simple calculation. □

The above shows that we have one realization of $\mathbf{Z}$ that is output stable and another that is input stable. We now show that $\mathbf{Z}$ is in $\mathbf{H}_\infty$ under the assumption A1.

THEOREM 6.2. *If $\Sigma(A,B,C,0)$ is system-stable and $\sigma(A)\cap i\mathbb{R}$ has measure zero, then $\mathbf{Z}\in\mathbf{H}_\infty(\mathcal{L}(U,Y))$ and with $\mathbf{K}_c(-s)=\mathbf{Z}(s)$ we have*

$$\|\mathbf{G}+\mathbf{K}_c\|_\infty\le\sigma.$$

Proof. It follows from Lemma 6.1, part 4, and Theorem 4.4, part 4, that for almost all $\omega\in\mathbb{R}$ we have

$$\mathfrak{K}_c(i\omega)=\mathfrak{V}_{21}(i\omega)^*\mathfrak{V}_{22}(i\omega)^{-*}. \tag{6.2}$$

From Lemma 5.1 we see that for almost all $\omega\in\mathbb{R}$ we have

$$\mathfrak{P}(i\omega)^*J_\sigma\mathfrak{P}(i\omega)=\mathfrak{X}(i\omega)J_1\mathfrak{X}(i\omega)^*. \tag{6.3}$$

From (6.2) we obtain for almost all $\omega\in\mathbb{R}$

$$\begin{bmatrix}\mathfrak{G}(i\omega)+\mathfrak{K}_c(i\omega)\\ I\end{bmatrix}=\mathfrak{P}(i\omega)\mathfrak{V}(i\omega)^*\begin{bmatrix}0\\ \mathfrak{V}_{22}(i\omega)^{-*}\end{bmatrix}.$$

So

$$
\begin{aligned}
&(\mathfrak{G}(i\omega)+\mathfrak{K}_c(i\omega))^*(\mathfrak{G}(\mathrm{i}\omega)+\mathfrak{K}_{\mathfrak{c}}(\mathrm{i}\omega))-\sigma^2 I\\
&=\begin{bmatrix}\mathfrak{G}(\mathrm{i}\omega)+\mathfrak{K}_{\mathfrak{c}}(\mathrm{i}\omega)\\ I\end{bmatrix}^* J_\sigma \begin{bmatrix}\mathfrak{G}(\mathrm{i}\omega)+\mathfrak{K}_{\mathfrak{c}}(\mathrm{i}\omega)\\ I\end{bmatrix}\\
&=\begin{bmatrix}0\\ \mathfrak{V}_{22}(i\omega)^{-*}\end{bmatrix}^* \mathfrak{V}(i\omega)\mathfrak{P}(i\omega)^* J_\sigma \mathfrak{P}(i\omega)\mathfrak{V}(i\omega)^* \begin{bmatrix}0\\ \mathfrak{V}_{22}(i\omega)^{-*}\end{bmatrix}\\
&=\begin{bmatrix}0\\ \mathfrak{V}_{22}(i\omega)^{-*}\end{bmatrix}^* \begin{bmatrix}I & 0\\ 0 & -I\end{bmatrix} \begin{bmatrix}0\\ \mathfrak{V}_{22}(i\omega)^{-*}\end{bmatrix}, \qquad (6.4)
\end{aligned}
$$

where in the last step we have used the spectral factorization (6.3). The above shows that

$$\|\mathfrak{G}(i\omega)+\mathfrak{K}_c(i\omega)\|^2-\sigma^2=-\|\mathfrak{V}_{22}(i\omega)^{-*}\|^2,$$

and so for almost all $\omega\in\mathbb{R}$

$$\|\mathfrak{G}(i\omega)+\mathfrak{K}_c(i\omega)\|\le\sigma,$$

which implies that $\mathfrak{G}+\mathfrak{K}_c\in\mathbf{L}_\infty(i\mathbb{R};\mathcal{L}(U,Y))$. Since $\mathbf{G}\in\mathbf{H}_\infty(\mathcal{L}(U,Y))$ and by Lemma 3.5 $\mathfrak{G}$ coincides almost everywhere with the boundary function of $\mathbf{G}$ on the imaginary axis, we have $\mathfrak{G}\in\mathbf{L}_\infty(i\mathbb{R};\mathcal{L}(U,Y))$ and thus $\mathfrak{K}_c\in\mathbf{L}_\infty(i\mathbb{R};\mathcal{L}(U,Y))$. From this it follows that $\mathfrak{Z}\in\mathbf{L}_\infty(i\mathbb{R};\mathcal{L}(U,Y))$. This, together with the output stability of the state linear system (6.1) from Lemma 6.1 using Lemmas 3.5 and 3.6, shows that $\mathbf{Z}\in\mathbf{H}_\infty(\mathcal{L}(U,Y))$. Thus with $\mathbf{K}_c(-s)=\mathbf{Z}(s)$ we obtain $\|\mathbf{G}+\mathbf{K}_c\|_\infty\le\sigma$. ∎

Our main result in this section is to show that under assumption A2, $\mathbf{Z}\in\mathbf{H}_\infty$ and $\mathbf{K}_c$ solves the suboptimal Nehari problem.

THEOREM 6.3. *Assume that $\Sigma(A,B,C,0)$ is system-stable and that $L_1=L_B$ and $L_2=L_C$, the controllability and observability gramians of the system $\Sigma(A,B,C,0)$, respectively, and that U and Y are finite-dimensional. Then $\mathbf{Z}\in\mathbf{H}_\infty(\mathcal{L}(U,Y))$ and with $\mathbf{K}_c(-s)=\mathbf{Z}(s)$ we have*

$$\|\mathbf{G}+\mathbf{K}_c\|_\infty\le\sigma.$$

Proof. The idea is to follow the lines of the proof of Theorem 6.2, replacing the characteristic functions by their corresponding transfer functions. So all we need to show is that the following two key properties hold for almost all $\omega\in\mathbb{R}$:

$$\mathbf{Z}(i\omega)=\mathbf{V}_{21}(-i\omega)^*\mathbf{V}_{22}^{-1}(-i\omega)^*, \qquad (6.5)$$

$$\mathbf{P}(i\omega)^*J_\sigma\mathbf{P}(i\omega)=\mathbf{X}(i\omega)J_1\mathbf{X}(i\omega)^*. \qquad (6.6)$$

Since (6.6) has already been shown in Lemma 5.4, it remains to show only (6.5). By Lemma 6.1, part 4, on some right half-plane we have $\mathbf{Z}(s)=\mathbf{V}_{21}(\bar{s})^*\mathbf{V}_{22}^{-1}(\bar{s})^*$. Using Lemma 6.1, part 1, Lemma 5.6, and Lemma 5.9 this equality holds on $\mathbb{C}_0^+$ (all functions are holomorphic on $\mathbb{C}_0^+$). Lemmas 3.5 and 5.10 now give (6.5). ∎

Under assumption A2 we can show that $\mathbf{Z}$ has a realization as a system-stable state linear system.

COROLLARY 6.4. *Assume that $\Sigma(A,B,C,0)$ is system-stable and that $L_1=L_B$ and $L_2=L_C$, the controllability and observability gramians of the system $\Sigma(A,B,C,0)$,*

respectively, and that U and Y are finite-dimensional. Then $\Sigma(A_W^*, L_2B, -\sigma^{-2}CW, 0)$ and $\Sigma(A_X^*, -\sigma^{-2}XB, CL_1, 0)$ are system-stable state linear systems.

Proof. We have already shown in Lemma 6.1 and Theorem 6.3 that $\Sigma(A_W^*, L_CB, -\sigma^{-2}CW, 0)$ is output and input-output stable. The input stability follows from the identity

$$B^*L_C(sI - A_W)^{-1} = B^*L_C(sI - A)^{-1} - \sigma^{-2}B^*L_C(sI - A_W)^{-1}WC^*C(sI - A)^{-1}$$

and the fact $B^*L_C(sI-A)^{-1}z \in \mathbf{H}_2(Y)$ (Lemma 3.14), that $B^*L_C(sI-A_W)^{-1}WC^* \in \mathbf{H}_\infty(\mathcal{L}(Y,U))$ (the input-output stability shown earlier), and that $C(sI - A)^{-1}z \in \mathbf{H}_2(Y)$ for all $z \in Z$. The proof for the other realization is similar. ∎

7. Parametrization of solutions. First we give a parameterization of a family of solutions to the Nehari problem in terms of the transfer function of (5.6) and an $\mathbf{H}_\infty$ parameter under assumption A2.

THEOREM 7.1. *Let $\Sigma(A, B, C, 0)$ be system-stable with transfer function $\mathbf{G}$ and let L_B and L_C be the controllability and observability gramians, respectively. Assume that U and Y are finite-dimensional. For $\sigma > r^{\frac{1}{2}}(L_BL_C)$ define*

$$\begin{bmatrix} \mathbf{R}_1(s) \\ \mathbf{R}_2(s) \end{bmatrix} = \mathbf{V}(\bar{s})^* \begin{bmatrix} \mathbf{Q}(-s) \\ I_U \end{bmatrix}, \tag{7.1}$$

where $\mathbf{Q}(-s) \in \mathbf{H}_\infty(\mathcal{L}(U,Y))$. If $\|\mathbf{Q}\|_\infty \le 1$, then $\mathbf{K}(-s) = \mathbf{R}_1(s)\mathbf{R}_2(s)^{-1} \in \mathbf{H}_\infty(\mathcal{L}(U,Y))$ and satisfies

$$\|\mathbf{K} + \mathbf{G}\|_\infty \le \sigma.$$

We prove this in a series of lemmas.

LEMMA 7.2. *Under the assumptions of Theorem 7.1 we have the following. For all $s \in \mathbb{C}_0^+$ and almost all $s \in i\mathbb{R}$ we have $\|\mathbf{V}_{12}(s)\mathbf{V}_{22}^{-1}(s)\| < 1$.*

Proof. Consider the state linear system $\Sigma(A, \sigma^{-1}B, CN_\sigma, 0)$, and denote its transfer function by $\mathbf{E}$. The control Riccati equation (4.1) corresponding to this system is precisely (4.9). From the proof of Theorem 4.2, specifically by (4.5), we have $\|[\mathbf{N}; \mathbf{M}]\|_\infty \le 1$, where $[\mathbf{N}; \mathbf{M}]$ is the transfer function of the closed-loop system $\Sigma(A_X, \sigma^{-1}B, [CN_\sigma, -\sigma^{-1}B^*X], [0; I])$. It is easily calculated that $\mathbf{V}_{12}(s)\mathbf{V}_{22}^{-1}(s) = \mathbf{N}(s)$. It also easily seen that $\mathbf{N}(s) = \mathbf{E}(s)\mathbf{M}(s)$ on some right half-plane. By stability this extends to the right half-plane and by taking nontangential limits to almost everywhere on the imaginary axis.

From the above inequality we obtain for almost all $\omega \in \mathbb{R}$ that $\|\mathbf{N}(i\omega)\| \le 1$. We show that, in fact, strict inequality holds. Suppose, on the contrary, that $\|\mathbf{N}(i\omega_0)\| = 1$. Then there would exist a sequence u_n with norm one such that $\|\mathbf{N}(i\omega_0)u_n\| \to 1$. From the above $\mathbf{H}_\infty$ bound we conclude that $\|\mathbf{M}(i\omega_0)u_n\| \to 0$. Since for almost all $\omega \in \mathbb{R}$ we have $\mathbf{N}(i\omega) = \mathbf{E}(i\omega)\mathbf{M}(i\omega)$, we obtain $\|\mathbf{N}(i\omega_0)u_n\| \to 0$. This gives the desired contradiction.

We now extend this inequality to $s \in \mathbb{C}_0^+$. From the above we know that $\|\mathbf{V}_{12}\mathbf{V}_{22}^{-1}\|_\infty = \|\mathbf{N}\|_\infty \le 1$. Suppose that there exists a point $s_0 \in \mathbb{C}_0^+$ such that $\|\mathbf{V}_{12}(s_0)\mathbf{V}_{22}^{-1}(s_0)\| = 1$. Then $\|\mathbf{V}_{12}(s)\mathbf{V}_{22}^{-1}(s)\|$ has a maximum in $\mathbb{C}_0^+$ and is therefore constant by the maximum modulus principle; see, e.g., [17, Theorem 3.13.1, p. 100]. This implies that $\|\mathbf{V}_{12}(s)\mathbf{V}_{22}^{-1}(s)\| = 1$ for all $s \in \mathbb{C}_0^+$. But then the boundary function would have norm one almost everywhere, and we have shown that this is not true. ∎

The next lemma ensures that $\mathbf{K}$ is well defined.

LEMMA 7.3. *Under the assumptions of Theorem* 7.1, $\mathbf{R}_2(s)$ *is invertible for all* $s \in \mathbb{C}_0^+$ *and almost all* $s \in i\mathbb{R}$.

Proof. Define $\mathbf{P}(s) := \mathbf{Q}(-\bar{s})^*$. Then $\mathbf{P} \in \mathbf{H}_\infty(\mathcal{L}(U,Y))$, and $\|\mathbf{P}\|_\infty \le 1$. Next, using Lemma 7.2 we obtain $\|\mathbf{P}(s)\mathbf{V}_{12}(s)\mathbf{V}_{22}^{-1}(s)\| < 1$. From this we see that $(I + \mathbf{P}(s)\mathbf{V}_{12}(s)\mathbf{V}_{22}^{-1}(s))^{-1} = \mathbf{V}_{22}(s)(\mathbf{V}_{22}(s) + \mathbf{P}(s)\mathbf{V}_{12}(s))^{-1}$ exists. Hence $\mathbf{T} := (\mathbf{V}_{22} + \mathbf{P}\mathbf{V}_{12})^{-1}$ exists. Since we have $\mathbf{R}_2(s)^{-1} = \mathbf{T}(\bar{s})^*$ we have that $\mathbf{R}_2(s)^{-1}$ exists. ∎

Next we prove Theorem 7.1 under the assumption $\|\mathbf{Q}\|_\infty < 1$.

LEMMA 7.4. *Theorem* 7.1 *is true for all* $\mathbf{Q}(-s) \in \mathbf{H}_\infty(\mathcal{L}(U,Y))$ *with* $\|\mathbf{Q}\|_\infty < 1$.

Proof. We show that $\mathbf{R}_2^{-1} \in \mathbf{H}_\infty(\mathcal{L}(U))$. This follows as the proof of Lemma 7.3 using that $\mathbf{H}_\infty$ is a Banach algebra: from $\|\mathbf{P}\mathbf{V}_{12}\mathbf{V}_{22}^{-1}\|_\infty < 1$ we conclude that $(I + \mathbf{P}\mathbf{V}_{12}\mathbf{V}_{22}^{-1})^{-1} \in \mathbf{H}_\infty$, and using that $\mathbf{V}_{22}^{-1} \in \mathbf{H}_\infty$ by Lemma 5.9 it follows that $\mathbf{R}_2^{-1} \in \mathbf{H}_\infty(\mathcal{L}(U))$.

$\mathbf{K}(-s) := \mathbf{R}_1(s)\mathbf{R}_2(s)^{-1}$ defines an $\mathbf{L}_\infty$ function which satisfies $\|\mathbf{G} + \mathbf{K}\|_\infty \le \sigma$. The proof is similar to that of Theorem 6.2.

Last we show that $\mathbf{K}(-s) \in \mathbf{H}_\infty(\mathcal{L}(U,Y))$.

Since for all $s \in \mathbb{C}_0^+$ we have $\mathbf{V}(s)\mathbf{X}(s) = I$, we obtain

$$\mathbf{X}_{11}(s)\mathbf{V}_{11}(s) + \mathbf{X}_{12}(s)\mathbf{V}_{21}(s) = I$$

and

$$\mathbf{X}_{11}(s)\mathbf{V}_{12}(s) + \mathbf{X}_{12}(s)\mathbf{V}_{22}(s) = 0,$$

from which we obtain

$$\mathbf{V}_{11}(s) = \mathbf{X}_{11}(s)^{-1} + \mathbf{V}_{12}(s)\mathbf{V}_{22}(s)^{-1}\mathbf{V}_{21}(s).$$

So

$$\begin{aligned}
\mathbf{R}_1(s) &= \mathbf{V}_{21}(\bar{s})^* + \mathbf{V}_{11}(\bar{s})^*\mathbf{Q}(-s) \\
&= \mathbf{X}_{11}(\bar{s})^{-*}\mathbf{Q}(-s) + \mathbf{V}_{21}(\bar{s})^*(I + \mathbf{V}_{22}(\bar{s})^{-*}\mathbf{V}_{12}(\bar{s})^*\mathbf{Q}(-s)) \\
&= \mathbf{X}_{11}(\bar{s})^{-*}\mathbf{Q}(-s) + \mathbf{V}_{21}(\bar{s})^*\mathbf{V}_{22}(\bar{s})^{-*}\mathbf{R}_2(s) \\
&= \mathbf{X}_{11}(\bar{s})^{-*}\mathbf{Q}(-s) + \mathbf{K}_c(-s)\mathbf{R}_2(s),
\end{aligned}$$

and so

$$\mathbf{K}(-s) = \mathbf{X}_{11}(\bar{s})^{-*}\mathbf{Q}(-s)\mathbf{R}_2(s)^{-1} + \mathbf{K}_c(-s).$$

Now $\mathbf{K}_c(-s) \in \mathbf{H}_\infty$ by Theorem 6.3, $\mathbf{X}_{11}^{-1} \in \mathbf{H}_\infty$ by Corollary 5.11, $\mathbf{R}_2(s)^{-1} \in \mathbf{H}_\infty$ as we proved above, and $\mathbf{Q}(-s) \in \mathbf{H}_\infty$ is given. So $\mathbf{K}(-s) \in \mathbf{H}_\infty(U,Y)$. ∎

The proof for the case $\|\mathbf{Q}\|_\infty \le 1$ follows the approach in [1].

LEMMA 7.5. *Theorem* 7.1 *is true for all* $\mathbf{Q}(-s) \in \mathbf{H}_\infty(\mathcal{L}(U,Y))$ *with* $\|\mathbf{Q}\|_\infty \le 1$.

Proof. We first note that $\mathbf{H}_\infty$ is weak-$*$ closed in $\mathbf{L}_\infty$. In the case of the unit disc and the unit circle instead of the right half-plane and the imaginary axis, this follows as in [27, p. 197]. The case we consider follows, using a Möbius transform.

Next we note that if a bounded sequence F_n in $\mathbf{L}_\infty$ converges pointwise to $F \in \mathbf{L}_\infty$, then F_n converges to F in the weak-$*$ topology. In the case of the unit disc and the unit circle instead of the right half-plane and the imaginary axis, this follows as in [1, Proposition 2.3]. The case we consider follows, using a Möbius transformation.

Using these two results, we prove the lemma. For $t \in (0,1)$ define $\mathbf{Q}_t := t\mathbf{Q}$. Then $\|\mathbf{Q}_t\| \le t < 1$. Define $\mathbf{K}_t$ in terms of $\mathbf{Q}_t$. Then by Lemma 7.4 we have $\mathbf{K}_t \in \mathbf{H}_\infty$. If $t \to 1$, then for almost all $\omega \in \mathbb{R}$ we have $\mathbf{K}_t(i\omega) \to \mathbf{K}(i\omega)$. Since $\mathbf{K}_t$ is bounded in norm by $\|\mathbf{G}\|_\infty + \sigma$, the limit function (which is well defined by Corollary 7.3) is in $\mathbf{L}_\infty$. By the above this implies that $\mathbf{K}_t$ converges to $\mathbf{K}$ in the weak-$*$ topology. Since $\mathbf{H}_\infty$ is closed in the weak-$*$ topology and $\mathbf{K}_t \in \mathbf{H}_\infty$, we obtain $\mathbf{K} \in \mathbf{H}_\infty$. □

Remark 7.6. It is clear from the results in the previous section and from the proof of Theorem 7.1 that the conclusions also hold under assumption A1: $\sigma(A) \cap i\mathbb{R}$ has measure zero. In this case L_B and L_C can be replaced by arbitrary bounded nonnegative solutions of the Lyapunov equations, and the assumption that U and Y should be finite-dimensional is redundant.

8. Well-posed linear systems and reciprocals. In this section we solve the suboptimal Nehari problem via the reciprocal system as in Curtain and Sasane [9].

First we briefly review the definitions of a well-posed linear system (see Weiss [32], Staffans [29]). Given an $\mathcal{L}(U,Y)$-valued function $\mathbf{G}$ that is holomorphic and uniformly bounded on some right half-plane, there exist operators A, B, C such that

• A is the infinitesimal generator of a strongly continuous semigroup $T(\cdot)$ on a separable Hilbert space Z;

• $C \in \mathcal{L}(D(A), Y)$ is an admissible observation operator with respect to $T(\cdot)$; i.e., given $\tau > 0$, there exists a $\gamma > 0$ such that

$$\int_0^\tau \|CT(t)z\|^2 \,\mathrm{d}t \le \gamma \|z\|^2 \quad \forall\ z \in D(A);$$

• $B^* \in \mathcal{L}(D(A^*), U)$, and B is an admissible control operator for $T(\cdot)$; i.e., for any $\tau > 0$, there exists a $\beta > 0$ such that for all $u \in \mathbf{L}_2(0,\tau;U)$

$$\left\| \int_0^\tau T(t-s)Bu(s)\,\mathrm{d}s \right\|^2 \le \beta \int_0^\tau \|u(t)\|^2 \,\mathrm{d}t;$$

• The operators A, B, C should be such that

$$\mathbf{G}(s) - \mathbf{G}(\alpha) = (\alpha - s)C(sI - A)^{-1}(\alpha I - A)^{-1}B \tag{8.1}$$

for any α, s larger than the growth bound of the semigroup T.

A triple A, B, C that satisfies the above conditions is called a *realization* of the function $\mathbf{G}$. Such a realization is not unique. A well-posed linear system is specified by operators A, B, C and a transfer function $\mathbf{G}$ that satisfy the above conditions.

The expression (8.1) is defined for all $s \in \rho(A)$, and as in section 2 to avoid confusion, we call this the *characteristic function* and denote it by $\mathfrak{G}$. If the admissibility definitions can be extended to $\tau = \infty$, then the term *infinite-time admissible* is used. Well-posed linear systems form a nice generalization of state linear systems, and the concepts of infinite-time admissibility are the natural extensions of input and output stability in Definition 2.3, and we shall use the terms input and output stability. In Grabowski [15] and Hansen and Weiss [16] it is shown that Lemma 2.6 generalizes perfectly to well-posed linear systems and in Curtain [11] that Lemma 3.4 also applies to well-posed linear systems. We call Σ a *system-stable well-posed linear system* if it is input stable and output stable and $\mathbf{G} \in \mathbf{H}_\infty(\mathcal{L}(U,Y))$.

The concept of a reciprocal system was introduced in [10].

Definition 8.1. *Suppose that the well-posed linear system Σ with generating operators A, B, C and transfer function $\mathbf{G}$ is such that $0 \in \rho(A)$. Its* reciprocal system *is the state linear system $\Sigma(A^{-1}, A^{-1}B, -CA^{-1}, \mathfrak{G}(0))$.*

The justification for this definition is the nice relationship between the well-posed linear system and its reciprocal system shown in [11, Lemma 3.2].

THEOREM 8.2. *Suppose that A, B, C are the generating operators of a well-posed linear system Σ with transfer function $\mathbf{G}$ and zero is in the resolvent set of A. Denote the characteristic function of its reciprocal system by $\mathfrak{G}_r$ and the transfer function of its reciprocal system by $\mathbf{G}_r$. Then the following hold:*

1. *$\mathfrak{G}(s) = \mathfrak{G}_r(\frac{1}{s})$ whenever s is in the resolvent set of A.*
2. *Σ is output stable if and only if its reciprocal system is output stable. If they are output stable, then their observability gramians are identical.*
3. *Σ is input stable if and only if its reciprocal system is input stable. If they are input stable, then their controllability gramians are identical.*
4. *The well-posed linear system is system-stable if and only if its reciprocal system is system-stable. In this case, we have $\mathbf{G}(s) = \mathbf{G}_r(\frac{1}{s})$ for $s \in \mathbb{C}_0^+$.*

The advantages of working with the reciprocal system are that all its generating operators are bounded and the close connections with the original well-posed linear system give us the following result.

THEOREM 8.3. *Let Σ be a system-stable well-posed linear system with generating operators A, B, C and transfer function $\mathbf{G}$ and assume that $0 \in \rho(A)$. Let $\mathbf{G}_r$ denote the transfer function of its reciprocal system $\Sigma(A^{-1}, A^{-1}B, -CA^{-1}, \mathfrak{G}(0))$. Then $\mathbf{K}(-s) \in \mathbf{H}_\infty(\mathcal{L}(U,Y))$ satisfies*

$$\| \mathbf{G} + \mathbf{K} \| \leq \sigma$$

if and only if $\mathbf{K}_r(-s) \in \mathbf{H}_\infty(\mathcal{L}(U,Y))$ satisfies

$$\| \mathbf{G}_r - \mathfrak{G}(0) + \mathbf{K}_r \| \leq \sigma, \tag{8.2}$$

where $\mathbf{K}(s) = \mathbf{K}_r(\frac{1}{s}) - \mathfrak{G}(0)$ for all $s \in \mathbb{C}_0^+$ and almost all $s \in i\mathbb{R}$.

Proof. From Theorem 8.2, part 4, we have $\mathbf{G}(s) = \mathbf{G}_-(\frac{1}{s})$ for $s \in \mathbb{C}_0^+$ and by input-output stability this extends to almost everywhere on $i\mathbb{R}$. So for all $s \in \mathbb{C}_0^+$ and almost all $s \in i\mathbb{R}$ we have

$$\mathbf{K}(-s) + \mathbf{G}(s) = \mathbf{K}_r\left(-\frac{1}{s}\right) + \left[\mathbf{G}_r\left(\frac{1}{s}\right) - \mathfrak{G}(0)\right].$$

Thus

$$\begin{aligned}\sup_{\mathbb{C}_0^+} \| \mathbf{K}(-s) + \mathbf{G}(s) \| &= \sup_{\mathbb{C}_0^+} \left\| \mathbf{K}_r\left(-\frac{1}{s}\right) + \mathbf{G}_r\left(\frac{1}{s}\right) - \mathfrak{G}(0) \right\| \\ &= \sup_{\mathbb{C}_0^+} \| \mathbf{K}_r(-s) + \mathbf{G}_r(s) - \mathfrak{G}(0) \|,\end{aligned}$$

which proves the claim. ∎

Remark 8.4. 1. Since $\mathbf{G}_r(\frac{1}{s}) - \mathfrak{G}(0)$ is the transfer function of the system-stable state linear system $\Sigma(A^{-1}, A^{-1}B, -CA^{-1}, 0)$, the results on state linear systems in this article generalize to well-posed linear systems in an obvious way. Note that the formulas that we so obtain are not the same as for the case of state linear systems but are in terms of the generating operators of the reciprocal system. The analogous formulas for the well-posed linear system need not be well defined.

2. Finally, we remark that the assumption in Theorem 8.3 that $0 \in \rho(A)$ can be relaxed. The arguments in this section can be adapted to the alternative assumption

that $i\omega \in \rho(A)$ for some real ω. Denoting $A_\omega = A - i\omega I$, we introduce the ω-reciprocal systems $\Sigma(A_\omega^{-1}, A_\omega^{-1}, -CA_\omega^{-1}, \mathbf{G}(i\omega))$ with transfer function $\mathbf{G}_r^\omega$. Noting that $\mathbf{G}(s + i\omega) = \mathbf{G}_r^\omega(\frac{1}{s})$, we can obtain connections between the Nehari problem for Σ and this new reciprocal system. By proving our results on state linear systems in discrete time and using the Cayley transform, one can even obtain Theorem 8.3 without any assumption on the spectrum.

9. Appendix. In this appendix we study real-analytic functions on a complex Banach space E following Dieudonné [12]. A function $f : \Omega \subset \mathbb{R}^2 \to E$ with Ω open is called real-analytic if at every point $\omega \in \Omega$ there exist vectors $c_{i,j} \in E$ such that

$$f(x, y) = \sum_{i,j \in \mathbb{N}^2} c_{i,j}(x - \omega_1)^i (y - \omega_2)^j$$

for all points (x, y) in a neighborhood of ω, the series converging absolutely in this neighborhood.

Consider a holomorphic function $h : \Omega \subset \mathbb{C} \to E$. It follows from Goursat's theorem [12, section 9.10, Problem 1] that at every point h can be expanded in an absolutely convergent power series in the complex variable z. Define $h_{\mathbb{R}} : \Omega \subset \mathbb{R}^2 \to E$ by $h_{\mathbb{R}}(x, y) := h(x + iy)$. It is easily seen that $h_{\mathbb{R}}$ is real-analytic: the series expansion in x and y follows from the series expansion in $x+iy$. We further note that if $g : \Omega \to \mathbb{C}$ is real-analytic, then so is $\bar{g}$.

COROLLARY 9.1. *Using the notation and assumptions of Lemma* 3.12*, we have that* $(x, y) \mapsto \langle \widehat{\mathcal{B}}(x + iy)u, L_C Bu\rangle$ *and* $(x, y) \mapsto \langle L_C Bu, \widehat{\mathcal{B}}(x + iy)u\rangle$ *are real-analytic on the right half-plane.*

Proof. With $h(s) = g(s) = \langle \widehat{\mathcal{B}}(s)u, L_C Bu\rangle$ this follows from the above discussion. ∎

We have the following characterization of real-analyticity.

LEMMA 9.2. *$f : \Omega \subset \mathbb{R}^2 \to E$ is real-analytic if and only if there exists an open set $\Omega_{\mathbb{C}} \subset \mathbb{C}^2$ such that $\Omega_{\mathbb{C}} \cap \mathbb{R}^2 = \Omega$ and a holomorphic function $f_{\mathbb{C}} : \Omega_{\mathbb{C}} \to E$ such that $f_{\mathbb{C}}|_\Omega = f$.*

Proof. This follows from [12, subsection 9.4.5, p. 209] and Goursat's theorem [12, section 9.10, Problem 1]. ∎

THEOREM 9.3. *If $f, g : \Omega \to H$ are real-analytic, then $\langle f, g\rangle$ is real-analytic.*

Proof. Since f is real-analytic, there exists a holomorphic function $f_{\mathbb{C}}$ of which f is the restriction. Since $\bar{g}$ is real-analytic, there exists a holomorphic function $\bar{g}_{\mathbb{C}}$ of which $\bar{g}$ is the restriction. We thus have that $(f_{\mathbb{C}}, \bar{g}_{\mathbb{C}})$ is holomorphic. We define a bilinear function by $B(h_1, h_2) = \langle h_1, \bar{h}_2\rangle$. Since the composition of holomorphic functions is holomorphic (and a bilinear function is holomorphic), we have that if h_1 and h_2 are holomorphic, then $B(h_1, h_2)$ is. We thus have that $\langle f_{\mathbb{C}}, \overline{\bar{g}_{\mathbb{C}}}\rangle$ is holomorphic. Restricted to Ω, this function equals $\langle f, g\rangle$. This shows that $\langle f, g\rangle$ is real-analytic. ∎

The above theorem in particular shows that the squared norm of a real-analytic function is real-analytic, which implies that the squared norm of a holomorphic function is real-analytic. This gives the following corollary.

COROLLARY 9.4. *Using the notation and assumptions of Lemma* 3.12*, we have that* $(x, y) \mapsto \|\mathbf{G}(x + iy)u\|^2$ *and* $(x, y) \mapsto 2x\ \|L_C^{1/2}\widehat{\mathcal{B}}(x + iy)u\|^2$ *are real-analytic.*

We quote the following identity theorem for real-analytic functions. This is used in Lemmas 3.12 and 5.4 with A the right half-plane and U some right half-plane.

LEMMA 9.5. *Let $A \subset \mathbb{R}^2$ be an open connected set, and let f and g be two real-analytic functions in A with values in E. If there is a nonempty open subset U of A such that $f(x) = g(x)$ in U, then $f(x) = g(x)$ for every $x \in A$.*

Proof. This follows from [12, subsection 9.4.2, p. 208]. ∎

Acknowledgments. The authors are grateful to Hans Zwart for lively discussions on Lemma 3.4, to Kalle Mikkola and Amol Sasane for their useful comments, and to Erik Thomas for a discussion on real-analytic functions. We are very grateful to the two anonymous reviewers for their painstaking work in reviewing our manuscript. Their efforts have improved this paper considerably.

REFERENCES

[1] J. A. Ball, K. M. Mikkola, and A. J. Sasane, *State-space formulas for the Nehari–Takagi problem for nonexponentially stable infinite-dimensional systems*, SIAM J. Control Optim., 44 (2005), pp. 531–563.

[2] R. F. Curtain and A. Ran, *Explicit formulas for Hankel norm approximations of infinite-dimensional systems*, Integral Equations Operator Theory, 13 (1989), pp. 455–469.

[3] R. F. Curtain and H. J. Zwart, *The Nehari problem for the Pritchard-Salamon class of infinite-dimensional linear systems: A direct approach*, Integral Equations Operator Theory, 18 (1994), pp. 130–153.

[4] R. F. Curtain and H. J. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1995.

[5] R. F. Curtain and A. Ichikawa, *The Nehari problem for infinite-dimensional systems of parabolic type*, Integral Equations Operator Theory, 26 (1996), pp. 29–45.

[6] R. F. Curtain and J. C. Oostveen, *The Nehari problem for nonexponentially stable systems*, Integral Equations Operator Theory, 31 (1998), pp. 307–320.

[7] R. F. Curtain and J. C. Oostveen, *Necessary and sufficient conditions for strong stability of distributed parameter systems*, Systems Control Lett., 37 (1999), pp. 11–18.

[8] R. F. Curtain and A. J. Sasane, *Compactness and nuclearity of the Hankel operator and internal stability of infinite-dimensional state linear systems*, Internat. J. Control, 74 (2001), pp. 1260–1270.

[9] R. F. Curtain and A. J. Sasane, *Hankel norm approximation for well-posed linear systems*, Systems Control Lett., 48 (2003), pp. 407–414.

[10] R. F. Curtain, *Regular linear systems and their reciprocals: Applications to Riccati equations*, Systems Control Lett., 49 (2003), pp. 81–89.

[11] R. F. Curtain, *Riccati equations for stable well-posed linear systems: The generic case*, SIAM J. Control Optim., 42 (2003), pp. 1681–1702.

[12] J. Dieudonné, *Foundations of Modern Analysis*, Academic Press, New York, 1969.

[13] C. Foias and A. Tannenbaum, *On the Nehari problem for a certain class of L_∞-functions appearing in control theory*, J. Funct. Anal., 74 (1987), pp. 146–159.

[14] K. Glover, R. F. Curtain, and J. R. Partington, *Realisation and approximation of linear infinite-dimensional systems with error bounds*, SIAM J. Control Optim., 26 (1988), pp. 863–898.

[15] P. Grabowski, *On the spectral-Lyapunov approach to parametric optimization of distributed parameter systems*, IMA J. Math. Control Inform., 7 (1991), pp. 317–338.

[16] S. Hansen and G. Weiss, *New results on the operator Carleson measure criterion*, IMA J. Math. Control Inform., 14 (1997), pp. 3–32.

[17] E. Hille and R. S. Phillips, *Functional Analysis and Semi-groups*, Amer. Math. Soc. Colloq. Publ., 31, AMS, Providence, RI, 1957.

[18] A. Kheifets, *Parametrization of solutions of the Nehari problem and nonorthogonal dynamics*, in Operator Theory and Interpolation (Bloomington, IN, 1996), Oper. Theory Adv. Appl. 115, Birkhäuser, Basel, 2000, pp. 213–233.

[19] K. Mikkola, *Infinite-Dimensional Linear Systems, Optimal Control, and Riccati Equations*, Ph.D. thesis, Helsinki University of Technology, Helsinki, Finland, 2002; also available at http://www.math.hut.fi/reports.

[20] N. K. Nikol'skiĭ, *Treatise on the Shift Operator*, Springer-Verlag, Berlin, 1986.

[21] J. Oostveen, *Strongly Stabilizable Distributed Parameter Systems*, Frontiers Appl. Math. 20, SIAM, Philadelphia, 2000.

[22] L. B. Page, *Bounded and compact vectorial Hankel operators*, Trans. Amer. Math. Soc., 150 (1970), pp. 529–539.

[23] V. V. Peller, *Hankel Operators and Their Applications*, Springer Monogr. Math., Springer-Verlag, New York, 2003.

[24] A. Ran, *Hankel norm approximation for infinite-dimensional systems and Wiener-Hopf factorization*, in Modelling, Robustness and Sensitivity Reduction in Control Systems, NATO, Adv. Sci. Inst. Ser. F Comput. Systems Sci. 34, R. F. Curtain, ed., Springer-Verlag, Berlin, 1986, pp. 57–69.

[25] M. Rosenblum and J. Rovnyak, *Hardy Classes and Operator Theory*, Oxford Math. Monogr., The Clarendon Press, Oxford University Press, New York, 1985.

[26] D. Salamon, *Realization theory in Hilbert space*, Math. Systems Theory, 21 (1989), pp. 147–164.

[27] D. Sarason, *Generalized interpolation in H^∞*, Trans. Amer. Math. Soc., 127 (1967), pp. 179–203.

[28] A. J. Sasane, *Hankel Norm Approximation for Infinite-Dimensional Systems*, Lecture Notes in Control and Inform. Sci. 277, Springer-Verlag, Berlin, 2002.

[29] O. J. Staffans, *Well-Posed Linear Systems*, Cambridge University Press, Cambridge, UK, 2005.

[30] O. J. Staffans, *Coprime factorizations and well-posed linear systems*, SIAM J. Control Optim., 36 (1998), pp. 1268–1292.

[31] M. Weiss and G. Weiss, *Optimal control of stable weakly regular linear systems*, Math. Control Signals Systems, 10 (1997), pp. 287–330.

[32] G. Weiss, *Regular linear systems with feedback*, Math. Control Signals Systems, 7 (1994), pp. 23–57.

[33] H.J. Zwart, *Transfer functions for infinite-dimensional systems*, Systems Control Lett., 52 (2004), pp. 247–255.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 3, pp. 1019–1039

SOLUTION OF FILTERING PROBLEM WITH NONLINEAR OBSERVATIONS*

STEPHEN S.-T. YAU† AND SHING-TUNG YAU‡

Abstract. For all known finite-dimensional filters, one always needs the conditon that the observation terms are degree one polynomial. On the other hand, in many practical examples, e.g., tracking problem, the observation terms may be nonlinear. Our new method in this paper can treat filtering problems with nonlinear observation terms in the first time, which includes Kalman–Bucy filter as a special case.

Key words. filtering problem, nonlinear observations, real time computation, DMZ equation, Kolmogorov equation

AMS subject classifications. 93E10, 93E11, 60G35

DOI. 10.1137/S0363012902411970

1. Introduction.

In 1961, Kalman–Bucy first established the finite-dimensional filters for linear filtering system with Gaussian initial distribution. In the sixties and early seventies, the basic approach to nonlinear filtering theory was via the "innovations method" originally proposed by Kailath and subsequently rigorously developed by Fujisaki, Kallianpur, and Kunita in 1972 [10]. As pointed out by Mitter [13], the difficulty with this approach is that the innovations process is not, in general, explicitly computable. In view of this weakness, Brockett [2] and Mitter [13] proposed, independently, the idea of using estimation algebras to construct finite-dimensional nonlinear filters. The idea is to imitate the Wei–Norman approach of using the Lie algebraic method to solve the DMZ equation, which the unnormalized conditional probability of the state must satisfy. Perhaps the most important merit of the Lie algebra approach is the following. As long as the estimation algebra is finite dimensional, not only the finite-dimensional filter can be constructed explicitly, but also the filter so constructed is universal in the sense of Chaleyat–Maurel and Michel [4]. In [23], [17], and [20] Yau proves that the number of sufficient statistics in the Lie algebra method, which is required in the computation of conditional probability density, is linear in n, where n is the dimension of the state space. Recently, Stephen Yau [17] and Tam, Wong, and Yau [14], [16], [5], [21], [20], and [6] have completely classified all finite-dimensional estimation algebras of maximal rank. In particular, they have proved that all the observation terms $h_i(x)$, $1 \leq i \leq m$ must be degree one polynomials.

However, in the Wei–Norman approach, one has to know explicitly the basis as vector space of the estimation algebra in order to reduce the DMZ equation to a finite system of ordinary differential equations, Kolmogorov equation, and several first-order linear partial differential equations. Classically, one knows the explicit

*Received by the editors July 24, 2002; accepted for publication (in revised form) December 8, 2004; published electronically September 20, 2005. Research partially supported by U.S. Army Research Office and NSF.

http://www.siam.org/journals/sicon/44-3/41197.html

†Department of Mathematics, Statistics and Computer Science (M/C 249), University of Illinois at Chicago, 851 South Morgan Street, Chicago, IL 60607-7045 (yau@uic.edu). Ze-Jiang Professor of East China Normal University.

‡Department of Mathematics, Harvard University, Cambridge, MA 02138 (yau@math.harvard.edu).

basis for the estimation algebra only in the case that it has maximal rank. Typically people assume that the linear system is controllable and observable. Recently, a new direct method was introduced to study the linear filtering and exact filtering systems with arbitrary initial condition for which f, g and h in (2.1) are independent of time (cf. [22], [23], [19], [18]). This approach offers several advantages. It is easy and the derivation no longer needs controllability and observability. Thus, the algorithm is universal for any linear filtering system. Furthermore, it eliminates the necessity of integrating n first-order linear partial differential equations, as was the case in the Lie algebra method. Finally, the number of sufficient statistics required to compute the conditional probability density of the state in this direct method is n. In all the direct methods in [22], [23], [18], and [19] they need to assume that all the observation terms $h_i(x)$, $1 \leq i \leq m$, are degree one polynomials.

In [26], we have proved the existence and decay estimates of the solution to the DMZ equation under the assumption that $f(x)$ and $h(x)$ in (2.1) have linear growth. In this paper, we use the theory developed in [26] to show that the real time computation of the DMZ equation can be reduced to numerical solution of Kolmogorov equation if $f(x)$ and $h(x)$ have linear growth. Similar results under a much stronger assumption that $f(x)$ and $h(x)$ are bounded functions were treated by various authors including Bensoussan, Glowinski, and Rascanu [1], Elliott and Glowinski [8], Florchinger and LeGland [9], Mikulevicious and Rozovskii [12]. Unlike our results, however, their results cannot cover Kalman–Bucy filters. Theorem 4.2 of this paper says that if the drifts ($f(x)$) are affine and the observation terms ($h(x)$) are nonlinear with linear growths, then the Kolmogorov equation can be solved in real time.

For all known finite-dimensional filters, one always needs the condition that the observation terms are degree one polynomial. On the other hand, in many practical examples, e.g., tracking problem, the observation terms may be nonlinear. Our new method in this paper can treat filtering problems with nonlinear observation terms in the first time, which includes Kalman–Bucy filter as a special case.

This paper is organized as follows. In section 2 we shall set up the notations and recall the basic filtering problem. In section 3, we shall show that real time computation of the DMZ equation can be reduced to off time computation of the Kolmogorov equation. An explicit algorithm of such a reduction is provided. In the appendix, we give a rigorous proof that the solution of our algorithm converges to the solution of the DMZ equation in pointwise and L^2 sense. In section 4, we show that if the drifts are linear and the observation terms are nonlinear with linear growths, then the Kolmogorov equation can be solved in real time via a system of ODEs. Consequently, the nonlinear filtering problem with linear drifts and nonlinear observations with linear growth can be solved in real time and memoryless manner. In section 5, we give a conclusion of this paper.

2. Basic filtering problem. The filtering problem considered here is based on the following signal observation model in Itô form:

$$\begin{cases} dx(t) = f(x(t))dt + g(x(t))dv(t) & x(0) = x_0 \\ dy(t) = h(x(t))dt + dw(t) & y(0) = 0 \end{cases} \tag{2.1}$$

in which x, v, y and w are, respectively, $\mathbb{R}^n, \mathbb{R}^p, \mathbb{R}^m$, and $\mathbb{R}^m$ valued processes and v and w independent, standard Brownian processes. We further assume that $n = p$; f, g, and h are vector-valued, orthogonal matrix-valued and vector-valued C^∞ smooth functions. We shall refer to $x(t)$ as the state and $y(t)$ as the observation at time t.

Let $\rho(t,x)$ denote the conditional probability density of the state given the observation $\{y(s) : 0 \le s \le t\}$. It is well known that $\rho(t,x)$ is given by normalizing a function $\sigma(t,x)$ that satisfies the following DMZ equation in Fisk–Stratonovich form:

$$\begin{cases} d\sigma(t,x) = L_0\sigma(t,x)dt + \sum_{i=1}^{m} L_i\sigma(t,x)dy_i(t) \\ \sigma(0,x) = \sigma_0(x), \end{cases} \tag{2.2}$$

where

$$L_0 = \frac{1}{2}\sum_{i=1}^{n}\frac{\partial^2}{\partial x_i^2} - \sum_{i=1}^{n} f_i(x)\frac{\partial}{\partial x_i} - \sum_{i=1}^{n}\frac{\partial f_i}{\partial x_i}(x) - \frac{1}{2}\sum_{i=1}^{m} h_i^2(x),$$

and for $i = 1, \dots, m$, L_i is the zero-degree differential operator of multiplication by h_i and σ_0 is the probability density of the initial point x_0.

Davis introduced a new unnormalized density

$$u(t,x) = \exp\left(-\sum_{i=1}^{m} h_i(x)y_i(t)\right)\sigma(t,x). \tag{2.3}$$

He reduced (2.2) to the following time-varying partial differential equation which is called the robust DMZ-equation:

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) = L_0u(t,x) + \sum_{i=1}^{m} y_i(t)[L_0, L_i]u(t,x) \\ \qquad\qquad + \dfrac{1}{2}\sum_{i,j=1}^{m} y_i(t)y_j(t)[[L_0, L_i], L_j]u(t,x) \\ u(0,x) = \sigma_0(x), \end{cases} \tag{2.4}$$

where $[\cdot\ ,\ \cdot]$ is the Lie bracket as described in [14]. It is easy to show [24] that (2.4) is equivalent to the following time-varying partial differential equation:

$$\begin{cases} \dfrac{\partial u}{\partial t}(t,x) = \dfrac{1}{2}\sum_{i=1}^{n}\dfrac{\partial^2 u}{\partial x_i^2}(t,x) + \sum_{i=1}^{n}\left(-f_i(x) + \sum_{j=1}^{m} y_j(t)\dfrac{\partial h_j}{\partial x_i}(x)\right)\dfrac{\partial u}{\partial x_i}(t,x) \\ \qquad - \left[\sum_{i=1}^{n}\dfrac{\partial f_i}{\partial x_i}(x) + \dfrac{1}{2}\sum_{i=1}^{m} h_i^2(x) - \dfrac{1}{2}\sum_{i=1}^{m} y_i(t)\Delta h_i(x)\right. \\ \qquad \left. + \sum_{i=1}^{m}\sum_{j=1}^{n} y_i(t)f_j(x)\dfrac{\partial h_i}{\partial x_j}(x) - \dfrac{1}{2}\sum_{i,j=1}^{m}\sum_{k=1}^{n} y_i(t)y_j(t)\dfrac{\partial h_i}{\partial x_k}(x)\dfrac{\partial h_j}{\partial x_k}(x)\right]u(t,x) \\ u(0,x) = \sigma_0(x). \end{cases} \tag{2.5}$$

In this paper we shall solve the filtering problem in the case $f_i(x)$, $1 \le i \le n$, are degree one polynomials and $h_j(x)$, $1 \le j \le m$, may be nonlinear with linear growth, i.e., $|h_j(x)| \le C(1 + |x|)$ for some constant C.

3. Reduction from robust DMZ equation to Kolmogorov equation. The fundamental problem of nonlinear filtering theory is how to solve the robust DMZ

equation (2.5) in real time and memoryless manner. In this section, we shall describe our algorithm which achieves this goal for a large class of filtering system with arbitrary initial distribution by reducing it to solve Kolmogorov equation. Our algorithm is based on the following proposition.

PROPOSITION 3.1. *For any τ_1, τ_2 with $\tau_1 < \tau_2$, $\widetilde{u}(t,x)$ satisfies the following Kolmogorov equation:*

(3.1)
$$\frac{\partial \widetilde{u}}{\partial t}(t,x) = \frac{1}{2}\Delta\widetilde{u}(t,x) - \sum_{i=1}^{n} f_i(x)\frac{\partial \widetilde{u}}{\partial x_i}(t,x) - \left(\sum_{i=1}^{n}\frac{\partial f_i}{\partial x_i}(x) + \frac{1}{2}\sum_{i=1}^{m} h_i^2(x)\right)\widetilde{u}(t,x)$$

for $\tau_1 \le t \le \tau_2$ if and only if

$$u(t,x) = e^{-\sum_{i=1}^{m} y_i(\tau_1)h_i(x)}\widetilde{u}(t,x)$$

satisfies the robust DMZ equation with observation being freezed at $y(\tau_1)$,

$$\begin{aligned}\frac{\partial u}{\partial t}(t,x) = {} & \frac{1}{2}\Delta u(t,x) + \sum_{i=1}^{n}\left(-f_i(x) + \sum_{j=1}^{m} y_j(\tau_1)\frac{\partial h_j}{\partial x_i}(x)\right)\frac{\partial u}{\partial x_i}(t,x) \\ & -\Bigg(\sum_{i=1}^{n}\frac{\partial f_i}{\partial x_i}(x) + \frac{1}{2}\sum_{i=1}^{m} h_i^2(x) - \frac{1}{2}\sum_{i=1}^{m} y_i(\tau_1)\Delta h_i(x) \\ & \quad + \sum_{i=1}^{m}\sum_{j=1}^{n} y_i(\tau_1) f_j(x)\frac{\partial h_i}{\partial x_j}(x) \\ & \quad - \frac{1}{2}\sum_{k=1}^{n}\sum_{i,j=1}^{m} y_i(\tau_1)y_j(\tau_1)\frac{\partial h_i}{\partial x_k}(x)\frac{\partial h_j}{\partial x_k}(x)\Bigg)u(t,x). \qquad (3.2)\end{aligned}$$

Proof. It is straightforward to show that

$$\begin{aligned} & e^{\sum_{i=1}^{m} y_i(\tau_1)h_i(x)}\Bigg[-\frac{\partial}{\partial t} + \frac{1}{2}\Delta + \sum_{i=1}^{n}\left(-f_i(x) + \sum_{j=1}^{m} y_j(\tau_1)\frac{\partial h_j}{\partial x_i}\right)\frac{\partial}{\partial x_i} \\ & \quad -\Bigg(\sum_{i=1}^{n}\frac{\partial f_i}{\partial x_i}(x) + \frac{1}{2}\sum_{i=1}^{m} h_i^2(x) - \frac{1}{2}\sum_{i=1}^{m} y_i(\tau_1)\Delta h_i(x) \\ & \quad\quad + \sum_{i=1}^{m}\sum_{j=1}^{n} y_i(\tau_1) f_j(x)\frac{\partial h_i}{\partial x_j}(x) \\ & \quad\quad - \frac{1}{2}\sum_{k=1}^{n}\sum_{i,j=1}^{m} y_i(\tau_1)y_j(\tau_1)\frac{\partial h_i}{\partial x_k}\frac{\partial h_j}{\partial x_k}\Bigg)\Bigg]u(t,x) \\ & = -\frac{\partial \widetilde{u}}{\partial t}(t,x) + \frac{1}{2}\Delta\widetilde{u}(t,x) - \sum_{i=1}^{n} f_i(x)\frac{\partial \widetilde{u}}{\partial x_i}(t,x) \\ & \quad -\left(\sum_{i=1}^{n}\frac{\partial f_i}{\partial x_i}(x) + \frac{1}{2}\sum_{i=1}^{m} h_i^2(x)\right)\widetilde{u}(t,x). \qquad (3.3)\end{aligned}$$

Proposition (3.1) follows immediately from (3.3). □

We remark that (3.2) is obtained from the robust DMZ equation by freezing the observation $y(t)$ to $y(\tau_1)$. Based on Proposition (3.1), we shall formulate our algorithm to solve the robust DMZ equation and we shall show in Appendices A and B that the solution of our algorithm approximates the solution of the robust DMZ equation very well in both pointwise and L^2-sense.

Suppose that $u(t,x)$ is the solution of the robust DMZ equation and we want to compute $u(\tau,x)$. Let $\mathcal{P}_k = \{0 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_k = \tau\}$ be a partition of $[0,\tau]$. Let $u_i(t,x)$ be a solution of the following partial differential equation for $\tau_{i-1} \le t \le \tau_i$:

$$
\begin{cases}
\dfrac{\partial u_i}{\partial t}(t,x) = \dfrac{1}{2}\Delta u_i(t,x) + \displaystyle\sum_{\ell=1}^{n}\left(-f_\ell(x) + \sum_{j=1}^{m} y_j(\tau_{i-1})\frac{\partial h_j}{\partial x_\ell}(x)\right)\frac{\partial u_i}{\partial x_\ell}(t,x) \\
\qquad\qquad -\left(\displaystyle\sum_{\ell=1}^{n}\frac{\partial f_\ell}{\partial x_\ell}(x) + \frac{1}{2}\sum_{\ell=1}^{m} h_\ell^2(x) - \frac{1}{2}\sum_{j=1}^{m} y_j(\tau_{i-1})\Delta h_j(x)\right. \\
\qquad\qquad +\displaystyle\sum_{j=1}^{m}\sum_{\ell=1}^{n} y_j(\tau_{i-1}) f_\ell(x)\frac{\partial h_j}{\partial x_\ell}(x) \\
\qquad\qquad \left. -\dfrac{1}{2}\displaystyle\sum_{p=1}^{n}\sum_{j,\ell=1}^{m} y_j(\tau_{i-1}) y_\ell(\tau_{i-1})\frac{\partial h_j}{\partial x_p}(x)\frac{\partial h_\ell}{\partial x_p}(x)\right) u_i(t,x) \\
u_i(\tau_{i-1},x) = u_{i-1}(\tau_{i-1},x).
\end{cases} \tag{3.4}
$$

Define the norm of the partition $\mathcal{P}_k$ by $|\mathcal{P}_k| = \sup_{1\le i\le k}\{|\tau_i - \tau_{i-1}|\}$. In Appendices A and B, we shall show that in both pointwise and L^2 sense

$$
u(\tau,x) = \lim_{|\mathcal{P}_k|\to 0} u_k(\tau,x). \tag{3.5}
$$

Therefore it remains to describe an algorithm to compute $u_k(\tau_k,x)$. By Proposition 3.1, $u_1(\tau_1,x)$ can be computed by $\widetilde{u}_1(\tau_1,x)$ where $\widetilde{u}_1(t,x)$ for $0 \le t \le \tau_1$ satisfies the following Kolmogorov equation:

$$
\begin{cases}
\dfrac{\partial \widetilde{u}_1}{\partial t}(t,x) = \dfrac{1}{2}\Delta\widetilde{u}_1(t,x) - \displaystyle\sum_{j=1}^{n} f_j(x)\frac{\partial \widetilde{u}_1}{\partial x_j}(x) - \left(\sum_{j=1}^{n}\frac{\partial f_j}{\partial x_j}(x) + \frac{1}{2}\sum_{i=1}^{m} h_j^2(x)\right)\widetilde{u}_1(t,x) \\
\widetilde{u}_1(0,x) = \sigma_0(x)e^{\sum_{j=0}^{m} y_j(0)h_j(x)} = \sigma_0(x).
\end{cases} \tag{3.6}
$$

In fact, by the uniqueness solution of the Kolmogorov equation, we have

$$
u_1(t,x) = \widetilde{u}_1(t,x), \qquad 0 \le t \le \tau_1. \tag{3.7}
$$

In general, Proposition 3.1 tells us that for $i \ge 2$, $u_i(\tau_i,x)$ can be computed by $\widetilde{u}_i(\tau_i,x)$, where $\widetilde{u}_i(t,x)$ for $\tau_{i-1} \le t \le \tau_i$ satisfies the following Kolmogorov equation:

$$
\begin{cases}
\dfrac{\partial \widetilde{u}_i}{\partial t}(t,x) = \dfrac{1}{2}\Delta\widetilde{u}_i(t,x) - \displaystyle\sum_{j=1}^{n} f_j(x)\frac{\partial \widetilde{u}_i}{\partial x_j}(t,x) - \left(\sum_{j=1}^{n}\frac{\partial f_j}{\partial x_j}(x) + \frac{1}{2}\sum_{j=1}^{m} h_j^2(x)\right)\widetilde{u}_i(t,x) \\
\widetilde{u}_i(\tau_{i-1},x) = e^{\sum_{j=1}^{m}(y_j(\tau_{i-1}) - y_j(\tau_{i-2}))h_j(x)}\,\widetilde{u}_{i-1}(\tau_{i-1},x),
\end{cases} \tag{3.8}
$$

where the last initial condition comes from

$$\begin{aligned}\widetilde{u}_i(\tau_{i-1},x) &= u_i(\tau_{i-1},x)e^{\sum_{j=1}^{m} y_j(\tau_{i-1})h_j(x)} = u_{i-1}(\tau_{i-1},x)e^{\sum_{j=1}^{m} y_j(\tau_{i-1})h_j(x)}\\ &= e^{\sum_{j=1}^{m}(y_j(\tau_{i-1})-y_j(\tau_{i-2}))h_j(x)}\widetilde{u}_{i-1}(\tau_{i-1},x).\end{aligned}$$

In fact, we have

$$u_i(\tau_i,x) = e^{-\sum_{j=1}^{m} y_j(\tau_{i-1})h_j(x)}\widetilde{u}_i(\tau_i,x). \tag{3.9}$$

In view of (2.3), (3.5), and (3.9), we have the following theorem.

THEOREM 3.2. *The unnormalized density σ can be computed via solution $\widetilde{u}$ of the Kolmogorov equation* (3.8). *More specifically,*

$$\sigma(\tau,x) = \lim_{|\mathcal{P}_k|\to 0}\widetilde{u}_k(\tau_k,x) \tag{3.10}$$

Proof.

$$\begin{aligned}\sigma(\tau,x) &= u(\tau,x)\exp\left(\sum_{i=1}^{m} h_i(x)y_i(\tau)\right) && \text{by (2.3)}\\ &= \lim_{|\mathcal{P}_k|\to 0} u_k(\tau,x)\exp\left(\sum_{i=1}^{m} h_i(x)y_i(\tau)\right), && \text{by (3.5)}\end{aligned}$$

where $\mathcal{P}_k = \{0=\tau_0<\tau_1<\cdots<\tau_k=\tau\}$.

In view of (3.9), we have

$$\begin{aligned}\sigma(\tau,x) &= \lim_{|\mathcal{P}_k|\to 0} e^{-\sum_{i=1}^{m} y_j(\tau_{k-1})h_j(x)}\widetilde{u}_k(\tau,x)e^{\sum_{i=1}^{m} h_i(x)y_i(\tau)}\\ &= \lim_{|\mathcal{P}_k|\to 0}\widetilde{u}_k(\tau,x). \qquad \square\end{aligned}$$

Observe that in our algorithm at step i (Lemma B.2), we only need the observation at time τ_{i-1} and τ_{i-2}. We do not need any other previous observation data. Observe also that the Kolmogorov equation (3.8) is uniform for all time steps and it depends on observation $y(t)$ only via initial condition.

4. Filtering problem with nonlinear observations. Consider the filtering system (2.1) with affine drift,

$$f_i(x) = \sum_{j=1}^{n}\ell_{ij}x_j + \ell_i, \qquad 1\le i\le n, \tag{4.1}$$

where ℓ_{ij}, ℓ_i are constants, and nonlinear observation

$$\sum_{i=1}^{m} h_i^2(x) = \sum_{i,j=1}^{n} q_{ij}x_ix_j + \sum_{i=1}^{n} q_ix_i + q_0, \tag{4.2}$$

where $q_{ij} = q_{ji}$, q_i, q_0 are constants.

We first remark that if $h_i(x)$, $1 \le i \le m$, are nonlinear observation with linear growths as follows:

$$h_i^2(x) \le m(1 + |x|^2), \qquad 1 \le i \le m-1, \tag{4.3}$$

where M is a constant, and

$$h_m^2(x) = (m-1)M(1+|x|^2) - \sum_{i=1}^{m-1} h_i^2(x), \tag{4.4}$$

then condition (4.2) is satisfied. The purpose of this section is to prove the following theorem.

THEOREM 4.1. *The unnormalized density of the filtering system* (2.1) *with affine drift* (4.1), *nonlinear observation* (4.2), *and Gaussian initial distribution can be computed in real time in a memoryless way.*

In view of Theorem 3.2, in order to solve the nonlinear filtering problem with conditions (4.1), (4.2) it suffices to solve the following Kolmogorov equation in real time. For $\tau_1 \le t \le \tau_2$,

$$\begin{cases} \dfrac{\partial \widetilde{u}}{\partial t}(t,x) = \dfrac{1}{2}\Delta \widetilde{u}(t,x) - \displaystyle\sum_{j=1}^{n} f_j(x)\frac{\partial \widetilde{u}}{\partial x_j}(t,x) - \left(\sum_{j=1}^{n}\frac{\partial f_j}{\partial x_j}(x) + \frac{1}{2}\sum_{j=1}^{m} h_j^2(x)\right)\widetilde{u}(t,x) \\ \widetilde{u}(0,x) = \phi(x). \end{cases} \tag{4.5}$$

It is well known that any $\phi(x)$ is well approximated by finite linear combination of Gaussians of the form $\alpha_1 G_1 + \cdots + \alpha_p G_p$, where α_is are real numbers and G_is are Gaussian distributions. Let $\widetilde{u}_i$ be the solution of (4.5) with initial distribution G_i. Since (4.5) is a linear partial differential equation, it follows that the solution of (4.5) is of the form $\alpha_1\widetilde{u}_1 + \cdots + \alpha_p\widetilde{u}_p$. Therefore it remains to solve (4.5) with Gaussian initial distribution. Theorem 4.2 gives an explicit solution of (4.5) with linear drift (4.1), nonlinear observation (4.2), and Gaussian initial distribution in terms of solutions of ODEs.

THEOREM 4.2. *Consider the filtering system* (2.1) *with linear drift* (4.1), *nonlinear observation* (4.2), *and Kolmogorov equation. For* $\tau_1 \le t \le \tau_2$,

$$\begin{cases} \dfrac{\partial \widetilde{u}}{\partial t}(t,x) = \dfrac{1}{2}\Delta \widetilde{u}(t,x) - \displaystyle\sum_{j=1}^{n} f_j(x)\frac{\partial \widetilde{u}}{\partial x_j}(t,x) - \left(\sum_{j=1}^{n}\frac{\partial f_j}{\partial x_j}(x) + \frac{1}{2}\sum_{j=1}^{m} h_j^2(x)\right)\widetilde{u}(t,x) \\ \widetilde{u}(\tau_1,x) = \exp[x^T A(\tau_1)x + B^T(\tau_1)x + C(\tau_1)], \end{cases} \tag{4.6}$$

where $A(\tau_1) = (A_{ij}(\tau_1))$ *is a* $n \times n$ *matrix,* $B^T(\tau_1) = (B_1(\tau_1), \ldots, B_n(\tau_1))$, $x^T = (x_1, \ldots, x_n)$ *are* $1 \times n$ *matrix and* $C(\tau_1)$ *is a scalar. Then the solution of* (4.6) *is of the following form:*

$$\widetilde{u}(t,x) = \exp(x^T A x + B^T x + C), \tag{4.7}$$

where $A = A^T = (A_{ij}(t))$ *is a* $n \times n$ *matrix valued function of* t, $B^T = (B_1(t), \ldots, B_n(t))$ *is a* $1 \times n$ *matrix valued function of* t, *and* $C(t)$ *is a scalar function of* t.

Moreover, $A(t)$, $B^T(t)$, and $C(t)$ satisfy the following system of nonlinear ODEs*:*

$$\begin{cases} \dfrac{dA}{dt}(t) = 2A^2(t) - A(t)L - L^T A(t) - \dfrac{1}{2}Q \\ A(t)|_{t=\tau_1} = A(\tau_1) \end{cases} \tag{4.8}$$

$$\begin{cases} \dfrac{dB^T}{dt}(t) = 2B^T(t)A(t) - B^T(t)L - 2\ell^T A(t) - \dfrac{1}{2}q \\ B^T(t)|_{t=\tau_1} = B^T(\tau_1) \end{cases} \tag{4.9}$$

$$\begin{cases} \dfrac{dC}{dt}(t) = tr\, A(t) + \dfrac{1}{2}B^T(t)B(t) - \ell^T B(t) - \dfrac{1}{2}q_0 - tr\, L \\ C(t)|_{t=\tau_1} = C(\tau_1), \end{cases} \tag{4.10}$$

where $L = (\ell_{ij})$, $Q = (q_{ij})$, $1 \le i, j \le n$, $\ell^T = (\ell_1, \dots, \ell_n)$, $q^T = (q_1, \dots, q_n)$ as in (4.1) *and* (4.2).

Proof. Differentiating (4.7) with respect to t and x, respectively, we get the following equations:

$$\begin{aligned} \frac{\partial \widetilde{u}}{\partial t} &= \left(x^T \frac{dA}{dt} x + \frac{dB^T}{dt} x + \frac{dC}{dt} \right) \widetilde{u} \\ \nabla \widetilde{u} &= [(A + A^T)x + B] e^{x^T Ax + B^T x + C} \\ \Delta \widetilde{u} &= \{ 2trA + [(A + A^T)x + B]^T [(A + A^T)x + B] \} e^{x^T Ax + B^T x + C} \\ &= [x^T (AA^T + A^T A + 2A^2)x + 2B^T (A + A^T)x + 2trA + B^T B]\widetilde{u} \\ \sum_{j=1}^{n} f_j(x) \frac{\partial \widetilde{u}}{\partial x_j} &= (Lx + \ell)^T \nabla \widetilde{u} \\ &= [x^T (A^T + A)Lx + (B^T L + \ell^T A + \ell^T A^T)x + \ell^T B]\widetilde{u}, \end{aligned} \tag{4.11}$$

where $L = (\ell_{ij}), \ell^T = (\ell_1, \dots, \ell_n)$

$$\left(\sum_{j=1}^{m} \frac{\partial f_j}{\partial x_j}(x) + \frac{1}{2} \sum_{j=1}^{m} h_j^2(x) \right) \widetilde{u}(t,x) = \left(\frac{1}{2} x^T Q x + \frac{1}{2} q^T x + \frac{1}{2} q_0 + trL \right) \widetilde{u}(t,x),$$

where $Q = (q_{ij}), q^T = (q_1, \dots, q_n)$.

Therefore the R.H.S. of (4.6) is given by

$$\begin{aligned} &\frac{1}{2}\Delta \widetilde{u}(t,x) - \sum_{j=1}^{n} f_j(x) \frac{\partial \widetilde{u}}{\partial x_j}(t,x) - \left(\sum_{j=1}^{n} \frac{\partial f_j}{\partial x_j}(x) + \frac{1}{2} \sum_{j=1}^{m} h_j^2(x) \right) \widetilde{u}(t,x) \\ &= \left[x^T \left(\frac{1}{2} AA^T + \frac{1}{2} A^T A + A^2 \right) x + B^T (A + A^T)x + trA + \frac{1}{2} B^T B \right] \widetilde{u} \\ &\quad - [x^T (A^T + A)Lx + (B^T L + \ell^T A + \ell^T A^T)x + \ell^T B]\widetilde{u} \\ &\quad - \left(\frac{1}{2} x^T Q x + \frac{1}{2} q^T x + \frac{1}{2} q_0 + trL \right) \widetilde{u}(t,x) \\ &= \left[x^T \left(\frac{1}{2} AA^T + \frac{1}{2} A^T A + A^2 - A^T L - AL - \frac{1}{2} Q \right) x + (B^T A + B^T A^T - B^T L \right. \\ &\quad \left. - \ell^T A - \ell^T A^T - \frac{1}{2} q^T)x + trA + \frac{1}{2} B^T B - \ell^T B - \frac{1}{2} q_0 - trL \right] \widetilde{u}. \end{aligned} \tag{4.12}$$

By comparing (4.11) and (4.12), we get (4.8), (4.9), and (4.10), which are necessary and sufficient conditions for (4.7) to be a solution of (4.6). □

5. Conclusion. All the known finite dimensional filters require observation terms linear in nature. In this paper we have solved the nonlinear filtering problem with linear drift and nonlinear observations in real time and memoryless manner. We first show that the solution of the DMZ equation can be obtained by solving the Kolmogorov equation. We also show that the Kolmogorov equation can be solved via solutions of systems of ODEs if the summation of observations is a quadratic polynomial (cf. (4.2)).

Appendix A: Pointwise Convergence of (3.5). By changing variables from x_i to $\sqrt{2}x_i$ and by letting

$$\overline{u}(t,x) = u\left(t, \frac{x}{\sqrt{2}}\right), \tag{A.1}$$

we get

$$\begin{aligned}
\frac{\partial \overline{u}}{\partial t}(t,x) &= \frac{\partial u}{\partial t}\left(t, \frac{x}{\sqrt{2}}\right),\\
\frac{\partial \overline{u}}{\partial x_i}(t,x) &= \frac{1}{\sqrt{2}}\frac{\partial u}{\partial x_i}\left(t, \frac{x}{\sqrt{2}}\right),\\
\frac{\partial^2 \overline{u}}{\partial x_i^2}(t,x) &= \frac{1}{2}\frac{\partial^2 u}{\partial x_i^2}\left(t, \frac{x}{\sqrt{2}}\right).
\end{aligned}$$

Hence the robust DMZ equation becomes

$$\frac{\partial\, \overline{u}}{\partial t}(t,x) = \Delta \overline{u}(t,x) + \sum_{i=1}^{m} \overline{f}_i(t,x)\frac{\partial\, \overline{u}}{\partial x_i}(t,x) - \overline{V}(t,x)\overline{u}(t,x), \tag{A.2}$$

where

$$\overline{f}_i(t,x) = \sqrt{2}\left[- f_i\left(\frac{x}{\sqrt{2}}\right) + \sum_{j=1}^{m} y_j(t)\frac{\partial h_j}{\partial x_i}\left(\frac{x}{\sqrt{2}}\right)\right] \tag{A.3}$$

$$\begin{aligned}
\overline{V}(t,x) = {} & \sum_{i=1}^{n}\frac{\partial f_i}{\partial x_i}\left(\frac{x}{\sqrt{2}}\right) + \frac{1}{2}\sum_{i=1}^{m} h_i^2\left(\frac{x}{\sqrt{2}}\right) - \sum_{i=1}^{m} y_i(t)\Delta h_i\left(\frac{x}{\sqrt{2}}\right)\\
& + \sum_{i=1}^{m}\sum_{j=1}^{n} y_i(t) f_j\left(\frac{x}{\sqrt{2}}\right)\frac{\partial h_i}{\partial x_j}\left(\frac{x}{\sqrt{2}}\right)\\
& - \frac{1}{2}\sum_{i,j=1}^{m}\sum_{k=1}^{n} y_i(t) y_j(t)\frac{\partial h_i}{\partial x_k}\left(\frac{x}{\sqrt{2}}\right)\frac{\partial h_j}{\partial x_k}\left(\frac{x}{\sqrt{2}}\right).
\end{aligned} \tag{A.4}$$

For any $\tau > 0$, we shall consider the following parabolic equations on $[0,\tau] \times \mathbb{R}^n$.

$$\begin{cases}
\dfrac{\partial\, \overline{u}}{\partial t}(t,x) = \Delta \overline{u}(t,x) + \displaystyle\sum_{i=1}^{n} \overline{f}_i(t,x)\frac{\partial\, \overline{u}}{\partial x_i}(t,x) - \overline{V}(t,x)\overline{u}(t,x)\\
\overline{u}(0,x) = \overline{\psi}(x)
\end{cases} \tag{A.5}$$

$$\text{(A.6)}\qquad \begin{cases} \dfrac{\partial \widetilde{u}}{\partial t}(t,x) = \Delta \widetilde{u}(t,x) + \displaystyle\sum_{i=1}^{n} \widetilde{f}_i(0,x)\frac{\partial \widetilde{u}}{\partial x_i}(t,x) - \widetilde{V}(0,x)\widetilde{u}(t,x) \\ \widetilde{u}(0,x) = \widetilde{\psi}(x), \end{cases}$$

where $\widetilde{f}_i(0,x) := \overline{f}_i(0,x)$ and $\widetilde{V}(0,x) := \overline{V}(0,x)$ are obtained from $\overline{f}_i(t,x)$ and $\overline{V}(t,x)$ by freezing the time variable at 0. For simplicity, we shall assume that the first, second, and third derivatives of $h(x)$ are bounded.

The goal of this appendix is to prove that if $\widetilde{\psi}(x)$ is close to $\overline{\psi}(x)$ uniformly in x, then $\widetilde{u}(\tau,x)$ is close to $\overline{u}(\tau,x)$ uniformly in x. From (A.5) and (A.6), we deduce that

(A.7)

$$\begin{aligned} \frac{\partial(\overline{u}-\widetilde{u})}{\partial t}(t,x) &= \Delta(\overline{u}-\widetilde{u})(t,x) + \sum_{i=1}^{n} \overline{f}_i(t,x)\frac{\partial(\overline{u}-\widetilde{u})}{\partial x_i}(t,x) - \overline{V}(t,x)(\overline{u}-\widetilde{u})(t,x) \\ &\quad + \sum_{i=1}^{n} (\overline{f}_i(t,x) - \widetilde{f}_i(0,x))\frac{\partial \widetilde{u}}{\partial x_i}(t,x) - (\overline{V}(t,x) - \widetilde{V}(0,x))\widetilde{u}(t,x) \\ &= (\Delta - \overline{V}(t,x))(\overline{u}-\widetilde{u})(t,x) + \sum_{i=1}^{n} \overline{f}_i(t,x)\frac{\partial(\overline{u}-\widetilde{u})}{\partial x_i}(t,x) + G_\tau(t,x), \end{aligned}$$

where

$$\text{(A.8)}\qquad G_\tau(t,x) = \sum_{i=1}^{n} (\overline{f}_i(t,x) - \widetilde{f}_i(0,x))\frac{\partial \widetilde{u}}{\partial x_i}(t,x) - (\overline{V}(t,x) - \widetilde{V}(0,x))\widetilde{u}(t,x).$$

LEMMA A.1. *There exists a nonnegative function $\alpha(t,x,y)$ such that*

$$\text{(A.9)}\qquad \begin{cases} \dfrac{\partial \alpha}{\partial t}(t,x,y) = \Delta_x \alpha(t,x,y) - \displaystyle\sum_{i=1}^{n} \overline{f}_i(\tau - t, x)\frac{\partial \alpha}{\partial x_i}(t,x,y) \\ \qquad\qquad - \left[\overline{V}(\tau - t, x) + \displaystyle\sum_{i=1}^{n} \frac{\partial \overline{f}_i}{\partial x_i}(\tau - t, x)\right]\alpha(t,x,y) \\ \alpha(0,x,y) = \delta_y(x), \qquad \int_x \alpha(0,x,y)dx = 1, \end{cases}$$

where $\int_x$ denotes the integration with respect to x variable.

Proof. Let $\beta_n(x,y)$ be a sequence of Gaussian with

$$\text{(A.10)}\qquad \int_x \beta_n(x,y)dx = 1 \quad \text{and} \quad \lim_{n\to\infty} \beta_n(x,y) = \delta_y(x).$$

In view of [26], there exists a solution $\alpha_n(t,x,y)$ with initial condition $\alpha_n(0,x,y) = \beta_n(x,y)$. By maximal principle, $\alpha(t,x,y) \geq 0$ for all $t \geq 0$. We shall take $\alpha(t,x,y) = \lim_{n\to\infty} \alpha_n(t,x,y)$. □

THEOREM A.2. *Let $w(t,x) = \overline{u}(t,x) - \widetilde{u}(t,x)$, where $\overline{u}$ and $\widetilde{u}$ are the solutions of the parabolic equations* (A.5) *and* (A.6)*, respectively. Let $\alpha(t,x,y)$ be the nonnegative function in* Lemma A.1. *Then*

$$w(\tau,y) = \int_x \alpha(\tau,x,y)w(0,x)dx + \int_0^\tau \int_x \alpha(t,x,y)G_\tau(t,x)dx,$$

where $G_\tau(t,x)$ is given in (A.8).

Proof.

$$\text{(A.11)}\quad \int_0^\tau \frac{d}{dt}\int_x \alpha(\tau-t,x,y)w(t,x)dx = -\int_0^\tau\int_x \frac{\partial\alpha}{\partial t}(\tau-t,x,y)w(t,x)dx + \int_0^\tau\int_x \alpha(\tau-t,x,y)\frac{\partial w}{\partial t}(t,x)dx$$

$$\text{L.H.S. of (A.11)} = w(\tau,y) - \int_x \alpha(\tau,x,y)w(0,x)dx$$

$$\begin{aligned}\text{R.H.S. of (A.11)} &= -\int_0^\tau\int_x \Delta_x\alpha(\tau-t,x,y)w(t,x)dx \\ &\quad + \int_0^\tau\int_x \sum_{i=1}^n \overline{f}_i(t,x)\frac{\partial\alpha}{\partial x_i}(\tau-t,x,y)w(t,x)dx \\ &\quad + \int_0^\tau\int_x \left[\overline{V}(t,x) + \sum_{i=1}^n \frac{\partial\overline{f}_i}{\partial x_i}(t,x)\right]\alpha(\tau-t,x,y)w(t,x)dx \\ &\quad + \int_0^\tau\int_x \alpha(\tau-t,x,y)\frac{\partial w}{\partial t}(t,x)dx \\ &= \int_0^\tau\int_x \alpha(\tau-t,x,y)\left[\frac{\partial w}{\partial t}(t,x) - \Delta w(t,x) - \sum_{i=1}^n \overline{f}_i(t,x)\frac{\partial w}{\partial x_i}(t,x)\right. \\ &\qquad\qquad \left. + \overline{V}(t,x)w(t,x)\right]dx \\ &= \int_0^\tau\int_x \alpha(\tau-t,x,y)G_\tau(t,x)dx. \qquad \text{by (A.7)}\end{aligned}$$

In the above computation, we have used the fact proved in [26] that $\alpha(t,x,y)$ has Gaussian decay in x. □

PROPOSITION A.3. *Let $\alpha(t,x,y)$ be the nonnegative function in* Lemma A.1. *Suppose that $\overline{V}(t,x) \geq -c_1$ for some positive constant c_1. Then*

$$\text{(A.12)}\qquad \int_x \alpha(\tau,x,y)dx \leq e^{c_1\tau}.$$

Proof.

$$\begin{aligned}e^{c_1t}\frac{d}{dt}\left(e^{-c_1t}\int_x \alpha(t,x,y)dx\right) &= -c_1\int_x \alpha(t,x,y)dx + \int_x \frac{\partial\alpha}{\partial t}(t,x,y)dx \\ &= -c_1\int_x \alpha(t,x,y)dx + \int_x \Delta_x\alpha(t,x,y)dx - \int_x\sum_{i=1}^n \overline{f}_i(\tau-t,x)\frac{\partial\alpha}{\partial x_i}(t,x,y)dx \\ &\quad - \int_x \left[\overline{V}(\tau-t,x) + \sum_{i=1}^n \frac{\partial\overline{f}_i}{\partial x_i}(\tau-t,x)\right]\alpha(t,x,y)dx \\ &= -c_1\int_x \alpha(t,x,y)dx - \int_x \overline{V}(\tau-t,x)\alpha(t,x,y)dx \\ &= -\int_x [\overline{V}(\tau-t,x)+c_1]\alpha(t,x,y)dx \leq 0.\end{aligned}$$

It follows that $e^{-c_1 t}\int_x \alpha(t,x,y)dx$ is a decreasing function of t and (A.12) follows. □

THEOREM A.4. *With the assumption of* Proposition A.3, *let* $w(t,x) = \overline{u}(t,x) - \widetilde{u}(t,x)$, *where* $\overline{u}$ *and* $\widetilde{u}$ *are the solutions of the parabolic equations* (A.5) *and* (A.6), *respectively. If* τ *is small and* $w(0,x)$ *is small uniformly in* τ, *then* $w(\tau,x)$ *is small uniformly in* x. *More precisely, we have*

$$\sup_{y\in\mathbb{R}^n}|w(\tau,y)| \le e^{c_1\tau}\sup_{x\in\mathbb{R}^n}|w(0,x)| + \tau e^{c_1\tau}\sup_{\substack{x\in\mathbb{R}^n\\ 0\le t\le\tau}}|G_\tau(t,x)|, \tag{A.13}$$

where $G_\tau(t,x)$ *is given in* (A.8).

Proof. In view of (A.3), (A.4), and (A.8), we have

$$\begin{aligned}
G_\tau(t,x) &= \sum_{i=1}^n (\overline{f}_i(t,x) - \widetilde{f}_i(0,x))\frac{\partial\widetilde{u}}{\partial x_i}(t,x) - (\overline{V}(t,x) - \widetilde{V}(0,x))\widetilde{u}(t,x)\\
&= \sum_{i=1}^n \sqrt{2}\sum_{j=1}^m (y_j(t) - y_j(0))\frac{\partial h_j}{\partial x_i}\left(\frac{x}{\sqrt{2}}\right)\frac{\partial\widetilde{u}}{\partial x_i}(t,x)\\
&\quad + \Bigg[-\sum_{i=1}^m (y_i(t) - y_i(0))\Delta h_i\left(\frac{x}{\sqrt{2}}\right) + \sum_{i=1}^m\sum_{j=1}^n (y_i(t) - y_i(0))f_j\left(\frac{x}{\sqrt{2}}\right)\frac{\partial h_i}{\partial x_j}\left(\frac{x}{\sqrt{2}}\right)\\
&\qquad -\frac{1}{2}\sum_{i,j=1}^m\sum_{k=1}^n (y_i(t)y_j(t) - y_i(0)y_j(0))\frac{\partial h_i}{\partial x_k}\left(\frac{x}{\sqrt{2}}\right)\frac{\partial h_j}{\partial x_k}\left(\frac{x}{\sqrt{2}}\right)\Bigg]\widetilde{u}(t,x).
\end{aligned}$$

Therefore if τ is small, then $G_\tau(t,x)$ is uniformly small in x for $0 \le t \le \tau$, because both $\widetilde{u}(t,x)$ and $\frac{\partial\widetilde{u}}{\partial x_i}(t,x)$ have Gaussian decay by [26]. The estimate (A.13) follows readily from Theorem A.2. □

Now we consider the global situation. For a fixed $T > 0$, we want to find the solution $\overline{u}(t,x)$ of the following parabolic equation on $[0,T]\times\mathbb{R}^n$:

$$\begin{cases} \dfrac{\partial\,\overline{u}}{\partial t}(t,x) = \Delta\overline{u}(t,x) + \displaystyle\sum_{j=1}^n \overline{f}_j(t,x)\frac{\partial\,\overline{u}}{\partial x_j}(t,x) - \overline{V}(t,x)\overline{u}(t,x)\\ \overline{u}(0,x) = \overline{\psi}(x). \end{cases} \tag{A.14}$$

Let $\{0 < \tau_1 < \tau_2 < \cdots < \tau_k = T\}$ be a partition of $[0,T]$. Let $\widetilde{u}_i(t,x)$ be the solution of the following parabolic equation on $[\tau_{i-1},\tau_i]\times\mathbb{R}^n$:

$$\begin{cases} \dfrac{\partial\widetilde{u}_i}{\partial t}(t,x) = \Delta\widetilde{u}_i(t,x) + \displaystyle\sum_{j=1}^n \widetilde{f}_j(\tau_{i-1},x)\frac{\partial\widetilde{u}_i}{\partial x_j}(t,x) - \widetilde{V}(\tau_{i-1},x)\widetilde{u}_i(t,x)\\ \widetilde{u}_i(\tau_{i-1},x) = \widetilde{u}_{i-1}(\tau_{i-1},x), \end{cases} \tag{A.15}$$

where $\widetilde{u}_1(0,x) = \overline{\psi}(x)$; $\widetilde{f}_j(\tau_{i-1},x)$ and $\widetilde{V}(\tau_{i-1},x)$ are functions independent of t and equal to $\overline{f}_j(\tau_{i-1},x)$ and $\overline{V}(\tau_{i-1},x)$, respectively.

LEMMA A.5. *Fix* T, *let* $G_{\tau_i}(t,x) = \sum_{j=1}^n(\overline{f}_j(t,x) - \widetilde{f}_j(\tau_{i-1},x))\frac{\partial\widetilde{u}_i}{\partial x_j}(t,x) - (\overline{V}(t,x) - \widetilde{V}(\tau_{i-1},x))\widetilde{u}_i(t,x)$. *For any given* $\epsilon > 0$, *we can choose* k *sufficiently large so that*

$$\sup_{1\le i\le n}\ \sup_{\tau_{i-1}\le t\le\tau_i}\ \sup_{x\in\mathbb{R}^n}|G_{\tau_i}(t,x)| \le \epsilon.$$

Proof. This follows from the proof of Theorem A.4. □

We are now ready to prove the main theorem in this appendix.

THEOREM A.6. *Let* $\overline{u}(t,x)$ *and* $\widetilde{u}_k(t,x)$ *be the solutions of* (A.14) *and* (A.15), *respectively. For any* $\epsilon > 0$, *let* k *be sufficiently large so that* Lemma A.5 *holds. Then*

$$|\overline{u}(T,x) - \widetilde{u}_k(T,x)| \le \epsilon T e^{c_1 T},$$

where c_1 *is the constant in* Proposition A.3.

Proof. In view of $\widetilde{u}_1(0,x) = \overline{\psi}(x) = \overline{u}(0,x)$ and Theorem A.4, we have

$$|\overline{u}(\tau_{1,x}) - \widetilde{u}_1(\tau_{1,x})| \le \tau_1 e^{c_1\tau_1} \sup_{\substack{x\in\mathbb{R}^n \\ 0\le t\le\tau_1}} |G_{\tau_1}(t,x)|.$$

By Theorem A.4 and induction, we have

$$\begin{aligned}
|\overline{u}(\tau_{2,x}) - \widetilde{u}_2(\tau_{2,x})| &\le \tau_1 e^{c_1\tau_1} e^{c_1(\tau_2-\tau_1)} \sup_{\substack{x\in\mathbb{R}^n \\ 0\le t\le\tau_1}} |G_{\tau_1}(t,x)| \\
&\quad + (\tau_2-\tau_1) e^{c_1(\tau_2-\tau_1)} \sup_{\substack{x\in\mathbb{R}^n \\ \tau_1\le t\le\tau_2}} |G_{\tau_2}(t,x)| \\
|\overline{u}(\tau_{k,x}) - \widetilde{u}_k(\tau_{k,x})| &\le \tau_1 e^{c_1\tau_k} \sup_{\substack{x\in\mathbb{R}^n \\ 0\le t\le\tau_1}} |G_{\tau_1}(t,x)| + (\tau_2-\tau_1) e^{c_1(\tau_k-\tau_1)} \sup_{\substack{x\in\mathbb{R}^n \\ \tau_1\le t\le\tau_2}} G_{\tau_2}(t,x)| \\
&\quad + \cdots + (\tau_i - \tau_{i-1}) e^{c_1(\tau_k-\tau_{i-1})} \sup_{\substack{x\in\mathbb{R}^n \\ \tau_{i-1}\le t\le\tau_i}} |G_{\tau_i}(t,x)| \\
&\quad + \cdots + (\tau_k - \tau_{k-1}) e^{c_1(\tau_k-\tau_{k-1})} \sup_{\substack{x\in\mathbb{R}^n \\ \tau_{k-1}\le t\le\tau_k}} |G_{\tau_k}(t,x)| \\
&\le \epsilon[\tau_1 e^{c_1\tau_k} + (\tau_2-\tau_1) e^{c_1(\tau_k-\tau_1)} + \cdots + (\tau_i - \tau_{i-1}) e^{c_1(\tau_k-\tau_{i-1})} \\
&\quad + \cdots + (\tau_k - \tau_{k-1}) e^{c_1(\tau_k-\tau_{k-1})}] \\
&\le \epsilon[\tau_1 + (\tau_2-\tau_1) + \cdots + (\tau_i - \tau_{i-1}) + \cdots + (\tau_n - \tau_{n-1})] e^{c_1 T} \\
&= \epsilon T e^{c_1 T}. \qquad \square
\end{aligned}$$

THEOREM A.7. *Fix* $T > 0$, *let* $\mathcal{P}_n = \{0 < \tau_1 < \tau_2 < \cdots < \tau_k = T\}$ *be a partition of* $[0,T]$. *Let* $\overline{u}(t,x)$ *be the solution of the following parabolic equation on* $[0,T]\times\mathcal{R}^n$:

$$\begin{cases} \dfrac{\partial \overline{u}}{\partial t}(t,x) = \Delta\overline{u}(t,x) + \displaystyle\sum_{j=1}^n \overline{f}_j(t,x)\frac{\partial \overline{u}}{\partial x_j}(t,x) - \overline{V}(t,x)\overline{u}(t,x) \\ \overline{u}(0,x) = \psi(x). \end{cases}$$

Let $\overline{u}_i(t,x)$ *be the solution of the following parabolic equation on* $[\tau_{i-1},\tau_i]\times\mathcal{R}^n$:

$$\begin{cases} \dfrac{\partial \widetilde{u}_i}{\partial t}(t,x) = \Delta\widetilde{u}_i(t,x) + \displaystyle\sum_{j=1}^n \widetilde{f}_j(\tau_{i-1},x)\frac{\partial \widetilde{u}_i}{\partial x_j}(t,x) - \widetilde{V}(\tau_{i-1},x)\widetilde{u}_i(t,x) \\ \widetilde{u}_i(\tau_{i-1},x) = \widetilde{u}_{i-1}(\tau_{i-1},x), \end{cases}$$

where $\widetilde{u}_i(0,x) = \psi(x)$ *and* $\widetilde{f}_j(\tau_{i-1},x) = \overline{f}_j(\tau_{i-1},x)$, $\widetilde{V}(\tau_{i-1},x) = \overline{V}(\tau_{i-1},x)$ *are obtained from* $\overline{f}_j(t,x)$ *and* $\overline{V}(t,x)$ *by freezing time variable at* τ_{i-1}. *Then*

$$\overline{u}(\tau,x) = \lim_{|\mathcal{P}_k|\to 0} \widetilde{u}_k(\tau_k,x) \text{ uniformly in } x.$$

Appendix B: L^2 Convergence of (3.5). In Appendix A we have shown that the solution $\widetilde{u}(t,x)$ of (A.6) is uniformly close to the solution $\overline{u}(t,x)$ of (A.5) for $0 \le t \le T$ if $\widetilde{\psi}(x) = \widetilde{u}(0,x)$ is uniformly close to $\overline{\psi}(x) = \overline{u}(0,x)$. In this section, we shall show that $\widetilde{u}(t,x)$ is also close to $\overline{u}(t,x)$ in L^2-sense, if $\overline{\psi}(x)$ is close to $\widetilde{\psi}(x)$ in L^2 sense. We first recall the following lemma.

LEMMA B.1. *If* $\frac{d\alpha}{dt}(t) \le c\alpha(t)+\beta(t)$, *where* c *is a constant, then* $e^{-ct}\alpha(t)-\alpha(0) \le \int_0^t e^{-cs}\beta(s)ds$.

Let $\overline{f}_{2R}$, $\widetilde{f}_{2R}$, $\overline{V}_{2R}$, and $\widetilde{V}_{2R}$ be the functions obtained by multiplying $\overline{f}, \widetilde{f}, \overline{V}$, and $\widetilde{V}$, respectively, by a cut off function σ which is equal to one in the ball of radius $R \ge 1$ and equal to zero outside a ball of radius $2R$. We can choose σ such that

$$|\nabla\sigma(x)| \le \frac{4}{1+|x|} \text{ and } |\Delta\sigma(x)| \le \frac{4}{1+|x|^2}. \tag{B.1}$$

Consider the following equations:

$$\frac{\partial \overline{u}_{2R}}{\partial t} = \Delta\overline{u}_{2R} + \sum_{i=1}^{n} (\overline{f}_{2R})_i \frac{\partial \overline{u}_{2R}}{\partial x_i} - \overline{V}_{2R}\overline{u}_{2R} \tag{B.2}$$

$$\frac{\partial \widetilde{u}_{2R}}{\partial t} = \Delta\widetilde{u}_{2R} + \sum_{i=1}^{n} (\widetilde{f}_{2R})_i \frac{\partial \widetilde{u}_{2R}}{\partial x_i} - \widetilde{V}_{2R}\widetilde{u}_{2R} \tag{B.3}$$

in the ball B_{2R} of radius $2R$ with the Neumann condition, where $(f_{2R})_i$ and $(\widetilde{f}_{2R})_i$ denote the ith components of f_{2R} and $\widetilde{f}_{2R}$, respectively. Let $\overline{\psi}_{2R}(x) = \overline{\psi}(x)\sigma(x)$ and $\widetilde{\psi}_{2R}(x) = \widetilde{\psi}(x)\sigma(x)$ to be the initial conditions of (B.2) and (B.3), respectively. Then (B.2) and (B.3) have unique solutions, respectively, for $t \in [0,\infty)$ with Neumann condition on $\partial B_{2R} \times (0,T]$.

LEMMA B.2. *Assume that* (4.1)–(4.3) *hold and the first, second, and third derivatives of* $h_i(x)$ *are bounded. Let* $\widetilde{c}$ *and* δ *be positive constants such that* $\widetilde{\widetilde{c}} := \widetilde{c}+\delta < \frac{5}{254}$. *Choose* τ *and* ϵ *suitably small with* $\tau+\epsilon < \delta$. *Then the following conclusions hold for any* $0 \le t \le \tau$ *for both* $\rho\epsilon\{\overline{\rho},\widetilde{\rho}\}$, $u \in \{\overline{u},\widetilde{u}\}$, *and where* $\overline{\rho}(t,x) = \frac{\widetilde{c}(1+|x|^2)}{t+\epsilon}$, $\widetilde{\rho}(t,x) = \frac{\widetilde{\widetilde{c}}(1+|x|^2)}{t+\epsilon}$:

(i) $$\int_{\{t\}\times B_{2R}} e^{\overline{\rho}}\overline{u}_{2R}^2 \le \int_{\{0\}\times B_{2R}} e^{\overline{\rho}}\overline{u}_{2R}^2$$

(ii) $$\int_{\{t\}\times B_{2R}} e^{\overline{\rho}}|\nabla\overline{u}_{2R}|^2 \le \int_{\{0\}\times B_{2R}} e^{\overline{\rho}}|\nabla\overline{u}_{2R}|^2 + \int_0^t \int_{B_{2R}} e^{\overline{\rho}(s,x)}|\overline{u}_{2R}(s,x)|^2$$

(iii) $$\int_{\{t\}\times B_{2R}} e^{\overline{\rho}}|\Delta\overline{u}_{2R}|^2 \le \int_{\{0\}\times B_{2R}} e^{\overline{\rho}}|\Delta\overline{u}_{2R}|^2$$

$$+O\bigg(\int_{[0,t]\times B_{2R}} e^{\overline{\rho}}|\nabla\overline{\rho}|^2|\overline{f}_{2R}|^2|\nabla\overline{u}_{2R}|^2+\int_{[0,t]\times B_{2R}} e^{\overline{\rho}}|\nabla\overline{f}_{2R}|^2|\nabla\overline{u}_{2R}|^2$$
$$+\int_{[0,t]\times B_{2R}} e^{\overline{\rho}}|\overline{f}_{2R}||\nabla\overline{u}_{2R}|^2|\Delta\overline{f}_{2R}|+\int_{[0,t]\times B_{2R}} e^{\overline{\rho}}|\overline{f}_{2R}|^4|\nabla\overline{u}_{2R}|^2$$
$$+\int_{[0,t]\times B_{2R}} e^{\overline{\rho}}|\nabla(\overline{V}_{2R}\overline{u}_{2R})|^2+\int_{[0,t]\times B_{2R}} e^{\overline{\rho}}|\nabla\overline{u}_{2R}|^2\left(\sum_{i=1}^n\frac{\partial(\overline{f}_{2R})_i}{\partial x_i}\right)^2\bigg).$$

Moreover, the following inequalities hold for both $\{\overline{\rho},\overline{f},\overline{V}\}$, *or* $\{\overline{\rho},\widetilde{f},\widetilde{V}\}$ *or* $\{\widetilde{\rho},\overline{f},\overline{V}\}$ *or* $\{\widetilde{\rho},\widetilde{f},\widetilde{V}\}$ *if* δ *is small enough,*

(iv) $$\frac{\partial\overline{\rho}}{\partial t}+2|\nabla\overline{\rho}|^2-\sum_{i=1}^n\overline{f}_i\frac{\partial\overline{\rho}}{\partial x_i}-\sum_{i=1}^n\frac{\partial\overline{f}_i}{\partial x_i}-2\overline{V}\le 0.$$

Proof. (i), (ii), and (iii) follow from Lemma 1.3 of [26] by setting $\epsilon_1=\frac{1}{5}$ in that lemma. In equality (iv), it follows from

$$\frac{\partial\widetilde{\rho}}{\partial t}+2|\nabla\widetilde{\rho}|^2-\sum_{i=1}^n\overline{f}_i\frac{\partial\widetilde{\rho}}{\partial x_i}-\sum_{i=1}^n\frac{\partial\overline{f}_i}{\partial x_i}-2\overline{V}$$
$$\le\left[\frac{-\,\widetilde{\widetilde{c}}\,(1-8\,\widetilde{\widetilde{c}})}{(t+\epsilon)^2}+\frac{2c\,\widetilde{\widetilde{c}}}{t+\epsilon}+(n+2)c\right](1+|x|)^2$$

as $1-8\,\widetilde{\widetilde{c}}\ge 0$. □

PROPOSITION B.3. *Consider the parabolic differential equations* (A.5) *and* (A.6). *Let* ϕ *be any smooth function defined on* $\mathbb{R}^n$ *with compact support contained in a domain* Ω. *Let* $\overline{\rho}$ *be any smooth function on* $\mathbb{R}_+\times\mathbb{R}^n$ *satisfying*

$$\frac{\partial\overline{\rho}}{\partial t}+2|\nabla\overline{\rho}|^2-\sum_{i=1}^n\overline{f}_i\frac{\partial\overline{\rho}}{\partial x_i}-\sum_{i=1}^n\frac{\partial\overline{f}_i}{\partial x_i}-2\overline{V}\le 0. \tag{B.4}$$

Then

$$\frac{d}{dt}\int_{\{t\}\times\Omega}\phi^2e^{\overline{\rho}}(\overline{u}-\widetilde{u})^2\le\int_{\{t\}\times\Omega}\phi^2e^{\overline{\rho}}(\overline{u}-\widetilde{u})^2+10\int_{\{t\}\times\Omega}e^{\overline{\rho}}(\overline{u}-\widetilde{u})^2|\nabla\phi|^2$$
$$+4\int_{\{t\}\times\Omega}e^{\overline{\rho}}\bigg|\sum_{i=1}^n\overline{f}_i\frac{\partial\phi}{\partial x_i}\bigg|^2(\overline{u}-\widetilde{u})^2+4\int_{\{t\}\times\Omega}e^{\overline{\rho}}\phi^2\bigg|\sum_{i=1}^n(\overline{f}_i-\widetilde{f}_i)\frac{\partial\overline{\rho}}{\partial x_i}\bigg|^2\widetilde{u}^2$$
$$+2\int_{\{t\}\times\Omega}e^{\overline{\rho}}\phi^2\widetilde{u}|\overline{f}-\widetilde{f}|^2+4\int_{\{t\}\times\Omega}e^{\overline{\rho}}\phi^2\widetilde{u}^2|\overline{V}-\widetilde{V}|^2$$
$$+2\int_{\{t\}\times\Omega}e^{\overline{\rho}}\phi^2|\overline{f}-\widetilde{f}|^2\widetilde{u}^2+4\int_{\{t\}\times\Omega}e^{\overline{\rho}}\phi^2\widetilde{u}^2\bigg|\sum_{i=1}^n\bigg(\frac{\partial\overline{f}_i}{\partial x_i}-\frac{\partial\widetilde{f}_i}{\partial x_i}\bigg)\bigg|^2. \tag{B.5}$$

Proof. From (A.5) and (A.6), we deduce that

$$\frac{\partial(\overline{u}-\widetilde{u})}{\partial t}=\Delta(\overline{u}-\widetilde{u})+\sum_{i=1}^n f_i\frac{\partial(\overline{u}-\widetilde{u})}{\partial x_i}-V(u-\widetilde{u})+\sum_{i=1}^n(\overline{f}_i-\widetilde{f}_i)\frac{\partial\widetilde{u}}{\partial x_i}-(\overline{V}-\widetilde{V})\widetilde{u}. \tag{B.6}$$

Then using (B.6) and integrating by part, we obtain

$$\begin{aligned}\frac{d}{dt}\int_{\{t\}\times\Omega}\phi^2(\overline{u}-\widetilde{u})^2e^{\overline{\rho}} \le &\int_{\{t\}\times\Omega}e^{\overline{\rho}}\phi^2(\overline{u}-\widetilde{u})^2\left(\frac{\partial\overline{\rho}}{\partial t}+2|\nabla\overline{\rho}|^2-\sum_{i=1}^n\overline{f}_i\overline{\rho}_i-\sum_{i=1}^n\frac{\partial\overline{f}_i}{\partial x_i}-2\overline{V}\right)\\ &-\frac{1}{2}\int_{\{t\}\times\Omega}e^{\overline{\rho}}\phi^2|\nabla(\overline{u}-\widetilde{u})|^2+8\int_{\{t\}\times\Omega}e^{\overline{\rho}}(\overline{u}-\widetilde{u})^2|\nabla\phi|^2\\ &-2\int_{\{t\}\times\Omega}\phi e^{\overline{\rho}}\left(\sum_{i=1}^n\overline{f}_i\phi_i\right)(\overline{u}-\widetilde{u})^2-4\int_{\{t\}\times\Omega}e^{\overline{\rho}}\phi\left[\sum_{i=1}^n(\overline{f}_i-\widetilde{f}_i)\phi_i\right]\widetilde{u}(\overline{u}-\widetilde{u})\\ &-2\int_{\{t\}\times\Omega}e^{\overline{\rho}}\phi^2\sum_{i=1}^n(\overline{f}_i-\widetilde{f}_i)\frac{\partial\overline{\rho}}{\partial x_i}\widetilde{u}(\overline{u}-\widetilde{u})+2\int_{\{t\}\times\Omega}e^{\overline{\rho}}\phi^2\widetilde{u}^2|f-\widetilde{f}|^2\\ &-2\int_{\{t\}\times\Omega}e^{\overline{\rho}}\phi^2(\overline{u}-\widetilde{u})\widetilde{u}\sum_{i=1}^n\left(\frac{\partial\overline{f}_i}{\partial x_i}-\frac{\partial\widetilde{f}_i}{\partial x_i}\right)-2\int_{\{t\}\times\Omega}e^{\overline{\rho}}\phi^2(\overline{u}-\widetilde{u})\widetilde{u}(\overline{V}-\widetilde{V}).\end{aligned}\tag{B.7}$$

In view of (B.4), (B.7) implies that

$$\begin{aligned}\frac{d}{dt}\int_{\{t\}\times\Omega}\phi^2e^{\overline{\rho}}(\overline{u}-\widetilde{u})^2 \le\ & 8\int_{\{t\}\times\Omega}e^{\overline{\rho}}(\overline{u}-\widetilde{u})^2|\nabla\phi|^2+4\int_{\{t\}\times\Omega}e^{\overline{\rho}}(\overline{u}-\widetilde{u})^2\left|\sum_{i-1}^n\overline{f}_i\frac{\partial\phi}{\partial x_i}\right|^2\\ &+\frac{1}{4}\int_{\{t\}\times\Omega}e^{\overline{\rho}}\phi^2(\overline{u}-\widetilde{u})^2+\frac{1}{4}\int_{\{t\}\times\Omega}e^{\overline{\rho}}\phi^2(\overline{u}-\widetilde{u})^2\\ &+4\int_{\{t\}\times\Omega}e^{\overline{\rho}}\phi^2\left|\sum_{i=1}^n(\overline{f}_i-\widetilde{f}_i)\frac{\partial\overline{\rho}}{\partial x_i}\right|^2\widetilde{u}^2+2\int_{\{t\}\times\Omega}e^{\overline{\rho}}\phi^2\widetilde{u}^2|\overline{f}-\widetilde{f}|^2\\ &+\frac{1}{4}\int_{\{t\}\times\Omega}e^{\overline{\rho}}\phi^2(\overline{u}-\widetilde{u})^2+4\int_{\{t\}\times\Omega}e^{\overline{\rho}}\phi^2\widetilde{u}^2|\overline{V}-\widetilde{V}|^2\\ &+4\left[\frac{1}{2}\int_{\{t\}\times\Omega}e^{\overline{\rho}}(\overline{u}-\widetilde{u})^2|\nabla\phi|^2+\frac{1}{2}\int_{\{t\}\times\Omega}e^{\overline{\rho}}\phi^2|\overline{f}-\widetilde{f}|^2\widetilde{u}^2\right]\\ &+2\left[\frac{1}{8}\int_{\{t\}\times\Omega}e^{\overline{\rho}}\phi^2(\overline{u}-\widetilde{u})^2+2\int_{\{t\}\times\Omega}e^{\overline{\rho}}\phi^2\widetilde{u}^2\left|\sum_{i=1}^n\left(\frac{\partial\overline{f}_i}{\partial x_i}-\frac{\partial\widetilde{f}_i}{\partial x_i}\right)\right|^2\right].\end{aligned}$$

Inequality (B.5) follows immediately. □

The following theorem states that when τ is sufficiently small and $\overline{\psi}(x)$ close to $\widetilde{\psi}(x)$ in L^2-sense, then the solution $\widetilde{u}(t,x)$ of (A.6) approximates the solution $\overline{u}(t,x)$ of (A.5) well in L^2-sense.

THEOREM B.4. *Consider the parabolic differential equation* (A.5) *and* (A.6). *Assume that* (4.1)–(4.3) *hold and the first, second, and third derivatives of* $h_i(x)$ *are bounded. Let* $\widetilde{c}$ *and* δ *be positive constants such that* $\widetilde{\widetilde{c}}:=\widetilde{c}+\delta<\frac{5}{254}$. *Let*

$$\overline{\rho}(t,x)=\frac{\widetilde{c}(1+|x|^2)}{t+\epsilon},\qquad \widetilde{\rho}(t,x)=\frac{\widetilde{\widetilde{c}}\ (1+|x|^2)}{t+\epsilon}.$$

Suppose that

$$\int_{\mathbb{R}^n}e^{\overline{\rho}(0,x)}(|\overline{\psi}(x)|^2+|\nabla\widetilde{\psi}(x)|^2+|\Delta\overline{\psi}(x)|^2)<\infty$$
$$\int_{\mathbb{R}^n}e^{\overline{\rho}(0,x)}(|\widetilde{\psi}(x)|^2+|\nabla\widetilde{\psi}(x)|^2+|\Delta\widetilde{\psi}(x)|^2)<\infty.$$

Choose τ and ϵ suitably small so that $\tau+\epsilon<\delta$ and the conclusions of Lemma B.2 *hold. Suppose that for $0\le t\le\tau$,*

$$|\overline{f}(t,x)-\widetilde{f}(t,x)|\le\widetilde{\epsilon}_1 c(1+|x|) \tag{B.8}$$

$$\left|\sum_{i=1}^n\left(\frac{\partial\overline{f}_i}{\partial x_i}(t,x)-\frac{\partial\widetilde{f}_i}{\partial x_i}(t,x)\right)\right|\le\widetilde{\epsilon}_1 c \tag{B.9}$$

$$|\overline{V}(t,x)-\widetilde{V}(t,x)|\le\widetilde{\epsilon}_1 c(1+|x|^2) \tag{B.10}$$

$$\int_{\mathbb{R}^n} e^{\overline{\rho}(0,x)}|\overline{\psi}(x)-\widetilde{\psi}(x)|^2\le\widetilde{\epsilon}_2. \tag{B.11}$$

Then

$$\begin{aligned}\int_{\{t\}\times\mathbb{R}^n} e^{\overline{\rho}}(\overline{u}-\widetilde{u})^2 &\le\widetilde{\epsilon}_2 e^t+16\widetilde{\epsilon}_1^2c^2\widetilde{c}^2\frac{t}{\epsilon(t+\epsilon)}e^t d_1+24t\widetilde{\epsilon}_1^2c^2e^t d_1\\ &\le\widetilde{\epsilon}_2e^\tau+\widetilde{\epsilon}_1^2\tau e^\tau c_1,\end{aligned}$$

where $d_1=\int_{\mathbb{R}^n}e^{\overline{\rho}(0,x)}(\widetilde{\psi}(x))^2$, $c_1=\frac{16c^2\widetilde{c}^2d_1}{\epsilon^2}+24c^2d_1$, and c is a constant for linear growth of $\nabla\overline{V}$ and $\nabla\widetilde{V}$, i.e., $|\nabla\overline{V}(t,x)|\le c(1+|x|)$, and $|\nabla\widetilde{V}(t,x)|\le c(1+|x|)$.

Proof. Let $R_0\ge1$ and $B_{R_0}^c=\{x\in\mathbb{R}^n:|x|>R_0\}$ and

$$\phi(x)=\begin{cases}1 & \text{for } |x|\le R_0\\ \frac{\log R-\log|x|}{\log R-\log R_0} & \text{for } R_0\le|x|\le R=2R_0\\ 0 & \text{for } |x|\ge R=2R_0.\end{cases}$$

Let Ω be defined as B_R in Proposition B.3. In view of Lemma B.1 and (B.5), we have

$$\begin{aligned}&e^{-t}\int_{\{t\}\times\Omega}\phi^2e^{\overline{\rho}}(\overline{u}-\widetilde{u})^2-\int_{\{0\}\times\Omega}\phi^2e^{\overline{\rho}}(\overline{u}-\widetilde{u})^2\\ &\le10\int_0^t e^{-s}\int_{\{s\}\times\Omega}e^{\overline{\rho}}(\overline{u}-\widetilde{u})^2|\nabla\phi|^2+4\int_0^t e^{-s}\int_{\{s\}\times\Omega}e^{\rho}\left|\sum_{i=1}^n\overline{f}_i\frac{\partial\phi}{\partial x_i}\right|^2(\overline{u}-\widetilde{u})^2\\ &+4\int_0^t e^{-s}\int_{\{s\}\times\Omega}e^{\overline{\rho}}\phi^2\left|\sum_{i=1}^n(\overline{f}_i-\widetilde{f}_i)\frac{\partial\rho}{\partial x_i}\right|^2\widetilde{u}^2+2\int_0^t e^{-s}\int_{\{s\}\times\Omega}e^{\overline{\rho}}\phi^2\widetilde{u}^2|\overline{f}-\widetilde{f}|^2\\ &+4\int_0^t e^{-s}\int_{\{s\}\times\Omega}e^{\overline{\rho}}\phi^2\widetilde{u}^2|\overline{V}-\widetilde{V}|^2+2\int_0^t e^{-s}\int_{\{s\}\times\Omega}e^{\overline{\rho}}\phi^2|\overline{f}-\widetilde{f}|^2\widetilde{u}^2\\ &+4\int_0^t e^{-s}\int_{\{s\}\times\Omega}e^{\overline{\rho}}\phi^2\widetilde{u}^2\left|\sum_{i=1}^n\left(\frac{\partial\overline{f}_i}{\partial x_i}-\frac{\partial\widetilde{f}_i}{\partial x_i}\right)\right|^2\\ &\le\frac{10e^{\overline{\rho}(0,R)}}{R_0^2(\log R-\log R_0)^2}\int_0^t e^{-s}\int_{\{s\}\times(B_{R_0}^C\cap B_R)}(\overline{u}-\widetilde{u})^2\\ &+\frac{4c^2(1+R)^2e^{\overline{\rho}(4R)}}{R_0^2(\log R-\log R_0)^2}\int_0^t e^{-s}\int_{\{s\}\times(B_{R_0}^c\cap B_R)}(\overline{u}-\widetilde{u})^2\end{aligned}$$

$$
\begin{aligned}
&+4\int_0^t e^{-s}\int_{\{s\}\times\Omega} e^{\overline{\rho}}\phi^2\left|\sum_{i=1}^n(\overline{f}_i-\widetilde{f}_i)\frac{\partial\overline{\rho}}{\partial x_i}\right|^2\widetilde{u}^2\\
&+4\int_0^t e^{-s}\int_{\{s\}\times\Omega} e^{\overline{\rho}}\phi^2\widetilde{u}^2|\overline{f}-\widetilde{f}|^2+4\int_0^t e^{-s}\int_{\{s\}\times\Omega} e^{\overline{\rho}}\phi^2\widetilde{u}^2|\overline{V}-\widetilde{V}|^2\\
&+4\int_0^t e^{-s}\int_{\{s\}\times\Omega} e^{\overline{\rho}}\phi^2\widetilde{u}^2\left|\sum_{i=1}^n\left(\frac{\partial\overline{f}_i}{\partial x_i}-\frac{\partial\widetilde{f}_i}{\partial x_i}\right)\right|^2. \qquad \text{(B.12)}
\end{aligned}
$$

Observe that (B.11) implies

$$
e^t\int_{\{0\}\times B_R} e^{\overline{\rho}}(\overline{u}-\widetilde{u})^2\le\widetilde{\epsilon}_2 e^t. \qquad \text{(B.13)}
$$

By Corollary 4.1 of [26], u and $\widetilde{u}$ decay like Gaussian in x variables. So we shall assume

$$
\max_{x\in\mathbb{R}^n}(|u|,|\widetilde{u}|)\le D_1e^{-D_2|x|^2}\text{ for }t\text{ small}, \qquad \text{(B.14)}
$$

for some $D_1,D_2>0$. In view of the proof of Corollary 4.1 of [26], we can take $D_2\ge\frac{4\widetilde{c}}{\epsilon}+1$ for sufficiently small t

$$
\begin{aligned}
&\frac{[10+4c^2(1+R)^2]e^{\overline{\rho}(0,R)}}{R_0^2(\log R-\log R_0)^2}\int_0^t e^{-s}\int_{\{s\}\times(B^c_{R_0}\cap B_R)}(\overline{u}-\widetilde{u})^2\\
&\le\frac{4[10+4c^2(1+R)^2]te^{\overline{\rho}(0,R)}}{R_0^2(\log R-\log R_0)^2}\int_{B^c_{R_0}\cap B_r}D_1e^{-D_2|x|^2}\\
&\le\frac{4D_1[10+4c^2(1+R)^2]te^{\overline{\rho}(0,R)}}{R_0^2(\log R-\log R_0)^2}\omega_0R^ne^{-D_2R_0^2}\\
&=\frac{4\omega_0D_1R^n[10+4c^2(1+R)^2]t}{R_0^2(\log R-\log R_0)^2}\exp\left(\frac{\overline{c}}{\epsilon}+\left(\frac{4\overline{c}}{\epsilon}-D_2\right)R_0^2\right)\\
&\le\frac{4\omega_0D_1R^n[10+4c^2(1+R)^2]t}{R_0^2(\log R-\log R_0)^2}\exp\left(-R_0^2+\frac{\overline{c}}{\epsilon}\right), \qquad \text{(B.15)}
\end{aligned}
$$

where ω_0 is the volume of the unit ball in $\mathbb{R}^n$. Recall that $|\nabla\overline{\rho}|^2=\frac{4\overline{c}^2|x|^2}{(t+\epsilon)^2}$. Hence (B.8) implies

$$
\begin{aligned}
&4\int_0^t e^{-s}\int_{\{s\}\times\Omega} e^{\overline{\rho}}\phi^2\left|\sum_{i=1}^n(\overline{f}_i-\widetilde{f}_i)\frac{\partial\overline{\rho}}{\partial x_i}\right|^2\widetilde{u}^2\\
&\le 4\int_0^t\int_{\{s\}\times\Omega} e^{\overline{\rho}}\phi^2|\overline{f}-\widetilde{f}|^2|\nabla\rho|^2\widetilde{u}^2\\
&\le 4\int_0^t\int_{\{s\}\times\Omega}\widetilde{\epsilon}_1^2c^2(1+|x|)^2\frac{4\overline{c}^2|x|^2}{(s+\epsilon)^2}e^{\overline{\rho}}\widetilde{u}^2\\
&\le 16\widetilde{\epsilon}_1^2c^2\widetilde{c}^2\int_0^t\frac{1}{(s+\epsilon)^2}\int_{\{s\}\times B_R}e^{\widetilde{\rho}}\widetilde{u}^2. \qquad \text{(B.16)}
\end{aligned}
$$

Similarly, we can prove that

$$4\int_0^t e^{-s}\int_{\{s\}\times\Omega} e^{\overline{\rho}}\phi^2\widetilde{u}^{\,2}|\overline{f}-\widetilde{f}|^2 \le 16\widetilde{\epsilon}_1^2c^2\int_0^t\int_{\{s\}\times B_R} e^{\widetilde{\rho}}\widetilde{u}^2 \tag{B.17}$$

$$4\int_0^t e^{-s}\int_{\{s\}\times\Omega} e^{\overline{\rho}}\phi^2\widetilde{u}^{\,2}|\overline{V}-\widetilde{V}|^2 \le 4\widetilde{\epsilon}_1^2c^2\int_0^t\int_{\{s\}\times B_R} e^{\widetilde{\rho}}\widetilde{u}^2 \tag{B.18}$$

$$4\int_0^t e^{-s}\int_{\{s\}\times\Omega} e^{\overline{\rho}}\phi^2\widetilde{u}^2\left|\sum_{i=1}^n\left(\frac{\partial\overline{f}_i}{\partial x_i}-\frac{\partial\widetilde{f}_i}{\partial x_i}\right)\right|^2 \le 4\widetilde{\epsilon}_1^2c^2\int_0^t\int_{\{s\}\times B_R} e^{\overline{\rho}}\widetilde{u}^2. \tag{B.19}$$

Hence (B.12)–(B.18) and Lemma B.2 imply that

$$\begin{aligned}
&\int_{\{t\}\times B_{R_0}} e^{\overline{\rho}}(\overline{u}-\widetilde{u})^2\\
&\quad\le \widetilde{\epsilon}_2e^t+\frac{4te^t\omega_0R^nD_1[10+4c^2(1+R)^2]}{R_0^2(\log R-\log R_0)^2}\exp\left(-R_0^2+\frac{\overline{c}}{\epsilon}\right)\\
&\quad+16e^t\widetilde{\epsilon}_1^2c^2\widetilde{c}^2\int_0^t\frac{1}{(s+\epsilon)^2}\int_{\{s\}\times B_R} e^{\widetilde{\rho}}\widetilde{u}^2+16e^t\widetilde{\epsilon}_1^2c^2\int_0^t\int_{\{s\}\times B_R} e^{\widetilde{\rho}}\widetilde{u}^2\\
&\quad+4e^t\widetilde{\epsilon}_1^2c^2\int_0^t\int_{\{s\}\times B_R} e^{\widetilde{\rho}}\widetilde{u}^2+4\widetilde{\epsilon}_1^2c^2e^t\int_0^t\int_{\{s\}\times B_R} e^{\widetilde{\rho}}\widetilde{u}^2\\
&\quad\le \widetilde{\epsilon}_2e^t+\frac{4te^t\omega_0D_1R^n[10+4c^2(1+R)^2]}{R_0^2(\log R-\log R_0)^2}\exp\left(-R_0+\frac{\overline{c}}{\epsilon}\right)\\
&\quad+16e^t\widetilde{\epsilon}_1^2c^2\widetilde{c}^2\int_0^t\frac{1}{(s+\epsilon)^2}\int_{\{0\}\times B_R} e^{\widetilde{\rho}}\widetilde{u}^2\\
&\quad+24e^t\widetilde{\epsilon}_1^2c^2\int_0^t\int_{\{0\}\times B_R} e^{\widetilde{\rho}}\widetilde{u}^2\\
&\quad\le \widetilde{\epsilon}_2e^t+\frac{16e^t\widetilde{\epsilon}_1^2c^2\widetilde{c}^{\,2}d_1t}{\epsilon(t+\epsilon)}+24te^t\widetilde{\epsilon}_1^2c^2d_1\\
&\quad+\frac{4te^t\omega_0D_1R^n[10+4c^2(1+R)^2]}{R_0^2(\log R-\log R_0)^2}\exp\left(-R_0^2+\frac{\overline{c}}{\epsilon}\right).
\end{aligned}\tag{B.20}$$

Let $R=2R_0$ go to infinity in (5.20), we obtain the estimate in the statement of Theorem B.4. □

Now we are ready to consider the global situation. For a fixed $T>0$, we want to find the solution $\overline{u}(t,x)$ of (A.5).

THEOREM B.5. *Let $\overline{u}(t,x)$ and $\widetilde{u}_i(t,x)$ be the solutions of* (A.14) *and* (A.15)*, respectively. For $\widetilde{\epsilon}_1>0$, let $|\mathcal{P}_k|=\sup_i\{|t_i-t_{i-1}|\}$ be sufficiently small so that the following estimates hold:*

$$|\overline{f}(t,x)-\widetilde{f}(\tau_{i-1},x)|\le\widetilde{\epsilon}_1c(1+|x|),\ \text{for}\ \ \tau_{i-1}\le t\le\tau_i, \tag{B.21}$$

$$\left|\sum_{j=1}^n\left(\frac{\partial\overline{f}_j}{\partial x_j}(t,x)-\frac{\partial\widetilde{f}_j}{\partial x_j}(\tau_{i-1},x)\right)\right|\le\widetilde{\epsilon}_1c \tag{B.22}$$

$$\left|\overline{V}(t,x)-\widetilde{V}(\tau_{i-1},x)\right|\le\widetilde{\epsilon}_1c(1+|x|^2). \tag{B.23}$$

Then

$$\int_{\mathbb{R}^n} e^{\overline{\rho}(T,x)}(\overline{u}(T,x)-\widetilde{u}_k(T,x))^2\le\widetilde{\epsilon}_1^2c_1k|\mathcal{P}_k|e^T\le\widetilde{\epsilon}_1^2c_1c_2(T),$$

where $\overline{\rho}(t,x) = \frac{\widetilde{c}(1+|x|^2)}{t+\epsilon}$ *so that the conclusion of* Theorem B.4 *holds,* c_1 *is the constant in* Theorem B.4*, and* $c_2(T)$ *is a constant that depends only on* T*.*

Proof. In view of $\widetilde{u}_1(0,x) = \psi(x) = \overline{u}(0,x)$ and Theorem B.4, we have

$$\int_{\{\tau_1\}\times\mathbb{R}^n} e^{\overline{\rho}}(\overline{u}-\widetilde{u})^2 \le \widetilde{\epsilon}_1^2 \tau_1 e^{\tau_1} c_1$$

$$\int_{\{\tau_2\}\times\mathbb{R}^n} e^{\overline{\rho}}(\overline{u}-\widetilde{u})^2 \le \widetilde{\epsilon}_1^2 c_1[\tau_1 e^{\tau_2} + (\tau_2-\tau_1)e^{\tau_2-\tau_1}].$$

By Theorem B.4 and induction, we have

$$\begin{aligned}\int_{\{\tau_k\}\times\mathbb{R}^n} e^{\overline{\rho}}(\overline{u}-\widetilde{u})^2 &= \widetilde{\epsilon}_1^2 c_1[\tau_1 e^{\tau_k} + (\tau_2-\tau_1)e^{\tau_k-\tau_1} + (\tau_3-\tau_2)e^{\tau_k-\tau_2} \\ &\qquad + \cdots + (\tau_k-\tau_{k-1})e^{\tau_k-\tau_{k-1}}] \\ &\le \widetilde{\epsilon}_1^2 c_1 k|\mathcal{P}_k|e^T \\ &\le \widetilde{\epsilon}_1^2 c_1 c_2(T). \qquad \square\end{aligned}$$

As a consequence of Theorem B.5, we have the following L^2-convergent theorem.

THEOREM B.6. *Fix* $T > 0$*, let* $\mathcal{P}_k = \{0 < \tau_1 < \tau_2 < \cdots < \tau_k = T\}$ *be a partition of* $[0,T]$*. Let* $\overline{u}(t,x)$ *be the solution of* (A.14) *on* $[0,T]\times\mathbb{R}^n$*. Let* $\widetilde{u}_i(t,x)$ *be the solution of* (A.15) *on* $[\tau_{i-1},\tau_i]\times\mathbb{R}^n$*. Let* $\overline{\rho}(t,x) = \frac{\widetilde{c}(1+|x|^2)}{t+\epsilon}$ *so that the conclusion of* Theorem B.5 *holds. Then*

$$\lim_{|\mathcal{P}_k|\to 0}\int_{\{T\}\times\mathbb{R}^n} \overline{\rho}(\overline{u}-\widetilde{u}_k)^2 = 0.$$

Acknowledgments. We gratefully acknowledge the referees for their careful reading of our paper and some useful suggestions of improving our presentation.

REFERENCES

[1] A. BENSOUSSAN, R. GLOWINSKI AND A. RASCANU, *Approximation of the Zakai equation by the splitting up method*, in Stochastic Systems and Optimization (Warsaw, 1988) (J. Zabczyk, ed.), Springer-Verlag, Berlin, (1989), pp. 257–265.

[2] R. W. BROCKETT, *Nonlinear systems and nonlinear estimation theory,* in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. C. Willems, eds. Reidel, Dordrecht, Boston, 1981.

[3] R. W. BROCKETT AND J. M. C. CLARK, *The geometry of the conditional density functions,* in Analysis and Optimization of Stochastic Systems, O. L. R. Jacobs, et al., eds., Academic Press, New York, 1980, pp. 299–309.

[4] M. CHALEYAT-MAUREL AND D. MICHEL, *Des resultats de non-existence de filtre de dimension finie*, Stochastics, 13 (1984), pp. 83–102.

[5] J. CHEN AND S. S.-T. YAU, *Finite dimensional filters with nonlinear drift VI: Linear structure of* Ω, Math. Control Signals Systems, 9 (1996), pp. 370–385.

[6] W. L. CHIOU AND S. S.-T. YAU, *Finite-dimensional filters with nonlinear drift II: Brockett's problem on classification of finite-dimensional estimation algebra*, SIAM J. Control Optim., 32 (1994), pp. 297–310.

[7] M. H. A. DAVID, *On a multiplicative functional transformation arising in nonlinear filtering theory,* Z. Wahrsch Verw. Gebiete, 54 (1980), pp. 125–139.

[8] R. J. ELLIOTT AND R. GLOWINSKI, *Approximations to solutions of the Zakai filtering equation*, Stochastic Anal. Appl., 7 (1988), pp. 145–168.

[9] P. FLORCHINGER AND F. LEGLAND, *Time-discretization of the Zakai equation for diffusion processes observed in correlated noise, Analysis and Optimization of Systems*, A. Bensoussan and J.J. Lions, eds., Lecture Notes in Inform. Sci. and Control, 144, Springer-Verlag, Berlin, pp. 228–237, Stoachastics Stochastics Rep., 35 (1991), pp. 233–256.

[10] M. Fujisaki, G. Kallianpur and H. Kunita, *Stochastic differential equations for the nonlinear filtering problem*, Osaka J. Math., 1 (1972), pp. 19–40.

[11] R. E. Kalman and R. S. Bucy, *New results in linear filtering and prediction theory,* Trans. ASME Ser. D. J. Basic Engrg., 83 (1961), pp. 95–108.

[12] R. Mikulevicious and B. L. Rozovskii, *Separation of observations and parameters in nonlinear filtering*, Proceedings of the 32nd IEEE Conference on Decision and Control, San Antonio, TX, 1993, pp. 1564–1559.

[13] S. K. Mitter, *On the analogy between mathematical problems of nonlinear filtering and quantum physics,* Ricerche Automat., 10 (1979), pp. 163–216.

[14] L.-F. Tam, W. S. Wong, and S. S.-T. Yau, *On a necessary and sufficient condition for finite dimensionality of estimation algebras,* SIAM J. Control Optim., 28 (1990), pp. 173–185.

[15] S. S.-T. Yau, *Recent results on nonlinear filtering: New class of finite dimensional filters,* in Proceedings of the 29th Conf. Decision and Control, Honolulu, 1990, pp. 231–233.

[16] S. S.-T. Yau, *Finite dimensional filters with nonlinear drift I: A class of filters including both Kalman-Bucy filters and Benes filters,* J. Math. Syst., Estimation and Control, 4(2) (1994), pp. 181–203.

[17] S. S.-T. Yau, *Brockett's problem on nonlinear filtering theory*, in Lectures on Systems, Control and Information, AMS/IP, Stud. Adv. Math., 17 (2000), pp. 177–212.

[18] S.-T. Yau and G.-Q. Hu, *Direct method without Riccati equation for Kalman-Bucy filtering system with arbitrary initial conditions,* in Proceedings of the 13th World Congress IFAC, vol. II, San Francisco, CA, 1996, pp. 469–474.

[19] S. S.-T. Yau and G. Q. Hu, *Finite dimensional filters with nonlinear drift. X: Explicit solution of DMZ equation*, IEEE Trans. Automat. Control, 46 (2001), pp. 142–148.

[20] S. S.-T. Yau and G. Q. Hu, *Finite dimensional filters with nonlinear drift XIV: Classification of finite-dimensional estimation algebras of maximal rank with arbitrary state space dimension and Mitter conjecture*, submitted.

[21] S. S.-T. Yau, X. Wu and W. S. Wong, *Hessian Matrix non-decomposition theorem*, Math. Res. Lett., 6 (1999), pp. 663–673.

[22] S. S.-T. Yau and S.-T. Yau, *Finite dimensional filters with nonlinear drift III: Duncan-Mortensen-Zakai equation with arbitrary initial condition for linear filtering system and the Benes filtering system,* IEEE Trans. Aerosp. Elec. Syst., 33, 1997, pp. 1277–1294.

[23] S. S.-T. Yau and S.-T. Yau, *New direct method for Kalman-Bucy filtering system with arbitrary initial condition,* in Proceedings of the 33rd Conf. Decision and Control, Lake Buena Vista, FL, 1994, pp. 1221–1225.

[24] S. S.-T. Yau and S.-T. Yau, *Explicit solution to a Kolmogorov equation,* Appl. Math. Optim., 34 (1996), pp. 231–266.

[25] S.-T. Yau and S. S.-T. Yau, *Real time solution of nonlinear filtering problem without memory, I*, Math. Res. Lett., 7 (2000), pp. 671–693.

[26] S.-T. Yau and S. S.-T. Yau, *Existence and uniqueness and decay estimates for the time dependent parabolic equation with application to Duncan-Mortensen-Zakai equation*, Asian J. Math., 2 (1998), pp. 1079–1149.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 3, pp. 1040–1066

LINEAR COMPLEMENTARITY SYSTEMS: ZENO STATES*

JINGLAI SHEN† AND JONG-SHI PANG†

Abstract. A linear complementarity system (LCS) is a hybrid dynamical system defined by a linear time-invariant ordinary differential equation coupled with a finite-dimensional linear complementarity problem (LCP). The present paper is the first of several papers whose goal is to study some fundamental issues associated with an LCS. Specifically, this paper addresses the issue of Zeno states and the related issue of finite number of mode switches in such a system. The cornerstone of our study is an expansion of a solution trajectory to the LCS near a given state in terms of an observability degree of the state. On the basis of this expansion and an inductive argument, we establish that an LCS satisfying the P-property has no strongly Zeno states. We next extend the analysis for such an LCS to a broader class of problems and provide sufficient conditions for a given state to be weakly non-Zeno. While related mode-switch results have been proved by Brunovsky and Sussmann for more general hybrid systems, our analysis exploits the special structure of the LCS and yields new results for the latter that are of independent interest and complement those by these two and other authors.

Key words. linear complementarity systems, Zeno states, P-matrix, complementarity problem

AMS subject classifications. 34A40, 90C33, 93B12, 93C10

DOI. 10.1137/040612270

1. Introduction.

A linear complementarity system (LCS) is a special dynamical system defined by an linear ordinary differential equation (ODE) involving an algebraic variable that is required to be a solution of a standard linear complementarity problem (LCP) [11]. While being a special instance of a differential variational inequality, which has recently been studied in great depth in [23], the LCS has itself received an extensive treatment in two excellent Ph.D. theses [5, 13] and in related articles [3, 8, 9, 15, 16, 17]. In addition, the LCS belongs to the broad framework of a hybrid system [18, 20, 30, 34, 35, 36, 26, 28], which is defined by a finite number of smooth ODEs, called *modes*, with transitions between the modes occurring along a state trajectory. Examples of dynamical systems in which the complementarity paradigm has played a prominent role include nonsmooth mechanical systems [2, 24] in general and multibody dynamics simulation under frictional contacts in particular [1, 21, 29, 31, 32], as well as switched electrical networks and switched control systems, e.g., relay systems and variable structure systems [6, 7, 14, 19, 38]. In addition, linear-quadratic dynamic Nash games with linear dynamics and control constraints naturally lead to LCSs with special boundary conditions. For an excellent state-of-the-art review of complementarity systems and their applications in engineering and economics, we refer to the excellent recent article by Schumacher [27].

The LCS occupies a fundamental role in the study of nonsmooth dynamical systems because it is arguably the simplest of such systems. Though seemingly simple, the analysis of the LCS in general is complicated by impulsive and multimodal behavior of its solutions. In the references cited above, such as in the two theses [5, 13],

*Received by the editors July 26, 2004; accepted for publication (in revised form) February 26, 2005; published electronically September 20, 2005. This work was supported by the National Science Foundation under a Focused Research Group grant DMS-0353216 and also partially by the grant CCR-0098013.

http://www.siam.org/journals/sicon/44-3/61227.html

†Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180 (shenj2@rpi.edu, pangj@rpi.edu).

the study of the LCS has employed many concepts and results from (constrained) linear systems theory; in particular, the concept of system passivity [37] has played a major role. In contrast to this system-theoretic approach, we feel that a better understanding of the LCS as a basic mathematical model can be achieved by considering the simplest, albeit nontrivial, instance of such a system. Motivated by this contrasting "mathematical programming" approach, we are led to consider an LCS satisfying the "P-property," i.e., where the underlying finite-dimensional LCP has a unique solution for all constant vectors. An immediate consequence of this property is that the LCS is globally equivalent to an ODE with a piecewise linear, thus Lipschitz continuous, right-hand side (which albeit is only implicitly defined). While this is a great simplification, the piecewise linear nature of the right-hand side renders the LCS a nonsmooth system and leads to many important system-theoretic and control issues that require careful study. Several of these topics are the main concern of this and accompanying papers. Extending the class of LCSs with the P-property, we will also consider a broader class of systems and study their (non-)Zeno states.

The organization of the rest of the paper is as follows. In section 2, we formally define the LCS, review some basic results of the LCP, and introduce two new LCP concepts that are useful for our study. Section 3 addresses the question of whether there can be infinitely many mode transitions in any finite time, i.e., the *Zeno behavior* of the LCS. Formal algebraic definitions of (non-)Zeno states and of mode switches that are tailored to the LCS are presented. An expansion based technique is developed to prove non-Zenoness, which is applicable to an LCS with the P-property. Extended Zeno results are presented in section 4. A special bimodal system is considered in section 5. The paper ends with some concluding remarks in the sixth and last section.

2. Preliminary discussion. Defined by a tuple of four matrices, $A \in \Re^{n\times n}$, $B \in \Re^{n\times m}$, $C \in \Re^{m\times n}$, and $D \in \Re^{m\times m}$, and a vector $x^0 \in \Re^n$, the goal of the LCS is to find trajectories $x(t) \in \Re^n$ and $u(t) \in \Re^m$ satisfying

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ 0 \le u \perp Cx + Du &\ge 0, \\ x(0) &= x^0, \end{aligned} \tag{1}$$

where $\dot{x} \equiv \frac{dx}{dt}$ denotes the time derivative of the trajectory $x(t)$ and $a \perp b$ means that the two vectors a and b are orthogonal, i.e., $a^T b = 0$. While it is in general possible for the above differential and complementarity conditions to hold only at almost all times t, in the present paper, the conditions that we will impose on the tuple (A, B, C, D) will ensure that the x-trajectory is continuously differentiable and the u-trajectory is well defined (albeit not necessarily continuous) on the time interval of interest.

It is clear that LCP theory has a major role to play in the study of the LCS. For this reason, we summarize in the next subsection some of the essential concepts from this theory that are relevant to the developments in this paper. Details of this review can be found in the monograph [11], and the results therein will be used freely. Two new LCP concepts are introduced in subsection 2.2.

2.1. LCP background. Formally, given a vector $q \in \Re^m$ and a matrix $M \in \Re^{m\times m}$, the aim of the LCP (q, M) is to find a vector $u \in \Re^m$ such that

$$0 \le u \perp w \equiv q + Mu \ge 0.$$

The solution set of this problem is denoted by $\mathrm{SOL}(q, M)$. Among all matrix classes in LCP theory, the most fundamental one is that of the P-matrices. Specifically, M

is a P-matrix if all its principal minors are positive. This is the class of matrices that will be the starting point in our study of the LCS (1). It is well known that M is a P-matrix if and only if $\mathrm{SOL}(q, M)$ is a singleton for all $q \in \Re^m$; moreover, the unique element of $\mathrm{SOL}(q, M)$, which we denote $u(q)$, is a piecewise linear function of $q \in \Re^m$. This implies in particular that $u(q)$ is (globally) Lipschitz continuous and directionally differentiable. The directional derivative, denoted $u'(q; dq)$, of the solution function $u(q)$ along the direction dq can be expressed as the unique solution of a certain mixed LCP. Specifically, define three fundamental index sets associated with $u(q)$:

$$\begin{aligned}
\alpha &\equiv \{\, i \,:\, u(q)_i > 0 = (\, q + Mu(q)\,)_i \,\}, \\
\beta &\equiv \{\, i \,:\, u(q)_i = 0 = (\, q + Mu(q)\,)_i \,\}, \\
\gamma &\equiv \{\, i \,:\, u(q)_i = 0 < (\, q + Mu(q)\,)_i \,\}.
\end{aligned}$$

It follows that $u'(q; dq)$ is the unique solution $\widehat{u}$ of the mixed LCP:

$$\begin{aligned}
0 &= (\, dq + M\widehat{u}\,)_\alpha, \\
0 \le \widehat{u}_\beta \perp (\, dq + M\widehat{u}\,)_\beta &\ge 0, \\
0 &= \widehat{u}_\gamma.
\end{aligned}$$

The solution set $\mathrm{SOL}(q, M)$ of an LCP is in general the union of finitely many polyhedra, each called a *piece* of this set. Indeed, we have

$$\mathrm{SOL}(q, M) = \bigcup_\alpha \left\{ u \in \Re^m : \begin{array}{ll} (\, q + Mu\,)_\alpha = 0, & u_\alpha \ge 0 \\ (\, q + Mu\,)_{\bar\alpha} \ge 0, & u_{\bar\alpha} = 0 \end{array} \right\},$$

where the union ranges over all subsets α of $\{1, \ldots, m\}$. The case where $\mathrm{SOL}(q, M)$ is convex for all $q \in \Re^m$ is particularly important. This case is characterized by the column sufficiency property of the matrix M. Specifically, a matrix M is *column sufficient* if $u \circ Mu \le 0 \Rightarrow u \circ Mu = 0$, where $\circ$ denotes the Hadamard product of two vectors. It is easy to see that the property of column sufficiency is inherited by the principal submatrices of M and also by the principal pivot transforms of M. That is, if M is column sufficient, then so is the principal submatrix $M_{\alpha\alpha}$ for all $\alpha \subseteq \{1, \ldots, m\}$; moreover, if $M_{\alpha\alpha}$ is nonsingular, then the matrix below, called the α-principal pivot transform of M,

$$\begin{bmatrix} (\, M_{\alpha\alpha}\,)^{-1} & -(\, M_{\alpha\alpha}\,)^{-1} M_{\alpha\bar\alpha} \\ M_{\bar\alpha\alpha}(\, M_{\alpha\alpha}\,)^{-1} & M_{\bar\alpha\bar\alpha} - M_{\bar\alpha\alpha}(\, M_{\alpha\alpha}\,)^{-1} M_{\alpha\bar\alpha} \end{bmatrix}, \tag{2}$$

is also column sufficient, where $\bar\alpha$ is the complement of α in $\{1, \ldots, m\}$. If M is column sufficient, then

$$\mathrm{SOL}(q, M) \equiv \left\{ u \in \Re^m : \begin{array}{ll} (\, q + Mu\,)_\alpha = 0, & u_\alpha \ge 0 \\ (\, q + Mu\,)_{\bar\alpha} \ge 0, & u_{\bar\alpha} = 0 \end{array} \right\},$$

where α is the set consisting of all indices i for which there exists a solution $u \in \mathrm{SOL}(q, M)$ with $u_i > 0$.

Another known property of the LCP that we need is the "semistability" of its solutions sets. Specifically, by [12, Proposition 5.5.5, Corollary 5.5.9], it follows that

for any matrix $M \in \Re^{m\times m}$ and every vector $q \in \Re^m$, there exist positive scalars c and ε such that

$$\| q' - q \| \leq \varepsilon \Rightarrow \mathrm{SOL}(q', M) \subseteq \mathrm{SOL}(q, M) + c \| q - q' \| \mathcal{B},$$

where $\mathcal{B}$ is the (closed) unit ball in $\Re^m$.

2.2. New LCP concepts. While we are unable to directly deal with the entire class of LCSs with a column sufficient matrix D, such matrices provide the motivation to introduce two new LCP concepts to be used later. The first new LCP concept is a broadening of the class of column sufficient matrices that addresses the convexity of the solution sets of the homogeneous problems only.

DEFINITION 1. *The matrix $M \in \Re^{m\times m}$ is said to be* weakly column sufficient *if for every triple of index sets (α, β, γ) that partition $\{1, \ldots, m\}$, the set of vectors $u \in \Re^m$ satisfying*

$$\begin{aligned} (Mu)_\alpha &= 0, \\ 0 \leq u_\beta \perp (Mu)_\beta &\geq 0, \\ u_\gamma &= 0 \end{aligned} \tag{3}$$

is convex, or equivalently, is polyhedral. □

We leave it to the reader to verify that the convexity of the solution set of (3) is equivalent to its polyhedrality. Besides the class of column sufficient matrices which must be weakly column sufficient, a "nondegenerate matrix," i.e., one whose principal minors are all nonzero, is also weakly column sufficient; indeed if M is nondegenerate, then the only solution to the system (3) is the zero vector. The following result summarizes several properties of weak column sufficiency.

PROPOSITION 2. *Let $M \in \Re^{m\times m}$ be weakly column sufficient. The following statements are valid:*

(a) *For every subset $\tilde{\alpha}$ of $\{1, \ldots, m\}$, the principal submatrix $M_{\tilde{\alpha}\tilde{\alpha}}$ is weakly column sufficient.*

(b) *For every subset $\tilde{\alpha}$ of $\{1, \ldots, m\}$ such that $M_{\tilde{\alpha}\tilde{\alpha}}$ is nonsingular, the $\tilde{\alpha}$-principal pivot transform (2) of M is weakly column sufficient.*

(c) *The solution set of the homogeneous LCP $(0, M)$ is polyhedral; in fact,*

$$SOL(0, M) \equiv \left\{ u \in \Re^m : \begin{array}{ll} (Mu)_\alpha = 0, & u_\alpha \geq 0 \\ (Mu)_{\bar{\alpha}} \geq 0, & u_{\bar{\alpha}} = 0 \end{array} \right\},$$

where α is the set consisting of all indices i for which there exists $u \in SOL(0, M)$ such that $u_i > 0$ and $\bar{\alpha}$ is the complement of α.

Proof. Let α', β', and γ' be three index sets partitioning the subset $\tilde{\alpha}$. A vector $u_{\tilde{\alpha}}$ satisfies the system

$$\begin{aligned} (M_{\tilde{\alpha}\tilde{\alpha}} u_{\tilde{\alpha}})_{\alpha'} &= 0, \\ 0 \leq u_{\beta'} \perp (M_{\tilde{\alpha}\tilde{\alpha}} u_{\tilde{\alpha}})_{\beta'} &\geq 0, \\ u_{\gamma'} &= 0 \end{aligned}$$

if and only if the vector $u \equiv (u_{\tilde{\alpha}}, 0)$ satisfies the system (3) with $(\alpha, \beta, \gamma) = (\alpha', \beta', \gamma' \cup \hat{\alpha})$, where $\hat{\alpha}$ is the complement of $\tilde{\alpha}$ in $\{1, \ldots, m\}$. Consequently, part (a) holds. To

prove (b), let $\widetilde{M}$ be the $\tilde{\alpha}$-principal pivot transform (2) of M. Let α', β', and γ' be three index sets partitioning the set $\{1, \ldots, m\}$. Consider the system

$$\begin{aligned} (\widetilde{M}\widetilde{u})_{\alpha'} &= 0, \\ 0 \leq \widetilde{u}_{\beta'} &\perp (\widetilde{M}\widetilde{u})_{\beta'} \geq 0, \\ \widetilde{u}_{\gamma'} &= 0. \end{aligned} \tag{4}$$

Let $\widetilde{w} \equiv \widetilde{M}\widetilde{u}$. By pivoting on $(M_{\tilde{\alpha}\tilde{\alpha}})^{-1}$ in $\widetilde{M}$, we can recover the original system $w = Mu$, where the variables u and w are related to $\widetilde{u}$ and $\widetilde{w}$ via the identities:

$$w \equiv \begin{pmatrix} w_{\tilde{\alpha}} \\ w_{\hat{\alpha}} \end{pmatrix}, \quad u \equiv \begin{pmatrix} u_{\tilde{\alpha}} \\ u_{\hat{\alpha}} \end{pmatrix},$$

$$\widetilde{w} = \begin{pmatrix} u_{\tilde{\alpha}} \\ w_{\hat{\alpha}} \end{pmatrix}, \quad \widetilde{u} = \begin{pmatrix} w_{\tilde{\alpha}} \\ u_{\hat{\alpha}} \end{pmatrix}.$$

Therefore, system (4) is equivalent to system (3) for some suitable triple (α, β, γ) that partitions $\{1, \ldots, m\}$. From this equivalence, the convexity of the solution set of the former system can be easily proved. This establishes (b). To prove (c), we note that the LCP $(0, M)$ is just the system (3) with $\beta = \{1, \ldots, m\}$. Hence the polyhedrality of $\mathrm{SOL}(0, M)$ follows from the weak column sufficiency of M. The representation of $\mathrm{SOL}(0, M)$ can be proved in the same way as in the case of a column sufficient matrix; see the proof in [11, Theorem 3.5.8] for details. □

To introduce the second new LCP concept, we note that if $q \neq 0$, then for every solution u of the LCP (q, M), there must exist at least one index i such that $u_i > 0$ or $w_i \equiv (q + Mu)_i > 0$. Define two sets of "identifiable indices":

$$\begin{aligned} \mathcal{I}_u &\equiv \{\, i : u_i > 0 \quad \forall\, u \in \mathrm{SOL}(q, M) \,\}, \\ \mathcal{I}_w &\equiv \{\, i : (q + Mu)_i > 0 \quad \forall\, u \in \mathrm{SOL}(q, M) \,\}, \end{aligned}$$

one, or both, of which may be empty in general.

DEFINITION 3. *The LCP (q, M), where $q \neq 0$, is* identifiable *if the following two conditions hold:*

(a) *$\mathcal{I}_u \cup \mathcal{I}_w \neq \emptyset$, and*

(b) *the principal submatrix $M_{\mathcal{I}_u \mathcal{I}_u}$ is nonsingular if $\mathcal{I}_u \neq \emptyset$. (By convention, this condition is vacuously true if $\mathcal{I}_u$ is empty.)* □

If the LCP (q, M), where $q \neq 0$, has a unique solution u, then the LCP is identifiable if $M_{\alpha\alpha}$ is nonsingular, where α is the (possibly empty) support of u. The following lemma asserts a positivity property of the "identifiable variables" of an LCP.

PROPOSITION 4. *For any pair (q, M) with $q \neq 0$, there exists a scalar $\sigma > 0$ such that $u_i \geq \sigma$ and $(q + Mu)_j \geq \sigma$ for all $u \in \mathrm{SOL}(q, M)$ and all $i \in \mathcal{I}_u$ and $j \in \mathcal{I}_w$.*

Proof. We prove the claim only for the u-variable. For each $i \in \mathcal{I}_u$, consider the optimization problem

$$\begin{aligned} \text{minimize} \quad & u_i \\ \text{subject to} \quad & u \in \mathrm{SOL}(q, M). \end{aligned}$$

Since the feasible set of this problem is the union of finitely many polyhedra and its objective function is linear and positive (hence bounded below) on this set, it follows from linear programming theory that the above problem attains a finite minimum objective value which must be positive. The desired claim follows readily. □

2.3. Back to the LCS. Returning to the LCS (1), assume that D is a P-matrix. In this case, (1) is equivalent to the ODE

$$\dot{x} = Ax + Bu(Cx), \quad x(0) = x^0, \tag{5}$$

where the right-hand side $Ax+Bu(Cx)$ is a piecewise linear function of x. (Note that $u(0) = 0$.) As such, (1) has a unique solution trajectory $(x(t), u(t))$ defined on $[0, \infty)$ with $x(t)$ being continuously differentiable and $u(t) \equiv u(Cx(t))$ being continuous. In contrast to the above representation (5) which involves the implicit function $u(Cx)$, the right-hand side of the LCS (1) can be represented explicitly using the complementarity cones associated with the matrix D. Specifically, for each index subset δ of $\{1, \ldots, m\}$ with complement $\bar{\delta}$, define the polyhedral cone

$$\mathcal{C}_\delta \equiv \left\{ q \in \Re^m : E_\delta \begin{pmatrix} q_\delta \\ q_{\bar{\delta}} \end{pmatrix} \geq 0 \right\}, \tag{6}$$

where

$$E_\delta \equiv \begin{bmatrix} -(D_{\delta\delta})^{-1} & 0 \\ -D_{\bar{\delta}\delta}(D_{\delta\delta})^{-1} & I \end{bmatrix} \in \Re^{m \times m}.$$

Since D is a P-matrix, it is clear that E_δ is well defined and nonsingular. Defining the matrix

$$K_\delta \equiv \begin{bmatrix} -(D_{\delta\delta})^{-1} & 0 \\ 0 & 0 \end{bmatrix},$$

we have $u(Cx) = K_\delta Cx$, provided that $Cx \in \mathcal{C}_\delta$. Consequently, the ODE (5) can be written equivalently as

$$\dot{x} = (A + BK_\delta C)x \quad \text{if } E_\delta Cx \geq 0,$$

whose right-hand side is now in an *explicit*, piecewise linear form.

Consider next the case where D is not a P-matrix but the submatrix $D_{\alpha\alpha}$ is nonsingular for some subset α of $\{1, \ldots, m\}$. We can define a system that is equivalent to (1) by "pivoting" on $D_{\alpha\alpha}$ as done for a standard LCP [11], i.e., by solving for the variable u_α in the equation

$$w_\alpha = C_{\alpha\cdot}x + D_{\alpha\alpha}u_\alpha + D_{\alpha\bar{\alpha}}u_{\bar{\alpha}}$$

in terms of the other variables w_α, x, and $u_{\bar{\alpha}}$, where $\bar{\alpha}$ is the complement of α in $\{1, \ldots, m\}$, and then substituting the resulting expression for u_α into the other conditions in (1). The equivalent LCS, which we call the *α-principal transform* of (1), is

$$\begin{aligned} \dot{x} &= \widetilde{A}x + \widetilde{B}\widetilde{u}, \\ 0 \leq \widetilde{u} &\perp \widetilde{C}x + \widetilde{D}\widetilde{u} \geq 0, \\ x(0) &= x^0, \end{aligned} \tag{7}$$

where $\tilde{u} = (w_\alpha, u_{\bar{\alpha}})$ and

$$
\begin{aligned}
\widetilde{A} &\equiv A - B_{\cdot\alpha}(\,D_{\alpha\alpha}\,)^{-1}C_{\alpha\cdot},\\
\widetilde{B} &\equiv \begin{bmatrix} B_{\cdot\alpha}(\,D_{\alpha\alpha}\,)^{-1} & B_{\cdot\bar{\alpha}} - B_{\cdot\alpha}(\,D_{\alpha\alpha}\,)^{-1}D_{\alpha\bar{\alpha}} \end{bmatrix},\\
\widetilde{C} &\equiv \begin{bmatrix} -(\,D_{\alpha\alpha}\,)^{-1}C_{\alpha\cdot} \\ C_{\bar{\alpha}\cdot} - D_{\bar{\alpha}\alpha}(\,D_{\alpha\alpha}\,)^{-1}C_{\alpha\cdot} \end{bmatrix},\\
\widetilde{D} &\equiv \begin{bmatrix} (\,D_{\alpha\alpha}\,)^{-1} & -(\,D_{\alpha\alpha}\,)^{-1}D_{\alpha\bar{\alpha}} \\ D_{\bar{\alpha}\alpha}(\,D_{\alpha\alpha}\,)^{-1} & D_{\bar{\alpha}\bar{\alpha}} - D_{\bar{\alpha}\alpha}(\,D_{\alpha\alpha}\,)^{-1}D_{\alpha\bar{\alpha}} \end{bmatrix}.
\end{aligned} \tag{8}
$$

We call the tuple $(\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D})$ the *α-principal transform* of (A, B, C, D). Of particular importance in the subsequent analysis is the following principal subsystem of this transform:

$$
\begin{aligned}
\dot{x} &= \widetilde{A}x + [\,B_{\cdot\bar{\alpha}} - B_{\cdot\alpha}(\,D_{\alpha\alpha}\,)^{-1}D_{\alpha\bar{\alpha}}\,]u_{\bar{\alpha}},\\
0 \le u_{\bar{\alpha}} &\perp [\,C_{\bar{\alpha}\cdot} - D_{\bar{\alpha}\alpha}(\,D_{\alpha\alpha}\,)^{-1}C_{\alpha\cdot}\,]x + [\,D_{\bar{\alpha}\bar{\alpha}} - D_{\bar{\alpha}\alpha}(\,D_{\alpha\alpha}\,)^{-1}D_{\alpha\bar{\alpha}}\,]u_{\bar{\alpha}} \ge 0.
\end{aligned} \tag{9}
$$

To illustrate the role of the latter subsystem, suppose that at some state $x(t_*) = x^*$, the corresponding algebraic vector $u(t_*) = u^*$ is a strongly regular solution [25] of the LCP (Cx^*, D). This means that the submatrix $D_{\alpha_*\alpha_*}$ is nonsingular and the Schur complement $D_{\beta_*\beta_*} - D_{\beta_*\alpha_*}(D_{\alpha_*\alpha_*})^{-1}D_{\alpha_*\beta_*}$ is a P-matrix, where

$$
\begin{aligned}
\alpha_* &\equiv \{\, i \,:\, u_i^* > 0 = (\,Cx^* + Du^*\,)_i \,\},\\
\beta_* &\equiv \{\, i \,:\, u_i^* = 0 = (\,Cx^* + Du^*\,)_i \,\},\\
\gamma_* &\equiv \{\, i \,:\, u_i^* = 0 < (\,Cx^* + Du^*\,)_i \,\}.
\end{aligned}
$$

If the solution trajectory $(x(t), u(t))$ is continuous near t_*, it then follows that for all t sufficiently near t_*, $(Cx(t) + Du(t))_i > 0$ for all $i \in \gamma_*$ and $u_i(t) > 0$ for all $i \in \alpha_*$. This implies that $u_i(t) = 0$ for all $i \in \gamma_*$ and $(Cx(t) + Du(t))_i = 0$ for all $i \in \alpha_*$. Hence, locally for t near t_*, the trajectory $(x(t), u(t))$ must satisfy the following mixed LCS obtained by fixing some variables at zero:

$$
\begin{aligned}
\dot{x} &= Ax + Bu,\\
0 &= (\,Cx + Du\,)_i \quad \forall\, i \in \alpha_*,\\
0 \le u_i &\perp (\,Cx + Du\,)_i \ge 0 \quad \forall\, i \in \beta_*,\\
0 &= u_i \quad \forall\, i \in \gamma_*.
\end{aligned} \tag{10}
$$

Since $D_{\alpha_*\alpha_*}$ is nonsingular, we can carry out the pivot operation as described above and deduce that (10) is equivalent to, for all t sufficiently near t_*,

$$
\begin{aligned}
\dot{x} &= \overline{A}x + \overline{B}u_\beta,\\
0 \le u_\beta &\perp \overline{C}x + \overline{D}u_\beta \ge 0,
\end{aligned} \tag{11}
$$

where

$$
\begin{aligned}
\overline{A} &\equiv A - B_{\cdot\alpha_*}(\,D_{\alpha_*\alpha_*}\,)^{-1}C_{\alpha_*}, & \overline{B} &\equiv B_{\cdot\beta_*} - B_{\cdot\alpha_*}(\,D_{\alpha_*\alpha_*}\,)^{-1}D_{\alpha_*\beta_*},\\
\overline{C} &\equiv C_{\beta_*\cdot} - D_{\beta_*\alpha_*}(\,D_{\alpha_*\alpha_*}\,)^{-1}C_{\alpha_*}, & \overline{D} &\equiv D_{\beta_*\beta_*} - D_{\beta\alpha_*}(\,D_{\alpha_*\alpha_*}\,)^{-1}D_{\alpha_*\beta_*}.
\end{aligned}
$$

The resulting LCS (11) has the P-property. We summarize this reduction in the following result, which we will later use to deduce an important consequence of the state x^*; see Corollary 15. For further discussion of the above reduction process, see [23, section 5.2].

PROPOSITION 5. *Let $(x(t), u(t))$ be a solution trajectory of the LCS (1) that is continuous near a time t_*. If $u(t_*)$ is a strongly regular solution of the LCP $(Cx(t_*), D)$, then $(x(t), u(t))$ must satisfy the reduced system (11) locally near t_*.* □

Finally, for any nonsingular constant matrix P, we can consider the change of variables $\bar{x} \equiv Px$ and obtain an LCS in the $\bar{x}$-variable that is equivalent to the original (1). This equivalent LCS is

$$\begin{aligned} \dot{\bar{x}} &= PAP^{-1}\bar{x} + PBu, \\ 0 \le u \perp CP^{-1}\bar{x} + Du &\ge 0, \\ \bar{x}(0) &= Px^0. \end{aligned}$$

Particularly useful for us later (see the proof of Lemma 12) is the transformation so that the pair (PAP^{-1}, CP^{-1}) is of a particular form satisfying a favorable observability condition.

3. Zeno states of an LCS. In what follows, we define two types of Zeno states of a general LCS; see Definition 6. Generally speaking, the presence of a Zeno state in a hybrid system could have an adverse effect on the numerical simulation of a solution trajectory to the system. This issue, which is closely tied to mode switches, has been dealt with extensively in the literature; see, e.g., [4, 9, 19, 33, 38]. In particular, for hybrid systems described by ODEs with piecewise real analytic right-hand sides, the results by Brunovsky [4] and Sussmann [33] show that there is a finite number of mode switches (defined in the sense in the cited references). While the latter results are in principle applicable to the LCS with the P-property, their treatment does not reveal the important complementarity nature of the LCS. (See [5, 9] for some special Zeno results for the LCS where the D matrix is positive definite or satisfies a passifiability assumption.) Because of the fundamental role of the LCS in hybrid system theory, it is useful to have a simplified approach that exploits the characteristics of the LCS. Most importantly, the Zeno concepts defined below are of a refined, algebraic nature that takes into account possible degeneracy of the solutions to the complementarity conditions. Analytically, our proofs of the main Zeno results, Theorems 9 and 21, are based on a local expansion of a solution trajectory to the LCS (Lemma 14) which is a new result by itself and enables us to study systems failing the P-property. Furthermore, this expansion reveals a local property of a solution trajectory of (1) in terms of an "observability degree" of a given state relative to the pair (C, A).

As is well known, an LCS is a special linear hybrid system with finitely many "modes," where a mode is a linear differential algebraic equation (LDAE) defined by a pair of disjoint index sets $(\alpha, \bar{\alpha})$ whose union is the index set $\{1, \ldots, m\}$; specifically, such an LDAE is as follows:

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ 0 &= (\, Cx + Du \,)_\alpha, \\ 0 &= u_{\bar{\alpha}}. \end{aligned} \tag{12}$$

Every solution trajectory of the LCS (1) must satisfy, at every time instant, the above LDAE for a certain pair $(\alpha, \bar{\alpha})$ that is dependent on the time. Conversely, if $(x(t), u(t))$

is a solution trajectory satisfying the latter LDAE, then $(x(t), u(t))$ satisfies the LCS (1) if $(Cx(t) + Du(t))_{\bar{\alpha}} \geq 0$ and $u(t)_{\alpha} \geq 0$. In general, it is possible for a solution trajectory $(x(t), u(t))$ of the LCS (1) to satisfy at any given time t the LDAE (12) for multiple pairs of index sets, due to degeneracy of the complementarity conditions. For every such trajectory, at each time t_* that is neither the initial nor the terminal time, there exist (i) an infinite sequence of times $\{t_k^-\}$ converging to t_* from the left and a pair $(\alpha_-, \bar{\alpha}_-)$ of index sets partitioning $\{1, \ldots, m\}$ such that $(x(t_k^-), u(t_k^-))$ satisfies the LDAE (12) corresponding to $(\alpha_-, \bar{\alpha}_-)$ for all k, and (ii) an infinite sequence of times $\{t_k^+\}$ converging to t_* from the right and a pair $(\alpha_+, \bar{\alpha}_+)$ of partitioning index sets such that $(x(t_k^+), u(t_k^+))$ satisfies the LDAE (12) corresponding to $(\alpha_+, \bar{\alpha}_+)$ for all k. An intuitive definition of a *mode switch* at time t_* is that the two pairs of index sets $(\alpha_-, \bar{\alpha}_-)$ and $(\alpha_+, \bar{\alpha}_+)$ are not equal. Roughly speaking, a *Zeno state* of an LCS is a state near which there are infinitely many modes switches.

We now formalize the above informal discussion by defining "weak" and "strong" Zeno states of a solution trajectory $(x(t), u(t))$ of the LCS (1). Both are local properties of a state. The strong Zeno concept is more refined than the weak Zeno concept and is defined in terms of the index sets

$$\begin{aligned}
\alpha(t) &\equiv \{\, i \,:\, u_i(t) \,>\, 0 \,=\, (\, Cx(t) + Du(t)\,)_i \,\}, \\
\beta(t) &\equiv \{\, i \,:\, u_i(t) \,=\, 0 \,=\, (\, Cx(t) + Du(t)\,)_i \,\}, \\
\gamma(t) &\equiv \{\, i \,:\, u_i(t) \,=\, 0 \,<\, (\, Cx(t) + Du(t)\,)_i \,\};
\end{aligned}$$

in contrast, the weak Zeno concept relaxes the strong concept by restricting to the two combined sets $\alpha(t) \cup \beta(t)$ and $\gamma(t) \cup \beta(t)$. The two concepts coincide if $u(t)$ is a nondegenerate solution of the LCP $(Cx(t), D)$ for all t near t_*, in which case, the degenerate set $\beta(t)$ is empty for all such t.

DEFINITION 6. *Let $(x(t), u(t))$ be a solution trajectory of* (1) *and let $x(t_*) = x^*$. We say that x^* is*

(a) strongly left non-Zeno relative to $(x(t), u(t))$ *if a scalar $\varepsilon_- > 0$ and a triple of index sets $(\alpha_-, \beta_-, \gamma_-)$ exist such that $(\alpha(t), \beta(t), \gamma(t)) = (\alpha_-, \beta_-, \gamma_-)$ for every $t \in [t_* - \varepsilon_-, t_*)$;*

(b) strongly right non-Zeno relative to $(x(t), u(t))$ *if a scalar $\varepsilon_+ > 0$ and a triple $(\alpha_+, \beta_+, \gamma_+)$ of index sets exist such that $(\alpha(t), \beta(t), \gamma(t)) = (\alpha_+, \beta_+, \gamma_+)$ for every $t \in (t_*, t_* + \varepsilon_+]$;*

(c) weakly left non-Zeno relative to $(x(t), u(t))$ *if a scalar $\varepsilon_- > 0$ and a pair of index sets α_- and $\bar{\alpha}_-$ partitioning $\{1, \ldots, m\}$ exist such that $(x(t), u(t))$ satisfies the LDAE* (12) *corresponding to $(\alpha_-, \bar{\alpha}_-)$ for all $t \in [\, t_* - \varepsilon_-, t_*\,)$;*

(d) weakly right non-Zeno relative to $(x(t), u(t))$ *if a scalar $\varepsilon_+ > 0$ and a pair of index sets α_+ and $\bar{\alpha}_+$ partitioning $\{1, \ldots, m\}$ exist such that $(x(t), u(t))$ satisfies the LDAE* (12) *corresponding to $(\alpha_+, \bar{\alpha}_+)$ for all $t \in (t_*, t_* + \varepsilon_+]$.*

The state $x^* \equiv x(t_*)$ is said to be left (right) Zeno of the first (second) kind relative to the trajectory $(x(t), u(t))$ *if it is* not *strongly (weakly) left (right) non-Zeno relative to the same trajectory. When x^* is strongly (weakly) left* and *right non-Zero, then we say that x^* is strongly (weakly) non-Zeno; when x^* is either left* or *right Zeno of the first (second) kind, then we say that x^* is Zeno of the first (second) kind. When the trajectory is clear from the context, we will omit the phrase "relative to the trajectory."* □

Definition 6 is applicable to both the initial and the terminal states of an LCS. Specifically, if x^* is the initial state x^0 of the LCS (1), then we are interested only in the right (non-)Zeno property of x^*; similarly, if x^* is the terminal state $x(T)$ of the LCS

(1) at a prescribed terminal time $T > 0$, then we are interested only in the left (non-) Zeno property of x^*. It is clear that the strongly (left or right) non-Zeno properties must imply the respective weakly (left or right) non-Zeno properties. Nevertheless, the converse is clearly not always true. In essence, the left Zeno properties of a state refer to its reachability from the left, and the right Zeno properties refer to the continuation from the state. Obviously, these properties have important numerical implications when the LCS is solved by a time-stepping method. For instance, the numerical methods discussed in [30] for solving ODEs with discontinuous right-hand sides are based on the presumed absence of Zeno states. Further discussion of such numerical matters is beyond the scope of this paper.

An important remark should be made for Definition 6: namely, this definition pertains to the two trajectories $x(t)$ and $u(t)$ jointly. This is distinct from the treatment in [4, 33] which is applicable to the implicit formulation (5) of the LCS in which the algebraic variable u is eliminated and treated only implicitly. It is not a straightforward task to directly apply the results in these cited references to analyze the Zeno properties of the LCS as described in Definition 6, where the u-trajectory plays a prominent role. In many realistic applications of the LCS (such as in contact mechanics), the role of the algebraic variable u is as important as the differential variable x; thus, an explicit treatment of the former, as emphasized herein, is warranted.

Zeno states are closely tied to mode switches, which we formally define next. For simplicity, we present the definition below only for a time that is neither the initial nor the terminal time of a trajectory. The triple of index sets $(\alpha(t), \beta(t), \gamma(t))$ in this definition are the fundamental index sets associated with the pair $(x(t), u(t))$.

DEFINITION 7. *Let $(x(t), u(t))$ be a solution trajectory of* (1) *and let t_* be an intermediate time of this trajectory. We say that t_* is a*

(a) switch time of the first kind *relative to $(x(t), u(t))$ if there exist two triples of index sets, $(\alpha_-, \beta_-, \gamma_-)$ and $(\alpha_+, \beta_+, \gamma_+)$, and two infinite sequences of times, $\{t_k^-\}$ and $\{t_k^+\}$, the former converging to t_* from the left and the latter converging to t_* from the right, such that, for all k,*

$$(\alpha(t_k^-), \beta(t_k^-), \gamma(t_k^-)) = (\alpha_-, \beta_-, \gamma_-) \neq (\alpha_+, \beta_+, \gamma_+) = (\alpha(t_k^+), \beta(t_k^+), \gamma(t_k^+));$$

(b) switch time of the second kind *relative to $(x(t), u(t))$ if there exist two infinite sequences of times, $\{t_k^-\}$ and $\{t_k^+\}$, the former converging to t_* from the left and the latter converging to t_* from the right, such that for* no *pair of index sets $(\alpha, \bar{\alpha}_-)$ partitioning $\{1, \ldots, m\}$, $(x(t_k^-), u(t_k^-))$ and $(x(t_k^+), u(t_k^+))$ both satisfy the LDAE* (12) *for all k.* □

The following result shows that the absence of Zeno states in a finite time of interval provides a sufficient condition for the finite number of switch times in the interval.

PROPOSITION 8. *Let $(x(t), u(t))$ be a solution trajectory of the LCS* (1) *defined on an open interval containing $[0, T]$. If the trajectory has no Zeno states of the first (second) kind, then there is a finite number of switch times of the first (second) kind relative to $(x(t), u(t))$ in $[0, T]$.*

Proof. We prove the result for the "second" kind only. If $(x(t), u(t))$ contains no Zeno states of the second kind, then for every $t \in [0, T]$, there exist a right neighborhood $\mathcal{N}_t^+ \equiv (t, t + \varepsilon_t)$ and a left neighborhood $\mathcal{N}_t^- \equiv (t - \varepsilon_t, t)$ of t, for some scalar $\varepsilon_t > 0$, and two pairs of index sets, $(\alpha_t^+, \bar{\alpha}_t^+)$ and $(\alpha_t^-, \bar{\alpha}_t^-)$, both partitioning $\{1, \ldots, m\}$, such that for all $t' \in \mathcal{N}_t^+$, the pair $(x(t'), u(t'))$ satisfies the LDAE (12) corresponding to $(\alpha_t^+, \bar{\alpha}_t^+)$, and that for all $t' \in \mathcal{N}_t^-$, the pair $(x(t'), u(t'))$ satisfies

the LDAE (12) corresponding to $(\alpha_t^-, \bar{\alpha}_t^-)$. The family

$$\{ (t - \varepsilon_t, t + \varepsilon_t) : t \in [0, T] \}$$

constitutes an open covering of the compact interval $[0, T]$. Hence there exists a finite sequence $\{t_0, t_1, \ldots, t_\ell\} \subset [0, T]$ such that

$$[0, T] \subset \bigcup_{i=0}^{\ell} [t_i - \varepsilon_{t_i}, t_i + \varepsilon_{t_i}].$$

By refining the partition on the right-hand side, we may assume without loss of generality that there exist a finite sequence of times $0 = t_0' < t_1' < \cdots < t_k' < t_{k+1}' = T$ and a corresponding sequence of index sets α_i of $\{1, \ldots, m\}$ with respective complements $\bar{\alpha}_i$ such that $(x(t), u(t))$ satisfies (12) corresponding to $(\alpha_i, \bar{\alpha}_i)$ for all $t \in (t_i', t_{i+1}')$, $i = 0, 1, \ldots, k$. Consequently, the only possible switch times of the second type in the interval $[0, T]$ are the times t_i' for $i = 0, 1, \ldots, k+1$. □

3.1. The P-matrix case. The following is the main Zeno result for an LCS with the P-property.

THEOREM 9. *If D is a P-matrix, then all states of the LCS* (1) *must be strongly non-Zeno.*

The proof of the above theorem is accomplished via several lemmas. The first such lemma, which is a global and time-invariant version of Proposition 5.3 in [26], gives a decay rate for a Lipschitz continuous system.

LEMMA 10. *Let $\dot{x} = f(x)$, $x(0) = x^0$ be a dynamical system on $\Re^n$, where $f(x)$ is globally Lipschitz continuous in x with Lipschitz constant $L \geq 0$. If $x = 0$ is an equilibrium of the system, i.e., $f(0) = 0$, then*

$$\| x^0 \|_2 \, e^{-Lt} \leq \| x(t) \|_2 \leq \|x^0\|_2 \, e^{Lt} \quad \forall \, t \geq 0.$$

The main proof of Theorem 9 is divided into two parts, depending on whether a state x^* in question is observable or unobservable with respect to the pair (C, A). The concept of observability is well known for a linear time-invariant system $\dot{x} = Ax + Bu$ and $y = Cx + Du$ and is briefly reviewed here. A state $x \in \Re^n$ is *unobservable* with respect to (C, A) if $Ce^{At}x \equiv 0$ for all t; otherwise it is called *observable* with respect to (C, A). Without confusion, we usually simply call a state observable/unobservable. The set of all unobservable states is a subspace of $\Re^n$, called the *unobservable subspace* of the pair (C, A). An equivalent condition for a state x being observable is that $CA^k x \neq 0$ for some $0 \leq k \leq n-1$. The linear system is *observable* if $x = 0$ is the only unobservable state. In such the case, we call (C, A) an observable pair.

The next lemma asserts that the LCS (1) with the P-property is trivial if the initial state x^0 is unobservable.

LEMMA 11. *Let D be a P-matrix. If x^0 is unobservable, then the unique solution trajectory of* (1) *is $(x(t), u(t)) = (e^{At}x^0, 0)$ for all $t \geq 0$. In this case, we have $\beta(t) = \{1, \ldots, m\}$ for all $t \geq 0$; hence, all states $x(t)$ are strongly non-Zeno.*

Proof. This follows easily from the uniqueness of the solution trajectory and the fact that $Ce^{At}x^0 = 0$ for all $t \geq 0$. □

Lemma 11 suggests that we may assume without loss of generality that x^0 is observable. The next lemma asserts that in this case, all states on the trajectory $x(t)$ are observable.

LEMMA 12. *Let D be a P-matrix. If $x^0 = x(0)$ is observable, then so is $x(t)$ for all $t \geq 0$.*

Proof. Lemma 10 is applicable to the implicit form (5) of the LCS. It follows that $\|x(t)\| \geq \|x^0\| e^{-Lt}$ for some constant $L \geq 0$. Since x^0 is observable, it is not zero. Hence $x(t) \neq 0$ for all $t \geq 0$. Consequently, we may assume without loss of generality that (C, A) is an unobservable pair. Let $\overline{O}(C, A)$ denote the unobservable subspace whose dimension is n_2 with $1 \leq n_2 \leq n$. According to linear system theory [10, p. 203], there exists a nonsingular matrix $P \in \Re^{n \times n}$ such that the change of variables $\bar{x} = Px$ transforms the original linear system (1) into the observable canonical form

$$\dot{\bar{x}} = \begin{pmatrix} \dot{\bar{x}}_o \\ \dot{\bar{x}}_{uo} \end{pmatrix} = \begin{bmatrix} \bar{A}_o & 0 \\ \bar{A}_{21} & \bar{A}_{uo} \end{bmatrix} \begin{pmatrix} \bar{x}_o \\ \bar{x}_{uo} \end{pmatrix} + \begin{bmatrix} \bar{B}_o \\ \bar{B}_{uo} \end{bmatrix} u,$$

$$Cx = \begin{bmatrix} \bar{C}_o & 0 \end{bmatrix} \begin{pmatrix} \bar{x}_o \\ \bar{x}_{uo} \end{pmatrix} = \bar{C}_o \bar{x}_o,$$

and $(\bar{C}_o, \bar{A}_o)$ is an observable pair, $\bar{x} \equiv \left(\begin{smallmatrix} \bar{x}_o \\ \bar{x}_{uo} \end{smallmatrix}\right)$ is the transformed state, $\bar{x}_o \in \Re^{n-n_2}$ and $\bar{x}_{uo} \in \Re^{n_2}$ correspond to the observable part and unobservable part of $\bar{x}$, respectively. Hence, the LCS (1) can be decomposed into the observable dynamics

$$\begin{aligned} \dot{\bar{x}}_o &= \bar{A}_o \bar{x}_o + \bar{B}_o u, \\ 0 \leq u &\perp \bar{C}_o \bar{x}_o + Du \geq 0, \end{aligned} \tag{13}$$

and the unobservable dynamics

$$\dot{\bar{x}}_{uo} = \bar{A}_{21} \bar{x}_o + \bar{A}_{uo} \bar{x}_{uo} + \bar{B}_{uo} u. \tag{14}$$

Moreover, any unobservable state $\widehat{x} \in \overline{O}(C, A)$ is transformed to the following form under the above transformation:

$$P\widehat{x} = \begin{pmatrix} 0 \\ \bar{x}_{uo} \end{pmatrix}.$$

This means that the observable part of an original unobservable state must be zero and the observable part of an original observable state must not be zero. Since (13) remains an LCS with the P-property, and since $\bar{x}_o(0) \neq 0$ because $x(0)$ is observable, it follows that $\bar{x}_o(t) \neq 0$ for all $t \geq 0$, which means that $x(t)$ must be observable. ∎

A noteworthy remark is that Lemma 12 can be proved using the reverse-time argument.[1] Letting $(x^r(t), u^r(t)) \equiv (x(-t), u(-t))$ for all $t \geq 0$, one easily sees that the pair $(x^r(t), u^r(t))$ satisfies a reverse-time LCS for all $t \geq 0$:

$$\begin{aligned} \dot{x}^r(t) &= -Ax^r(t) - Bu^r(t), \\ 0 \leq u^r(t) &\perp Cx^r(t) + Du^r(t) \geq 0. \end{aligned}$$

Hence, the reverse-time LCS $(-A, -B, C, D)$ preserves the P-property and its solution pair is unique as well. Suppose $x(0)$ is observable but $x(t)$ is not at some $t \geq 0$. Then using the reverse-time LCS and Lemma 11, $x(0) = e^{-At} x(t)$, which is unobservable as well. However, the decomposition of the LCS into the observable dynamics (13)

[1] We thank an anonymous reviewer for bringing this remark to our attention.

and the unobservable dynamics (14) in Lemma 12 has its own interest. Therefore, we present it for this purpose.

Combining the above two lemmas, we have therefore proved Theorem 9 for the unobservable states. We formally state this conclusion in the following corollary, which requires no proof.

COROLLARY 13. *Any unobservable state of an LCS with the P-property is strongly non-Zeno.* □

We next turn our attention to the observable states. The cornerstone of the treatment of these states is an expansion of the solution trajectory near any given time t_*. We will make use of the (unique) solution $u(\pm CA^k x)$ of the LCPs $(\pm CA^k x, D)$. In general, except for the fact that both are nonnegative, the two vectors $u^{k+}(x) \equiv u(CA^k x^*)$ and $u^{k-}(x) \equiv u(-CA^k x^*)$ have very little to do with each other. Since D is a P-matrix, we can speak of the directional derivative of the solution function $u'(q; dq)$ of the LCP (q, D) at the vectors $q \equiv \pm CA^k x$ along the directions $dq \equiv \pm(CA^{k+1}x + CBu^{k\pm}(x))$. Specifically, we will use the following directional derivatives:

$$u'(CA^k x; CA^{k+1}x + CBu^{k+}(x)) = \lim_{\tau \downarrow 0} \frac{u(CA^k x + \tau(CA^{k+1}x + CBu^{k+}(x))) - u(CA^k x)}{\tau},$$

$$u'(CA^k x; -(CA^{k+1}x + CBu^{k+}(x))) = \lim_{\tau \downarrow 0} \frac{u(CA^k x - \tau(CA^{k+1}x + CBu^{k+}(x))) - u(CA^k x)}{\tau},$$

$$u'(-CA^k x; CA^{k+1}x - CBu^{k-}(x)) = \lim_{\tau \downarrow 0} \frac{u(-CA^k x + \tau(CA^{k+1}x - CBu^{k+}(x))) - u(-CA^k x)}{\tau}.$$

The reason for distinguishing these derivatives will be evident from the next result. In this result, we use the standard notation $o(f(t))$ to mean a function such that $\lim_{0 \neq t \to 0} \frac{o(f(t))}{t} = 0$; the notation $O(f(t))$ also has the standard meaning.

LEMMA 14. *Let D be a P-matrix. Let $x^* = x(t_*)$ be an arbitrary state of the solution trajectory $(x(t), u(t))$ such that $CA^j x^* = 0$ for all $j = 0, 1, \ldots, k-1$ for some integer $k \geq 0$. The following two statements hold:*

(a) *For all $t > t_*$,*

$$x(t) = \sum_{j=0}^{k+2} \frac{(t - t_*)^j}{j!} A^j x^* + \frac{(t - t_*)^{k+1}}{(k+1)!} Bu(CA^k x^*) + \frac{(t - t_*)^{k+2}}{(k+2)!} Bu'(CA^k x^*; CA^{k+1}x^* + CBu(CA^k x^*)) + o(|t - t_*|^{k+2}),$$

$$u(t) = \frac{(t - t_*)^k}{k!} u(CA^k x^*) + \frac{(t - t_*)^{k+1}}{(k+1)!} u'(CA^k x^*; CA^{k+1}x^* + CBu(CA^k x^*)) + o(|t - t_*|^{k+1}).$$

(b) *For all* $t < t_*$,

$$x(t) = \sum_{j=0}^{k+2} \frac{(t-t_*)^j}{j!} A^j x^* + \frac{(t-t_*)^{k+1}}{(k+1)!} Bu(CA^k x^*) + \frac{(t-t_*)^{k+2}}{(k+2)!} Bu'(CA^k x^*; -CA^{k+1}x^* - CBu(CA^k x^*)) + o(|t-t_*|^{k+2}),$$

$$u(t) = \frac{(t-t_*)^k}{k!} u(CA^k x^*) + \frac{|t-t_*|^{k+1}}{(k+1)!} u'(CA^k x^*; -CA^{k+1}x^* - CBu(CA^k x^*)) + o(|t-t_*|^{k+1})$$

if k *is even; and if* k *is odd,*

$$x(t) = \sum_{j=0}^{k+2} \frac{(t-t_*)^j}{j!} A^j x^* - \frac{(t-t_*)^{k+1}}{(k+1)!} Bu(-CA^k x^*) + \frac{(t-t_*)^{k+2}}{(k+2)!} Bu'(-CA^k x^*; CA^{k+1}x^* - CBu(-CA^k x^*)) + o(|t-t_*|^{k+2}),$$

$$u(t) = \frac{|t-t_*|^k}{k!} u(-CA^k x^*) + \frac{(t-t_*)^{k+1}}{(k+1)!} u'(-CA^k x^*; CA^{k+1}x^* - CBu(-CA^k x^*)) + o(|t-t_*|^{k+1}).$$

Proof. Define

$$z(t) = x(t) - \sum_{j=0}^{k} (t-t_*)^j \frac{A^j x^*}{j!}. \tag{15}$$

Hence, $z(t_*) = 0$ and $z(t)$ satisfies

$$\dot{z}(t) = Az(t) + \frac{(t-t_*)^k}{k!} A^{k+1}x^* + Bu(t), \tag{16}$$

where $u(t)$ satisfies

$$0 \le u(t) \perp Cz(t) + \frac{(t-t_*)^k}{k!} CA^k x^* + Du(t) \ge 0. \tag{17}$$

Since D is a P-matrix, it follows that there exists a constant $\eta > 0$ such that for all t,

$$\| u(t) \| \le \eta [\, \| z(t) \| + |t-t_*|^k \,].$$

By (16), we deduce the existence of positive constants λ and μ such that for all $t < t_*$,

$$\| z(t) \| \le \lambda (t_* - t)^{k+1} + \mu \int_t^{t_*} \| z(s) \| \, ds.$$

Thus by Gronwall–Bellman inequality, we obtain, for some constant $\mu' > 0$,

$$\begin{aligned} \| z(t) \| &\le \lambda (t_* - t)^{k+1} + \lambda \mu \int_t^{t_*} (t_* - \tau)^{k+1} e^{\mu(t_* - \tau)} d\tau \\ &\le \lambda (t_* - t)^{k+1} + \mu' (t_* - t)^{k+2} \end{aligned}$$

for all $t < t_*$ sufficiently near t_*. A similar bound can be derived for $\|z(t)\|$ for all $t > t_*$ sufficiently near t_*. Consequently, we deduce the existence of a scalar $\lambda' > 0$ such that for all t near t_*,

$$\| z(t) \| \leq \lambda' | t - t_* |^{k+1}. \tag{18}$$

Suppose that k is odd. For all $t < t_*$, (17) implies

$$0 \leq v(t) \perp k! C \frac{z(t)}{| t - t_* |^k} - CA^k x^* + Dv(t) \geq 0, \tag{19}$$

where $v(t) \equiv k!u(t)/|t - t_*|^k$. Since D is a P-matrix, it follows from the Lipschitz continuity of the solutions to the LCP (q, D) that, for some constant $L > 0$, we have

$$\| v(t) - u(-CA^k x^*) \| \leq L \frac{\| z(t) \|}{| t - t_* |^k}$$

or equivalently

$$\left\| u(t) - \frac{| t - t_* |^k}{k!} u(-CA^k x^*) \right\| \leq L k! \| z(t) \| \quad \forall t < t_* \text{ sufficiently near } t_*. \tag{20}$$

For k odd and for $t < t_*$, we can write (16) as

$$\begin{aligned} \dot{z}(t) &= Az(t) + \frac{(t - t_*)^k}{k!}[A^{k+1}x^* - Bu(-CA^k x^*)] + B\left[u(t) - \frac{| t - t_* |^k}{k!} u(-CA^k x^*)\right] \\ &= \frac{(t - t_*)^k}{k!}[A^{k+1}x^* - Bu(-CA^k x^*)] + B\left[u(t) - \frac{| t - t_* |^k}{k!} u(-CA^k x^*)\right] \\ &\quad + O(|t - t_*|^{k+1}), \end{aligned}$$

where the last inequality is due to (18). Integrating the above and using (20) and (18), we deduce

$$z(t) = z(t_*) + \int_{t_*}^{t} \dot{z}(s)\, ds = \frac{(t - t_*)^{k+1}}{(k+1)!} [A^{k+1}x^* - Bu(-CA^k x^*)] + O(|t - t_*|^{k+2})$$

for all $t < t_*$ sufficiently near t_*. Substituting into (19), we obtain

$$0 \leq v(t) \perp \frac{t_* - t}{k+1} C[A^{k+1}x^* - Bu(-CA^k x^*)] + O(|t - t_*|^2) - CA^k x^* + Dv(t) \geq 0.$$

Hence, we deduce

$$v(t) = u(-CA^k x^*) + \frac{t_* - t}{k+1} u'(-CA^k x^*; CA^{k+1}x^* - CBu(-CA^k x^*)) + o(|t - t_*|),$$

which yields

$$\begin{aligned} u(t) = \frac{| t - t_* |^k}{k!} u(-CA^k x^*) + \frac{(t_* - t)^{k+1}}{(k+1)!} u'(-CA^k x^*; CA^{k+1}x^* - CBu(-CA^k x^*)) \\ + o(|t - t_*|^{k+1}). \end{aligned}$$

Substituting this into (16), we deduce

$$\dot{z}(t) = Az(t) + \frac{(t-t_*)^k}{k!}\,[\,A^{k+1}x^* - Bu(-CA^kx^*)\,] + \frac{(t-t_*)^{k+1}}{(k+1)!}\,Bu'(-CA^kx^*; CA^{k+1}x^* - CBu(-CA^kx^*)) + o(|t-t_*|^{k+1}).$$

Integrating the above equation and recalling $x(t) = z(t) + \sum_{j=0}^{k}(t-t_*)^j A^j x^*/j!$, we obtain the desired expansion for $x(t)$ when k is odd and $t < t_*$.

Next, assume that k is even and consider $t < t_*$ sufficiently near t_*. In this case, instead of (19), we have

$$0 \le v(t) \perp k!\,C\,\frac{z(t)}{(t-t_*)^k} + CA^kx^* + Dv(t) \ge 0;$$

moreover, (20) is replaced by

$$\left\| u(t) - \frac{(t-t_*)^k}{k!}\,u(CA^kx^*) \right\| \le L\,k!\,\|\,z(t)\,\| \quad \forall\, t < t_* \text{ sufficiently near } t_*.$$

Proceeding as above, we obtain from (16)

$$\dot{z}(t) = \frac{(t-t_*)^k}{k!}\,[\,A^{k+1}x^* + Bu(CA^kx^*)\,] + B\left[u(t) - \frac{(t-t_*)^k}{k!}\,u(CA^kx^*)\right] + O(|t-t_*|^{k+1}).$$

At this point, we can repeat the above proof and obtain the desired expansion for $(x(t), u(t))$ in this case where k is even. Finally, the proof for statement (a) is similar and therefore omitted. □

On the basis of Lemma 14, we can complete the proof of Theorem 9 by defining the *observability degree* of an observable state x with respect to the pair (C, A), which is defined as the first nonnegative integer k such that $CA^kx \neq 0$.

Proof of Theorem 9. We use induction on m, the dimension of the input variable u. The case $m = 0$ is trivial. Inductively, assume that the theorem is valid for an integer $m \ge 0$. Consider the LCS (1) where the algebraic variable u is of dimension $m+1 \ge 1$. Let $x^* = x(t_*)$ be an arbitrary state. We first prove that x^* is strongly right non-Zeno. By the above arguments, we may assume without loss of generality that x^* is observable. Let $k \ge 0$ be the observability degree of x^*; thus $Cx^* = \cdots = CA^{k-1}x^* = 0$ and $CA^kx^* \neq 0$. The expansion in part (a) of Lemma 14 holds for the trajectory $(x(t), u(t))$ in a small interval $[t_*, t_* + \varepsilon_+]$ for some $\varepsilon_+ > 0$. Since $CA^kx^* \neq 0$, there must exist an index i such that either $u_i(CA^kx^*) > 0$ or $[CA^kx^* + Du(CA^kx^*)]_i > 0$. By part (a) of Lemma 14, this implies that if $u_i(CA^kx^*) > 0$ for some index i, then $u_i(t) > 0$ for all $t > t_*$ sufficiently near t_*, which implies, by complementarity, that $[Cx(t) + Du(t)]_i = 0$ for all such t. Hence, we can solve for $u_i(t)$ from this equation, obtaining

$$u_i(t) = d_{ii}^{-1}\left[C_{i\cdot}x(t) + \sum_{j\neq i} d_{ij}\,u_j(t)\right],$$

which we can then substitute into the remaining conditions in (1). This substitution

results in an LCS:

$$\begin{aligned} \dot{x}(t) &= \widehat{A}x(t) + \widehat{B}\widehat{u}(t), \\ 0 \leq \widehat{u}(t) &\perp \widehat{C}x(t) + \widehat{D}\widehat{u}(t), \\ x(t_*) &= x^*, \end{aligned} \tag{21}$$

where the matrix $\widehat{D}$ is the Schur complement of the diagonal entry d_{ii} in D and the algebraic variable $\widehat{u}$ is of dimension m, which is one less than that of the original variable u. The original trajectory $(x(t), u(t))$ with the variable u_i removed must satisfy (21) in a small interval $(t_*, t_* + \varepsilon'_+]$ for some $\varepsilon'_+ > 0$. Since $\widehat{D}$ remains a P-matrix, by the induction hypothesis, there exist an index set $(\alpha'_+, \beta'_+, \gamma'_+)$ such that $(\widehat{\alpha}(t), \widehat{\beta}(t), \widehat{\gamma}(t)) = (\alpha'_+, \beta'_+, \gamma'_+)$ for all $t > t_*$ sufficiently near t_*, where $(\widehat{\alpha}(t), \widehat{\beta}(t), \widehat{\gamma}(t))$ are the three fundamental index sets associated with the solution trajectory $(x(t), \widehat{u}(t))$. Clearly, we have $(\alpha(t), \beta(t), \gamma(t)) = (\alpha'_+ \cup \{i\}, \beta'_+, \gamma'_+)$ for all $t > t_*$ sufficiently near t_*.

Next, consider the case where $[CA^k x^* + Du(CA^k x^*)]_i > 0$ for some index i. We then have

$$[\, Cx(t) + Du(t)\,]_i = \frac{(\,t - t_*\,)^k}{k\,!}\,[\, CA^k x^* + Du(CA^k x^*)\,]_i + o(|t - t_*|^{k+1}),$$

which implies that $[\, Cx(t) + Du(t)\,]_i > 0$, and thus $u_i(t) = 0$ by complementarity for all $t > t_*$ sufficiently near t_*. Setting this variable equal to zero and dropping the ith column of B and D and the ith row of C and D, we obtain a principal linear complementarity subsystem of (1) that is satisfied by the trajectory $(x(t), u(t))$ for all $t > t_*$ sufficiently near t_*. The induction hypothesis can be applied to the resulting subsystem whose algebraic variable is of one less dimension than that of the original u. Finally, we can apply part (b) of Lemma 14 to deal with $t < t_*$ and employ similar reductions to complete the inductive proof. □

4. Extended Zeno results. Theorem 9 can be easily extended to the mixed LCS

$$\begin{aligned} \dot{x} &= Ax + B^1 u^1 + B^2 u^2, \\ 0 &= C^1 x + D^{11} u^1 + D^{12} u^2, \\ 0 \leq u^2 &\perp C^2 x + D^{21} u^1 + D^{22} u^2 \geq 0, \\ x(0) &= x^0, \end{aligned}$$

provided that the matrix D^{11} is nonsingular and the Schur complement

$$D^{22} - D^{21}(\,D^{11}\,)^{-1} D^{12}$$

is a P-matrix. Instead of presenting the details of this easy extension, we consider a local version of the extension that pertains to an LCS with a "strongly regular" state but which is not of the P-type. Specifically, we call x^* a *strongly regular state* of the LCS (1) if the LCP (Cx^*, D) has a strongly regular solution. The following corollary of Theorem 9 shows that any such state must be strongly non-Zeno. For simplicity, we treat the case where x^* is neither an initial nor a terminal state of a solution trajectory.

COROLLARY 15. *Any strongly regular state of the LCS* (1) *is strongly non-Zeno relative to a continuous solution trajectory of the system. In fact, if* $(x(t), u(t))$ *is such a trajectory defined on an open interval containing* t_* *and if the LCP* $(Cx(t_*), D)$ *has a strongly regular solution, then* $(x(t), u(t))$ *is the unique continuous solution trajectory passing through* $x^* \equiv x(t_*)$ *for all* t *sufficiently near* t_**, and* x^* *is strongly non-Zeno relative to this trajectory.*

Proof. We first establish the uniqueness of $(x(t), u(t))$. Suppose that $(\widetilde{x}(t), \widetilde{u}(t))$ is another continuous solution trajectory of (1) passing through x^* and defined on the same interval as $(x(t), u(t))$. By the strong regularity of u^*, a neighborhood $\mathcal{V}$ of Cx^*, a neighborhood $\mathcal{U}$ of u^*, and a Lipschitz continuous function $\widehat{u} : \mathcal{V} \to \mathcal{U}$ exist such that for every $q \in \mathcal{V}$, $\widehat{u}(q)$ is the unique solution of the LCP (q, D) in $\mathcal{U}$. Since $(x(t), u(t))$ and $(\widetilde{x}(t), \widetilde{u}(t))$ are both continuous near t_*, it follows that for all t sufficiently near t_*, $(Cx(t), u(t))$ and $(C\widetilde{x}(t), \widetilde{u}(t))$ both belongs to $\mathcal{V} \times \mathcal{U}$. Hence we have $u(t) = \widehat{u}(Cx(t))$ and $\widetilde{u}(t) = \widehat{u}(C\widetilde{x}(t))$. Moreover, a constant $L > 0$ exists such that for all t sufficiently near t_*, we have

$$\|u(t) - \widehat{u}(t)\| \le L \, \|x(t) - \widehat{x}(t)\|. \tag{22}$$

Since

$$\frac{d\,(x(t) - \widehat{x}(t))}{dt} = A(x(t) - \widehat{x}(t)) + B(u(t) - \widehat{u}(t)),$$

(22) implies that the right-hand side is a Lipschitz function of $x(t) - \widehat{x}(t)$. Since $x(t_*) = \widehat{x}(t_*) = x^*$, it follows that the two trajectories $x(t)$ and $\widehat{x}(t)$, and thus the two trajectories, $u(t)$ and $\widehat{u}(t)$, must coincide in a sufficiently small open interval containing t_*. The uniqueness of $(x(t), u(t))$ therefore follows.

By Proposition 5, it follows that for all t sufficiently near t_*, the trajectory $(x(t), u(t))$ must satisfy the reduced system (11). Since $\overline{D}$, being the Schur complement of $D_{\alpha_*\alpha_*}$ in a principal submatrix of D, remains a P-matrix, it follows that $x(t_*)$ is a strongly non-Zeno state of (11) relative to the trajectory $(x(t), u_{\beta_*}(t))$. Therefore, a scalar $\varepsilon > 0$ and two triples of index sets $(\alpha_{0+}, \beta_{0+}, \gamma_{0+})$ and $(\alpha_{0-}, \beta_{0-}, \gamma_{0-})$ exist such that

$$\left.\begin{aligned} \alpha_0(t) &\equiv \{i \in \beta_* : u_i(t) > 0 = (\,\overline{C}x(t) + \overline{D}u_{\beta_*}(t)\,)_i\} = \alpha_{0+} \\ \beta_0(t) &\equiv \{i \in \beta_* : u_i(t) = 0 = (\,\overline{C}x(t) + \overline{D}u_{\beta_*}(t)\,)_i\} = \beta_{0+} \\ \gamma_0(t) &\equiv \{i \in \beta_* : u_i(t) = 0 < (\,\overline{C}x(t) + \overline{D}u_{\beta_*}(t)\,)_i\} = \gamma_{0+} \end{aligned}\right\} \quad \forall\, t \in (t_*, t_* + \varepsilon]$$

and

$$\left.\begin{aligned} \alpha_0(t) &= \alpha_{0-} \\ \beta_0(t) &= \beta_{0-} \\ \gamma_0(t) &= \gamma_{0-} \end{aligned}\right\} \quad \forall\, t \in [\,t_* - \varepsilon, t_*\,).$$

Since $0 < u_{\alpha_*}(t) = -(D_{\alpha_*\alpha_*})^{-1}(C_{\alpha_*}.x(t) + D_{\alpha_*\beta_*}u_{\beta_*}(t))$ and $0 = u_{\gamma_*}(t)$, it follows that

$$\overline{C}x(t) + \overline{D}u_{\beta_*}(t) = (\,Cx(t) + Du(t)\,)_{\beta_*}$$

for all t sufficiently near t_*. Consequently,

$$\left.\begin{aligned} \alpha(t) &= \alpha_* \cup \alpha_{0+} \\ \beta(t) &= \beta_{0+} \\ \gamma(t) &= \gamma_* \cup \gamma_{0+} \end{aligned}\right\} \quad \forall\, t \in (\,t_*, t_* + \varepsilon\,]$$

and

$$\left.\begin{array}{l}\alpha(t) = \alpha_* \cup \alpha_{0-} \\ \beta(t) = \beta_{0-} \\ \gamma(t) = \gamma_* \cup \gamma_{0-}\end{array}\right\} \quad \forall\, t \in [\, t_* - \varepsilon, t_* \,).$$

This shows that x^* is a strongly non-Zeno state of (1). □

One important consequence of the P-property is that the u-trajectory must necessarily be unique. In what follows, we present an extended treatment of the Zeno issue for an LCS with nonunique u-trajectories. Specifically, we make several alternative assumptions on the tuple (A, B, C, D), the first of which ensures the existence and uniqueness of a continuously differentiable solution trajectory $x(t)$ corresponding to various subsystems of (1).

(A) For every $x \in \Re^n$ and every triple of index sets (α, β, γ) partitioning $\{1, \ldots, m\}$ with $\beta \neq \emptyset$, the mixed LCP

$$\begin{aligned} 0 &= [\, Cx + Du \,]_\alpha, \\ 0 \leq u_\beta \perp [\, Cx + Du \,]_\beta &\geq 0, \\ 0 &= u_\gamma \end{aligned} \tag{23}$$

has a solution $u \in \Re^m$; moreover, $Bu^1 = Bu^2$ for any two such solutions u^1 and u^2.

The fundamental role of the above assumption is described in the following result.

Proposition 16. *Under assumption* (A)*, for every triple of index sets* (α, β, γ) *partitioning* $\{1, \ldots, m\}$ *with* $\beta \neq \emptyset$*, the system*

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ 0 &= [\, Cx + Du \,]_\alpha, \\ 0 \leq u_\beta \perp [\, Cx + Du \,]_\beta &\geq 0, \\ 0 &= u_\gamma, \\ x(0) &= x^0 \end{aligned} \tag{24}$$

has a unique solution trajectory $x(t)$ *for all* $t \in [0, T]$ *for any* $T > 0$*; moreover,* $x(t)$ *is continuously differentiable on its domain.*

Proof. Let $\mathcal{S}(x) \subset \Re^m$ denote the solution set of (23). As a multifunction, the map $\mathcal{S} : x \mapsto \mathcal{S}(x)$ is a polyhedral multifunction; i.e., its graph is the union of finitely many polyhedra. Under assumption (A), the mapping

$$\widehat{B} \,:\, x \in \Re^n \,\mapsto\, B\mathcal{S}(x)$$

is a single-valued function whose graph is the union of finitely many polyhedra. As such, by a result due originally to Gowda (see [12, Exercise 5.6.14]), it follows that $\widehat{B}$ is a (globally) Lipschitz continuous function on $\Re^n$. In terms of this mapping, the system (24) can be equivalently stated as

$$\dot{x} = Ax + \widehat{B}(x), \quad x(0) = x^0.$$

Since the right-hand side of the ODE is Lipschitz continuous, the existence and uniqueness and the continuous differentiability of a solution trajectory $x(t)$ follows from classical ODE theory. □

Before proceeding further, we make several remarks about Proposition 16 and assumption (A). First, if one is interested in the existence of a unique x-trajectory to the single LCS (1), then it suffices to assume that $B\mathrm{SOL}(Cx, D)$ is a singleton for all $x \in \Re^n$. Second, while the x-trajectory is necessarily unique in the proposition, no such uniqueness is asserted for the u-trajectory; no continuity of the u-trajectory is asserted either. This is a significant departure from the P-property under which the u-trajectory exists and is both unique and continuous. Nevertheless, it can be shown that in the case where C has full row rank, assumption (A) implies that D must be a P-matrix. Hence, condition (A) is most interesting when C is deficient in row rank. A class of triples (B, C, D) satisfying assumption (A) with D being non-P is presented in section 5. It should be noted that condition (A) is different from the passifiability property of the triple (B, C, D) used in [5]; the latter property, along with a minimality assumption on (A, B, C, D), yields the existence and uniqueness of a continuous x-trajectory and an $\mathcal{L}_2$ u-trajectory of the LCS (1). Examples of the class of triples (B, C, D) from section 5 can easily be constructed which fail the passifiability condition; conversely, any triple $(B, C, 0)$ with CB symmetric positive definite is passifiable but fails condition (A).

Another difference between assumption (A) and the P-matrix assumption is that (A) does not imply any apparent determinant properties of the principal matrix

$$\begin{bmatrix} D_{\alpha\alpha} & D_{\alpha\beta} \\ D_{\beta\alpha} & D_{\beta\beta} \end{bmatrix};$$

in particular, it could be singular. Assumption (A) does imply

$$\left.\begin{array}{c} 0 = (Du)_\alpha \\ 0 \le u_\beta \perp (Du)_\beta \ge 0 \\ 0 = u_\gamma \end{array}\right\} \Rightarrow Bu = 0,$$

and is implied by the following more restrictive condition:

$$u \circ Du \le 0 \Rightarrow Bu = 0.$$

For our purpose, the following invariance properties of assumption (A) are important for the extension of the previous inductive argument to an LCS not satisfying the P-property.

PROPOSITION 17. *Suppose that (B, C, D) satisfies condition* (A). *The same condition holds for the following triples of index sets:*

(a) $(B_{\cdot\tilde{\alpha}}, C_{\tilde{\alpha}\cdot}, D_{\tilde{\alpha}\tilde{\alpha}})$ *for every subset* $\tilde{\alpha} \subseteq \{1, \ldots, m\}$;

(b) *the triple* $(\widetilde{B}, \widetilde{C}, \widetilde{D})$ *associated with the* $\tilde{\alpha}$*-principal transform* (7) *of* (B, C, D) *for every subset* $\tilde{\alpha} \subseteq \{1, \ldots, m\}$ *such that* $D_{\tilde{\alpha}\tilde{\alpha}}$ *is nonsingular;*

(c) (PB, CP^{-1}, D) *for every nonsingular matrix* P.

Proof. Let α', β', and γ' be any three index sets partitioning $\tilde{\alpha}$, with $\beta' \neq \emptyset$. Let $x \in \Re^n$ be arbitrary. We need to show that the system

$$\begin{aligned} 0 &= [Cx + Du]_{\alpha'}, \\ 0 \le u_{\beta'} &\perp [Cx + Du]_{\beta'} \ge 0, \\ 0 &= u_{\gamma'} \end{aligned}$$

has a solution $u_{\tilde{\alpha}}$. Moreover, if $u^1_{\tilde{\alpha}}$ and $u^2_{\tilde{\alpha}}$ are any two such solutions, we must have $B_{\cdot\tilde{\alpha}}u^1_{\tilde{\alpha}} = B_{\cdot\tilde{\alpha}}u^2_{\tilde{\alpha}}$. But this is clear from condition (A) with the choice of $(\alpha, \beta, \gamma) = (\alpha', \beta', \gamma' \cup (\{1, \ldots, m\} \backslash \tilde{\alpha}))$. To prove (b), let $(\alpha', \beta', \gamma')$ be a triple of index sets partitioning $\{1, \ldots, m\}$ with $\beta' \neq \emptyset$ and consider the system

$$\begin{aligned} 0 &= [\,\widetilde{C}x + \widetilde{D}\widetilde{u}\,]_{\alpha'}, \\ 0 \leq \widetilde{u}_{\beta'} &\perp [\,\widetilde{C}x + \widetilde{D}\widetilde{u}\,]_{\beta'} \geq 0, \\ 0 &= \widetilde{u}_{\gamma'}. \end{aligned} \tag{25}$$

Letting $\widetilde{w} \equiv \widetilde{C}x + \widetilde{D}\widetilde{u}$ and "pivoting" on $(D_{\tilde{\alpha}\tilde{\alpha}})^{-1}$, we obtain $w \equiv Cx + Du$, where the relation between the pairs (w, u) and $(\widetilde{w}, \widetilde{u})$ is as follows:

$$\begin{aligned} w &\equiv \begin{pmatrix} w_{\tilde{\alpha}} \\ \\ w_{\bar{\alpha}} \end{pmatrix}, \quad u \equiv \begin{pmatrix} u_{\tilde{\alpha}} \\ \\ u_{\bar{\alpha}} \end{pmatrix}, \\ \widetilde{w} &= \begin{pmatrix} u_{\tilde{\alpha}} \\ \\ w_{\bar{\alpha}} \end{pmatrix}, \quad \widetilde{u} = \begin{pmatrix} w_{\tilde{\alpha}} \\ \\ u_{\bar{\alpha}} \end{pmatrix}, \end{aligned} \tag{26}$$

where $\bar{\alpha}$ is the complement of $\tilde{\alpha}$ in $\{1, \ldots, m\}$. Therefore, system (25) is equivalent to system (23) for some suitable triple (α, β, γ) that is derived from $(\alpha', \beta', \gamma')$. Therefore, the existence of a solution to (25) follows from condition (A) on the original triple (B, C, D). Suppose that $\widetilde{u}^1$ and $\widetilde{u}^2$ are any two solutions satisfying (25). Corresponding to $\widetilde{u}^i$, for $i = 1, 2$, let $\widetilde{w}^i \equiv \widetilde{C}x + \widetilde{D}\widetilde{u}^i$ and (w^i, u^i) be defined accordingly. It follows that $Bu^1 = Bu^2$. We have

$$\begin{aligned} \widetilde{B}\widetilde{u}^i &= B_{\cdot\tilde{\alpha}}(\,D_{\tilde{\alpha}\tilde{\alpha}}\,)^{-1}\widetilde{u}^i_{\tilde{\alpha}} + [\,B_{\cdot\bar{\alpha}} - B_{\cdot\alpha}(\,D_{\tilde{\alpha}\tilde{\alpha}}\,)^{-1}D_{\tilde{\alpha}\bar{\alpha}}\,]\widetilde{u}^i_{\bar{\alpha}} \\ &= B_{\cdot\tilde{\alpha}}(\,D_{\tilde{\alpha}\tilde{\alpha}}\,)^{-1}w^i_{\tilde{\alpha}} + [\,B_{\cdot\bar{\alpha}} - B_{\cdot\tilde{\alpha}}(\,D_{\tilde{\alpha}\tilde{\alpha}}\,)^{-1}D_{\tilde{\alpha}\bar{\alpha}}\,]u^i_{\bar{\alpha}} \\ &= B_{\cdot\tilde{\alpha}}(\,D_{\tilde{\alpha}\tilde{\alpha}}\,)^{-1}[\,Cx + Du^i\,]_{\tilde{\alpha}} + [\,B_{\cdot\bar{\alpha}} - B_{\cdot\tilde{\alpha}}(\,D_{\tilde{\alpha}\tilde{\alpha}}\,)^{-1}D_{\tilde{\alpha}\bar{\alpha}}\,]u^i_{\bar{\alpha}} \\ &= B_{\cdot\tilde{\alpha}}(\,D_{\tilde{\alpha}\tilde{\alpha}}\,)^{-1}C_{\tilde{\alpha}\cdot}x + Bu^i; \end{aligned}$$

hence, $\widetilde{B}\widetilde{u}^1 = \widetilde{B}\widetilde{u}^2$. This proves (b). Finally (c) is obvious. ∎

To motivate the following discussion, consider an unobservable state $x^* = x(t_*)$. In this case, $(x(t), u(t)) = (e^{A(t-t_*)}x^*, 0)$ is trivially an admissible solution trajectory to (1) for $t > t_*$. Moreover, under assumption (A), the trajectory $x(t) = e^{A(t-t_*)}x^*$ is unique for $t > t_*$. Nevertheless, if D is not an R_0-matrix, i.e., if the homogeneous LCP $(0, D)$ has a nonzero solution, then it is very difficult, if not impossible, to ascertain the Zeno properties of $(x(t), u(t))$ jointly. The reason is very simple: the LCP $(0, D)$, which must be satisfied by the u-trajectory in this case, is totally unaffected by the x-trajectory. Consequently, if one expects an unobservable state x^* to be (right) non-Zeno, one must restrict oneself to the class of matrices D for which the LCP $(0, D)$ has a polyhedral solution set; this is the principal motivation to introduce the class of weakly column sufficient matrices (Definition 1).

The following result extends the key expansion Lemma 14 and is applicable to an arbitrary tuple (A, B, C, D) satisfying condition (A).

LEMMA 18. *Suppose that (A, B, C, D) satisfy condition* (A). *Let $x^* = x(t_*)$ be a given state of the solution trajectory $(x(t), u(t))$ such that $CA^j x^* = 0$ for all $j = 0, 1, \ldots, k-1$ for some integer $k \geq 0$. The following two statements hold:*

(a) *For each $t > t_*$ sufficiently near t_*, there exists $u^{t+} \in \mathrm{SOL}(CA^k x^*, D)$ such that*

$$x(t) = \sum_{j=0}^{k+1} \frac{(t-t_*)^j}{j!} A^j x^* + \frac{(t-t_*)^{k+1}}{(k+1)!} B u^{t+} + O(|t-t_*|^{k+2}),$$

$$u(t) = \frac{(t-t_*)^k}{k!} u^{t+} + O(|t-t_*|^{k+1}).$$

(b) *If k is even, then for each $t < t_*$ sufficiently near t_*, there exists $u^{t+} \in \mathrm{SOL}(CA^k x^*, D)$ such that the expansion in part* (a) *remains valid. If k is odd, then for each $t < t_*$, there exists $u^{t-} \in \mathrm{SOL}(-CA^k x^*, D)$ such that*

$$x(t) = \sum_{j=0}^{k+1} \frac{(t-t_*)^j}{j!} A^j x^* - \frac{(t-t_*)^{k+1}}{(k+1)!} B u^{t-} + O(|t-t_*|^{k+2}),$$

$$u(t) = \frac{|t-t_*|^k}{k!} u^{t-} + O(|t-t_*|^{k+1}).$$

Proof. We prove only statement (a). Proceeding as in the proof of Lemma 14, we define $z(t)$ by (15) and note that (16) and (17) must hold. By the same result due of Gowda that we used in the proof of Proposition 16, we can deduce the existence of a constant $\eta > 0$ such that

$$\| Bu(t) \| \leq \eta [\| z(t) \| + | t - t_* |^k]$$

for all t. Consequently, it follows that (18) holds. The vector $v(t) \equiv k!u(t)/|t-t_*|^k$ satisfies, for all $t > t_*$,

$$0 \leq v(t) \perp k! C \frac{z(t)}{| t - t_* |^k} + CA^k x^* + Dv(t) \geq 0.$$

Since $\|z(t)\|$ is of order $|t-t_*|^{k+1}$, by the semistability of the LCP $(CA^k x^*, D)$, it follows that for every $t > t_*$ sufficiently near t_*, there exists $u^{t+} \in \mathrm{SOL}(CA^k x^*, D)$ such that $\|v(t) - u^{t+}\|$ is of order $O(|t-t_*|)$. The expansion for $u(t)$ in part (a) thus follows readily. Substituting this expansion into the differential equation

$$\dot{z}(t) = Az(t) + \frac{(t-t_*)^k}{k!} A^{k+1} x^* + Bu(t)$$

$$= Az(t) + \frac{(t-t_*)^k}{k!} [A^{k+1} x^* + Bu^{t+}] + O(| t - t_* |^{k+1})$$

using the fact that Bu^{t+} is independent of t (because $B\mathrm{SOL}(CA^k x^*, D)$ is a singleton), and integrating, we can deduce the desired expansion for $x(t)$. The details are not repeated. □

On the basis of the concept of an identifiable LCP, we introduce the following.

Definition 19. *A state x^* is said to be* identifiable *with respect to the triple (A, C, D) if, for each subset α of $\{1, \ldots, m\}$, if x^* is observable with degree k with respect to the pair $(C_{\alpha\cdot}, A)$, then the LCPs $(\pm C_{\alpha\cdot} A^k x^*, D_{\alpha\alpha})$ are identifiable. The state x^* is said to be* totally identifiable *with respect to the tuple (A, B, C, D) if x^* is*

identifiable with respect to all triples $(\widehat{A}, \widehat{C}, \widehat{D})$*, where*

$$\widehat{A} \equiv A - B_{\cdot\alpha}(\, D_{\alpha\alpha}\,)^{-1} C_{\alpha\cdot},$$

$$\widehat{C} \equiv C_{\bar{\alpha}\cdot} - D_{\bar{\alpha}\alpha}(\, D_{\alpha\alpha}\,)^{-1} C_{\alpha\cdot},$$

$$\widehat{D} \equiv D_{\bar{\alpha}\bar{\alpha}} - D_{\bar{\alpha}\alpha}(\, D_{\alpha\alpha}\,)^{-1} D_{\alpha\bar{\alpha}},$$

and α*, with complement* $\bar{\alpha}$*, ranges over all subsets of* $\{1, \ldots, m\}$ *for which* $D_{\alpha\alpha}$ *is nonsingular.* □

The next lemma asserts that the above identifiability property is inherited by the principal (sub)transforms of a given tuple.

LEMMA 20. *Suppose that* x^* *is totally identifiable with respect to* (A, B, C, D)*. Then* x^* *is also totally identifiable with respect to the following tuples:*

(a) $(A, B_{\cdot\alpha}, C_{\alpha\cdot}, D_{\alpha\alpha})$ *for all subsets* α *of* $\{1, \ldots, m\}$*;*

(b) *the principal subtuples* $(\widehat{A}, \widehat{B}, \widehat{C}, \widehat{D})$ *associated with all legitimate principal pivot transforms of* (A, B, C, D)*, where*

$$\widehat{A} \equiv A - B_{\cdot\alpha}(\, D_{\alpha\alpha}\,)^{-1} C_{\alpha\cdot}, \qquad \widehat{B} \equiv B_{\cdot\bar{\alpha}} - B_{\cdot\alpha}(\, D_{\alpha\alpha}\,)^{-1} D_{\alpha\bar{\alpha}},$$

$$\widehat{C} \equiv C_{\bar{\alpha}\cdot} - D_{\bar{\alpha}\alpha}(\, D_{\alpha\alpha}\,)^{-1} C_{\alpha\cdot}, \quad \widehat{D} \equiv D_{\bar{\alpha}\bar{\alpha}} - D_{\bar{\alpha}\alpha}(\, D_{\alpha\alpha}\,)^{-1} D_{\alpha\bar{\alpha}},$$

and α*, with complement* $\bar{\alpha}$*, ranges over all subsets of* $\{1, \ldots, m\}$ *for which* $D_{\alpha\alpha}$ *is nonsingular.*

Moreover, Px^* *is totally identifiable with respect to the triple* $(PAP^{-1}, PB, CP^{-1}, D)$ *for any nonsingular matrix* P*.*

Proof. Statement (a) is obvious. The validity of statement (b) is based on the observation that a tuple $(\widehat{\widehat{A}}, \widehat{\widehat{C}}, \widehat{\widehat{D}})$, where

$$\widehat{\widehat{A}} \equiv \widehat{A} - \widehat{B}_{\cdot\hat{\alpha}}(\, \widehat{D}_{\hat{\alpha}\hat{\alpha}}\,)^{-1} \widehat{C}_{\hat{\alpha}\cdot},$$

$$\widehat{\widehat{C}} \equiv \widehat{C}_{\bar{\hat{\alpha}}\cdot} - D_{\bar{\hat{\alpha}}\hat{\alpha}}(\, D_{\hat{\alpha}\hat{\alpha}}\,)^{-1} \widehat{C}_{\hat{\alpha}\cdot},$$

$$\widehat{\widehat{D}} \equiv \widehat{D}_{\bar{\hat{\alpha}}\bar{\hat{\alpha}}} - \widehat{D}_{\bar{\hat{\alpha}}\hat{\alpha}}(\, \widehat{D}_{\hat{\alpha}\hat{\alpha}}\,)^{-1} D_{\hat{\alpha}\bar{\hat{\alpha}}},$$

and $\hat{\alpha}$, with complement $\bar{\hat{\alpha}}$, is a subset of $\bar{\alpha}$ for which $\widehat{D}_{\hat{\alpha}\hat{\alpha}}$ is nonsingular, can be shown to be a principal subtuple associated with the $(\alpha \cup \hat{\alpha})$-principal pivot transform of (A, B, C, D). Hence (b) holds. The last assertion follows easily from the identity $CP^{-1}(PAP^{-1})^k = CA^kP^{-1}$. □

Our extended Zeno result for an LCS without the P-property is the following. The statement of the theorem assumes that x^* is neither the initial nor the terminal state of the solution trajectory so that we do not need to pay attention to the one-sidedness of these special states.

THEOREM 21. *Let* D *be a weakly column sufficient matrix. Suppose that condition* (A) *holds for the tuple* (A, B, C, D)*. The following two statements hold for any state* $x^* = x(t_*)$ *and any u-trajectory:*

(a) *If* x^* *is unobservable with respect to* (C, A)*, then* x^* *is weakly non-Zeno.*

(b) *If* x^* *is totally identifiable with respect to the tuple* (A, B, C, D)*, then* x^* *is weakly non-Zeno.*

Proof. We follow the proof of Theorem 9. Suppose that the initial state x^0 is unobservable with respect to (C, A). In this case, the unique x-trajectory is $x(t) = e^{At}x^0$

and we have $Cx(t) = 0$ for all t. Hence $u(t) \in \mathrm{SOL}(0, D)$ for all t. The weak column sufficiency of D then completes the proof. So we assume that x^0 is observable. The proof of Lemma 12 shows that all subsequent states are observable. This establishes part (a). We use induction on m to prove that if x^* is totally identifiable with respect to (A, B, C, D), then x^* is weakly right non-Zeno; the proof of weakly left non-Zenoness is similar and therefore omitted. Since the LCP $(CA^k x^*, D)$ is identifiable, where k is the observability degree of x^* with respect to the pair (C, A), either one of the two index sets $\mathcal{I}_u$ or $\mathcal{I}_w$ is nonempty. Without loss of generality, assume that the former is so. By Proposition 4, there exists a scalar $\sigma > 0$ such that $u_i \geq \sigma$ for all $u \in \mathrm{SOL}(CA^k x^*, D)$ and all $i \in \mathcal{I}_u$. Consequently, by the expansion of $u(t)$ near t_* as described in part (a) of Lemma 18 it follows that $u_i(t) > 0$ for all $i \in \mathcal{I}_u$ and all $t > t_*$ sufficiently near t_*. Moreover, $D_{\mathcal{I}_u \mathcal{I}_u}$ is nonsingular by the identifiability assumption. Consequently, the trajectory $(x(t), u(t))$ must satisfy the principal subtransform (9) with $\alpha \equiv \mathcal{I}_u$ for all $t > t_*$ sufficiently near t_*. The induction hypothesis then completes the proof. □

5. A special bimodal system. As an illustration of another application of the expansion Lemma 18, we consider a special bimodal system which has $D \equiv ff^T$, $B \equiv bf^T$, and $C \equiv fc^T$ for some m-vector f and n-vectors b and c. To avoid trivialities, we assume that f has no zero components. It is easy to see that condition (A) holds for the triple $(B, C, D) \equiv (bf^T, fc^T, ff^T)$. Notice that the LCS (1) with this triple remains an MIMO (multiple input, multiple output) system; nevertheless, it is a bimodal system because of the lemma below.

LEMMA 22. *The LCP*

$$0 \leq u \perp fc^T x + ff^T u \geq 0$$

has a solution for all $x \in \Re^n$; moreover, for any such solution u, $f^T u = 0$ if $fc^T x \geq 0$, and $c^T x + f^T u = 0$ otherwise. Consequently,

$$\mathrm{SOL}(fc^T x, ff^T) = \begin{cases} \{\, u \geq 0 : f^T u = 0 \,\} & \text{if } fc^T x \geq 0, \\ \{\, u \geq 0 : c^T x + f^T u = 0 \,\} & \text{otherwise.} \end{cases}$$

Proof. If $fc^T x \geq 0$, then $u = 0$ is a solution of the LCP. Since $f^T u$ is a constant on the solution set of this LCP, it follows that $f^T u = 0$ for all such solutions in this case. If $fc^T x \not\geq 0$, then $f^T u \neq 0$ for all solutions of the LCP. For any such solution u, we have

$$0 = (f^T u)(c^T x) + (f^T u)^2,$$

which yields $c^T x + f^T u = 0$ as claimed. The representation of $\mathrm{SOL}(fc^T x, ff^T)$ is easy to establish. □

In view of the above lemma, it follows that the LCS (1) is of the following bimodal kind:

$$\dot{x} = \begin{cases} Ax & \text{if } fc^T x \geq 0, \\ (A - bc^T)x & \text{otherwise.} \end{cases}$$

Since $f \neq 0$, it is clear that x^* is an observable state of the pair (C, A) if and only if the scalar $c^T A^k x^* \neq 0$ for some integer $k \geq 0$. If $c^T x^* = \cdots = c^T A^{k-1} x^* = 0 \neq c^T A^k x^*$, Lemma 18 implies that, for $t > t_*$,

$$c^T x(t) = \frac{(t - t_*)^k}{k!} c^T A^k x^* + O(|t - t_*|^{k+1}).$$

Since $c^T A^k x^*$ is a nonzero scalar, it follows that $c^T x(t)$ is nonzero and of one sign for all $t > t_*$ sufficiently near t_*. Since f is a constant vector, it follows that either $fc^T x(t) \geq 0$ for all $t > t_*$ sufficiently near t_*, which implies $f^T u(t) = 0$ for all $u(t) \in \mathrm{SOL}(fc^T x(t), ff^T)$, or $c^T x(t) + f^T u(t) = 0$ for all such t. In the latter case, it follows that x^* is a weakly right non-Zeno state with respect to the trajectory $(x(t), u(t))$. In the former case, f must be either a positive or a negative vector depending on whether $c^T A^k x^* > 0$ or $c^T A^k x^* < 0$; in either case, we must have $u(t) = 0$ for all $t > t_*$ sufficiently near t_* because $f^T u(t) = 0$ and $u(t) \geq 0$. This is enough to show that x^* is a weakly right non-Zeno state. A similar argument will establish that x^* is also a weakly left non-Zeno state. We have, therefore, proved the following result for an observable state. The proof of the result for an unobservable state is the same as before and is not repeated.

THEOREM 23. *Let $(B, C, D) \equiv (bf^T, fc^T, ff^T)$, where f has no zero component. The LCS* (1) *has no Zeno states of the second kind.* □

6. Concluding remarks. In this paper, via a basic expansion of the solution trajectory near a given time, we have shown that an LCS with the P-property has no Zeno states of the first kind, that the totally identifiable states of an LCS with the weakly column sufficient property are weakly non-Zeno, and that a certain bimodal LCS has no Zeno states of the second kind. Subsequently to the completion to this work, we have extended the results in several directions, in particular, to a special LCS of the "positive semidefinite plus" type [12] and to a strongly regular nonlinear complementarity system [22]. An interesting extension that we have *not* yet resolved is the case where $D = 0$ and CB is positive definite (but not symmetric). Such an LCS is not necessarily passifiable. Lastly, in the paper [22], we use the results established herein to study the "local observability" of an LCS.

Acknowledgments. The authors wish to thank Professor Hans Schumacher and Dr. Kanat Çamlibel for extensive discussion on the topic of this paper and for sharing their invaluable insights that contribute to our understanding of the LCS. They are also grateful to the associate editor, the referees, and David Stewart, who have made a number of constructive comments that have helped to improve the presentation of the paper.

REFERENCES

[1] M. ANITESCU, *Optimization-Based Simulation for Nonsmooth Rigid Multibody Dynamics*, Technical report ANL/MCS-P1161-0504, Division of Mathematics and Computer Science, Argonne National Laboratory, Argonne, IL, 2004.

[2] B. BROGLIATO, *Nonsmooth Mechanics*, 2nd ed., Springer-Verlag, London, 1999.

[3] B. BROGLIATO, *Some perspectives on analysis and control of complementarity systems*, IEEE Trans. Automat. Control, 48 (2003), pp. 918–935.

[4] P. BRUNOVSKY, *Regular synthesis for the linear-quadratic optimal control problem with linear control constraints*, J. Differential Equations, 38 (1980), pp. 344–360.

[5] M. K. ÇAMLIBEL, *Complementarity Methods in the Analysis of Piecewise Linear Dynamical Systems*, Ph.D. thesis, Center for Economic Research, Tilburg University, The Netherlands, 2001.

[6] M. K. ÇAMLIBEL, W. P. M. H. HEEMELS, A. J. VAN DER SCHAFT, AND J. M. SCHUMACHER, *Switched networks and complementarity*, IEEE Trans. Circuits Syst. I, 50 (2003), pp. 1036–1046.

[7] M. K. CAMLIBEL, W. P. M. H. HEEMELS, AND J. M. SCHUMACHER, *Well-posedness of a class of linear network with ideal diodes*, in Proceedings of the 14th International Symposium of Mathematical Theory of Networks and Systems, 2000.

[8] M. K. Camlibel, W. P. M. H. Heemels, and J. M. Schumacher, *Stability and controllability of planar bimodal linear complementarity systems*, in Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, HI, 2003, pp. 1651–1656.
[9] M. K. Camlibel and J. M. Schumacher, *On the Zeno behavior of linear complementarity systems*, in Proceedings of the 40th IEEE Conference on Decision and Control, Orlando, FL, 2001, pp. 346–351.
[10] C. T. Chen, *Linear System Theory and Design*, Oxford University Press, Oxford, 1984.
[11] R. W. Cottle, J. S. Pang, and R. E. Stone, *The Linear Complementarity Problem*, Academic Press, Cambridge, 1992.
[12] F. Facchinei and J. S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Spring-Verlag, New York, 2003.
[13] W. P. H. Heemels, *Linear Complementarity Systems: A Study in Hybrid Dynamics*, Ph.D. thesis, Department of Electrical Engineering, Eindhoven University of Technology, 1999.
[14] W. P. M. H. Heemels, M. K. Çamlibel, A. J. van der Schaft, and J. M. Schumacher, *Modelling, well-posedness, and stability of switched electrical networks*, in Hybrid Systems: Computation and Control (Proceedings of the 6th International Workshop, HSCC2003, Prague, 2003), O. Maler and A. Pnueli, eds., Lecture Notes in Comput. Sci. 2623, Springer, Berlin, 2003, pp. 249–266.
[15] W. P. M. H. Heemels, J. M. Schumacher, and S. Weiland, *Well-posedness of linear complementarity systems*, in Proceedings of the 38th IEEE Conference on Decision and Control, Phoenix, AZ, 1999, pp. 3037–3042.
[16] W. P. M. H. Heemels, J. M. Schumacher, and S. Weiland, *The rational linear complementarity systems*, Linear Algebra Appl., 294 (1999), pp. 93–135.
[17] W. P. M. H. Heemels, J. M. Schumacher, and S. Weiland, *Linear complementarity systems*, SIAM J. Appl. Math., 60 (2000), pp. 234–1269.
[18] J. Imura and A. van der Schaft, *Characterization of well-posedness of piecewise-linear systems*, IEEE Trans. Automat. Control, 45 (2000), pp. 1600–1619.
[19] K. H. Johansson, M. Egersted, J. Lygeros, and S. Sastry, *On the regularization of Zeno hybrid automata*, Systems Control Lett., 38 (1999), pp. 141–150.
[20] D. M. W. Leenaerts and W. M. G. van Bokhoven, *Piecewise Linear Modelling and Analysis*, Kluwer Academic, Dordrecht, The Netherlands, 1998.
[21] J. S. Pang, V. Kumar, and P. Song, *Convergence of Time-Stepping Methods for Initial and Boundary Value Frictional Compliant Contact Problems*, manuscript, Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY, 2004.
[22] J. S. Pang and J. L. Shen, *Strongly differential variational inequalities*, IEEE Trans. Automat. Control, submitted.
[23] J. S. Pang and D. E. Stewart, *Differential variational inequalities*, Math. Program. Ser. A, submitted.
[24] F. Pfeiffer and Ch. Glocker, *Multibody Dynamics with Unilateral Contacts*, John Wiley, New York, 1996.
[25] S. M. Robinson, *Strongly regular generalized equations*, Math. Oper. Res., 5 (1980), pp. 43–62.
[26] S. Sastry, *Nonlinear Systems: Analysis, Stability and Control*, Springer-Verlag, New York, 1999.
[27] J. M. Schumacher, *Complementarity systems in optimization*, Math. Program. Ser. B, 101 (2004), pp. 263–296.
[28] S. N. Simic, K. H. Johansson, J. Lygeros, and S. Sastry, *Toward a geometric theory of hybrid systems*, Dyn. Contin. Discrete Impuls. Syst. Ser. A, to appear.
[29] P. Song, J. S. Pang, and V. Kumar, *Semi-implicit time-stepping models for frictional compliant contact problems*, Internat. J. Numer. Methods Engrg., 60 (2004), pp. 2231–2261.
[30] D. E. Stewart, *High Accuracy Numerical Methods for Ordinary Differential Equations with Discontinuous Right-Hand Side*, Doctoral thesis, Department of Mathematics, University of Queensland, Australia, 1990.
[31] D. Stewart, *Rigid-body dynamics with friction and impact*, SIAM Rev., 42 (2000), pp. 3–39.
[32] D. Stewart and J. Trinkle, *An implicit time-stepping scheme for rigid-body dynamics with inelastic collisions and coulomb friction*, Internat. J. Numer. Methods Engrg., 39 (1996), pp. 2673–2691.
[33] H. J. Sussmann, *Bounds on the number of switchings for trajectories of piecewise analytic vector fields*, J. Differential Equations, 43 (1982), pp. 399–418.
[34] A.J. van der Schaft and J. M. Schumacher, *The complementarity-slackness class of hybrid systems*, Math. Control Signals Systems, 9 (1996), pp. 266–301.
[35] A. J. van der Schaft and J. M. Schumacher, *Complementarity modeling of hybrid systems*, IEEE Trans. Automat. Control, 43 (1998), pp. 483–490.

[36] A. J. van der Schaft and J. M. Schumacher, *An Introduction to Hybrid Dynamical Systems*, Springer-Verlag, London, 2000.
[37] J. C. Willems, *Dissipative dynamical systems*, Arch. Ration. Mech. Anal., 45 (1972), pp. 321–393.
[38] J. Zhang, K. H. Johansson, J. Lygeros, and S. Sastry, *Zeno hybrid systems*, Internat. J. Robust Nonlinear Control, 11 (2001), pp. 435–451.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 3, pp. 1067–1080

OPTIMAL CONTROL OF OBSTACLE FOR QUASI-LINEAR ELLIPTIC VARIATIONAL BILATERAL PROBLEMS*

QIHONG CHEN†, DELIN CHU‡, AND ROGER C. E. TAN‡

Abstract. This paper is concerned with an optimal control problem for quasi-linear elliptic variational inequality in which the bilateral obstacles are the control. The cost functional of this optimal control problem is of Lagrange type in which the pth power of Laplacian of the control appears. This feature leads to the fact that it is hard to derive the optimality system for the underlying problem. In this paper, the optimality system is established by utilizing the special structure of the approximate optimality system including the monotonicity of the leading differential operator.

Key words. obstacle optimal control, quasi-linear elliptic equations, bilateral variational inequality, optimality system

AMS subject classifications. 47J20, 49J20, 49K20

DOI. 10.1137/S0363012904443075

1. Introduction.

The variational inequalities and related optimal control problems have been studied extensively in the literature; see [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20, 23, 24, 25, 26, 27, 28, 29, 30] and the reference therein. Recently, various control problems for bilateral variational inequalities have been considered in [7, 14, 15, 16]. When the governing system is an obstacle variational inequality, the obstacle can also be regarded as the control. Such a case is referred to as an obstacle optimal control problem. To the best of our knowledge, such a problem was first studied in [1]. For the homogeneous case, an optimal obstacle control problem for an elliptic variational inequality is considered there. By virtue of the properties of the superharmonic functions, the existence and uniqueness as well as characterizations of the optimal pair are established in [1]. To study the obstacle optimal control problem for more general systems, an indirect obstacle control model is suggested in [11, 12, 13]. Motivated by [1] and [11], the regularity of the obstacle control problem has been investigated in [25] and [26]. The work in [1] has been extended by adding a nonzero source term to the right-hand side of the state equation and it has been found in [2] that such an extension is not trivial.

In this paper, we consider an obstacle optimal control problem for the quasi-linear case. The main feature of our problem is that the state satisfies a quasi-linear elliptic bilateral variational inequality and the input control is the pair of upper and lower obstacles, as stated as follows:

$$\begin{cases} \varphi \le y \le \psi, \\ -\mathrm{div} A(x, \nabla y)(y - \varphi) \le 0, \\ -\mathrm{div} A(x, \nabla y)(y - \psi) \le 0, \end{cases} \tag{1.1}$$

*Received by the editors April 17, 2004; accepted for publication (in revised form) March 2, 2005; published electronically September 20, 2005.

http://www.siam.org/journals/sicon/44-3/44307.html

†Department of Mathematics, Hangzhou Dianzi University, Hangzhou 310018, China (chenqih@yahoo.com). The work of this author was supported partly by FANEDD grant 200218, NSFC grant 10171059, and NUS grant R-146-000-047-112.

‡Department of Mathematics, National University of Singapore, Singapore 117543, Singapore (matchudl@nus.edu.sg, scitance@nus.edu.sg). The work of these authors was supported by NUS grant R-146-000-047-112.

where $A(x,\cdot)$ is nonlinear.

In practice, there are many real physical or geometrical problems related to obstacle variational inequalities (cf. [28]). Except for some ideal cases, the governing equations are usually quasi-linear or nonlinear. A typical example is that in the study of non-Newtonian fluids [17]

$$A(\nabla y) = |\nabla y|^{p-2}\nabla y.$$

Another example is that in the study of a minimal surface with obstacle [23, 28]

$$A(\nabla y) = (1+|\nabla y|^2)^{-\frac{1}{2}}\nabla y.$$

Similar cases occur in evolutionary problems. We found that the techniques developed in [1, 2] do not work when $A(x,\cdot)$ is nonlinear. Hence, from both theoretical and practical points of view, it is necessary to study the case with $A(x,\cdot)$ nonlinear.

Suppose $\Omega \subset \mathcal{R}^n$ is a bounded domain with a $C^{1,1}$ boundary $\partial\Omega$. Let $z_d \in L^2(\Omega)$ be a given target profile. For any $\varphi, \psi \in W_0^{1,p}(\Omega)$, we define

$$K(\varphi,\psi) = \{z \in W_0^{1,p}(\Omega) | \varphi \le z \le \psi \quad \text{a.e. } x \in \Omega\}$$

and consider a weak formulation of quasi-linear elliptic bilateral obstacle problem (1.1):

$$\begin{cases} y \in K(\varphi,\psi), \\ \displaystyle\int_\Omega A(x,\nabla y)\cdot\nabla(z-y)dx \ge 0 \qquad \forall z \in K(\varphi,\psi), \end{cases} \tag{1.2}$$

where

$$A(x,\eta) = (a_1(x,\eta),\dots,a_n(x,\eta)). \tag{1.3}$$

Given $\varphi,\psi \in W_0^{1,p}(\Omega)$, under some further mild assumptions on $A(x,\eta)$ (see (H_1) and (H_2) below), the variational inequality (1.2) is uniquely solvable [22]. We will denote the unique solution of (1.2) corresponding to (φ,ψ) by $y = T(\varphi,\psi)$.

Let

$$W = W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega)$$

and

$$U_{ad} = \{(\varphi,\psi) \in W \times W | \varphi \le \psi \quad \text{a.e. } \Omega\}.$$

We seek a pair of $(\bar\varphi,\bar\psi) \in U_{ad}$ so that the corresponding state $\bar y = T(\bar\varphi,\bar\psi)$ is close to the desired target profile z_d and the norm of $(\bar\varphi,\bar\psi)$ is not too large in $W \times W$. Consequently, we take our objective functional as

$$J(\varphi,\psi) = \int_\Omega \left\{ \frac{1}{2}(T(\varphi,\psi)-z_d)^2 + \frac{1}{p}[|\Delta\varphi|^p + |\Delta\psi|^p] \right\} dx, \tag{1.4}$$

which we try to minimize. More precisely, in this paper we study the following optimal control problem.

Problem (P). Find a pair of control $(\bar\varphi,\bar\psi) \in U_{ad}$ such that

$$J(\bar\varphi,\bar\psi) = \inf_{U_{ad}} J(\varphi,\psi). \tag{1.5}$$

One may also consider the case in which state constraints are presented. We will treat the problem with state constraints in a separate paper.

Related to Problem (P), we make the following two assumptions on $A(x,\eta)$ in (1.3):

(H_1) For any $\eta=(\eta_1,\dots,\eta_n)\in\mathcal{R}^n$, $a_j(\cdot,\eta)$ is a measurable function on Ω with $a_j(\cdot,0)=0$ and for any $x\in\Omega$, $a_j(x,\cdot)$ belongs to $C^1(\mathcal{R}^n)$, $j=1,\dots,n$.

(H_2) For any $p\geq 2$, $x\in\Omega$, and all $\xi,\eta\in\mathcal{R}^n$

$$\sum_{i,j=1}^{n}\frac{\partial a_j}{\partial\eta_i}(x,\eta)\xi_i\xi_j\geq\Lambda_1(k+|\eta|)^{p-2}|\xi|^2,$$

$$\sum_{i,j=1}^{n}\left|\frac{\partial a_j}{\partial\eta_i}(x,\eta)\right|\leq\Lambda_2|\eta|^{p-2},$$

where $k\in(0,1]$, and Λ_1 and Λ_2 are some positive constants.

The following lemma is an immediate consequence of assumptions (H_1) and (H_2).

LEMMA 1 (cf. [10]). *Under assumptions* (H_1)–(H_2), *there are positive constants* k_1 *and* k_2 *depending only on* n, Λ_1, *and* Λ_2 *such that for any* $x\in\Omega$, $\eta=(\eta_1,\dots,\eta_n)\in\mathcal{R}^n$, *and* $\eta'=(\eta_1',\dots,\eta_n')\in\mathcal{R}^n$

(a)

$$\sum_{j=1}^{n}(a_j(x,\eta)-a_j(x,\eta'))(\eta_j-\eta_j')\geq k_1|\eta-\eta'|^p.$$

(b)

$$\sum_{j=1}^{n}|a_j(x,\eta)|\leq k_2|\eta|^{p-1}.$$

In Problem (P) the cost functional is of Lagrange-type in which the Laplacian of the control (obstacle) appears. Such a term provides certain compactness of the control. As a direct result, the existence of the optimal control is almost routine. Consequently, the following existence theorem for Problem (P) can be obtained by some standard ideas with some suitable variational inequality techniques.

THEOREM 2. *Under assumptions* (H_1) and (H_2), *there exists an optimal solution* $(\bar\varphi,\bar\psi)$ *to Problem* (P).

In this paper, our main purpose is to establish the optimality system for Problem (P). In general, when the penalty on control is stronger (in the current case, the second derivative appears in the cost functional), the proof on the existence of optimal control becomes easier because of the stronger convergence property in the minimizing sequence. However, at the same time, the derivation of necessary conditions will become harder, since the corresponding duality relation has to be established in a bigger space. In addition, when the state equation is a variational inequality, one has to prove the convergence of the approximate adjoint equation, which is difficult since the approximate adjoint equation is defined in a very weak sense, and then some measure term will occur in the limit. As far as the optimality system is concerned, it is incomplete as long as the measure which intervenes in the adjoint equation has not been precisely described. To derive the optimality system for Problem (P), we have adopted the $W^{2,p}$ framework, which, however, causes some further difficulties due to

the lack of weak continuity of the leading differential operator in the approximate optimality systems. We have overcome the encountered difficulties by making full use of the special structure of the approximate optimality systems, including the monotonicity of the leading differential operator.

The rest of the paper is organized as follows. In section 2 some necessary support results are derived by studying an approximate optimal control problem. Then in section 3 the optimality system for Problem (P) is established, in which the support of the measure (mentioned above) has been well described. Finally, some conclusions are given in section 4.

2. An approximate problem and its convergence property. In this section we derive some necessary support results by studying an approximate problem related to Problem (P).

Let $\epsilon > 0$ and $(\varphi, \psi) \in U_{ad}$. We consider the following quasi-linear elliptic equation:

$$\begin{cases} -\mathrm{div} A(x, \nabla y) + \dfrac{1}{\epsilon}[\beta(y-\varphi) + \gamma(y-\psi)] = 0 \quad \text{in } \Omega, \\ y|_{\partial\Omega} = 0, \end{cases} \tag{2.1}$$

where

$$\beta(r) = \begin{cases} 0, & r > 0, \\ -r^2, & -\dfrac{1}{2} \le r \le 0, \\ r + \dfrac{1}{4}, & r < -\dfrac{1}{2}, \end{cases}$$

and

$$\gamma(r) = \begin{cases} 0, & r < 0, \\ r^2, & 0 \le r \le \dfrac{1}{2}, \\ r - \dfrac{1}{4}, & r > \dfrac{1}{2}. \end{cases}$$

As $\beta(\cdot)$ and $\gamma(\cdot)$ are nondecreasing, it is known that the above equation is uniquely solvable and the solution will be denoted by $y^\epsilon = T^\epsilon(\varphi, \psi)$.

The following lemma gives the Gâteaux-derivative of the operator T^ϵ.

LEMMA 3. *Let $(\xi, \zeta) \in U_{ad}$. Then we have*

$$\frac{T^\epsilon(\varphi + t\xi, \psi + t\zeta) - T^\epsilon(\varphi, \psi)}{t} \xrightarrow{w} w^\epsilon \quad \textit{in } W_0^{1,p}(\Omega) \qquad (t \to 0^+), \tag{2.2}$$

where w^ϵ satisfies

$$\begin{cases} -\mathrm{div}\left(\dfrac{\partial A}{\partial \eta}(x, \nabla y^\epsilon)\nabla w^\epsilon\right) + \dfrac{1}{\epsilon}[\beta'(y^\epsilon - \varphi) + \gamma'(y^\epsilon - \psi)]w^\epsilon \\ \quad = \dfrac{1}{\epsilon}[\beta'(y^\epsilon - \varphi)\xi + \gamma'(y^\epsilon - \psi)\zeta] \quad \textit{in } \Omega, \\ w^\epsilon|_{\partial\Omega} = 0 \end{cases} \tag{2.3}$$

with $y^\epsilon = T^\epsilon(\varphi, \psi)$.

Proof. First note that for any $t > 0$ and (φ, ψ), $(\xi, \zeta) \in U_{ad}$, we always get $(\varphi + t\xi, \psi + t\zeta) \in U_{ad}$.

The rest of the proof of Lemma 3 is similar to that of [10, Theorem 3.1] and is omitted here. □

Let $(\bar{\varphi}, \bar{\psi})$ be an optimal solution to Problem (P) and let $\bar{y} = T(\bar{\varphi}, \bar{\psi})$.

For any $\epsilon > 0$, we define

$$J^\epsilon(\varphi, \psi) = \int_\Omega \left\{ \frac{1}{2}(T^\epsilon(\varphi, \psi) - z_d)^2 + \frac{1}{p}[|\Delta\varphi|^p + |\Delta\psi|^p + |\varphi - \bar{\varphi}|^p + |\psi - \bar{\psi}|^p] \right\} dx$$

and investigate the following approximate optimal control problem.

Problem (P^ϵ). Find a pair of control $(\varphi^\epsilon, \psi^\epsilon) \in U_{ad}$ such that

$$J^\epsilon(\varphi^\epsilon, \psi^\epsilon) = \inf_{U_{ad}} J^\epsilon(\varphi, \psi).$$

Similar to Theorem 2, we can prove the existence for the approximate optimal control problem (P^ϵ).

THEOREM 4. *Problem* (P^ϵ) *has (at least) an optimal solution.*

In the following we derive the optimality system for Problem (P^ϵ).

THEOREM 5. *Assume* $(\varphi^\epsilon, \psi^\epsilon)$ *is an optimal solution to Problem* (P^ϵ) *and* $y^\epsilon = T^\epsilon(\varphi^\epsilon, \psi^\epsilon)$. *Then, there exists* $p^\epsilon \in H_0^1(\Omega)$ *such that the following optimality system is satisfied:*

$$\begin{cases} -\mathrm{div} A(x, \nabla y^\epsilon) + \dfrac{1}{\epsilon}[\beta(y^\epsilon - \varphi^\epsilon) + \gamma(y^\epsilon - \psi^\epsilon)] = 0 \quad \text{in } \Omega, \\ y^\epsilon|_{\partial\Omega} = 0, \end{cases} \tag{2.4}$$

$$\begin{cases} -\mathrm{div}\left(\dfrac{\partial A}{\partial \eta}(x, \nabla y^\epsilon)^T \nabla p^\epsilon\right) + \dfrac{1}{\epsilon}[\beta'(y^\epsilon - \varphi^\epsilon) + \gamma'(y^\epsilon - \psi^\epsilon)]p^\epsilon = y^\epsilon - z_d \quad \text{in } \Omega, \\ p^\epsilon|_{\partial\Omega} = 0, \end{cases} \tag{2.5}$$

and

$$\begin{aligned} &\int_\Omega \left[\frac{1}{\epsilon}\beta'(y^\epsilon - \varphi^\epsilon)p^\epsilon + |\varphi^\epsilon - \bar{\varphi}|^{p-2}(\varphi^\epsilon - \bar{\varphi})\right](\varphi - \varphi^\epsilon)dx \\ &\quad + \int_\Omega \left[\frac{1}{\epsilon}\gamma'(y^\epsilon - \psi^\epsilon)p^\epsilon + |\psi^\epsilon - \bar{\psi}|^{p-2}(\psi^\epsilon - \bar{\psi})\right](\psi - \psi^\epsilon)dx \\ &\quad + \int_\Omega [|\Delta\varphi^\epsilon|^{p-2}\Delta\varphi^\epsilon\Delta(\varphi - \varphi^\epsilon) + |\Delta\psi^\epsilon|^{p-2}\Delta\psi^\epsilon\Delta(\psi - \psi^\epsilon)]dx \\ &\quad \geq 0 \qquad \forall (\varphi, \psi) \in U_{ad}. \end{aligned} \tag{2.6}$$

Proof. As $(\varphi^\epsilon, \psi^\epsilon)$ is a solution to Problem (P^ϵ), we have

$$\forall (\varphi, \psi) \in U_{ad}, \qquad \liminf_{t \to 0^+} \frac{J^\epsilon(\varphi^\epsilon + t(\varphi - \varphi^\epsilon), \psi^\epsilon + t(\psi - \psi^\epsilon)) - J^\epsilon(\varphi^\epsilon, \psi^\epsilon)}{t} \geq 0.$$

This gives

$$\begin{aligned} &\int_\Omega [w^\epsilon(y^\epsilon - z_d) + |\Delta\varphi^\epsilon|^{p-2}\Delta\varphi^\epsilon\Delta(\varphi - \varphi^\epsilon) + |\Delta\psi|^{p-2}\Delta\psi^\epsilon\Delta(\psi - \psi^\epsilon)]dx \\ &\quad + \int_\Omega [|\varphi^\epsilon - \bar{\varphi}|^{p-2}(\varphi^\epsilon - \bar{\varphi})(\varphi - \varphi^\epsilon) + |\psi^\epsilon - \bar{\psi}|^{p-2}(\psi^\epsilon - \bar{\psi})(\psi - \psi^\epsilon)]dx \geq 0, \end{aligned} \tag{2.7}$$

where $w^\epsilon \in H_0^1(\Omega)$ satisfies

$$\begin{cases} -\mathrm{div}\left(\dfrac{\partial A}{\partial \eta}(x, \nabla y^\epsilon)\nabla w^\epsilon\right) + \dfrac{1}{\epsilon}[\beta'(y^\epsilon - \varphi^\epsilon) + \gamma'(y^\epsilon - \psi^\epsilon)]w^\epsilon \\ \quad = \dfrac{1}{\epsilon}[\beta'(y^\epsilon - \varphi^\epsilon)(\varphi - \varphi^\epsilon) + \gamma'(y^\epsilon - \psi^\epsilon)(\psi - \psi^\epsilon)] \quad \text{in } \Omega, \\ w^\epsilon|_{\partial\Omega} = 0. \end{cases}$$

Let p^ϵ be the solution of linear equation (2.5); then we obtain

$$\begin{aligned} &\int_\Omega w^\epsilon(y^\epsilon - z_d)dx \\ &\quad = \int_\Omega \left\{\nabla w^\epsilon \cdot \frac{\partial A}{\partial \eta}(x, \nabla y^\epsilon)^T \nabla p^\epsilon + \frac{1}{\epsilon}[\beta'(y^\epsilon - \varphi^\epsilon) + \gamma'(y^\epsilon - \psi^\epsilon)]p^\epsilon w^\epsilon\right\}dx \\ &\quad = \int_\Omega \frac{1}{\epsilon}[\beta'(y^\epsilon - \varphi^\epsilon)(\varphi - \varphi^\epsilon) + \gamma'(y^\epsilon - \psi^\epsilon)(\psi - \psi^\epsilon)]p^\epsilon dx. \end{aligned}$$

This leads to (2.6). □

Remark 1. Using the so-called monotonicity inequality (cf. [21])

$$(|A|^{p-2}A - |B|^{p-2}B) \cdot (A - B) \geq 0 \qquad (A, B \in \mathcal{R}^n)$$

we can deduce from (2.6) that

$$\begin{aligned} &\int_\Omega \left[\frac{1}{\epsilon}\beta'(y^\epsilon - \varphi^\epsilon)p^\epsilon + |\varphi^\epsilon - \bar\varphi|^{p-2}(\varphi^\epsilon - \bar\varphi)\right](\varphi - \varphi^\epsilon)dx \\ &\quad + \int_\Omega \left[\frac{1}{\epsilon}\gamma'(y^\epsilon - \psi^\epsilon)p^\epsilon + |\psi^\epsilon - \bar\psi|^{p-2}(\psi^\epsilon - \bar\psi)\right](\psi - \psi^\epsilon)dx \\ &\quad + \int_\Omega [|\Delta\varphi|^{p-2}\Delta\varphi\Delta(\varphi - \varphi^\epsilon) + |\Delta\psi|^{p-2}\Delta\psi\Delta(\psi - \psi^\epsilon)]dx \\ &\quad \geq 0 \qquad \forall (\varphi, \psi) \in U_{ad}. \end{aligned} \tag{2.8}$$

Remark 2. Inequality (2.6) implies

$$\begin{aligned} &\int_\Omega \left[\frac{1}{\epsilon}\beta'(y^\epsilon - \varphi^\epsilon)p^\epsilon + |\varphi^\epsilon - \bar\varphi|^{p-2}(\varphi^\epsilon - \bar\varphi)\right]wdx \\ &\quad + \int_\Omega \left[\frac{1}{\epsilon}\gamma'(y^\epsilon - \psi^\epsilon)p^\epsilon + |\psi^\epsilon - \bar\psi|^{p-2}(\psi^\epsilon - \bar\psi)\right]wdx \\ &\quad + \int_\Omega [|\Delta\varphi^\epsilon|^{p-2}\Delta\varphi^\epsilon + |\Delta\psi^\epsilon|^{p-2}\Delta\psi^\epsilon]\Delta wdx = 0 \qquad \forall w \in W. \end{aligned} \tag{2.9}$$

The above two remarks will play an important role in deriving the optimality system for the original Problem (P).

Next we consider the convergence of solution of the approximate problem—Problem (P^ϵ) as ϵ goes to 0^+.

LEMMA 6. *Assume* $y^\epsilon = T^\epsilon(\varphi^\epsilon, \psi^\epsilon)$. *Then*

$$\int_\Omega A(x, \nabla y^\epsilon) \cdot \nabla y^\epsilon dx \leq \int_\Omega A(x, \nabla y^\epsilon) \cdot \nabla z dx \qquad \forall z \in K(\varphi^\epsilon, \psi^\epsilon).$$

Proof. From the approximate equation (2.1), we have

$$\int_\Omega A(x,\nabla y^\epsilon)\cdot\nabla u dx+\frac{1}{\epsilon}\int_\Omega[\beta(y^\epsilon-\varphi^\epsilon)+\gamma(y^\epsilon-\psi^\epsilon)]udx=0 \qquad \forall u\in W_0^{1,p}(\Omega). \tag{2.10}$$

For any $z\in K(\varphi^\epsilon,\psi^\epsilon)$, substituting $u=y^\epsilon-z$ in (2.10), we obtain

$$\int_\Omega A(x,\nabla y^\epsilon)\cdot\nabla(y^\epsilon-z)dx+\frac{1}{\epsilon}\int_\Omega[\beta(y^\epsilon-\varphi^\epsilon)+\gamma(y^\epsilon-\psi^\epsilon)](y^\epsilon-z)dx=0. \tag{2.11}$$

Note that $\beta(y^\epsilon(x)-\varphi^\epsilon(x))$ differs from 0 only when $y^\epsilon(x)<\varphi^\epsilon(x)$, and $\gamma(y^\epsilon(x)-\psi^\epsilon(x))$ differs from 0 only when $y^\epsilon(x)>\psi^\epsilon(x)$. In any case we have, for any $z\in K(\varphi^\epsilon,\psi^\epsilon)$,

$$[\beta(y^\epsilon-\varphi^\epsilon)+\gamma(y^\epsilon-\psi^\epsilon)](y^\epsilon-z)\geq 0 \quad \text{a.e. in } \Omega.$$

Thus, we get

$$\int_\Omega A(x,\nabla y^\epsilon)\cdot\nabla(y^\epsilon-z)dx\leq 0 \qquad \forall z\in K(\varphi^\epsilon,\psi^\epsilon). \qquad \square$$

THEOREM 7. *Let $(\bar\varphi,\bar\psi)$ be an optimal solution to Problem* (P) *and $(\varphi^\epsilon,\psi^\epsilon)$ any optimal solution to Problem* (P^ϵ) *for any $\epsilon>0$. Then for $p>n$,*

$$\begin{aligned}
(\varphi^\epsilon,\psi^\epsilon)&\xrightarrow{w}(\bar\varphi,\bar\psi) && \text{in } W^{2,p}(\Omega)\times W^{2,p}(\Omega),\\
(\varphi^\epsilon,\psi^\epsilon)&\xrightarrow{s}(\bar\varphi,\bar\psi) && \text{in } W_0^{1,p}(\Omega)\times W_0^{1,p}(\Omega),\\
y^\epsilon=T^\epsilon(\varphi^\epsilon,\psi^\epsilon)&\xrightarrow{s}y=T(\bar\varphi,\bar\psi) && \text{in } W_0^{1,p}(\Omega),
\end{aligned} \tag{2.12}$$

and

$$\lim_{\epsilon\to 0^+}J^\epsilon(\varphi^\epsilon,\psi^\epsilon)=J(\bar\varphi,\bar\psi). \tag{2.13}$$

Proof. First, we note that

$$\frac{1}{p}\int_\Omega(|\Delta\varphi^\epsilon|^p+|\Delta\psi^\epsilon|^p)dx\leq J^\epsilon(\varphi^\epsilon,\psi^\epsilon)\leq J^\epsilon(0,0)=\frac{1}{2}\int_\Omega|z_d|^2dx+\frac{1}{p}\int_\Omega(|\bar\varphi|^p+|\bar\psi|^p)dx.$$

Thus $(\Delta\varphi^\epsilon,\Delta\psi^\epsilon)$ is bounded in $W^{2,p}(\Omega)\times W^{2,p}(\Omega)$. We may assume, extracting some subsequence if necessary,

$$\begin{aligned}
(\varphi^\epsilon,\psi^\epsilon)&\xrightarrow{w}(\varphi,\psi) \quad \text{in } W^{2,p}(\Omega)\times W^{2,p}(\Omega),\\
(\varphi^\epsilon,\psi^\epsilon)&\xrightarrow{s}(\varphi,\psi) \quad \text{in } W_0^{1,p}(\Omega)\times W_0^{1,p}(\Omega).
\end{aligned}$$

Applying Lemmas 1 and 6 and Hölder's inequality we get

$$\begin{aligned}
k_1\int_\Omega|\nabla y^\epsilon|^p dx&\leq\int_\Omega A(x,\nabla y^\epsilon)\cdot\nabla y^\epsilon dx\\
&\leq\int_\Omega A(x,\nabla y^\epsilon)\cdot\nabla\varphi^\epsilon dx\\
&\leq k_2\|\nabla y^\epsilon\|_p^{p-1}\|\nabla\varphi^\epsilon\|_p.
\end{aligned}$$

This implies that y^ϵ is bounded in $W_0^{1,p}(\Omega)$. Hence, for some subsequence, we have

$$y^\epsilon \xrightarrow{w} y \quad \text{in } W_0^{1,p}(\Omega),$$
$$y^\epsilon \xrightarrow{s} y \quad \text{in } L^p(\Omega).$$

Moreover, from Lemma 1(b), we can deduce that

$$\|A(x, \nabla y^\epsilon)\|_{p'} \le C. \tag{2.14}$$

For any $\eta \in W_0^{1,p}(\Omega)$, it follows from (2.10) that

$$\int_\Omega [\beta(y^\epsilon - \varphi^\epsilon) + \gamma(y^\epsilon - \psi^\epsilon)]\eta dx = -\epsilon \int_\Omega A(x, \nabla y^\epsilon) \cdot \nabla \eta dx \to 0$$

because the integral on the right-hand side is bounded. Then, by Lebesgue's dominated convergence theorem, we have

$$\int_\Omega [\beta(y - \varphi) + \gamma(y - \psi)]\eta dx = 0.$$

This implies that

$$\beta(y - \varphi) + \gamma(y - \psi) = 0 \quad \text{a.e. in } \Omega$$

due to the arbitrariness of η. By the definition of $\beta(\cdot)$ and $\gamma(\cdot)$, we have

$$\varphi(x) \le y(x) \le \psi(x) \quad \text{a.e. in } \Omega,$$

i.e., $y \in K(\varphi, \psi)$.

We further claim that y^ϵ strongly converges to y in $W_0^{1,p}(\Omega)$. In fact, let $v^\epsilon = \inf(\sup(y, \varphi^\epsilon), \psi^\epsilon)$; then $v^\epsilon \in K(\varphi^\epsilon, \psi^\epsilon)$ and v^ϵ strongly converges to y in $W_0^{1,p}(\Omega)$. By Lemma 6, we have

$$\int_\Omega A(x, \nabla y^\epsilon) \cdot \nabla y^\epsilon dx \le \int_\Omega A(x, \nabla y^\epsilon) \cdot \nabla v^\epsilon dx. \tag{2.15}$$

From Lemma 1, (2.14), and (2.15), we can further deduce that

$$\begin{aligned} 0 &\le k_1 \int_\Omega |\nabla(y^\epsilon - y)|^p dx \\ &\le \int_\Omega (A(x, \nabla y^\epsilon) - A(x, \nabla y)) \cdot (\nabla y^\epsilon - \nabla y) dx \\ &\le \int_\Omega A(x, \nabla y^\epsilon) \cdot (\nabla v^\epsilon - \nabla y) dx - \int_\Omega A(x, \nabla y) \cdot (\nabla y^\epsilon - \nabla y) dx \to 0. \end{aligned}$$

Therefore,

$$\nabla y^\epsilon \xrightarrow{s} \nabla y \quad \text{in } L^p(\Omega) \tag{2.16}$$

and by (H_1)

$$A(x, \nabla y^\epsilon) \to A(x, \nabla y) \quad \text{a.e. in } \Omega. \tag{2.17}$$

By (2.14) and (2.17), we conclude that

$$A(x,\nabla y^\epsilon) \overset{w}{\rightharpoonup} A(x,\nabla y) \quad \text{in } L^{p'}(\Omega). \tag{2.18}$$

It remains to prove $y = T(\varphi,\psi)$. For every $z \in K(\varphi,\psi)$, let $z^\epsilon = \inf(\sup(z,\varphi^\epsilon),\psi^\epsilon)$; then $z^\epsilon \in K(\varphi^\epsilon,\psi^\epsilon)$ and z^ϵ strongly converges to z in $W_0^{1,p}(\Omega)$. Again by Lemma 6, we have

$$\int_\Omega A(x,\nabla y^\epsilon)\cdot\nabla y^\epsilon dx \le \int_\Omega A(x,\nabla y^\epsilon)\cdot\nabla z^\epsilon dx. \tag{2.19}$$

Then, with (2.16) and (2.18), we may pass to the limit in (2.19) and get

$$\int_\Omega A(x,\nabla y)\cdot\nabla y dx \le \int_\Omega A(x,\nabla y)\cdot\nabla z dx. \tag{2.20}$$

This gives $\int_\Omega A(x,\nabla y)\cdot\nabla(z-y)dx \ge 0$ for every $z \in K(\varphi,\psi)$ and, therefore, $y = T(\varphi,\psi)$.

As U_{ad} is closed, then $(\varphi,\psi) \in U_{ad}$. Using the weak lower semicontinuity of the L^p-norm, we obtain

$$\begin{aligned} J(\varphi,\psi) + \frac{1}{p}[\|\varphi-\bar\varphi\|_p^p + \|\psi-\bar\psi\|_p^p] &\le \liminf_{\epsilon\to 0^+} J^\epsilon(\varphi^\epsilon,\psi^\epsilon) \\ &\le \limsup_{\epsilon\to 0^+} J^\epsilon(\varphi^\epsilon,\psi^\epsilon) \\ &\le \lim_{\epsilon\to 0^+} J^\epsilon(\bar\varphi,\bar\psi) = J(\bar\varphi,\bar\psi) \\ &\le J(\varphi,\psi). \end{aligned} \tag{2.21}$$

This yields $\|\varphi-\bar\varphi\|_p^p + \|\psi-\bar\psi\|_p^p \le 0$; so $\varphi = \bar\varphi$, $\psi = \bar\psi$, and hence $y = \bar y = T(\bar\varphi,\bar\psi)$. In addition, we see that

$$\lim_{\epsilon\to 0^+} J^\epsilon(\varphi^\epsilon,\psi^\epsilon) = J(\bar\varphi,\bar\psi). \tag{2.22}$$

Finally, the uniqueness of the limit point implies the convergence of the whole sequence of $(\varphi^\epsilon,\psi^\epsilon)$ and the whole sequence of y^ϵ as well. □

3. Optimality system for Problem (P). Now, we are in a position to establish the optimality system for the original Problem (P).

THEOREM 8. *Assume* (H_1) *and* (H_2). *Let* $(\bar y,\bar\varphi,\bar\psi)$ *be an optimal triple to Problem* (P). *Then there exist* $\bar p \in H_0^1(\Omega)$ *and* $\bar\mu \in H^{-1}(\Omega)\cap\mathcal{M}(\bar\Omega)$ *satisfying*

$$\begin{cases} -\operatorname{div}\left(\dfrac{\partial A}{\partial\eta}(x,\nabla\bar y)^T\nabla\bar p\right) + \bar\mu = \bar y - z_d & \text{in } \Omega, \\ \bar p|_{\partial\Omega} = 0, \end{cases} \tag{3.1}$$

and

$$\operatorname{supp}\bar\mu \subset \{x\in\Omega|\quad \bar y(x) = \bar\varphi(x) \quad \text{or} \quad \bar y(x) = \bar\psi(x)\} \qquad \left(\text{as } p > \frac{n}{2}\right) \tag{3.2}$$

such that

$$\Delta(|\Delta\bar\varphi|^{p-2}\Delta\bar\varphi + |\Delta\bar\psi|^{p-2}\Delta\bar\psi) + \bar\mu = 0 \qquad \text{in } \Omega, \tag{3.3}$$

where $\mathcal{M}(\bar{\Omega})$ is the set of all regular signed measures on $\bar{\Omega}$.

Proof. First consider the approximate problems (P^ϵ) related to $(\bar{y}, \bar{\varphi}, \bar{\psi})$. Let $(y^\epsilon, \varphi^\epsilon, \psi^\epsilon)$ be any optimal triple to Problem (P^ϵ). Then, by Theorem 5, the optimality condition (2.6) followed by (2.8) and (2.9) holds for some $p^\epsilon \in H_0^1(\Omega)$ satisfying (2.5) or equivalently that, for any $u \in H_0^1(\Omega)$,

(3.4)
$$\int_\Omega \nabla u^T \frac{\partial A}{\partial \eta}(x, \nabla y^\epsilon)^T \nabla p^\epsilon dx + \int_\Omega \frac{1}{\epsilon}[\beta'(y^\epsilon - \varphi^\epsilon) + \gamma'(y^\epsilon - \psi^\epsilon)]p^\epsilon u dx = \int_\Omega (y^\epsilon - z_d) u dx.$$

We put $u = p^\epsilon$ in (3.4) to obtain

(3.5)
$$\int_\Omega (\nabla p^\epsilon)^T \frac{\partial A}{\partial \eta} \nabla p^\epsilon dx + \int_\Omega \frac{1}{\epsilon}[\beta'(y^\epsilon - \varphi^\epsilon) + \gamma'(y^\epsilon - \psi^\epsilon)](p^\epsilon)^2 dx = \int_\Omega (y^\epsilon - z_d) p^\epsilon dx.$$

By (H_2), we get

$$\Lambda_1 \int_\Omega (k + |\nabla y^\epsilon|)^{p-2} |\nabla p^\epsilon|^2 dx + \int_\Omega \frac{1}{\epsilon}[\beta'(y^\epsilon - \varphi^\epsilon) + \gamma'(y^\epsilon - \psi^\epsilon)](p^\epsilon)^2 dx \le C\|p^\epsilon\|_2.$$

Then

$$\|p^\epsilon\|_{H_0^1(\Omega)} \le C,$$

where C is independent of ϵ. This implies that, for some subsequence,

$$p^\epsilon \overset{w}{\rightharpoonup} \bar{p} \quad \text{in } H_0^1(\Omega). \tag{3.6}$$

Denote

$$\mu_\varphi^\epsilon = \frac{1}{\epsilon}\beta'(y^\epsilon - \varphi^\epsilon)p^\epsilon, \quad \mu_\psi^\epsilon = \frac{1}{\epsilon}\gamma'(y^\epsilon - \psi^\epsilon)p^\epsilon; \quad \mu^\epsilon = \mu_\varphi^\epsilon + \mu_\psi^\epsilon.$$

From (3.4) we get

$$\begin{aligned} \left|\int_\Omega \mu^\epsilon u dx\right| &\le \left|\int_\Omega (y^\epsilon - z_d) u dx\right| + \left|\int_\Omega \nabla u^T \left(\frac{\partial A}{\partial \eta}\right)^T \nabla p^\epsilon dx\right| \\ &\le C\|u\|_{H_0^1(\Omega)} \qquad \forall u \in H_0^1(\Omega), \end{aligned} \tag{3.7}$$

i.e.,

$$\|\mu^\epsilon\|_{H^{-1}(\Omega)} \le C,$$

where C is independent of ϵ.

Furthermore, let $S_\tau(r) \in C^1(\mathcal{R})(\tau > 0)$ be a family of smooth approximations to *sign* r, satisfying the following:

$$S_\tau'(r) \ge 0 \qquad \forall r \in \mathcal{R}$$

and

$$S_\tau(r) = \begin{cases} 1 & \text{if } r > \tau, \\ 0 & \text{if } r = 0, \\ -1 & \text{if } r < -\tau. \end{cases}$$

Putting $u = S_\tau(p^\epsilon)$ in (3.4), we get

$$\int_\Omega S'_\tau(p^\epsilon)(\nabla p^\epsilon)^T \frac{\partial A}{\partial \eta} \nabla p^\epsilon dx + \int_\Omega \mu^\epsilon S_\tau(p^\epsilon) dx = \int_\Omega (y^\epsilon - z_d) S_\tau(p^\epsilon) dx.$$

This gives

$$\int_\Omega \mu^\epsilon S_\tau(p^\epsilon) dx \le C.$$

Letting $\tau \to 0^+$, we have

$$\|\mu^\epsilon\|_{L^1(\Omega)} \, (= \|\mu^\epsilon_\varphi\|_{L^1(\Omega)} + \|\mu^\epsilon_\psi\|_{L^1(\Omega)}) \le C.$$

Thus, extracting some subsequence if necessary, we may let

$$\begin{aligned} \mu^\epsilon_\varphi &\overset{w}{\rightharpoonup} \bar{\mu}_\varphi \quad \text{in } \mathcal{M}(\bar{\Omega}), \\ \mu^\epsilon_\psi &\overset{w}{\rightharpoonup} \bar{\mu}_\psi \quad \text{in } \mathcal{M}(\bar{\Omega}), \\ \mu^\epsilon &\overset{w}{\rightharpoonup} \bar{\mu} \quad \text{in } H^{-1}(\Omega) \cap \mathcal{M}(\bar{\Omega}), \end{aligned} \tag{3.8}$$

with $\bar{\mu} = \bar{\mu}_\varphi + \bar{\mu}_\psi$.

Passing to the limit in (2.8), we get

$$\begin{aligned} &\langle \bar{\mu}_\varphi, \varphi - \bar{\varphi} \rangle + \langle \bar{\mu}_\psi, \psi - \bar{\psi} \rangle \\ &\quad + \int_\Omega [|\Delta\varphi|^{p-2}\Delta\varphi\Delta(\varphi - \bar{\varphi}) + |\Delta\psi|^{p-2}\Delta\psi\Delta(\psi - \bar{\psi})] dx \\ &\quad \ge 0 \qquad \forall (\varphi, \psi) \in U_{ad} \end{aligned} \tag{3.9}$$

and consequently

$$\begin{aligned} &\langle \bar{\mu}, w \rangle + \int_\Omega [|\Delta(\bar{\varphi} + w)|^{p-2}\Delta(\bar{\varphi} + w) + |\Delta(\bar{\psi} + w)|^{p-2}\Delta(\bar{\psi} + w)]\Delta w dx \\ &\quad \ge 0 \qquad \forall w \in W. \end{aligned} \tag{3.10}$$

Noting that

$$\||\Delta\varphi^\epsilon|^{p-2}\Delta\varphi^\epsilon + |\Delta\psi^\epsilon|^{p-2}\Delta\psi^\epsilon\|_{p'} \le \|\Delta\varphi^\epsilon\|_p^{p-1} + \|\Delta\psi^\epsilon\|_p^{p-1} \le C$$

we may assume that, for some subsequence,

$$|\Delta\varphi^\epsilon|^{p-2}\Delta\varphi^\epsilon + |\Delta\psi^\epsilon|^{p-2}\Delta\psi^\epsilon \overset{w}{\rightharpoonup} F \quad \text{in } L^{p'}(\Omega).$$

Taking the limit in (2.9), we obtain

$$\langle \bar{\mu}, w \rangle + \int_\Omega F \Delta w dx = 0 \qquad \forall w \in W. \tag{3.11}$$

To prove (3.3), we need only to clarify that

$$F = |\Delta\bar{\varphi}|^{p-2}\Delta\bar{\varphi} + |\Delta\bar{\psi}|^{p-2}\Delta\bar{\psi}. \tag{3.12}$$

Combining (3.10) and (3.11), we have

$$\int_\Omega F\Delta w dx \le \int_\Omega [|\Delta(\bar{\varphi} + w)|^{p-2}\Delta(\bar{\varphi} + w) + |\Delta(\bar{\psi} + w)|^{p-2}\Delta(\bar{\psi} + w)]\Delta w dx \quad \forall w \in W. \tag{3.13}$$

For any given $\chi \in W$, choosing $w = \delta\chi$ ($\delta > 0$) in (3.13), then dividing it by δ and sending $\delta \to 0^+$, we get

$$\int_\Omega F\Delta\chi dx \le \int_\Omega [|\Delta\bar{\varphi}|^{p-2}\Delta\bar{\varphi} + |\Delta\bar{\psi}|^{p-2}\Delta\bar{\psi}]\Delta\chi dx. \tag{3.14}$$

On the other hand, choosing $w = \delta\chi$ ($\delta < 0$) in (3.13), then dividing it by δ and sending $\delta \to 0^-$, we get

$$\int_\Omega F\Delta\chi dx \ge \int_\Omega [|\Delta\bar{\varphi}|^{p-2}\Delta\bar{\varphi} + |\Delta\bar{\psi}|^{p-2}\Delta\bar{\psi}]\Delta\chi dx. \tag{3.15}$$

Thus, (3.12) follows.

To verify that $\bar{p}$ satisfies (3.1), it suffices to show that for any $u \in H_0^1(\Omega)$

$$\int_\Omega \nabla u^T \frac{\partial A}{\partial \eta}(x, \nabla y^\epsilon)^T \nabla p^\epsilon dx \to \int_\Omega \nabla u^T \frac{\partial A}{\partial \eta}(x, \nabla \bar{y})^T \nabla \bar{p} dx \quad (\epsilon \to 0^+). \tag{3.16}$$

Obviously,

$$\begin{aligned} &\int_\Omega \nabla u^T \left(\frac{\partial A}{\partial \eta}(x, \nabla y^\epsilon)^T \nabla p^\epsilon - \frac{\partial A}{\partial \eta}(x, \nabla \bar{y})^T \nabla \bar{p} \right) dx \\ &= \int_\Omega \nabla u^T \left(\frac{\partial A}{\partial \eta}(x, \nabla y^\epsilon)^T - \frac{\partial A}{\partial \eta}(x, \nabla \bar{y})^T \right) \nabla p^\epsilon dx \\ &\quad + \int_\Omega \nabla u^T \frac{\partial A}{\partial \eta}(x, \nabla \bar{y})^T (\nabla p^\epsilon - \nabla \bar{p}) dx. \end{aligned} \tag{3.17}$$

From (3.6), we get

$$\int_\Omega \nabla u^T \frac{\partial A}{\partial \eta}(x, \nabla \bar{y})^T (\nabla p^\epsilon - \nabla \bar{p}) dx \to 0 \qquad (\epsilon \to 0^+). \tag{3.18}$$

By (H_2), we know that $\sum_{i,j=1}^n \frac{\partial a_j}{\partial \eta_i}(x, \eta)$ is bounded. Then by Lebesgue's dominated convergence theorem

$$\int_\Omega \left| \nabla u^T \left(\frac{\partial A}{\partial \eta}(x, \nabla y^\epsilon)^T - \frac{\partial A}{\partial \eta}(x, \nabla \bar{y})^T \right) \right|^2 dx \to 0 \qquad (\epsilon \to 0^+). \tag{3.19}$$

Thus, using Hölder's inequality, we have

$$\begin{aligned} &\left| \int_\Omega \nabla u^T \left(\frac{\partial A}{\partial \eta}(x, \nabla y^\epsilon)^T - \frac{\partial A}{\partial \eta}(x, \nabla \bar{y})^T \right) \nabla p^\epsilon dx \right| \\ &\le \left(\int_\Omega \left| \nabla u^T \left(\frac{\partial A}{\partial \eta}(x, \nabla y^\epsilon)^T - \frac{\partial A}{\partial \eta}(x, \nabla \bar{y})^T \right) \right|^2 dx \right)^{1/2} \left(\int_\Omega |\nabla p^\epsilon|^2 dx \right)^{1/2} \\ &\to 0 \quad (\epsilon \to 0^+). \end{aligned} \tag{3.20}$$

Hence, in view of (3.17), (3.18), and (3.20), (3.16) holds.

Finally, we prove (3.2), which is understood as the following: for any $\chi \in C(\bar{\Omega})$ with $\operatorname{supp}\chi \subset \Omega' = \{x \in \Omega |\ \bar{\varphi}(x) < \bar{y}(x) < \bar{\psi}(x)\}$,

$$\langle \bar{\mu}, \chi \rangle_{\mathcal{M}(\bar{\Omega}), C(\bar{\Omega})} = 0. \tag{3.21}$$

In fact, for $p > \frac{n}{2}$, the $W^{2,p}$-bounded subset is relatively compact in $C^{\alpha}(\bar{\Omega})$ for some $\alpha \in (0,1)$. Then, for any $\chi \in C(\bar{\Omega})$ with supp $\chi \subset \Omega'$, the uniform convergence of the approximate optimal triple of control and state (cf. Theorem 7), combined with the compactness of supp χ, ensures that, for some $\epsilon_0 > 0$,

$$\varphi^{\epsilon}(x) < y^{\epsilon}(x) < \psi^{\epsilon}(x) \qquad \forall x \in \operatorname{supp} \chi, \qquad 0 < \epsilon < \epsilon_0,$$

which yields

$$\begin{aligned} \langle \bar{\mu}, \chi \rangle_{\mathcal{M}(\bar{\Omega}), C(\bar{\Omega})} &= \lim_{\epsilon \to 0^+} \int_{\Omega} \frac{1}{\epsilon} [\beta'(y^{\epsilon} - \varphi^{\epsilon}) + \gamma'(y^{\epsilon} - \psi^{\epsilon})] p^{\epsilon} \chi dx \\ &= \lim_{\epsilon \to 0^+} \int_{\operatorname{supp}\ \chi} \frac{1}{\epsilon} [\beta'(y^{\epsilon} - \varphi^{\epsilon}) + \gamma'(y^{\epsilon} - \psi^{\epsilon})] p^{\epsilon} \chi dx \\ &= 0. \end{aligned}$$

Thus, (3.2) holds.

Similarly, we can further prove that

$$\operatorname{supp}\ \bar{\mu}_{\varphi} \subset \{x \in \Omega | \quad \bar{y}(x) = \bar{\varphi}(x)\}$$

and

$$\operatorname{supp}\ \bar{\mu}_{\psi} \subset \{x \in \Omega | \quad \bar{y}(x) = \bar{\psi}(x)\}.$$

The proof is complete. □

4. Conclusions. In this paper we have studied the obstacle optimal control problem—Problem (P). The main contribution of the present work is Theorem 8, in which the optimality system for Problem (P) is obtained.

The present paper basically treated the problem without state constraints. One may consider the case in which state constraints are presented. However, different techniques are needed for solving the problem with different state constraints. Hence, the obstacle optimal control problem with state constraints are worthy of further investigation.

Acknowledgment. The authors are grateful to the anonymous referees for their invaluable comments and suggestions.

REFERENCES

[1] D.R. Adams, S.M. Lenhart, and J. Yong, *Optimal control of obstacle for elliptic variational inequality*, Appl. Math. Optim., 38 (1998), pp. 121–140.
[2] D.R. Adams and S.M. Lenhart, *An obstacle control problem with a source term*, Appl. Math. Optim., 47 (2002), pp. 59–78.
[3] V. Barbu, *Necessary conditions for nonconvex distributed control problems governed by elliptic variational inequalities*, J. Math. Anal. Appl., 80 (1981), pp. 566–597.
[4] V. Barbu, *Optimal Control of Variational Inequalities*, Pitman, London, 1984.
[5] V. Barbu, *Analysis and Control of Nonlinear Infinite Dimensional Systems*, Academic Press, New York, 1993.
[6] M. Bergounioux and S.M. Lenhart, *Optimal control of the obstacle in semilinear variational inequalities*, Positivity J., 8 (2004), pp. 229–242.
[7] M. Bergounioux and S.M. Lenhart, *Optimal control of bilateral obstacle problems*, SIAM J. Control Optim., 43 (2004), pp. 240–255.
[8] M. Bergounioux, *Optimal control of semilinear elliptic obstacle problem*, J. Nonlinear Convex Anal., 3 (2002), pp. 25–39.

[9] J.F. Bonnans and D. Tiba, *Pontryagin's principle in the control of semilinear elliptic variational inequalities*, Appl. Math. Optim., 23 (1991), pp. 299–312.
[10] E. Casas and L.A. Fernandez, *Distributed control of systems governed by a general class of quasilinear elliptic equations*, J. Differential Equations, 104 (1993), pp. 20–47.
[11] Q. Chen, *Indirect obstacle control problem for semilinear elliptic variational inequalities*, SIAM J. Control Optim., 38 (1999), pp. 138–158.
[12] Q. Chen, *Indirect obstacle optimal control for evolutionary variational inequalities with state constraints*, Sci. China (Ser. E), 43 (2000), pp. 653–669.
[13] Q. Chen, *Indirect obstacle minimax control for elliptic variational inequalities*, J. Optim. Theory Appl., 110 (2001), pp. 337–359.
[14] Q. Chen, *Optimal control of semilinear elliptic variational bilateral problem*, Acta Math. Sin., 16 (2000), pp. 123–140.
[15] Q. Chen, *Optimal control for semilinear evolutionary variational bilateral problem*, J. Math. Anal. Appl., 277 (2003), pp. 303–323.
[16] Q. Chen, *Minimax control for elliptic variational bilateral problem*, ANZIAM J., 44 (2003), pp. 539–559.
[17] J.I. Diaz, *Nonlinear Partial Differential Equations and Free Boundaries, Vol.* I: *Elliptic Equations*, Research Notes in Math. 106, Pitman, London, 1985.
[18] A. Friedman, *Variational Principles and Free-Boundary Problems*, John Wiley and Sons, New York, 1982.
[19] A. Friedman, *Optimal control for variational inequalities*, SIAM J. Control Optim., 24 (1986), pp. 439–451.
[20] A. Friedman, *Optimal control for parabolic variational inequalities*, SIAM J. Control Optim., 25 (1987), pp. 482–497.
[21] M. Fuchs and N. Fusco, *Partial regularity results for vector valued functions which minimize certain functions having nonquadratic growth under smooth side conditions*, J. Reine Angew. Math., 390 (1988), pp. 67–78.
[22] J. Heinonen, T. Kilpeläinen, and O. Martio, *Nonlinear Potential Theory of Second Order Degenerate Elliptic Partial Differential Equations*, Oxford University Press, Oxford, 1993.
[23] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1980.
[24] X. Li and J. Yong, *Optimal Control Theory for Infinite Dimensional Systems*, Birkhäuser Boston, Boston, 1995.
[25] H. Lou, *On the regularity of an obstacle control problem*, J. Math. Anal. Appl., 258 (2001), pp. 32–51.
[26] H. Lou, *An optimal control problem governed by quasi-linear variational inequalities*, SIAM J. Control Optim., 41 (2002), pp. 1229–1253.
[27] F. Mignot, *Contrôle dans les inequations variationelles elliptiques*, *J. Funct. Anal.*, 22 (1976), pp. 130–185.
[28] J.F. Rodrigues, *Obstacle Problems in Mathematical Physics*, North-Holland Math. Stud. 134, Elsevier Science, Amsterdam, 1987.
[29] G.M. Troianiello, *Elliptic Differential Equations and Obstacle Problems*, Plenum Press, New York, 1987.
[30] J. Yong, *Pontryagin maximum principle for semilinear second order elliptic partial differential equations and variational inequalities with state constraints*, Differential Integral Equations, 5 (1992), pp. 1307–1334.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 3, pp. 1081–1104

STATE AND MODE ESTIMATION FOR DISCRETE-TIME JUMP MARKOV SYSTEMS*

ROBERT J. ELLIOTT†, FRANCOIS DUFOUR‡, AND W. P. MALCOLM§

Abstract. In this article we compute new state and mode estimation algorithms for discrete-time Gauss–Markov models whose parameter sets switch according to a known Markov law. An important feature of our algorithms is that they are based upon the exact filter dynamics computed in [R. J. Elliott, F. Dufour, and D. Sworder, *IEEE Trans. Automat. Control*, 41 (1996), pp. 1807–1810].

The fundamental and well-known obstacle in estimation of jump Markov systems is managing the geometrically growing history of candidate hypotheses. In our scheme, we address this issue by proposing an extension of an idea due to Viterbi. Our scheme maintains a fixed number of candidate paths in a history, each identified by an optimal subset of estimated mode probabilities.

We compute finite-dimensional suboptimal filters and smoothers, which estimate the hidden state process and the mode probability. Our smoothers are based upon a duality between forward and backward dynamics. Further, our smoothing algorithms are general and can be configured into the standard forms of fixed point, fixed lag, and fixed interval smoothers. A computer simulation is included to demonstrate performance.

Key words. reference probability, jump Markov systems, hybrid dynamics, Viterbi algorithm, filtering, smoothing

AMS subject classifications. 93E11, 93E14, 60G35

DOI. 10.1137/S0363012904442628

1. Introduction.

In this article the reference probability method is used to compute state and mode estimation schemes for a discrete-time stochastic hybrid dynamical system.

Gauss–Markov jump linear systems arise quite naturally in many practical settings, for example, tracking a maneuvering object observed through radar. Here, no single set of dynamics will encapsulate all classes of motion, so one is led naturally to a hybrid collection of dynamics as a basic model. The estimation task for such models is significantly complicated by the need to jointly estimate a hidden state variable and the current model in effect. Currently, many of the standard techniques to solve this problem are ad hoc and not based upon the exact hybrid filter dynamics, which were presented in [5]. In contrast to this situation, our new filters and smoothers for Gauss–Markov jump linear systems are developed from the exact hybrid dynamics. Using a general result (see Lemma 2 in section 3), we propose a new suboptimal algorithm which provides an exact hypothesis management scheme, circumventing geometric growth in algorithmic complexity. Our approach is based, in part, upon approximating probability densities by finite Gaussian mixtures and is justified by the basic results given in [15].

In a simulation study, a comparison is given between the single extended Kalman filter (EKF), the IMM, and our algorithm. We also compute a general smoothing

*Received by the editors March 29, 2004; accepted for publication (in revised form) February 19, 2005; published electronically October 3, 2005.

http://www.siam.org/journals/sicon/44-3/44262.html

†Haskayne School of Business, University of Calgary, 2500 University Drive NW, Calgary AB, T2N 14N, Canada (relliott@ucalgary.ca).

‡Mathematiques Appliquees de Bordeaux Universite Bordeaux, 351 Cours de la Liberation, 33405 Talence Cedex, France (dufour@math.u-bordeaux.fr).

§National ICT Australia, Locked Bag 8001, Canberra ACT 2601, Australia (paul.malcolm@nicta.com.au).

algorithm for discrete-time hybrid dynamical system. To compute this algorithm, we exploit a duality between forward and backward (dual) dynamics. An interesting feature of our smoother is that it provides a new degree of freedom, that is, the product decomposition of the smoother density is approximated by mutually independent Gaussian mixtures, where the chosen accuracy of "the past" (influencing the smoother density) is independent of the chosen accuracy of "the future" (influencing the smoother density).

This paper is organized as follows. In section 2 we define the hybrid system dynamics and the reference probability measure. In section 3, we compute an exact filter for a hybrid dynamical system. Here, we suppose that our filter probability densities can be represented as finite Gaussian mixtures. This leads to a suboptimal filter whose memory requirements are fixed in time. Identities are given for the filter conditional mean estimate of the state and the filter conditional mean estimate of the corresponding state error covariance. In section 4 we compute an exact smoother for a hybrid system. In this section, we again suppose that certain functions can be represented as Gaussian mixtures. By using the same techniques as in section 3, we compute a suboptimal smoothing algorithm whose memory requirements are fixed in time. Identities are given for the smoother conditional mean estimate of the state and the smoother conditional mean estimate of the corresponding state error covariance. Finally, in section 5, we present a simulation study comparing the IMM and the extended Kalman filter to the new filter presented in section 3.

2. Dynamics and reference probability. Initially we suppose that all processes are defined on a fixed probability space $(\Omega, \mathcal{F}, P)$. For the class of jump Markov systems considered, we will require three sets of dynamics. These are the Markov chain dynamics for the process whose value determines the model parameters, the indirectly observed state process, and the observation process.

2.1. Markov chain dynamics. To model parameter switching we consider a time homogeneous discrete-time discrete-state Markov chain Z. If the cardinality of the state space is m, it is convenient, without loss of generality, to identify the state space of Z with an orthonormal basis indicator functions, which we denote by $\mathcal{L}$, that is,

$$\mathcal{L} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m\} = \left\{ \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \ldots, \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \right\} \subseteq \mathbb{R}^m. \tag{2.1}$$

The dynamics for the process Z may be written as

$$Z_k = \Pi Z_{k-1} + L_k \in \mathbb{R}^m. \tag{2.2}$$

Here, $\Pi = [\pi_{(j,i)}]_{\substack{1 \le j \le m \\ 1 \le i \le m}}$ is the transition matrix of Z, with elements

$$\pi_{(j,i)} \triangleq P(Z_k = \boldsymbol{e}_j \mid Z_{k-1} = \boldsymbol{e}_i) \tag{2.3}$$

for all $k \in \mathbb{N}$. The process L is a $(P, \sigma\{Z\})$-martingale increment, and we suppose $E[Z_0] = p_0$.

2.2. State process dynamics. We suppose that the indirectly observed state vector $x \in \mathbb{R}^n$ has dynamics

$$x_k = \sum_{j=1}^{m} \langle Z_k, \boldsymbol{e}_j \rangle A_j x_{k-1} + \sum_{j=1}^{m} \langle Z_k, \boldsymbol{e}_j \rangle B_j w_k. \tag{2.4}$$

Here w is a vector-valued Gaussian process with $w \sim N(0, \boldsymbol{I}_n)$. A_j and B_j are $n \times n$ matrices and, for each $j \in \{1, 2, \ldots, m\}$, are nonsingular. This condition can be relaxed [9].

2.3. Observation process dynamics. Consider a vector-valued observation process with values in $\mathbb{R}^d$ and dynamics

$$y_k = \sum_{j=1}^{m} \langle Z_k, \boldsymbol{e}_j \rangle C_j x_k + \sum_{j=1}^{m} \langle Z_k, \boldsymbol{e}_j \rangle D_j v_k. \tag{2.5}$$

Here v is a vector-valued Gaussian process with $v \sim N(0, \boldsymbol{I}_d)$. We suppose the matrices $D_j \in \mathbb{R}^{d \times d}$, for each $j \in \{1, 2, \ldots, m\}$, are nonsingular. The systems we shall consider in this article are described by the dynamics (2.2), (2.4), and (2.5). The three stochastic processes Z, x, and y are mutually statistically independent. Taken together, these dynamics form a triply stochastic system, with random inputs due to the processes Z, x, and y. For example, if the Markov chain Z is in the state $\boldsymbol{e}_j$, then the dynamical model with state x and observation y is defined by the parameters set $\{A_j, B_j, C_j, D_j\}$.

Remark 1. At the cost of more complicated notation, we could consider observations of the form

$$y_k = \sum_{j=1}^{m} \langle Z_k, \boldsymbol{e}_j \rangle \big(C_j x_k + H_j\big) + \sum_{j=1}^{m} \langle Z_k, \boldsymbol{e}_j \rangle D_j v_k. \tag{2.6}$$

This formulation includes scenarios where the Markov chain Z is observed directly. However, in this article we restrict our attention to cases where $H_j = 0$ for all $j \in \{1, 2, \ldots, m\}$. We define our filtrations as follows:

$$\mathbb{F}_k = \{\mathcal{F}_\ell\}_{0 \le \ell \le k}, \quad \text{where} \quad \mathcal{F}_k = \sigma\{x_\ell, 0 \le \ell \le k\}, \tag{2.7}$$

$$\mathbb{Z}_k = \{\mathcal{Z}_\ell\}_{0 \le \ell \le k}, \quad \text{where} \quad \mathcal{Z}_k = \sigma\{Z_\ell, 0 \le \ell \le k\}, \tag{2.8}$$

$$\mathbb{Y}_k = \{\mathcal{Y}_\ell\}_{0 \le \ell \le k}, \quad \text{where} \quad \mathcal{Y}_k = \sigma\{y_\ell, 0 \le \ell \le k\}, \tag{2.9}$$

$$\mathbb{G}_k = \{\mathcal{G}_\ell\}_{0 \le \ell \le k}, \quad \text{where} \quad \mathcal{G}_k = \sigma\{Z_\ell, x_\ell, y_\ell, 0 \le \ell \le k\}. \tag{2.10}$$

2.4. Reference probability. The dynamics given in (2.2), (2.4), and (2.5) are each defined on a measurable space $(\Omega, \mathcal{F})$ under a measure P. However, consider a new measure $P^\dagger$, under which the dynamics for the processes Z, x, and y are, respectively,

$$P^\dagger \quad \begin{cases} Z_k & = \Pi Z_{k-1} + L_k, \\ x_k & \text{are i.i.d. and } N(0, \boldsymbol{I}_n), \\ y_k & \text{are i.i.d. and } N(0, \boldsymbol{I}_d). \end{cases} \tag{2.11}$$

Notation. The symbol $\Phi(\cdot)$ will be used to denote the zero mean normal density on $\mathbb{R}^d$:

$$\Phi(\xi) = (2\pi)^{-d/2} \exp\bigl(-\tfrac{1}{2}\xi'\xi\bigr). \tag{2.12}$$

Similarly we shall also use the symbol $\Psi(\cdot)$ to denote a standardized Gaussian density. The space dimension on which these densities are defined will be clear by context. To avoid cumbersome notation with matrices, we sometimes denote the inverse of a matrix A by $\mathrm{inv}(A)$. Further, we denote the space of all $m \times n$ matrices by $\mathbb{M}^{m\times n}$. To compute the filter dynamics we now define the measure P by setting the restriction of its Radon–Nikodým derivative to $\mathcal{G}_k$ to

$$\Lambda_{0,k} \triangleq \frac{dP}{dP^\dagger}|_{\mathcal{G}_k} = \prod_{\ell=0}^{k} \lambda_\ell, \tag{2.13}$$

where

$$\lambda_\ell = \sum_{j=1}^{m} \langle Z_\ell, \boldsymbol{e}_j\rangle \frac{\Phi\bigl(D_j^{-1}(y_\ell - C_j x_\ell)\bigr)}{|D_j|\Phi(y_\ell)} \times \frac{\Psi\bigl(B_j^{-1}(x_\ell - A_j x_{\ell-1})\bigr)}{|B_j|\Psi(x_\ell)}.$$

The existence of P follows from the Kolmogorov extension theorem (see [16, Theorem 4, p. 166]). We quote the following form of Bayes' theorem (see [8]).

THEOREM 1. *Suppose $\gamma = \{\gamma_\ell, 0 \le \ell \le k\}$ is an integrable $\mathcal{G}$-adapted process. Then*

$$E[\gamma_k|\mathcal{Y}_k] = \frac{E^\dagger[\Lambda_{0,k}\gamma_k|\mathcal{Y}_k]}{E^\dagger[\Lambda_{0,k}|\mathcal{Y}_k]}. \tag{2.14}$$

As in [9] we can then show the following.

LEMMA 1. *Under the measure P, the dynamics for the Markov process Z are unchanged and x and y have dynamics given by* (2.4) *and* (2.5)*, respectively.*

3. Hybrid filter dynamics.

3.1. Exact filter dynamics. The following lemma is critical in what follows.

LEMMA 2. *Suppose the random vector $\xi \in \mathbb{R}^n$ is normally distributed with $\xi \sim N(\mu, \Sigma)$. Further, suppose A is any matrix in $\mathbb{M}^{m\times n}$, y is a vector in $\mathbb{R}^n$, and the matrix $B \in \mathbb{M}^{m\times m}$ is nonsingular. With $p(\xi)$ denoting the Gaussian density function for ξ, the following identity holds:*

$$\begin{aligned}
&\int_{\mathbb{R}^n} \Psi\bigl(B^{-1}(y - A\xi)\bigr)p(\xi)d\xi \\
&\quad = (2\pi)^{-n/2} E\bigl[\exp\bigl\{-\tfrac{1}{2}(y - A\xi)'\,\mathrm{inv}(BB')(y - A\xi)\bigr\}\bigr] \\
&\quad = (2\pi)^{-n/2}|B||BB' + A\Sigma A'|^{-\frac{1}{2}} \exp\bigl\{-\tfrac{1}{2}(y - A\mu)'(BB' + A\Sigma A')^{-1}(y - A\mu)\bigr\}. \qquad (3.1)
\end{aligned}$$

A proof of Lemma 2 is given in [6]. To compute a filter jointly estimating the density of the state vector x and the state of the chain Z, consider the expectation

$$E[\langle Z_k, \boldsymbol{e}_j\rangle f(x_k)|\mathcal{Y}_k] = \frac{E^\dagger[\Lambda_{0,k}\langle Z_k, \boldsymbol{e}_j\rangle f(x_k)|\mathcal{Y}_k]}{E^\dagger[\Lambda_{0,k}|\mathcal{Y}_k]}. \tag{3.2}$$

Here the function $f(\cdot)$ is an arbitrary bounded measurable real-valued test function. Write the numerator as

$$\sigma(\langle Z_k, \boldsymbol{e}_j\rangle f(x_k)) \triangleq E^\dagger[\Lambda_{0,k}\langle Z_k, \boldsymbol{e}_j\rangle f(x_k)|\mathcal{Y}_k]. \tag{3.3}$$

Suppose we choose the test function $f(x) = 1$. Then

$$\sum_{j=1}^m \sigma(\langle Z_k, \boldsymbol{e}_j\rangle) = \sigma\bigg(\sum_{j=1}^m \langle Z_k, \boldsymbol{e}_j\rangle\bigg) = \sigma(1) = E^\dagger[\Lambda_{0,k}|\mathcal{Y}_k]. \tag{3.4}$$

Therefore, if the numerator (3.3) can be evaluated for any such f and all j, the denominator of (3.2) can be found. For each $j \in \{1, 2, \ldots, m\}$, the quantity $\sigma\big(\langle Z_k, \boldsymbol{e}_j\rangle f(x_k)\big)$, defined at (3.3), is a continuous linear functional on the space of continuous functions and so defines a unique measure (see [14, Theorem 2.14, p. 40]). Further, suppose that there exists a corresponding unique, unnormalized density function $q_k^j(x)$ such that

$$\sigma(\langle Z_k, \boldsymbol{e}_j\rangle f(x_k)) \triangleq \int_{\mathbb{R}^n} f(\eta) q_k^j(\eta) d\eta. \tag{3.5}$$

Then the normalized conditional density is

$$P(x \in dx, Z_k = \boldsymbol{e}_j|\mathcal{Y}_k) = \frac{q_k^j(x)dx}{\int_{\mathbb{R}^n} q_k^j(\xi)d\xi}. \tag{3.6}$$

THEOREM 2. *The unnormalized probability density $q_k^j(x)$ satisfies the following integral-equation recursion:*

$$q_k^j(x) = \frac{\Phi\big(D_j^{-1}(y_k - C_j x)\big)}{\Phi(y_k)|D_j||B_j|} \sum_{r=1}^m \pi_{(j,r)} \int_{\mathbb{R}^n} \Psi\big(B_j^{-1}(x - A_j\xi)\big) q_{k-1}^r(\xi) d\xi. \tag{3.7}$$

Proof. Recalling the definition at (3.5), note that

$$\begin{aligned}
\int_{\mathbb{R}^n} f(\eta) q_k^j(\eta) d\eta &= \sigma\big(\langle Z_k, \boldsymbol{e}_j\rangle f(x_k)\big) \\
&= E^\dagger[\Lambda_{0,k}\langle Z_k, \boldsymbol{e}_j\rangle f(x_k)|\mathcal{Y}_k] \\
&= E^\dagger[\lambda_k \Lambda_{0,k-1}\langle Z_k, \boldsymbol{e}_j\rangle f(x_k)|\mathcal{Y}_k] \\
&= E^\dagger\bigg[\Lambda_{0,k-1}\langle Z_k, \boldsymbol{e}_j\rangle \frac{\Phi\big(D_j^{-1}(y_k - C_j x_k)\big)}{|D_j|\Phi(y_k)} \\
&\quad \times \frac{\Psi\big(B_j^{-1}(x_k - A_j x_{k-1})\big)}{|B_j|\Psi(x_k)} f(x_k)|\mathcal{Y}_k\bigg] \\
&= \frac{1}{\Phi(y_k)|D_j||B_j|} E^\dagger\bigg[\Lambda_{0,k-1}\langle \Pi Z_{k-1} + L_k, \boldsymbol{e}_j\rangle \\
&\quad \times \Phi\big(D_j^{-1}(y_k - C_j x_k)\big) \frac{\Psi\big(B_j^{-1}(x_k - A_j x_{k-1})\big)}{\Psi(x_k)} f(x_k)|\mathcal{Y}_k\bigg] \\
&= \frac{1}{\Phi(y_k)|D_j||B_j|} \sum_{r=1}^m \pi_{(j,r)} E^\dagger\bigg[\Lambda_{k-1}\langle Z_{k-1}, \boldsymbol{e}_r\rangle \\
&\quad \times \Phi\big(D_j^{-1}(y_k - C_j x_k)\big) \frac{\Psi\big(B_j^{-1}(x_k - A_j x_{k-1})\big)}{\Psi(x_k)} f(x_k)|\mathcal{Y}_{k-1}\bigg].
\end{aligned} \tag{3.8}$$

The final equality follows because under $P^\dagger$, the y_k are independent and $N(0, I_d)$. Further, since the expectations here are with respect to the measure $P^\dagger$, under which the random variables Z_k, x_k, y_k, and x_{k-1} are (mutually) independent, the expected value with respect to x_k is obtained by multiplying by the density $\Psi(x_k)$ and integrating; that is, with

$$
\begin{aligned}
H(x_{k-1};\Theta) &\triangleq \int_{\mathbb{R}^n} \left\{ \Phi\big(D_j^{-1}(y_k - C_j\eta)\big) \frac{\Psi\big(B_j^{-1}(\eta - A_j x_{k-1})\big)}{\Psi(\eta)} f(\eta) \right\} \Psi(\eta) d\eta \\
&= \int_{\mathbb{R}^n} \Phi\big(D_j^{-1}(y_k - C_j\eta)\big) \Psi\big(B_j^{-1}(\eta - A_j x_{k-1})\big) f(\eta) d\eta.
\end{aligned} \tag{3.9}
$$

Then

$$
\begin{aligned}
\int_{\mathbb{R}^n} f(\eta) q_k^j(\eta) d\eta &= \frac{1}{\Phi(y_k)|D_j||B_j|} \sum_{r=1}^m \pi_{(j,r)} E^\dagger[\Lambda_{k-1} \langle Z_{k-1}, \boldsymbol{e}_r \rangle H(x_{k-1};\Theta) | \mathcal{Y}_{k-1}] \\
&= \frac{1}{\Phi(y_k)|D_j||B_j|} \sum_{r=1}^m \pi_{(j,r)} \int_{\mathbb{R}^n} H(\xi;\Theta) q_{k-1}^r(\xi) d\xi \\
&= \frac{1}{\Phi(y_k)|D_j||B_j|} \sum_{r=1}^m \pi_{(j,r)} \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \Phi\big(D_j^{-1}(y_k - C_j\eta)\big) \\
&\quad \times \Psi\big(B_j^{-1}(\eta - A_j\xi)\big) q_{k-1}^r(\xi) f(\eta) d\eta d\xi.
\end{aligned} \tag{3.10}
$$

Finally, as (3.10) holds for any suitable test function f, it follows that

$$
q_k^j(x) = \frac{\Phi\big(D_j^{-1}(y_k - C_j x)\big)}{\Phi(y_k)|D_j||B_j|} \sum_{r=1}^m \pi_{(j,r)} \int_{\mathbb{R}^n} \Psi\big(B_j^{-1}(x - A_j\xi)\big) q_{k-1}^r(\xi) d\xi. \qquad \square \tag{3.11}
$$

The recursion given in (3.11) is an exact filter; it is expressed as a density which is in general infinite-dimensional. However, Gaussian densities are determined by their mean and variance. In what follows we will consider a finite Gaussian mixture representation of $q_{k-1}^r(x)$. The use of Gaussian sums is justified because any density can be approximated by the sum of Gaussian densities (see [15]). Further, the noise in the state and the observation processes is assumed to be Gaussian. In turn, this implies that sums of Gaussian densities are introduced. For example, if the initial state is known exactly, the conditional state at the next time is described by a sum of Gaussian densities, and the number of terms in the sum increases at each time step by a factor equal to the number of states in the chain.

We can then compute the integrals in (3.11) by using Lemma 2.

Suppose $\Sigma_\alpha, \Sigma_\mu \in \mathbb{M}^{n\times n}$ are covariance matrices and $x, \alpha, \mu \in \mathbb{R}^n$. The following identity will be useful in what follows:

$$
\begin{aligned}
&\exp\Big\{-\tfrac{1}{2}(x-\alpha)'\Sigma_\alpha(x-\alpha)\Big\} \exp\Big\{-\tfrac{1}{2}(x-\mu)'\Sigma_\mu(x-\mu)\Big\} \\
&\quad = \exp\Big\{-\tfrac{1}{2}\Big(\alpha'\Sigma_\alpha\alpha + \mu'\Sigma_\mu\mu\Big)\Big\} \exp\Big\{\tfrac{1}{2}\Big(\Sigma_\alpha\alpha + \Sigma_\mu\mu\Big)' \operatorname{inv}(\Sigma_\alpha + \Sigma_\mu)\Big(\Sigma_\alpha\alpha + \Sigma_\mu\mu\Big)\Big\} \\
&\qquad \times \exp\Big\{-\tfrac{1}{2}\Big(x - \operatorname{inv}(\Sigma_\alpha + \Sigma_\mu)(\Sigma_\alpha\alpha + \Sigma_\mu\mu)\Big)' (\Sigma_\alpha + \Sigma_\mu) \\
&\qquad\quad \times \Big(x - \operatorname{inv}(\Sigma_\alpha + \Sigma_\mu)(\Sigma_\alpha\alpha + \Sigma_\mu\mu)\Big)\Big\}.
\end{aligned} \tag{3.12}
$$

THEOREM 3. *Suppose the unnormalized probability density $q^r_{k-1}(\xi)$ (as it appears under the integral in* (3.11)*) can be written as a finite weighted Gaussian mixture with $M^q \in \mathbb{N}$ components. That is, for $k \in \{1, 2, \dots\}$, we suppose*

$$q^r_{k-1}(\xi) = \sum_{s=1}^{M^q} p^{r,s}_{k-1} \frac{1}{(2\pi)^{n/2}|\Sigma^{r,s}_{k-1|k-1}|^{\frac{1}{2}}} \times \exp\Big\{-\tfrac{1}{2}(\xi - \alpha^{r,s}_{k-1|k-1})' \operatorname{inv}(\Sigma^{r,s}_{k-1|k-1})(\xi - \alpha^{r,s}_{k-1|k-1})\Big\}. \tag{3.13}$$

Here $\Sigma^{r,s}_{k-1|k-1} \in \mathbb{M}^{n\times n}$ and $\alpha^{r,s}_{k-1|k-1} \in \mathbb{R}^n$, are both $\mathcal{Y}_{k-1}$-measurable functions for all pairs $(r, s) \in \{1, 2, \dots, m\} \times \{1, 2, \dots, M^q\}$. Using the Gaussian mixture (3.13)*, the recursion for the optimal unnormalized density process has the form*

$$q^j_k(x) \triangleq \frac{1}{(2\pi)^{(d+n)/2}\Phi(y_k)} \sum_{r=1}^{m}\sum_{s=1}^{M^q} K^q_{k,k-1}(j, r, s) \times \exp\Big\{-\tfrac{1}{2}\Big(x - \operatorname{inv}(\sigma^{j,r,s}_{k-1})\delta^{j,r,s}_{k,k-1}\Big)' \sigma^{j,r,s}_{k-1}\Big(x - \operatorname{inv}(\sigma^{j,r,s}_{k-1})\delta^{j,r,s}_{k,k-1}\Big)\Big\}. \tag{3.14}$$

The means of the exponential densities in (3.14) *are computed by the update equations*

$$\operatorname{inv}(\sigma^{j,r,s}_{k-1})\delta^{j,r,s}_{k,k-1} = A_r\alpha^{r,s}_{k-1|k-1} + \overline{\Sigma}^{j,r,s}_{k-1|k-1}C_r' \operatorname{inv}\Big(C_r\overline{\Sigma}^{j,r,s}_{k-1|k-1}C_r' + D_rD_r'\Big)\big(y_k - C_rA_r\alpha^{r,s}_{k-1|k-1}\big). \tag{3.15}$$

Here

$$\overline{\Sigma}^{j,r,s}_{k-1|k-1} \triangleq B_jB_j' + A_j\Sigma^{r,s}_{k-1|k-1}A_j' \in \mathbb{R}^{n\times n}, \tag{3.16}$$

$$\widetilde{u}^{j,r,s}_{k-1|k-1} \triangleq A_j\alpha^{r,s}_{k-1|k-1} \in \mathbb{R}^n, \tag{3.17}$$

$$\sigma^{j,r,s}_{k-1} \triangleq C_r' \operatorname{inv}(D_rD_r')C_r + \operatorname{inv}(\overline{\Sigma}^{j,r,s}_{k-1|k-1}), \tag{3.18}$$

$$\delta^{j,r,s}_{k,k-1} \triangleq \operatorname{inv}(\overline{\Sigma}^{j,r,s}_{k-1|k-1})\widetilde{u}^{j,r,s}_{k-1|k-1} + C_r' \operatorname{inv}(D_rD_r')y_k, \tag{3.19}$$

$$K^q_{k,k-1}(j, r, s) \triangleq \frac{\pi_{(j,r)}p^{r,s}_{k-1}}{|\overline{\Sigma}^{j,r,s}_{k-1|k-1}|^{\frac{1}{2}}|D_j|} \exp\Big\{\tfrac{1}{2}(\delta^{j,r,s}_{k,k-1})' \operatorname{inv}(\sigma^{j,r,s}_{k-1})\delta^{j,r,s}_{k,k-1}\Big\} \times \exp\Big\{-\tfrac{1}{2}\big[y_k' \operatorname{inv}(D_rD_r')y_k + (\widetilde{u}^{j,r,s}_{k-1|k-1})' \operatorname{inv}(\overline{\Sigma}^{j,r,s}_{k-1|k-1})\widetilde{u}^{j,r,s}_{k-1|k-1}\big]\Big\} \tag{3.20}$$

(Note that the square matrices $\overline{\Sigma}^{j,r,s}_{k-1|k-1}$ and $\sigma^{j,r,s}_{k-1}$ are symmetric.)

Proof. To prove Theorem 3, we apply the identity given in Lemma 2 to the recursion at (3.7). In the first application, we eliminate the integral in (3.7). Using the finite-mixture representation (3.13), we write the recursion (3.7) for the function $q^j(x)$ as

$$q^j_k(x) = \frac{\Phi\big(D_j^{-1}(y_k - C_jx)\big)}{\Phi(y_k)|D_j||B_j|} \sum_{r=1}^{m} \pi_{(j,r)} \Bigg\{\sum_{s=1}^{M^q} p^{r,s}_{k-1} \Bigg[\int_{\mathbb{R}^n} \Psi\big(B_j^{-1}(x - A_j\xi)\big) \times \frac{1}{(2\pi)^{n/2}|\Sigma^{r,s}_{k-1|k-1}|^{\frac{1}{2}}} \exp\Big\{-\tfrac{1}{2}\Big(\xi - \alpha^{r,s}_{k-1|k-1}\Big)' \times \operatorname{inv}(\Sigma^{r,s}_{k-1|k-1})\Big(\xi - \alpha^{r,s}_{k-1|k-1}\Big)\Big\} d\xi\Bigg]\Bigg\}. \tag{3.21}$$

Recalling Lemma 2, we see that

(3.22)
$$\begin{aligned}
&\int_{\mathbb{R}^n} \Psi\big(B_j^{-1}(x - A_j\xi)\big) \frac{1}{(2\pi)^{n/2}|\Sigma^{r,s}_{k-1|k-1}|^{\frac{1}{2}}} \\
&\qquad \times \exp\big\{-\tfrac{1}{2}\big(\xi - \alpha^{r,s}_{k-1|k-1}\big)' \operatorname{inv}\big(\Sigma^{r,s}_{k-1|k-1}\big)\big(\xi - \alpha^{r,s}_{k-1|k-1}\big)\big\} d\xi \\
&\quad = \frac{1}{(2\pi)^{n/2}} |B_j| |B_jB_j' + A_j\Sigma^{r,s}_{k-1|k-1}A_j'|^{-\frac{1}{2}} \\
&\qquad \times \exp\big\{-\tfrac{1}{2}\big(x - A_j\alpha^{r,s}_{k-1|k-1}\big)'\big(B_jB_j' + A_j\Sigma^{r,s}_{k-1|k-1}A_j'\big)^{-1}\big(x - A_j\alpha^{r,s}_{k-1|k-1}\big)\big\}.
\end{aligned}$$

Using this calculation and the definitions for the quantities $\overline{\Sigma}^{j,r,s}_{k-1|k-1}$ and $\widetilde{u}^{j,r,s}_{k-1|k-1}$, we can write

$$\begin{aligned}
q_k^j(x) = \sum_{r=1}^{m}\sum_{s=1}^{M^q} &\frac{\pi_{(j,r)}p^{r,s}_{k-1}}{(2\pi)^{n/2}|\overline{\Sigma}^{j,r,s}_{k-1|k-1}|^{\frac{1}{2}}\Phi(y_k)|D_j|}\Phi\big(D_j^{-1}(y_k - C_jx)\big) \\
&\times \exp\Big\{-\tfrac{1}{2}\Big(x - \widetilde{u}^{j,r,s}_{k-1|k-1}\Big)' \operatorname{inv}(\overline{\Sigma}^{j,r,s}_{k-1|k-1})\Big(x - \widetilde{u}^{j,r,s}_{k-1|k-1}\Big)\Big\}.
\end{aligned} \tag{3.23}$$

The products of exponential functions in this equation can be simplified with the identity (3.12), resulting in

$$\begin{aligned}
q_k^j(x) = \frac{1}{(2\pi)^{(d+n)/2}\Phi(y_k)} \sum_{r=1}^{m}\sum_{s=1}^{M^q} &\frac{\pi_{(j,r)}p^{r,s}_{k-1}}{|\overline{\Sigma}^{j,r,s}_{k-1|k-1}|^{\frac{1}{2}}|D_j|} \exp\Big\{\tfrac{1}{2}(\delta^{j,r,s}_{k,k-1})' \operatorname{inv}(\sigma^{j,r,s}_{k-1})\delta^{j,r,s}_{k,k-1}\Big\} \\
&\times \exp\Big\{-\tfrac{1}{2}\Big[y_k' \operatorname{inv}(D_rD_r')y_k + (\widetilde{u}^{j,r,s}_{k-1|k-1})' \operatorname{inv}(\overline{\Sigma}^{j,r,s}_{k-1|k-1})\widetilde{u}^{j,r,s}_{k-1|k-1}\Big]\Big\} \\
&\times \exp\Big\{-\tfrac{1}{2}\Big(x - \operatorname{inv}(\sigma^{j,r,s}_{k-1})\delta^{j,r,s}_{k}\Big)' \sigma^{j,r,s}_{k-1}\Big(x - \operatorname{inv}(\sigma^{j,r,s}_{k-1})\delta^{j,r,s}_{k}\Big)\Big\}.
\end{aligned} \tag{3.24}$$

Finally, (3.15) is computed by the matrix inversion lemma. □

Remark 2. It is interesting to note the similarity of Theorem 3 to the Kalman filter. The mean update equation given by (3.15) is precisely the Kalman filter state update equation with Kalman gain $\overline{\Sigma}^{j,r,s}_{k-1|k-1}C_r' \operatorname{inv}(C_r\overline{\Sigma}^{j,r,s}_{k-1|k-1}C_r' + D_rD_r')$.

COROLLARY 1. *The estimated conditional mode probability for model j is given by*

$$p_k^j \triangleq P(Z_k = \boldsymbol{e}_j|\mathcal{Y}_k) = \frac{\int_{\mathbb{R}^n} q_k^j(\xi)d\xi}{\sum_{\ell=1}^m \int_{\mathbb{R}^n} q_k^\ell(\xi)d\xi} = \frac{q_k^j}{\sum_{\ell=1}^m q_k^\ell}. \tag{3.25}$$

Here the estimated unnormalized mode probability q_k^j is computed by the double summation

$$q_k^j = \frac{1}{(2\pi)^{d/2}\Phi(y_k)} \sum_{r=1}^{m}\sum_{s=1}^{M^q} \zeta^q_{k,k-1}(j,r,s), \tag{3.26}$$

where

$$\zeta^q_{k,k-1}(j,r,s) \triangleq K^q_{k,k-1}(j,r,s)|\sigma^{j,r,s}_{k-1}|^{-\frac{1}{2}}. \tag{3.27}$$

Proof. The proof of Corollary 1 is immediate.

$$
\begin{aligned}
q_k^j &\triangleq \int_{\mathbb{R}^n} q_k^j(\xi)d\xi \\
&= \frac{1}{(2\pi)^{(d+n)/2}\Phi(y_k)} \sum_{r=1}^{m}\sum_{s=1}^{M^q} K_{k,k-1}^q(j,r,s) \\
&\quad \times \int_{\mathbb{R}^n} \exp\Big\{-\tfrac{1}{2}\Big(\xi - \mathrm{inv}(\sigma_{k-1}^{j,r,s})\delta_{k,k-1}^{j,r,s}\Big)' \sigma_{k-1}^{j,r,s}\Big(\xi - \mathrm{inv}(\sigma_{k-1}^{j,r,s})\delta_{k,k-1}^{j,r,s}\Big)\Big\}d\xi \\
&= \frac{1}{(2\pi)^{d/2}\Phi(y_k)} \sum_{r=1}^{m}\sum_{s=1}^{M^q} K_{k,k-1}^q(j,r,s)|\sigma_{k-1}^{j,r,s}|^{-\frac{1}{2}}. \qquad \square
\end{aligned} \tag{3.28}
$$

Remark 3. The scalar-valued quantities $\zeta_{k,k-1}^q(j,r,s)$ are all nonnegative weights. These quantities form the double sum (3.26), whose value is the estimated mode probability q_k^j. In what follows, we shall exploit the form of (3.27) to fix the memory requirements of our suboptimal filters. Note that if $\sigma_{k-1}^{j,r,s}$ is relatively large, then $|\sigma_{k-1}^{j,r,s}|^{-\frac{1}{2}}$ makes a small contribution to (3.26). That is, larger (error) covariances give smaller weights in the sum (3.26). Accordingly, a subset of the terms $\zeta_{k,k-1}^q(j,r,s)$ will be identified from the set of all such $m \times M^q$ quantities. The elements of this subset will be identified through their relative magnitudes, as the most substantial contributors to the sum (3.26) and therefore the quantity q_k^j. Ultimately, this subset of quantities will be used to identify, at every time k, the most significant components of the Gaussian mixture for the density $q_k(x)$.

3.2. Suboptimal filter dynamics. In this section we develop a suboptimal recursive filter by extending an idea due to Viterbi [18]. The motivation to develop a suboptimal filter is immediate from the dynamics of the unnormalized density (3.14). Suppose at time $k = 1$, these dynamics involve $m \times M^q$ densities. Then the next time, $k = 2$, these dynamics require $m \times m \times M^q$ densities. It is clear that the demand on memory requirements is exponential in time, with the number of densities required at time k being $m^k \times M^q$. What we wish to do is to circumvent this growth by identifying a subset of candidate mixture densities, from which we construct a suboptimal density with fixed (in-time) memory requirements.

3.2.1. Hypothesis management. Write

$$
\Gamma^q \triangleq \{1,2,\ldots,m\} \times \big\{1,2,\ldots,M^q\big\}, \tag{3.29}
$$

$$
\widetilde{S}_{k,k-1}^q(j,r,s) \triangleq \big\{\zeta_{k,k-1}^q(j,r,s)\big\}_{(r,s)\in\Gamma^q}. \tag{3.30}
$$

To remove the growth in memory requirements, we propose to identify, at each time k, the M^q-best candidate densities for each suboptimal density $q_k^j(x)$, using the corresponding estimated mode probabilities. The key to this idea is to identify M^q optimal densities, that is, the M^q components in the Gaussian mixture, through their corresponding set of estimated mode probabilities q^j, $j \in \{1,2,\ldots,m\}$.

Since the estimated mode probabilities, given by (3.26), are formed by a summation over nonnegative quantities, we can identify the M^q largest contributors to this sum and then use the corresponding indexes to identify the M^q-best Gaussian

densities. This maximization procedure is as follows:

$$\zeta^q_{k,k-1}(j, r^*_{k,1}, s^*_{k,1}) \triangleq \max_{(r,s)\,\in\,\Gamma^q} \widetilde{S}^q_{k,k-1}(j, r, s), \tag{3.31}$$

$$\zeta^q_{k,k-1}(j, r^*_{k,2}, s^*_{k,2}) \triangleq \max_{(r,s)\,\in\,\Gamma^q\setminus(r^*_{k,1}, s^*_{k,1})} \widetilde{S}^q_{k,k-1}(j, r, s), \tag{3.32}$$

$$\vdots \qquad\qquad \vdots$$

$$\zeta^q_{k,k-1}(j, r^*_{k,M^q}, s^*_{k,M^q}) \triangleq \max_{(r,s)\,\in\,\Gamma^q\setminus\{(r^*_{k,1}, s^*_{k,1}),\ldots,(r^*_{k,M^q-1}, s^*_{k,M^q-1})\}} \widetilde{S}^q_{k,k-1}(j, r, s). \tag{3.33}$$

Note that we are not directly interested in the quantities $\zeta^q_{k,k-1}(j, r^*_{k,\ell}, s^*_{k,\ell})$, but rather in the indexes that locate these quantities. The optimal index set, for the density of model j, is

$$\mathcal{I}^j_k \triangleq \big\{(r^*_{k,1}, s^*_{k,1}), (r^*_{k,2}, s^*_{k,2}), \ldots, (r^*_{k,M^q}, s^*_{k,M^q})\big\}. \tag{3.34}$$

Using these indexes, we approximate the suboptimal unnormalized density, of order M^q, corresponding to the density $q^j_k(x)$ as

(3.35)

$$q^j_k(x) \triangleq \frac{1}{(2\pi)^{(d+n)/2}\Phi(y_k)} \sum_{\ell=1}^{M^q} K^q_{k,k-1}(j, r^*_{k,\ell}, s^*_{k,\ell})$$
$$\times \exp\Big\{-\tfrac{1}{2}\Big(x - \operatorname{inv}(\sigma^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k-1})\delta^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k,k-1}\Big)' \sigma^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k-1} \Big(x - \operatorname{inv}(\sigma^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k-1})\delta^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k,k-1}\Big)\Big\}.$$

Further, by Corollary 1, the unnormalized mode probability corresponding to the expectation $E[Z_k = \boldsymbol{e}_j|\mathcal{Y}_k]$ is approximated by

$$q^j_k \triangleq \frac{1}{(2\pi)^{d/2}\Phi(y_k)} \sum_{\ell=1}^{M^q} \zeta^q_{k,k-1}(j, r^*_{k,\ell}, s^*_{k,\ell}). \tag{3.36}$$

To normalize the function $q^j_k(x)$, we need the sum of all terms $\{q^1_k, \ldots, q^m_k\}$ and so write

$$\varphi_{k|k} \triangleq \sum_{j=1}^{m} q^j_k. \tag{3.37}$$

3.2.2. Filter state statistics. Densities such as $q^j_k(x)$ provide all the information available. However, what is often required in state estimation (filtering) is an expression for the state estimate of x_k at time k given $\mathcal{Y}_k$, that is,

$$\widehat{x}_{k|k} \triangleq E\big[x_k|\mathcal{Y}_k\big]. \tag{3.38}$$

PROPOSITION 1. *The suboptimal state estimate $\widehat{x}_{k|k}$, for a Gaussian mixture of order M^q, has the representation*

$$\widehat{x}_{k|k} = \frac{1}{(2\pi)^{d/2}\Phi(y_k)\varphi_{k|k}} \sum_{j=1}^{m}\sum_{\ell=1}^{M^q} \frac{K^q_{k,k-1}(j, r^*_{k,\ell}, s^*_{k,\ell})}{|\sigma^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k-1}|^{\frac{1}{2}}}$$
$$\times\Big[A_{r^*_{k,\ell}}\alpha^{r^*_{k,\ell},s^*_{k,\ell}}_{k-1|k-1} + \overline{\Sigma}^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k-1|k-1} C'_{r^*_{k,\ell}} \operatorname{inv}\Big(D_{r^*_{k,\ell}} D'_{r^*_{k,\ell}} + C_{r^*_{k,\ell}} \overline{\Sigma}^{j,r,s}_{k-1|k-1} C'_{r^*_{k,\ell}}\Big)$$
$$\times\big(y_k - C_{r^*_{k,\ell}} A_{r^*_{k,\ell}} \alpha^{r^*_{k,\ell},s^*_{k,\ell}}_{k-1|k-1}\big)\Big]. \tag{3.39}$$

Here

$$\varphi_{k|k} = \frac{1}{(2\pi)^{d/2}\Phi(y_k)} \sum_{j=1}^{m} \sum_{\ell=1}^{M^q} K^q_{k,k-1}(j, r^*_{k,\ell}, s^*_{k,\ell}) |\sigma^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k-1}|^{-\frac{1}{2}}. \tag{3.40}$$

Proof.

$$\begin{aligned}
E\big[x_k|\mathcal{Y}_k\big] &= \frac{1}{\varphi_{k|k}} \int_{\mathbb{R}^n} \xi \Big[\sum_{j=1}^{m} q^j_k(\xi)\Big] d\xi \\
&= \frac{1}{\varphi_{k|k}} \sum_{j=1}^{m} \Big[\int_{\mathbb{R}^n} \xi\, q^j_k(\xi) d\xi\Big] \\
&= \frac{1}{\varphi_{k|k}} \sum_{j=1}^{m} \int_{\mathbb{R}^n} \xi \Bigg[\frac{1}{(2\pi)^{(d+n)/2}\Phi(y_k)} \sum_{\ell=1}^{M^q} K^q_{k,k-1}(j, r^*_{k,\ell}, s^*_{k,\ell}) \\
&\qquad \times \exp\Big\{-\tfrac{1}{2}\Big(\xi - \mathrm{inv}(\sigma^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k-1})\delta^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k,k-1}\Big)' \sigma^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k-1} \\
&\qquad\qquad \times \Big(\xi - \mathrm{inv}(\sigma^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k-1})\delta^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k,k-1}\Big)\Big\}\Bigg] d\xi \\
&= \frac{1}{(2\pi)^{(d+n)/2}\Phi(y_k)\varphi_{k|k}} \sum_{j=1}^{m} \sum_{\ell=1}^{M^q} K^q_{k,k-1}(j, r^*_{k,\ell}, s^*_{k,\ell}) \\
&\qquad \times \int_{\mathbb{R}^n} \xi \Bigg[\exp\Big\{-\tfrac{1}{2}\Big(\xi - \mathrm{inv}(\sigma^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k-1})\delta^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k,k-1}\Big)' \sigma^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k-1} \\
&\qquad\qquad \times \Big(\xi - \mathrm{inv}(\sigma^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k-1})\delta^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k,k-1}\Big)\Big\}\Bigg] d\xi \\
&= \frac{1}{(2\pi)^{d/2}\Phi(y_k)\varphi_{k|k}} \sum_{j=1}^{m} \sum_{\ell=1}^{M^q} \frac{K^q_{k,k-1}(j, r^*_{k,\ell}, s^*_{k,\ell})}{|\sigma^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k-1}|^{\frac{1}{2}}} \mathrm{inv}(\sigma^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k-1})\delta^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k,k-1}.
\end{aligned} \tag{3.41}$$

The proof is completed by applying the matrix inversion lemma to the term $\mathrm{inv}(\sigma^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k-1})\delta^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k,k-1}$. □

It is also of practical interest to have a relation for the corresponding error covariance, that is,

$$\begin{aligned}
\Sigma_{k|k} &\triangleq E\big[(x_k - \widehat{x}_{k|k})(x - \widehat{x}_{k|k})'|\mathcal{Y}_k\big] \\
&= \frac{1}{\varphi_{k|k}} \int_{\mathbb{R}^n} (\xi - \widehat{x}_{k|k})(\xi - \widehat{x}_{k|k})' \left(\sum_{j=1}^{m} q^j_k(\xi)\right) d\xi = \widehat{\Sigma}_{k|k}, \text{ say.}
\end{aligned} \tag{3.42}$$

PROPOSITION 2. *The filter state error covariance $\widehat{\Sigma}_{k|k}$, for a Gaussian mixture of order M^q, has the following representation:*

$$\begin{aligned}
\widehat{\Sigma}_{k|k} = \frac{1}{(2\pi)^{d/2}\Phi(y_k)\varphi_{k|k}} \sum_{j=1}^{m} \sum_{\ell=1}^{M^q} \frac{K^q_{k,k-1}(j, r^*_{k,\ell}, s^*_{k,\ell})}{|\sigma^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k-1}|^{\frac{1}{2}}} \Big[\mathrm{inv}(\sigma^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k-1}) \\
+ \Big(\mathrm{inv}(\sigma^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k-1})\delta^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k,k-1} - \widehat{x}_{k|k}\Big)\Big(\mathrm{inv}(\sigma^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k-1})\delta^{j,r^*_{k,\ell},s^*_{k,\ell}}_{k,k-1} - \widehat{x}_{k|k}\Big)'\Big].
\end{aligned} \tag{3.43}$$

Proof. Recalling the approximate unnormalized density (3.35), we see that

$$\begin{aligned}\widehat{\Sigma}_{k|k} &= \frac{1}{(2\pi)^{(d+n)/2}\Phi(y_k)\varphi_{k|k}} \sum_{j=1}^{m}\sum_{\ell=1}^{M^q} K^q_{k,k-1}(j, r^*_{k,\ell}, s^*_{k,\ell}) \int_{\mathbb{R}^n} \big(\xi - \widehat{x}_{k|k}\big)\big(\xi - \widehat{x}_{k|k}\big)' \\ &\quad \times \Bigg[\exp\Big\{ -\tfrac{1}{2}\Big(\xi - \mathrm{inv}(\sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}})\delta_{k,k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}}\Big)' \sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}} \\ &\qquad \times \Big(\xi - \mathrm{inv}(\sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}})\delta_{k,k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}}\Big)\Big\}\Bigg] d\xi \\ &= \frac{1}{(2\pi)^{d/2}\Phi(y_k)\varphi_{k|k}} \sum_{j=1}^{m}\sum_{\ell=1}^{M^q} \frac{K^q_{k,k-1}(j, r^*_{k,\ell}, s^*_{k,\ell})}{|\sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}}|^{\frac{1}{2}}} \int_{\mathbb{R}^n} \big(\xi - \widehat{x}_{k|k}\big)\big(\xi - \widehat{x}_{k|k}\big)' \\ &\quad \times \Bigg[\frac{|\sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}}|^{\frac{1}{2}}}{(2\pi)^{n/2}} \exp\Big\{ -\tfrac{1}{2}\Big(\xi - \mathrm{inv}(\sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}})\delta_{k,k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}}\Big)' \sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}} \\ &\qquad \times \Big(\xi - \mathrm{inv}(\sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}})\delta_{k,k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}}\Big)\Big\}\Bigg] d\xi. \end{aligned} \tag{3.44}$$

Writing

$$\big(\xi - \widehat{x}_{k|k}\big) = \Big(\xi - \mathrm{inv}(\sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}})\delta_{k,k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}} + \mathrm{inv}(\sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}})\delta_{k,k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}} - \widehat{x}_{k|k}\Big), \tag{3.45}$$

(3.43) is the sum of the integrals

(3.46)
$$\begin{aligned}&\frac{|\sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}}|^{\frac{1}{2}}}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \Big(\xi - \mathrm{inv}(\sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}})\delta_{k,k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}}\Big)\Big(\xi - \mathrm{inv}(\sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}})\delta_{k,k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}}\Big)' \\ &\quad \times \exp\Big\{ -\tfrac{1}{2}\Big(\xi - \mathrm{inv}(\sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}})\delta_{k,k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}}\Big)' \sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}} \\ &\qquad \times \Big(\xi - \mathrm{inv}(\sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}})\delta_{k,k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}}\Big)\Big\} d\xi = \mathrm{inv}\big(\sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}}\big), \end{aligned}$$

(3.47)
$$\begin{aligned}&\frac{|\sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}}|^{\frac{1}{2}}}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \Big(\xi - \mathrm{inv}(\sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}})\delta_{k,k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}}\Big) \\ &\quad \times \exp\Big\{ -\tfrac{1}{2}\Big(\xi - \mathrm{inv}(\sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}})\delta_{k,k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}}\Big)' \sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}} \\ &\qquad \times \Big(\xi - \mathrm{inv}(\sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}})\delta_{k,k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}}\Big)\Big\} d\xi (\mathrm{inv}(\sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}})\delta_{k,k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}} - \widehat{x}_{k|k})' = 0, \end{aligned}$$

(3.48)
$$\begin{aligned}&\frac{|\sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}}|^{\frac{1}{2}}}{(2\pi)^{n/2}} \Big(\mathrm{inv}(\sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}})\delta_{k,k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}} - \widehat{x}_{k|k}\Big)' \int_{\mathbb{R}^n} \Big(\xi - \mathrm{inv}(\sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}})\delta_{k,k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}}\Big) \\ &\quad \times \exp\Big\{ -\tfrac{1}{2}\Big(\xi - \mathrm{inv}(\sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}})\delta_{k,k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}}\Big)' \sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}} \\ &\qquad \times \Big(\xi - \mathrm{inv}(\sigma_{k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}})\delta_{k,k-1}^{j,r^*_{k,\ell},s^*_{k,\ell}}\Big)\Big\} d\xi = 0, \end{aligned}$$

and

(3.49)
$$\begin{aligned}
&\frac{|\sigma_{k-1}^{j,r_{k,\ell}^*,s_{k,\ell}^*}|^{\frac{1}{2}}}{(2\pi)^{n/2}}\Big(\operatorname{inv}(\sigma_{k-1}^{j,r_{k,\ell}^*,s_{k,\ell}^*})\delta_{k,k-1}^{j,r_{k,\ell}^*,s_{k,\ell}^*}-\widehat{x}_{k|k}\Big)\Big(\operatorname{inv}(\sigma_{k-1}^{j,r_{k,\ell}^*,s_{k,\ell}^*})\delta_{k,k-1}^{j,r_{k,\ell}^*,s_{k,\ell}^*}-\widehat{x}_{k|k}\Big)'\\
&\quad\times\int_{\mathbb{R}^n}\exp\Big\{-\tfrac{1}{2}\Big(\xi-\operatorname{inv}(\sigma_{k-1}^{j,r_{k,\ell}^*,s_{k,\ell}^*})\delta_{k,k-1}^{j,r_{k,\ell}^*,s_{k,\ell}^*}\Big)'\sigma_{k-1}^{j,r_{k,\ell}^*,s_{k,\ell}^*}\\
&\qquad\qquad\times\Big(\xi-\operatorname{inv}(\sigma_{k-1}^{j,r_{k,\ell}^*,s_{k,\ell}^*})\delta_{k,k-1}^{j,r_{k,\ell}^*,s_{k,\ell}^*}\Big)\Big\}d\xi\\
&\quad=\Big(\operatorname{inv}(\sigma_{k-1}^{j,r_{k,\ell}^*,s_{k,\ell}^*})\delta_{k,k-1}^{j,r_{k,\ell}^*,s_{k,\ell}^*}-\widehat{x}_{k|k}\Big)\Big(\operatorname{inv}(\sigma_{k-1}^{j,r_{k,\ell}^*,s_{k,\ell}^*})\delta_{k,k-1}^{j,r_{k,\ell}^*,s_{k,\ell}^*}-\widehat{x}_{k|k}\Big)'.
\end{aligned}$$

Finally, combining the above calculations the result follows. □

Remark 4. It is interesting to note that (3.43) is essentially in the form of an additive decomposition of two components, that is, a weighted sum of the two terms $\operatorname{inv}(\sigma_{k-1}^{j,r_{k,\ell}^*,s_{k,\ell}^*})$ and $(\operatorname{inv}(\sigma_{k-1}^{j,r_{k,\ell}^*,s_{k,\ell}^*})\delta_{k,k-1}^{j,r_{k,\ell}^*,s_{k,\ell}^*}-\widehat{x}_{k|k})(\operatorname{inv}(\sigma_{k-1}^{j,r_{k,\ell}^*,s_{k,\ell}^*})\delta_{k,k-1}^{j,r_{k,\ell}^*,s_{k,\ell}^*}-\widehat{x}_{k|k})'$. This suggests that the estimation error arises from two sources: the approximation by a finite mixture and the standard error in the estimate of x. The matrix $\operatorname{inv}(\sigma_{k-1}^{j,r_{k,\ell}^*,s_{k,\ell}^*})$ is related, for example, to the covariance $\Sigma_{k-1|k-1}^{r,s}$, which contributes to component s, in the finite mixture representation of the rth unnormalized density (recall (3.13)). The matrix $(\operatorname{inv}(\sigma_{k-1}^{j,r_{k,\ell}^*,s_{k,\ell}^*})\delta_{k,k-1}^{j,r_{k,\ell}^*,s_{k,\ell}^*}-\widehat{x}_{k|k})(\operatorname{inv}(\sigma_{k-1}^{j,r_{k,\ell}^*,s_{k,\ell}^*})\delta_{k,k-1}^{j,r_{k,\ell}^*,s_{k,\ell}^*}-\widehat{x}_{k|k})'$, however, computes the error covariance between the state estimate $\widehat{x}_{k|k}$ and the approximate mean of the state x. This decomposition could provide another method to determine the appropriate number of components for the Gaussian mixtures.

4. Hybrid smoother dynamics. Smoothing is a term used for a form of off-line estimation where one has information (from the observation process) beyond the current time. Suppose one has a set of observations generated by the dynamics (2.5), for example, $\{y_1, y_2, \dots, y_T\}$ and we wish to estimate

$$E\big[x_\ell|\mathcal{Y}_{0,T}\big], \qquad 0\le\ell\le T. \tag{4.1}$$

Currently, all smoothing schemes for Gauss–Markov jump linear systems are ad hoc and are not based upon the exact hybrid filter dynamics; for example, see [3] and [4].

To compute smoothers, we exploit a duality between the forward function $q_k^j(x)$ which evolves forward in time and a related function $v_k^j(x)$ which evolves backward in time (see [12] and [10]). To construct our smoothing algorithm, we compute a finite memory time-reversed recursion for the functions $v_k^j(x)$. A suboptimal v is then defined using an extension of the idea of Viterbi, by replacing a summation with a maximization.

4.1. Exact smoother dynamics. Recalling the form of Bayes' rule (2.14), we note that

$$E\big[\langle Z_k,\boldsymbol{e}_j\rangle f(x_k)|\mathcal{Y}_{0,T}\big]=\frac{E^\dagger\big[\Lambda_{0,T}\langle Z_k,\boldsymbol{e}_j\rangle f(x_k)|\mathcal{Y}_{0,T}\big]}{E^\dagger\big[\Lambda_{0,T}|\mathcal{Y}_T\big]}. \tag{4.2}$$

Write

$$\widetilde{\mathcal{F}}_k\triangleq\sigma\big\{x_\ell,Z_\ell,0\le\ell\le k\big\}. \tag{4.3}$$

For reasons identical to those given in section 3.1, we need consider only the numerator in the quotient of (4.2):

(4.4)
$$\begin{aligned} E^\dagger\big[\Lambda_{0,T}\langle Z_k, \boldsymbol{e}_j\rangle f(x_k)|\mathcal{Y}_{0,T}\big] &= E^\dagger\big[\Lambda_{0,k}\Lambda_{k+1,T}\langle Z_k, \boldsymbol{e}_j\rangle f(x_k)|\mathcal{Y}_{0,T}\big] \\ &= E^\dagger\Big[E^\dagger\big[\Lambda_{0,k}\Lambda_{k+1,T}\langle Z_k, \boldsymbol{e}_j\rangle f(x_k)|\widetilde{\mathcal{F}}_k \vee \mathcal{Y}_{0,T}\big]|\mathcal{Y}_{0,T}\Big] \\ &= E^\dagger\Big[\Lambda_{0,k}\langle Z_k, \boldsymbol{e}_j\rangle f(x_k) E^\dagger\big[\Lambda_{k+1,T}|\widetilde{\mathcal{F}}_k \vee \mathcal{Y}_{0,T}\big]|\mathcal{Y}_{0,T}\Big]. \end{aligned}$$

As our processes are Markov,

$$E^\dagger\big[\Lambda_{k+1|T}|\widetilde{\mathcal{F}}_k \vee \mathcal{Y}_{0,T}\big] = E^\dagger\big[\Lambda_{k+1|T}|Z_k, x_k, \mathcal{Y}_{0,T}\big]. \tag{4.5}$$

Write

$$v_k^j(x) \triangleq E^\dagger\big[\Lambda_{k+1,T}|Z_k = \boldsymbol{e}_j, x_k = x, \mathcal{Y}_{0,T}\big]. \tag{4.6}$$

Then, as in [13],

$$E^\dagger\big[\Lambda_{0,k}\langle Z_k, \boldsymbol{e}_j\rangle f(x_k) v_k^j(x_k)|\mathcal{Y}_{0,T}\big] = \int_{\mathbb{R}^n}\Big\{q_k^j(\xi)v_k^j(\xi)\Big\}f(\xi)d\xi. \tag{4.7}$$

We now compute a backward recursion for the function $v_k^j(x)$, similar in form to the forward recursion for the function $q_k^j(x)$, given in (3.7).

THEOREM 4. *The unnormalized function $v_k^j(x)$ satisfies the backward recursion*

(4.8)
$$v_k^j(x) = \sum_{r=1}^m \frac{\pi_{(r,j)}}{|D_r||B_r|\Phi(y_{k+1})}\int_{\mathbb{R}^n}\Phi\big(D_r^{-1}(y_{k+1} - C_r\xi)\big)\Psi(B_r^{-1}(\xi - A_r x))v_{k+1}^r(\xi)d\xi.$$

Proof. From definition (4.6), we again use repeated conditioning to write

(4.9)
$$\begin{aligned} v_k^j(x) &= E^\dagger[\Lambda_{k+1,T}|Z_k = \boldsymbol{e}_j, x_k = x, \mathcal{Y}_{0,T}\big] \\ &= E^\dagger[\lambda_{k+1}E^\dagger[\Lambda_{k+2,T}|Z_{k+1}, x_{k+1}, Z_k = \boldsymbol{e}_j, x_k = x, \mathcal{Y}_{0,T}]| \\ &\qquad Z_k = \boldsymbol{e}_j, x_k = x, \mathcal{Y}_{0,T}] \\ &= \sum_{r=1}^m E^\dagger[\lambda_{k+1}\langle Z_{k+1}, \boldsymbol{e}_r\rangle E^\dagger\big[\Lambda_{k+2,T}|Z_{k+1} = \boldsymbol{e}_r, x_{k+1}, Z_k = \boldsymbol{e}_j, x_k = x, \mathcal{Y}_{0,T}]| \\ &\qquad Z_k = \boldsymbol{e}_j, x_k = x, \mathcal{Y}_{0,T}] \\ &= \sum_{r=1}^m E^\dagger[\lambda_{k+1}\langle Z_{k+1}, \boldsymbol{e}_r\rangle E^\dagger\big[\Lambda_{k+2,T}|Z_{k+1} = \boldsymbol{e}_r, x_{k+1}, \mathcal{Y}_{0,T}]| \\ &\qquad Z_k = \boldsymbol{e}_j, x_k = x, \mathcal{Y}_{0,T}] \end{aligned}$$

(since Z is a Markov chain, and under $P^\dagger$ the process x is i.i.d.)

$$\begin{aligned} &= \sum_{r=1}^m E^\dagger\Big[\langle Z_{k+1}, \boldsymbol{e}_r\rangle \frac{\Phi\big(D_r^{-1}(y_{k+1} - C_r x_{k+1})\big)}{|D_r|\Phi(y_{k+1})}\frac{\Psi\big(B_r^{-1}(x_{k+1} - A_r x_k)\big)}{|B_r|\Psi(x_{k+1})}v_{k+1}^r(x_{k+1})| \\ &\qquad\qquad Z_k = \boldsymbol{e}_j, x_k = x, \mathcal{Y}_{0,T}\Big] \\ &= \sum_{r=1}^m \frac{\pi_{(r,j)}}{|D_r||B_r|\Phi(y_{k+1})}\int_{\mathbb{R}^n}\Phi\big(D_r^{-1}(y_{k+1} - C_r\xi)\big)\Psi(B_r^{-1}(\xi - A_r x))v_{k+1}^r(\xi)d\xi. \end{aligned}$$

∎

THEOREM 5. *Suppose the unnormalized function $v_{k+1}^r(\xi)$ (as it appears under the integral in (4.8)) may be approximated as a finite weighted Gaussian mixture with $M^v \in \mathbb{N}$ components. That is, suppose*

(4.10)
$$v_{k+1}^r(\xi) = \sum_{s=1}^{M^v} \rho_{k+1}^{r,s} \frac{1}{(2\pi)^{n/2}|\Sigma_{k+1|T}^{r,s}|^{\frac{1}{2}}} \exp\Big\{-\tfrac{1}{2}\big(\xi - \alpha_{k+1|T}^{r,s}\big)' \operatorname{inv}\big(\Sigma_{k+1|T}^{r,s}\big)\big(\xi - \alpha_{k+1|T}^{r,s}\big)\Big\}.$$

Here $\Sigma_{k+1|T}^{r,s} \in \mathbb{M}^{n\times n}$ and $\alpha_{k+1|T}^{r,s} \in \mathbb{R}^n$ are both $\mathcal{Y}_{k+1,T}$-measurable functions for all pairs $(r,s) \in \{1,2,\dots,m\} \times \{1,2,\dots,M^v\}$. Using the Gaussian mixture (4.10), the recursion for $v_k^j(x)$, at times $k \in \{1,2,\dots,T-1\}$, has the form

$$\begin{aligned} v_k^j(x) \triangleq{}& \frac{1}{(2\pi)^{(n+d)/2}\Phi(y_{k+1})} \sum_{r=1}^{m}\sum_{s=1}^{M^v} K_{k+1,T}^v(j,r,s) \\ &\times \exp\Big\{-\tfrac{1}{2}\Big(x - \operatorname{inv}(S_{k+1|T}^{r,s})\tau_{k+1|T}^{r,s}\Big)' S_{k+1,T}^{r,s}\Big(x - \operatorname{inv}(S_{k+1|T}^{r,s})\tau_{k+1|T}^{r,s}\Big)\Big\}. \end{aligned} \tag{4.11}$$

At the final time T for all $j \in \{1,2,\dots,m\}$,

$$v_T^j(x) \triangleq 1. \tag{4.12}$$

Here

$$\begin{aligned} K_{k+1,T}^v(j,r,s) \triangleq{}& \frac{\pi_{(r,j)}\rho_{k+1}^{r,s}}{|D_r||B_r||\widetilde{\Sigma}_{k+1|T}^{r,s}|^{\frac{1}{2}}\Big|\gamma_r + \operatorname{inv}(\Sigma_{k+1|T}^{r,s})\Big|^{\frac{1}{2}}} \\ &\times \exp\Big\{-\tfrac{1}{2}\Big((\alpha_{k+1|T}^{r,s})' \operatorname{inv}(\Sigma_{k+1|T}^{r,s})\Big[I - \widetilde{\Sigma}_{k+1|T}^{r,s}\operatorname{inv}(\Sigma_{k+1|T}^{r,s})\Big]\alpha_{k+1|T}^{r,s}\Big\} \\ &\times \exp\Big\{-\tfrac{1}{2}y_{k+1}'\operatorname{inv}(D_rD_r')yk + 1\Big\} \\ &\times \exp\Big\{-\tfrac{1}{2}\Big(\widetilde{\Sigma}_{k+1|T}^{r,s}\operatorname{inv}(\Sigma_{k+1|T}^{r,s})\alpha_{k+1|T}^{r,s}\Big)' \operatorname{inv}(\widetilde{\Sigma}_{k+1|T}^{r,s}) \\ &\times \Big(\widetilde{\Sigma}_{k+1|T}^{r,s}\operatorname{inv}\big(\Sigma_{k+1|T}^{r,s}\big)\alpha_{k+1|T}^{r,s}\Big)\Big\} \\ &\times \exp\Big\{\tfrac{1}{2}\Big(C_r\operatorname{inv}(D_rD_r')y_{k+1}\operatorname{inv}(\Sigma_{k+1|T}^{r,s})\alpha_{k+1|T}^{r,s}\Big)' \\ &\times \operatorname{inv}\big(\widetilde{\Sigma}_{k+1|T}^{r,s}\big)\Big(C_r\operatorname{inv}(D_rD_r')y_{k+1} + \operatorname{inv}(\Sigma_{k+1|T}^{r,s})\alpha_{k+1|T}^{r,s}\Big)\Big\} \\ &\times \exp\Big\{\tfrac{1}{2}(\tau_{k+1|T}^{r,s})'\operatorname{inv}(S_{k+1|T}^{r,s})\tau_{k+1|T}^{r,s}\Big\} \in \mathbb{R}, \end{aligned} \tag{4.13}$$

$$S_{k+1|T}^{r,s} \triangleq A_r'\operatorname{inv}(B_rB_r')A_r - A_r'\operatorname{inv}(B_rB_r')\widetilde{\Sigma}_{k+1|T}^{r,s}\operatorname{inv}(B_rB_r')A_r \in \mathbb{R}^{n\times n}, \tag{4.14}$$

$$\begin{aligned} \tau_{k+1|T}^{r,s} \triangleq{}& A_r'\operatorname{inv}(B_rB_r')\widetilde{\Sigma}_{k+1|T}^{r,s} \\ &\times \big(C_r'\operatorname{inv}(D_rD_r')y_{k+1} + \operatorname{inv}(\Sigma_{k+1|T}^{r,s})\alpha_{k+1|T}^{r,s}\big) \in \mathbb{R}^n, \end{aligned} \tag{4.15}$$

$$\gamma_r \triangleq \operatorname{inv}(B_rB_r') + C_r'\operatorname{inv}(D_rD_r')C_r \in \mathbb{R}^{n\times n}, \tag{4.16}$$

$$\mu_{k+1}^r \triangleq C_r'\operatorname{inv}(D_rD_r')y_{k+1} + \operatorname{inv}(B_rB_r')A_rx \in \mathbb{R}^n, \tag{4.17}$$

$$\widetilde{\Sigma}_{k+1|T}^{r,s} \triangleq \operatorname{inv}\big(\gamma_r^{-1} + \Sigma_{k+1|T}^{r,s}\big) \in \mathbb{R}^{n\times n}. \tag{4.18}$$

(Note that the square matrices γ^r and $\widetilde{\Sigma}_{k+1|T}^{r,s}$ are symmetric.)

Proof. To prove Theorem 5, we first evaluate the integral in the recursion (4.8).

To complete the proof, the result of this calculation is written as a weighted sum of Gaussian densities.

Write

$$v_k^j(x) = \sum_{r=1}^{m} \frac{\pi_{(r,j)}}{|D_r||B_r|\Phi(y_{k+1})} I_a, \tag{4.19}$$

where

$$I_a \triangleq \int_{\mathbb{R}^n} \Phi\big(D_r^{-1}(y_{k+1} - C_r\xi)\big)\Psi(B_r^{-1}(\xi - A_r x))v_{k+1}^r(\xi)d\xi. \tag{4.20}$$

Completing the square of the exponentials in I_a, we see that

$$\begin{aligned} I_a = {} & \frac{1}{(2\pi)^{(n+d)/2}} \exp\Big\{\tfrac{1}{2}(\mu_{k+1}^r)' \operatorname{inv}(\gamma^r)\mu_{k+1}^r\Big\} \exp\Big\{-\tfrac{1}{2}y_{k+1}' \operatorname{inv}(D_r D_r')y_{k+1}\Big\} \\ & \times \exp\Big\{-\tfrac{1}{2}x' A_r' \operatorname{inv}(B_r B_r')A_r x\Big\} I_b. \end{aligned} \tag{4.21}$$

Here

$$I_b = \int_{\mathbb{R}^n} \exp\Big\{-\tfrac{1}{2}\big(\xi - \operatorname{inv}(\gamma^r)\mu_{k+1}^r\big)'\gamma_r\big(\xi - \operatorname{inv}(\gamma^r)\mu_{k+1}^r\big)\Big\} v_{k+1}^r(\xi)d\xi, \tag{4.22}$$

$$\gamma_r \triangleq C_r' \operatorname{inv}(D_r D_r')C_r + \operatorname{inv}(B_r B_r'), \tag{4.23}$$

$$\mu_{k+1}^r \triangleq C_r' \operatorname{inv}(D_r D_r')y_{k+1} + \operatorname{inv}(B_r B_r')A_r x. \tag{4.24}$$

Since $v_{k+1}^r(\xi)$ is a Gaussian mixture, the integrand of I_b is a product of Gaussian densities and can be evaluated with directly, that is,

$$\begin{aligned} I_b = {} & \sum_{s=1}^{M^v} \rho_{k+1}^{r,s} \int_{\mathbb{R}^n} \exp\Big\{-\tfrac{1}{2}\big(\xi - \operatorname{inv}(\gamma^r)\mu_{k+1}^r\big)'\gamma_r\big(\xi - \operatorname{inv}(\gamma^r)\mu_{k+1}^r\big)\Big\} \\ & \times \frac{1}{(2\pi)^{n/2}|\Sigma_{k+1|T}^{r,s}|^{\frac{1}{2}}} \exp\Big\{-\tfrac{1}{2}\big(\xi - \alpha_{k+1|T}^{r,s}\big)' \operatorname{inv}(\Sigma_{k+1|T}^{r,s})\big(\xi - \alpha_{k+1|T}^{r,s}\big)\Big\} d\xi \\ = {} & \sum_{s=1}^{M^v} \rho_{k+1}^{r,s} \Bigg[\frac{\big|\operatorname{inv}\big(\gamma^r + \operatorname{inv}(\Sigma_{k+1|T}^{r,s})\big)\big|^{\frac{1}{2}}}{\big|\Sigma_{k+1|T}^{r,s}\big|^{\frac{1}{2}}} \exp\big\{-\tfrac{1}{2}(\mu_{k+1}^r)' \operatorname{inv}(\gamma^r)\mu_{k+1}^r\big\} \\ & \times \exp\Big\{-\tfrac{1}{2}(\alpha_{k+1|T}^{r,s})' \operatorname{inv}(\Sigma_{k+1|T}^{r,s})\alpha_{k+1|T}^{r,s}\Big\} \\ & \times \exp\Big\{\tfrac{1}{2}\big(\mu_{k+1}^r + \operatorname{inv}\big(\Sigma_{k+1|T}^{r,s}\big)\alpha_{k+1|T}^{r,s}\big)' \operatorname{inv}\big(\gamma^r + \operatorname{inv}(\Sigma_{k+1|T}^{r,s})\big) \\ & \times \big(\mu_{k+1}^r + \operatorname{inv}(\Sigma_{k+1|T}^{r,s})\alpha_{k+1|T}^{r,s}\big)\Big\}\Bigg]. \end{aligned} \tag{4.25}$$

Recalling (4.19) and the terms I_a and I_b, we see that

$$
\begin{aligned}
v_k^j(x) = {} & \frac{1}{(2\pi)^{(d+n)/2}\,\Phi(y_{k+1})} \sum_{r=1}^{m}\sum_{s=1}^{M^v} \frac{\pi_{(r,j)}\rho_{k+1}^{r,s}\big|\big(\gamma^r + \operatorname{inv}(\Sigma_{k+1|T}^{r,s})\big)\big|^{\frac{1}{2}}}{|D_r||B_r||\Sigma_{k+1|T}^{r,s}|^{\frac{1}{2}}} \\
& \times \exp\Big\{-\tfrac{1}{2}x' A_r' \operatorname{inv}(B_r B_r') A_r x\Big\} \exp\Big\{-\tfrac{1}{2} y_{k+1}' \operatorname{inv}(D_r D_r') y_{k+1}\Big\} \\
& \times \exp\Big\{-\tfrac{1}{2}(\alpha_{k+1|T}^{r,s})' \operatorname{inv}(\Sigma_{k+1|T}^{r,s})\alpha_{k+1|T}^{r,s}\Big\} \\
& \times \exp\Big\{\tfrac{1}{2}(\alpha_{k+1|T}^{r,s})' \operatorname{inv}(\Sigma_{k+1|T}^{r,s})\widetilde{\Sigma}_{k+1|T}^{r,s}\Sigma_{k+1|T}^{r,s}\alpha_{k+1|T}^{r,s}\Big\} \\
& \times \exp\Big\{\tfrac{1}{2}\Big[(\mu_{k+1}^r)'\widetilde{\Sigma}_{k+1|T}^{r,s}\mu_{k+1}^r + (\mu_{k+1}^r)'\widetilde{\Sigma}_{k+1|T}^{r,s} \operatorname{inv}(\Sigma_{k+1|T}^{r,s})\alpha_{k+1|T}^{r,s} \\
& + (\alpha_{k+1|T}^{r,s})' \operatorname{inv}(\Sigma_{k+1|T}^{r,s})\widetilde{\Sigma}_{k+1|T}^{r,s}\mu_{k+1}^r\Big]\Big\}. \qquad (4.26)
\end{aligned}
$$

Consequently

$$
\begin{aligned}
v_k^j(x) = {} & \frac{1}{(2\pi)^{(d+n)/2}\Phi(y_{k+1})} \sum_{r=1}^{m}\sum_{s=1}^{M^v} \frac{\pi_{(r,j)}\rho_{k+1}^{r,s}\big|\operatorname{inv}\big(\gamma^r + \operatorname{inv}(\Sigma_{k+1|T}^{r,s})\big)\big|^{\frac{1}{2}}}{|D_r||B_r||\Sigma_{k+1|T}^{r,s}|^{\frac{1}{2}}} \\
& \times \exp\Big\{-\tfrac{1}{2}\big(\alpha_{k+1|T}^{r,s}\big)' \operatorname{inv}(\Sigma_{k+1|T}^{r,s})\big[I - \widetilde{\Sigma}_{k+1|T}^{r,s} \operatorname{inv}(\Sigma_{k+1|T}^{r,s})\big]\alpha_{k+1|T}^{r,s}\Big\} \\
& \times \exp\Big\{-\tfrac{1}{2} y_{k+1}' \operatorname{inv}(D_r D_r') y_{k+1}\Big\} \exp\Big\{-\tfrac{1}{2}\big(\widetilde{\Sigma}_{k+1|T}^{r,s} \operatorname{inv}(\Sigma_{k+1|T}^{r,s})\alpha_{k+1|T}^{r,s}\big)' \\
& \times \operatorname{inv}(\widetilde{\Sigma}_{k+1|T}^{r,s})\big(\widetilde{\Sigma}_{k+1|T}^{r,s} \operatorname{inv}(\Sigma_{k+1|T}^{r,s})\alpha_{k+1|T}^{r,s}\big)\Big\} \\
& \times \exp\Big\{\tfrac{1}{2}\big(C_r \operatorname{inv}(D_r D_r') y_{k+1} + \operatorname{inv}(\Sigma_{k+1|T}^{r,s})\alpha_{k+1|T}^{r,s}\big)' \\
& \times \operatorname{inv}(\widetilde{\Sigma}_{k+1|T}^{r,s})\big(C_r \operatorname{inv}(D_r D_r') y_{k+1} + \operatorname{inv}(\Sigma_{k+1|T}^{r,s})\alpha_{k+1|T}^{r,s}\big)\Big\} \\
& \times \exp\Big\{\tfrac{1}{2}(\tau_{k+1|T}^{r,s})' \operatorname{inv}(S_{k+1|T}^{r,s})\tau_{k+1|T}^{r,s}\Big\} \\
& \times \exp\Big\{-\tfrac{1}{2}\Big(x - \operatorname{inv}(S_{k+1|T}^{r,s})\tau_{k+1|T}^{r,s}\Big)' S_{k+1|T}^{r,s}\Big(x - \operatorname{inv}(S_{k+1|T}^{r,s})\tau_{k+1|T}^{r,s}\Big)\Big\}. \qquad (4.27)
\end{aligned}
$$

□

COROLLARY 2. *Write*

$$
v_k^j \triangleq \int_{\mathbb{R}^n} v_k^j(\xi)\,d\xi. \qquad (4.28)
$$

The scalar-valued quantity v_k^j is computed by the double sum

$$
v_k^j = \frac{1}{(2\pi)^{d/2}\Phi(y_{k+1})} \sum_{r=1}^{m}\sum_{s=1}^{M^v} \vartheta_{k+1,T}^v(j,r,s). \qquad (4.29)
$$

Here

$$
\vartheta_{k+1,T}^v(j,r,s) \triangleq K_{k+1,T}^v(j,r,s)|S_{k+1|T}^{r,s}|^{-\frac{1}{2}}. \qquad (4.30)
$$

The proof of Corollary 2 is immediate.

4.2. Suboptimal smoother dynamics.

4.2.1. Hypothesis management. Write

$$\Gamma^v \triangleq \{1,2,\dots,m\} \times \{1,2,\dots,M^v\}, \tag{4.31}$$

$$\widetilde{S}^v_{k+1,T}(j,r,s) \triangleq \big\{\vartheta^v_{k+1,T}(j,r,s)\big\}_{(r,s)\,\in\,\Gamma^v}. \tag{4.32}$$

As before, we propose, at each time k, to identify the M^v-best candidate functions (components in the Gaussian mixture) for each function $v^j_k(x)$. This maximization procedure is as follows:

$$\vartheta^v_{k+1,T}(j,r^*_1,s^*_1) \triangleq \max_{(r,s)\,\in\,\Gamma^v} \widetilde{S}^v_{k+1,T}(j,r,s), \tag{4.33}$$

$$\vartheta^v_{k+1,T}(j,r^*_2,s^*_2) \triangleq \max_{(r,s)\,\in\,\Gamma^v\setminus(r^*_1,s^*_1)} \widetilde{S}^v_{k+1,T}(j,r,s), \tag{4.34}$$

$$\vdots \qquad\qquad \vdots$$

$$\vartheta^v_{k+1,T}(j,r^*_{M^v},s^*_{M^v}) \triangleq \max_{(r,s)\,\in\,\Gamma^v\setminus\{(r^*_1,s^*_1),\dots,(r^*_{M^v-1},s^*_{M^v-1})\}} \widetilde{S}^v_{k+1,T}(j,r,s). \tag{4.35}$$

The optimal index set, for function $v^j_k(x)$, is

$$\mathcal{I}_k(j) \triangleq \big\{(r^*_{k,1},s^*_{k,1}),(r^*_{k,2},s^*_{k,2}),\dots,(r^*_{k,M^v},s^*_{k,M^v})\big\}. \tag{4.36}$$

Using these indexes, the order M^v equation for $v^j_k(x)$, whose memory requirements are fixed in time, has the form

$$\begin{aligned} v^j_k(x) \triangleq{}& \frac{1}{(2\pi)^{(d+n)/2}\Phi(y_{k+1})} \sum_{\ell=1}^{M^v} K^v_{k+1,T}(j,r^*_{k+1,\ell},s^*_{k+1,\ell}) \\ &\times \exp\Big\{-\tfrac{1}{2}\Big(x-\operatorname{inv}(S^{r^*_{k+1,\ell},s^*_{k+1,\ell}}_{k+1|T})\tau^{r^*_{k+1,\ell},s^*_{k+1,\ell}}_{k+1|T}\Big)' S^{r^*_{k+1,\ell},s^*_{k+1,\ell}}_{k+1,T} \\ &\times\Big(x-\operatorname{inv}(S^{r^*_{k+1,\ell},s^*_{k+1,\ell}}_{k+1|T})\tau^{r^*_{k+1,\ell},s^*_{k+1,\ell}}_{k+1|T}\Big)\Big\}. \end{aligned} \tag{4.37}$$

4.2.2. Smoother state statistics. Write, for the smoother density normalization constant,

$$\varphi_{k|T} \triangleq \sum_{j=1}^{m}\Big(\int_{\mathbb{R}^n}\big[q^j_k(\xi)v^j_k(\xi)\big]d\xi\Big). \tag{4.38}$$

Combining the functions $q^j_k(x)$ and $v^j_k(x)$, we see that the fixed interval smoothed estimate of state may be written as

$$\begin{aligned} \widehat{x}_{k|T} \triangleq E\big[x_t \mid \mathcal{Y}_{0,T}\big] &= \frac{\int_{\mathbb{R}^n}\xi\Big[\sum_{j=1}^m q^j_k(\xi)v^j_k(\xi)\Big]d\xi}{\int_{\mathbb{R}^n}\Big[\sum_{j=1}^m q^j_k(\xi)v^j_k(\xi)\Big]d\xi} \\ &= \frac{1}{\varphi_{k|T}}\int_{\mathbb{R}^n}\xi\left[\sum_{j=1}^m q^j_k(\xi)v^j_k(\xi)\right]d\xi \\ &= \frac{1}{\varphi_{k|T}}\sum_{j=1}^m\left[\int_{\mathbb{R}^n}\xi\Big[q^j_k(\xi)v^j_k(\xi)\Big]d\xi\right]. \end{aligned} \tag{4.39}$$

Since the functions $q_k(x)$ and $v_k(x)$ are each weighted Gaussian mixtures, the integrals in (4.39) can be evaluated exactly.

PROPOSITION 3. *The smoothed state estimate $\widehat{x}_{k|T}$, for Gaussian mixtures with orders M^q and M^v, has the following representation:*

$$\begin{aligned}
\widehat{x}_{k|T} = \frac{1}{\varphi_{k|T}} \sum_{j=1}^{m} \sum_{\ell=1}^{M^q} \sum_{i=1}^{M^v} \Bigg[& K^v_{k,k-1}(j, r^*_{k,l}, s^*_{k,\ell}) K^v_{k+1,T}(j, r^*_{k,i}, s^*_{k,i}) \\
& \times \left| \operatorname{inv}\left(\sigma_{k-1}^{r^*_{k,\ell}, s^*_{k,\ell}} + S_{k+1|T}^{r^*_{k+1,\ell}, s^*_{k+1,\ell}}\right) \right|^{\frac{1}{2}} \exp\left\{ \tfrac{1}{2} \left(\sigma_k^{r^*_{k,\ell}, s^*_{k,\ell}} + \tau_{k+1|T}^{r^*_{k+1,i} s^*_{k+1,i}} \right)' \right. \\
& \times \operatorname{inv}\left(\sigma_{k-1}^{r^*_{k,\ell}, s^*_{k,\ell}} + S_{k+1|T}^{r^*_{k+1,\ell}, s^*_{k+1,\ell}}\right) \left(\delta_k^{r^*_{k,\ell}, s^*_{k,\ell}} + \tau_{k+1|T}^{r^*_{k+1,i}, s^*_{k+1,i}} \right) \Bigg\} \\
& \times \exp\Big\{ -\tfrac{1}{2} \Big[(\tau_{k+1|T}^{r^*_{k+1,\ell}, s^*_{k+1,\ell}})' \operatorname{inv}(S_{k+1|T}^{r^*_{k+1,\ell}, s^*_{k+1,\ell}}) \tau_{k+1|T}^{r^*_{k+1,i}, s^*_{k+1,i}} + (\delta_k^{r^*_{k,\ell}, s^*_{k,\ell}})' \\
& \times \operatorname{inv}(\sigma_{k-1}^{r^*_{k,\ell}, s^*_{k,\ell}}) \delta_k^{r^*_{k,\ell}, s^*_{k,\ell}} \Big] \Big\} \\
& \times \operatorname{inv}\Big(\sigma_{k-1}^{r^*_{k,\ell}, s^*_{k,\ell}} + S_{k+1|T}^{r^*_{k+1,\ell}, s^*_{k+1,\ell}}\Big) \Big(\delta_k^{r^*_{k,\ell}, s^*_{k,\ell}} + \tau_{k+1|T}^{r^*_{k+1,i}, s^*_{k+1,i}} \Big) \Bigg]. \qquad (4.40)
\end{aligned}$$

Here

$$\begin{aligned}
\varphi_{k|T} = \sum_{j=1}^{m} \sum_{\ell=1}^{M^q} \sum_{i=1}^{M^v} \Bigg[& K^q_{k,k-1}(j, r^*_{k,\ell}, s^*_{k,\ell}) K^v_{k+1,T}(j, r^*_{k,i}, s^*_{k,i}) \\
& \times \exp\left\{ \tfrac{1}{2} \left(\delta_k^{r^*_{k,\ell}, s^*_{k,\ell}} + \tau_{k+1|T}^{r^*_{k+1,i}, s^*_{k+1,i}} \right) \operatorname{inv}\left(\sigma_{k-1}^{r^*_{k,\ell}, s^*_{k,\ell}} + S_{k+1|T}^{r^*_{k+1,\ell}, s^*_{k+1,\ell}} \right) \right. \\
& \times \left(\delta_k^{r^*_{k,\ell}, s^*_{k,\ell}} + \tau_{k+1|T}^{r^*_{k+1,i}, s^*_{k+1,i}} \right) \Big\} \exp\Big\{ -\tfrac{1}{2} \Big[\left(\tau_{k+1|T}^{r^*_{k+1,i}, s^*_{k+1,i}} \right)' \operatorname{inv}\left(S_{k+1|T}^{r^*_{k+1,\ell}, s^*_{k+1,\ell}} \right) \\
& \times \tau_{k+1|T}^{r^*_{k+1,i}, s^*_{k+1,i}} + \left(\delta_k^{r^*_{k,\ell}, s^*_{k,\ell}} \right)' \operatorname{inv}\left(\sigma_{k-1}^{r^*_{k,\ell}, s^*_{k,\ell}} \right) \delta_k^{r^*_{k,i}, s^*_{k,i}} \Big] \Big\} \\
& \times \left| \operatorname{inv}\left(\sigma_{k-1}^{r^*_{k,\ell}, s^*_{k,\ell}} + S_{k+1|T}^{r^*_{k+1,\ell}, s^*_{k+1,\ell}} \right) \right|^{\frac{1}{2}} \Bigg] \in \mathbb{R}. \qquad (4.41)
\end{aligned}$$

Proof.

$$\begin{aligned}
\widehat{x}_{k|T} &= \frac{1}{\varphi_{k|T}} \sum_{j=1}^{m} \left[\int_{\mathbb{R}^n} \xi \left[q_k^j(\xi) v_k^j(\xi) \right] d\xi \right] \\
&= \frac{1}{\varphi_{k|T}} \sum_{j=1}^{m} \Bigg[\int_{\mathbb{R}^n} \xi \Bigg(\frac{1}{(2\pi)^{(d+n)/2} \Phi(y_k)} \sum_{\ell=1}^{M^q} K^q_{k,k-1}(j, r^*_{k,\ell}, s^*_{k,\ell}) \\
&\quad \times \exp\left\{ -\tfrac{1}{2} \left(\xi - \operatorname{inv}(\sigma_{k-1}^{r^*_{k,\ell}, s^*_{k,\ell}}) \delta_k^{r^*_{k,\ell}, s^*_{k,\ell}} \right)' \sigma_{k-1}^{r^*_{k,\ell}, s^*_{k,\ell}} \right. \\
&\quad \times \left. \left(\xi - \operatorname{inv}(\sigma_{k-1}^{r^*_{k,\ell}, s^*_{k,\ell}}) \delta_k^{r^*_{k,\ell}, s^*_{k,\ell}} \right) \right\} \Bigg) \\
&\quad \times \Bigg(\frac{1}{(2\pi)^{(d+n)/2} \Phi(y_{k+1})} \sum_{i=1}^{M^v} K^v_{k+1,T}(j, r^*_{k,i}, s^*_{k,i}) \\
&\quad \times \exp\left\{ -\tfrac{1}{2} \left(\xi - \operatorname{inv}(S_{k+1|T}^{r^*_{k+1,\ell}, s^*_{k+1,\ell}}) \tau_{k+1|T}^{r^*_{k+1,\ell}, s^*_{k+1,\ell}} \right)' \right.
\end{aligned}$$

$$
\begin{aligned}
&\times S_{k+1,T}^{r^*_{k+1,\ell},s^*_{k+1,\ell}}\Big(\xi-\operatorname{inv}(S_{k+1|T}^{r^*_{k+1,\ell},s^*_{k+1,\ell}})\tau_{k+1|T}^{r^*_{k+1,\ell},s^*_{k+1,\ell}}\Big)\Big\}\Big)d\xi\Big]\\
&=\frac{1}{(2\pi)^{(d+n)}\Phi(y_k)\Phi(y_{k+1})\varphi_{k|T}}\sum_{j=1}^{m}\sum_{\ell=1}^{M^q}\sum_{i=1}^{M^v}\Big[K^q_{k,k-1}(j,r^*_{k,\ell},s^*_{k,\ell})\\
&\times K^v_{k+1,T}(j,r^*_{k,i},s^*_{k,i})\int_{\mathbb{R}^n}\xi\Big(\exp\Big\{-\tfrac{1}{2}\Big(\xi-\operatorname{inv}(\sigma_{k-1}^{r^*_{k,\ell},s^*_{k,\ell}})\delta_k^{r^*_{k,\ell},s^*_{k,\ell}}\Big)'\sigma_{k-1}^{r^*_{k,\ell},s^*_{k,\ell}}\\
&\times\Big(\xi-\operatorname{inv}(\sigma_{k-1}^{r^*_{k,\ell},s^*_{k,\ell}})\delta_k^{r^*_{k,\ell},s^*_{k,\ell}}\Big)\Big\}\\
&\times\exp\Big\{-\tfrac{1}{2}\Big(\xi-\operatorname{inv}(S_{k+1|T}^{r^*_{k+1,\ell},s^*_{k+1,\ell}})\tau_{k+1|T}^{r^*_{k+1,\ell},s^*_{k+1,\ell}}\Big)'S_{k+1,T}^{r^*_{k+1,\ell},s^*_{k+1,\ell}}\\
&\times\Big(\xi-\operatorname{inv}(S_{k+1|T}^{r^*_{k+1,i},s^*_{k+1,i}})\tau_{k+1|T}^{r^*_{k+1,i},s^*_{k+1,i}}\Big)\Big\}\Big)d\xi\Big]\\
&=\frac{1}{(2\pi)^{(d+n)}\Phi(y_k)\Phi(y_{k+1})\varphi_{k|T}}\sum_{j=1}^{m}\sum_{\ell=1}^{M^q}\sum_{i=1}^{M^v}\Big[K^q_{k,k-1}(j,r^*_{k,\ell},s^*_{k,\ell})\\
&\times K^v_{k+1,T}(j,r^*_{k,i},s^*_{k,i})\exp\Big\{\tfrac{1}{2}\Big(\delta_k^{r^*_{k,\ell},s^*_{k,\ell}}+\tau_{k+1|T}^{r^*_{k+1,i},s^*_{k+1,i}}\Big)'\\
&\times\operatorname{inv}\Big(\sigma_{k-1}^{r^*_{k,\ell},s^*_{k,\ell}}+S_{k+1|T}^{r^*_{k+1,\ell},s^*_{k+1,\ell}}\Big)\Big(\delta_k^{r^*_{k,\ell},s^*_{k,\ell}}+\tau_{k+1|T}^{r^*_{k+1,i},s^*_{k+1,i}}\Big)\Big\}\\
&\times\exp\Big\{-\tfrac{1}{2}\Big[\big(\tau_{k+1|T}^{r^*_{k+1,i},s^*_{k+1,i}}\big)'\operatorname{inv}\big(S_{k+1|T}^{r^*_{k+1,\ell},s^*_{k+1,\ell}}\big)\tau_{k+1|T}^{r^*_{k+1,i},s^*_{k+1,i}}+\big(\delta_k^{r^*_{k,\ell},s^*_{k,\ell}}\big)'\\
&\times\operatorname{inv}(\sigma_{k-1}^{r^*_{k,\ell},s^*_{k,\ell}})\delta_k^{r^*_{k,\ell},s^*_{k,\ell}}\Big]\Big\}(2\pi)^{n/2}\Big|\operatorname{inv}\Big(\sigma_{k-1}^{r^*_{k,\ell},s^*_{k,\ell}}+S_{k+1|T}^{r^*_{k+1,\ell},s^*_{k+1,\ell}}\Big)\Big|^{\frac{1}{2}}\\
&\times\operatorname{inv}\Big(\sigma_{k-1}^{r^*_{k,\ell},s^*_{k,\ell}}+S_{k+1|T}^{r^*_{k+1,\ell},s^*_{k+1,\ell}}\Big)\Big(\delta_k^{r^*_{k,\ell},s^*_{k,\ell}}+\tau_{k+1|T}^{r^*_{k+1,i},s^*_{k+1,i}}\Big)\Big].\qquad\square
\end{aligned}
\tag{4.42}
$$

Remark 5. It is interesting to note that in the formulation of the smoother defined by (4.42), the integers M^q and M^v are mutually independent. This feature suggests that one is afforded an extra degree of freedom; that is, the practitioner can explicitly impose a desired accuracy in the contribution of the past (choice of M^q) and the contribution of the future (choice of M^v).

The smoother error covariance is defined as

$$
\begin{aligned}
\Sigma_{k|T}&\triangleq E\big[\big(x_k-\widehat{x}_{k|T}\big)\big(x_k-\widehat{x}_{k|T}\big)'|\mathcal{Y}_T\big]\\
&=\frac{1}{\varphi_{k|T}}\int_{\mathbb{R}^n}(\xi-\widehat{x}_{k|T})(\xi-\widehat{x}_{k|T})'\Bigg[\Bigg(\frac{1}{(2\pi)^{(d+n)/2}\Phi(y_k)}\sum_{\ell=1}^{M^q}K^q_{k,k-1}(j,r^*_{k,\ell},s^*_{k,\ell})\\
&\times\exp\Big\{-\tfrac{1}{2}\Big(\xi-\operatorname{inv}(\sigma_{k-1}^{r^*_{k,\ell},s^*_{k,\ell}})\delta_k^{r^*_{k,\ell},s^*_{k,\ell}}\Big)'\sigma_{k-1}^{r^*_{k,\ell},s^*_{k,\ell}}\Big(\xi-\operatorname{inv}(\sigma_{k-1}^{r^*_{k,\ell},s^*_{k,\ell}})\delta_k^{r^*_{k,\ell},s^*_{k,\ell}}\Big)\Big\}\Bigg)\\
&\times\Bigg(\frac{1}{(2\pi)^{(d+n)/2}\Phi(y_{k+1})}\sum_{i=1}^{M^v}K^v_{k+1,T}(j,r^*_{k,i},s^*_{k,i})\\
&\times\exp\Big\{-\tfrac{1}{2}\Big(\xi-\operatorname{inv}(S_{k+1|T}^{r^*_{k+1,\ell},s^*_{k+1,\ell}})\tau_{k+1|T}^{r^*_{k+1,\ell},s^*_{k+1,\ell}}\Big)'\\
&\times S_{k+1,T}^{r^*_{k+1,\ell},s^*_{k+1,\ell}}\Big(\xi-\operatorname{inv}(S_{k+1|T}^{r^*_{k+1,\ell},s^*_{k+1,\ell}})\tau_{k+1|T}^{r^*_{k+1,\ell},s^*_{k+1,\ell}}\Big)\Big\}\Bigg)\Bigg]d\xi=\widehat{\Sigma}_{k|T}.
\end{aligned}
\tag{4.43}
$$

PROPOSITION 4. *The smoother state error covariance $\widehat{\Sigma}_{k|T}$, for Gaussian mixtures with orders M^q and M^v, has the following representation:*

$$
\begin{aligned}
\widehat{\Sigma}_{k|T} = \frac{1}{\varphi_{k|T}} \sum_{j=1}^{m} \sum_{\ell=1}^{M^q} \sum_{i=1}^{M^v} & \\
& \times \Bigg[K^q_{k,k-1}(j, r^*_{k,\ell}, s^*_{k,\ell}) K^v_{k+1,T}(j, r^*_{k,i}, s^*_{k,i}) \Big| \operatorname{inv}\Big(\sigma_{k-1}^{r^*_{k,\ell}, s^*_{k,\ell}} + S_{k+1|T}^{r^*_{k+1,\ell}, s^*_{k+1,\ell}} \Big) \Big|^{\frac{1}{2}} \\
& \times \exp\Big\{ \tfrac{1}{2} \Big(\big(\tau_{k+1|T}^{r^*_{k+1,i}, s^*_{k+1,i}} \big)' \operatorname{inv}\big(S_{k+1|T}^{r^*_{k+1,\ell}, s^*_{k+1,\ell}} \big) \tau_{k+1|T}^{r^*_{k+1,i}, s^*_{k+1,i}} \\
& \times \big(\delta_k^{r^*_{k,\ell}, s^*_{k,\ell}} \big)' \operatorname{inv}\big(\sigma_{k-1}^{r^*_{k,\ell}, s^*_{k,\ell}} \big) \delta_k^{r^*_{k,\ell}, s^*_{k,\ell}} \Big) \Big\} \\
& \times \Big[\operatorname{inv}\Big(\sigma_{k-1}^{r^*_{k,\ell}, s^*_{k,\ell}} + S_{k+1|T}^{r^*_{k+1,\ell}, s^*_{k+1,\ell}} \Big) \\
& + \Big(\operatorname{inv}\big(\sigma_{k-1}^{r^*_{k,\ell}, s^*_{k,\ell}} + S_{k+1|T}^{r^*_{k+1,\ell}, s^*_{k+1,\ell}} \big) \big(\delta_k^{r^*_{k,\ell}, s^*_{k,\ell}} + \tau_{k+1|T}^{r^*_{k+1,i}, s^*_{k+1,i}} \big) - \widehat{x}_{k|T} \Big) \\
& \times \Big(\operatorname{inv}\big(\sigma_{k-1}^{r^*_{k,\ell}, s^*_{k,\ell}} + S_{k+1|T}^{r^*_{k+1,\ell}, s^*_{k+1,\ell}} \big) \big(\delta_k^{r^*_{k,\ell}, s^*_{k,\ell}} + \tau_{k+1|T}^{r^*_{k+1,i}, s^*_{k+1,i}} \big) - \widehat{x}_{k|T} \Big)' \Big] \Bigg].
\end{aligned} \tag{4.44}
$$

The proof of Proposition 4 is similar to the proof of Proposition 2, and so it is omitted.

5. Example. In this section we illustrate the performance of the new hybrid filter state estimator presented in section 3.2.2. Two scenarios are considered. In the first example, the Markov chain transition matrix is fixed and simulations are computed for five different values of signal-to-noise ratio (SNR). In the second example, we consider two different transition matrices. For each of our examples, comparisons are provided against the IMM algorithm (with standard Kalman filters). All the examples consider the same scalar-valued state process x and we set $m = 3$. Further, the order of the Gaussian mixtures is fixed at $M^q = 5$. For our first example, we set

$$
H_1 \triangleq \big\{ A_{H_1} = 1, B_{H_1} = 0.05, C_{H_1} = 1 \big\}, \tag{5.1}
$$

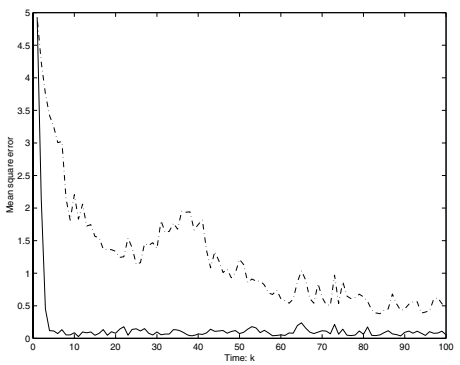
$$
H_2 \triangleq \big\{ A_{H_2} = 0.9, B_{H_2} = 0.01, C_{H_2} = 2 \big\}, \tag{5.2}
$$

$$
H_3 \triangleq \big\{ A_{H_3} = 1.1, B_{H_3} = 0.1, C_{H_3} = 1.5 \big\} \tag{5.3}
$$

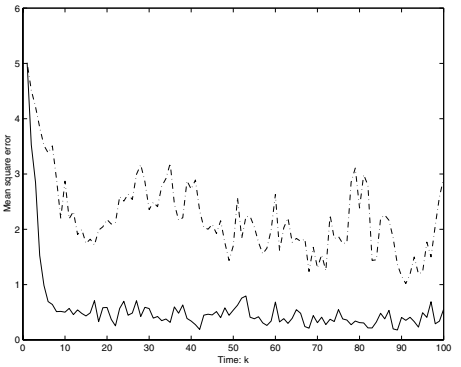
and

$$
\Pi \triangleq \begin{bmatrix} 0.8 & 0.2 & 0.2 \\ 0.1 & 0.6 & 0.2 \\ 0.1 & 0.2 & 0.6 \end{bmatrix}. \tag{5.4}
$$

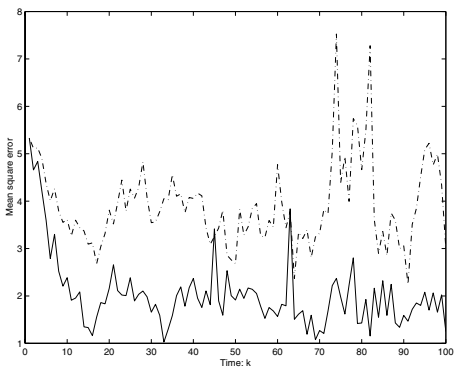
The initial state x_0 is a Gaussian random variable, with mean 10 and covariance 10. Each of the five Figures 5.1(a)–(e) shows simulations at different levels of SNR, which was varied only through the noise gain on the state dynamics; that is, SNR Case 1 $\big\{ D_{H_1} = 0.5, D_{H_2} = 0.25, D_{H_3} = 0.5 \big\}$, SNR Case 2 $\big\{ D_{H_1} = 1, D_{H_2} = 0.5, D_{H_3} = 1 \big\}$, SNR Case 3 $\big\{ D_{H_1} = 2, D_{H_2} = 1, D_{H_3} = 2 \big\}$, SNR Case 4 $\big\{ D_{H_1} = 4, D_{H_2} = 2, D_{H_3} = 4 \big\}$, SNR Case 5 $\big\{ D_{H_1} = 8, D_{H_2} = 4, D_{H_3} = 8 \big\}$. For these five scenarios, Monte Carlo simulations were computed, consisting of 100 trials and discrete time indexed from $k = 1$ to $k = 100$. The corresponding mean square errors for the IMM and the new hybrid filter are given in Figures 5.1(a)–(e).



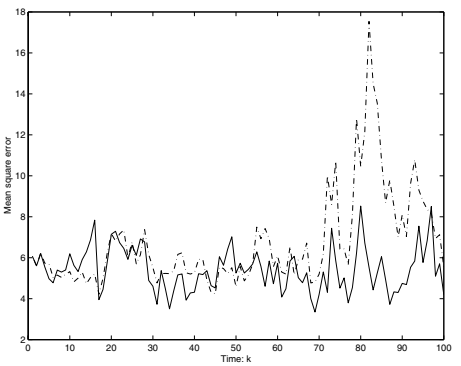
(a) Case 1



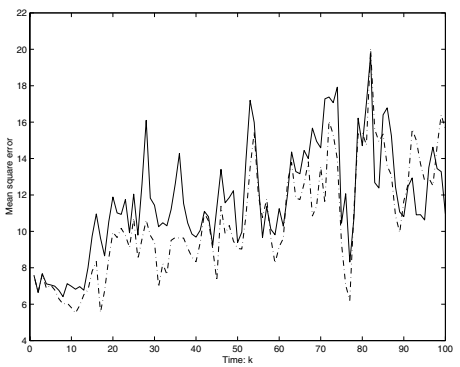
(b) Case 2



(c) Case 3

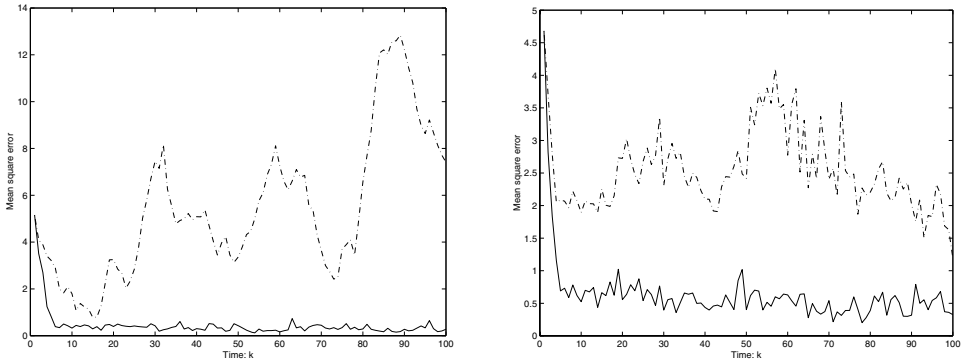


(d) Case 4



(e) Case 5

FIG. 5.1. *SNR performance of the IMM (dash-dotted line) and the new hybrid filter (solid line) with* $M^q = 5$.



(a) Case 6 (b) Case 7

FIG. 5.2. *Fast and slow switching Markov chain comparison of the IMM (dash-dotted line) and the new hybrid filter (solid line) with* $M^q = 5$.

In our second example, two different values of the parameter Π were considered, with

$$\text{(5.5)} \quad \Pi \text{ Case 6} \qquad \Pi \triangleq \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}, \qquad \Pi \text{ Case 7} \qquad \Pi \triangleq \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.3 & 0.4 & 0.3 \\ 0.3 & 0.3 & 0.4 \end{bmatrix}.$$

In this example, the same model parameters from the first example were used, that is, H_1, H_2, and H_3, with the state noise gains set at $\{D_{H_1} = 1, D_{H_2} = 0.5, D_{H_3} = 1\}$. In all these cases, the new hybrid filter significantly outperforms the IMM. In Cases 4 and 5 (see Figures 5.1(d) and (e)), when the SNR is very low the performances of these two algorithms are comparable. For the Cases 1 to 3, when the SNR is standard or high, the suboptimal filter is more efficient than the IMM (see Figures 5.1(a)–(c)). The same conclusion can be made when the Markov chain Z jumps slowly or rapidly (see Cases 6 and 7 and Figures 5.2(a) and (b)).

REFERENCES

[1] G. ACKERSON AND K. FU, *On state estimation in switching environments,* IEEE Trans. Automat. Control, 15 (1970), pp. 10–17.

[2] H. BLOM, *An efficient filter for abruptly changing systems,* in Proceedings of the 23rd IEEE Conference on Decision and Control, Las Vegas, NV, 1984, pp. 656–658.

[3] B. CHEN AND J. K. TUGNAIT, *An interacting multiple model fixed-lag smoothing algorithm for Markovian switching systems,* in Proceedings of the 37th IEEE Conference on Decision and Control, Tampa, FL, 1998, pp. 269–274.

[4] B. CHEN AND J. K. TUGNAIT, *Interacting multiple model fixed-lag smoothing algorithm for Markovian switching systems,* IEEE Trans. Aerosp. Electron. Syst., 36 (2000), pp. 243–250.

[5] R. J. ELLIOTT, F. DUFOUR, AND D. SWORDER, *Exact hybrid filters in discrete time,* IEEE Trans. Automat. Control, 41 (1996), pp. 1807–1810.

[6] R. J. ELLIOTT AND W. P. MALCOLM, *Reproducing Gaussian densities and linear Gaussian detection,* Systems Control Lett., 40 (2000), pp. 133–138.

[7] R. J. ELLIOTT, *Stochastic Calculus and Its Applications,* Springer-Verlag, New York, 1982.

[8] R. J. ELLIOTT, L. AGGOUN, AND J. B. MOORE, *Hidden Markov Models: Estimation and Control,* Appl. Math. (N.Y.) 29, Springer-Verlag, New York, 1995.

[9] R. J. Elliott and V. Krishnamurthy, *New finite-dimensional filters for parameter estimation of discrete-time linear Gaussian models*, IEEE Trans. Automat. Control, 44 (1999), pp. 938–951.
[10] R. J. Elliott and W. P. Malcolm, *Robust smoother dynamics for Poisson processes driven by an Itô diffusion,* in Proceedings of the IEEE Conference on Decision and Control, Orlando, FL, 2001, pp. 376–381.
[11] R. J. Elliott, *A general recursive discrete time filter,* J. Appl. Probab., 30 (1993), pp. 575–588.
[12] W. P. Malcolm and R. J. Elliott, *A general smoothing equation for Poisson observations,* in Proceedings of the IEEE Conference on Decision and Control, Phoenix, AZ, 1999, pp. 4106–4110.
[13] E. Pardoux, *Equations du filtrage nonlineare de la predictions et du lissage,* Stochastics, 6 (1982), pp. 193–231.
[14] W. Rudin, *Real and Complex Analysis,* McGraw-Hill, New York, 1966.
[15] H. W. Sorenson and D. L. Alspach, *Recursive Bayesian estimation using Gaussian sums,* Automatica, 7 (1971), pp. 465–479.
[16] A. N. Shiryaev, *Probability,* 2nd ed., Springer-Verlag, New York, 1996.
[17] J. K. Tugnait, *Adaptive estimation and identification for discrete systems with Markov jump parameters,* IEEE Trans. Automat. Control, AC-27 (1982), pp. 1054–1065.
[18] A. J. Viterbi, *Error bounds for the white Gaussian and other very noisy memoryless channels with generalized decision regions,* IEEE Trans. Inform. Theory, 15 (1969), pp. 279–287.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 3, pp. 1105–1147

ASYMPTOTIC STABILIZABILITY BY STATIONARY FEEDBACK OF THE TWO-DIMENSIONAL EULER EQUATION: THE MULTICONNECTED CASE*

OLIVIER GLASS†

Abstract. We construct a feedback law which allows us to asymptotically stabilize the Euler system for incompressible inviscid fluids in two dimensions, in the case of a multiconnected bounded domain, by means of a control localized on a part of the boundary that meets every connected component of the boundary. This generalizes a result of Coron [*SIAM J. Control Optim.*, 37 (1999), pp. 1874–1896] concerning simply connected domains.

Key words. incompressible inviscid fluids, stabilization by feedback control

AMS subject classifications. 93D15, 35B37, 76B75, 93C20

DOI. 10.1137/S0363012903431153

1. Introduction.

1.1. Statement of the problem. In this paper, we are concerned with the null asymptotic stabilization by means of a stationary feedback of the Euler system for inviscid incompressible fluids in two space dimensions, namely, the following system:

$$\begin{cases} \partial_t v(t,x) + (v(t,x).\nabla)v(t,x) + \nabla p(t,x) = 0 \text{ for } (t,x) \text{ in } [0,T^*) \times \Omega, \\ \operatorname{div} v(t,x) = 0 \text{ for } (t,x) \text{ in } [0,T^*) \times \Omega. \end{cases} \tag{1.1}$$

In the above equation, t is the time (the problem under consideration is formulated for $T^* = +\infty$), and x is the position in the domain Ω. The function $v : [0,T^*) \times \Omega \to \mathbb{R}^2$ is the velocity field and $p : [0,T^*) \times \Omega \to \mathbb{R}$ is the pressure. The domain Ω is two-dimensional (2-D), bounded, regular and nonsimply connected (let us agree that the boundary $\partial\Omega$ is decomposed into $\partial\Omega = \Gamma_0 \cup \cdots \cup \Gamma_g$, where the components Γ_i are nonempty, connected, and disjoint).

The initial-boundary problem for equation (1.1) has been studied by Yudovich (see [10]). Given initial data

$$v_{|t=0} = v_0 \text{ in } \Omega, \tag{1.2}$$

where $v_0 : \overline{\Omega} \to \mathbb{R}^2$ is a divergence-free vector field, and appropriate boundary conditions, the system is well-posed. The boundary conditions can be taken as the following data:

- the normal component of the velocity $v(t,x) \cdot n(x)$ on the whole boundary $\partial\Omega$ for any time ($n(x)$ is the unit outward normal on $\partial\Omega$), which has to satisfy

$$\int_{\partial\Omega} v(t,x) \cdot n(x) dx = 0 \quad \forall t \in [0,T^*),$$

*Received by the editors July 10, 2003; accepted for publication (in revised form) November 28, 2004; published electronically October 3, 2005.
http://www.siam.org/journals/sicon/44-3/43115.html
†Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, Boîte courrier 187, 75252 Paris Cedex 05, France (glass@ann.jussieu.fr).

- the vorticity $\omega(t,x) := \operatorname{curl} v(t,x)$ at the points of $\partial\Omega$ which, moved by the velocity flow, enter inside Ω, namely, the points in

$$\Sigma^-_{T^*} := \{(t,x) \in [0,T^*) \times \partial\Omega \ / \ v(t,x) \cdot n(x) < 0\}\,.$$

Let us emphasize that certain compatibility conditions between the solution and the boundary data must be taken into account in order to obtain a solution with suitable regularity.

In this paper, the boundary conditions (or a part of them) will be considered as a control (that is, a parameter to be determined, that we choose to influence the system). More precisely, we fix Σ an open part of the boundary; the part Σ of the boundary is the zone where we can choose the boundary conditions, whereas $\partial\Omega \setminus \Sigma$ represents a wall that cannot be crossed. In other words, we will consider the Euler system with

- the constraint

$$v(t,x) \cdot n(x) = 0 \quad \text{on } [0,T^*) \times (\partial\Omega\backslash\Sigma); \tag{1.3}$$

that is, the fluid cannot enter or quit the domain through $\partial\Omega\backslash\Sigma$ (it must slip on it),

- the boundary condition on $[0,T^*) \times \Sigma$, which is the control to be chosen.

In this setting, the problem of controllability (i.e., to steer a prescribed initial state v_0 to a prescribed final state v_1 in an arbitrary time by choosing a relevant control) was answered affirmatively by Coron in [2], under the necessary condition that Σ meets each connected component of the boundary.

Here we are interested in the problem of asymptotic stabilizability of the equilibrium $v \equiv 0$ by means of a stationary feedback. In other words, we want to find a continuous function f of the state $\mathcal{S}(t)$ of the system at time t such that if the control $\mathcal{C}(t)$ is given at each time by $\mathcal{C}(t) = f(\mathcal{S}(t))$, then the resulting closed system makes 0 globally asymptotically stable in the sense that

- any solution defined on $[0,T^*) \times \overline{\Omega}$ with $T^* < +\infty$ can be extended for $t \geq T^*$;
- for any neighborhood $\mathcal{U}$ of 0, one can find another neighborhood $\mathcal{V}$ of 0 such that any solution of the closed system beginning in $\mathcal{V}$ is in $\mathcal{U}$ for any $t \geq 0$;
- any solution tends to 0 as $t \to +\infty$.

The above-mentioned problem was solved by Coron in the case of a simply connected domain; see [4].

Remark 1. As was already the case in the controllability problem, the condition that Σ meets any connected component of the boundary is a necessary condition to solve the problem. For instance, the vorticity around any "uncontrolled" connected component just slips on it and cannot be "modified." Another obstruction is the Kelvin law which states that the velocity circulation around any uncontrolled connected component of the boundary is constant. Throughout this paper, we will suppose that Σ meets every connected component of the boundary.

1.2. Mathematical setting. We have to specify which data will be the state of the system, and also the precise structure of the control. A natural state to consider would be the whole velocity field $v(t,\cdot)$ in $\overline{\Omega}$, but if we chose $\mathcal{S}$ this way, then (as we will consider solutions that are continuous up to the boundary) it would completely determine the choice of the control (since the boundary conditions described above would be given by the normal component of the trace of v on $[0,T^*) \times \Sigma$ and by the trace of $\operatorname{curl} v$ on $\Sigma^-_{T^*}$).

To avoid such a problem, as suggested in [3], we shall consider the following data as the state of the system:

$$\mathcal{S}(t) = (\omega(t,\cdot), \lambda_1(t), \dots, \lambda_g(t)), \tag{1.4}$$

where $\omega(t,\cdot) : \overline{\Omega} \to \mathbb{R}$ is the vorticity field (which is scalar in two dimensions), that is,

$$\omega(t,x) := \operatorname{curl} v(t,x),$$

and where λ_i, for $i = 1, \dots, g$, is the velocity circulation around the component Γ_i of $\partial\Omega$, that is,

$$\lambda_i(t) := \int_{\Gamma_i} v(t,x) \cdot \vec{\tau}(x) dx.$$

Here $\vec{\tau}(x)$ is the unit tangent vector field on $\partial\Omega$, chosen so that $(\vec{\tau}, n)$ should be direct. (Let us remark that consequently $\vec{\tau}$ endows the curve Γ_i with an orientation that is positive if the curve is an inner component of $\partial\Omega$ and negative in the case of the outer component.)

Remark 2. Of course, only g circulations of v around Γ_i are needed among $(g+1)$ available, since the sum of all these circulations is related to ω by Green's formula.

Once given the state $\mathcal{S}(t)$ and $v(t,x) \cdot n(x)$ on Σ (which is a part of the control, say $\mathcal{C}_1$), one can reconstruct $v(t,\cdot)$ in $\overline{\Omega}$, for each $t \in [0, T^*)$, by means of the following system:

$$\begin{cases} \operatorname{curl} v(t,\cdot) = \omega(t,\cdot) \quad \text{in } \Omega, \\ \operatorname{div} v(t,\cdot) = 0 \quad \text{in } \Omega, \\ v(t,\cdot) \cdot n(\cdot) = \mathcal{C}_1(t) \quad \text{on } \Sigma, \\ v(t,\cdot) \cdot n(\cdot) = 0 \quad \text{on } \partial\Omega \setminus \Sigma, \\ \displaystyle\int_{\Gamma_i} v(t,x).\vec{\tau}(x) dx = \lambda_i(t) \quad \text{for } i = 1, \dots, g. \end{cases} \tag{1.5}$$

The Euler equation can be written in terms of the state $\mathcal{S}(t)$: as is well known, the vorticity in two dimensions satisfies

$$\partial_t \omega + (v \cdot \nabla)\omega = 0 \quad \text{in } (0, T^*) \times \Omega, \tag{1.6}$$

or, equivalently,

$$\partial_t \omega + \operatorname{div}(\omega v) = 0 \quad \text{in } (0, T^*) \times \Omega, \tag{1.7}$$

and the velocity circulations satisfy

$$\lambda_i(t) - \lambda_i(0) = \int_0^t \int_{\Gamma_i} v(s,x) \cdot n(x) \omega(s,x) dx ds. \tag{1.8}$$

One easily sees that the group composed of (1.5), (1.7), and (1.8) is equivalent to (1.1).

We still need to specify the exact structure of the control that we use. The first part of the control is the normal component of the velocity on Σ, which we call $\mathcal{C}_1$. We must stipulate the other part of the control, which concerns the entering vorticity.

Since $\omega(t,\cdot)$ is now a part of the state, it seems inappropriate to consider ω on $\Sigma^-_{T^*}$ as the second part of the control (as will be specified below, $\omega(t,\cdot)$ is continuous

up to the boundary in our problem). Thus a natural control to consider would be the following:

$$\mathcal{C}(t) = \begin{cases} v(t,x)\cdot n(x) & \text{on } \Sigma, \\ \partial_t \omega(t,x) & \text{on } \Sigma^-, \end{cases} \tag{1.9}$$

with Σ^- given by

$$\Sigma^- := \{x \in \partial\Omega \ / \ v(t,x)\cdot n(x) < 0\}. \tag{1.10}$$

(To simplify the notation, we omit the dependence of Σ^- on t and $\mathcal{C}_1$—besides, essentially, Σ^- will be constant in what follows.)

Let us point out that the control described in (1.9) was the one used in [4]. However, for technical reasons that will be explained in section 2.1, we will have to consider the following control of *mixed type*:

$$\mathcal{C}(t) = (\mathcal{C}_1(t), \mathcal{C}_2(t), \mathcal{C}_3(t)), \tag{1.11}$$

with

$$\mathcal{C}_1 = v(t,x)\cdot n(x) \quad \text{on } \Sigma, \tag{1.12}$$

$$\mathcal{C}_2 = \partial_t \omega_1(t,x) \quad \text{on } \Sigma^-, \tag{1.13}$$

$$\mathcal{C}_3 = \omega_2(t,x) \quad \text{on } \Sigma^-, \tag{1.14}$$

and the boundary condition for the entering vorticity obtained as

$$\omega(t,x) = \omega_1(t,x) + \omega_2(t,x) \quad \text{on } \Sigma^-. \tag{1.15}$$

In other words, the entering vorticity ω is the sum of two terms: ω_1, whose time derivative we control, and ω_2, which we control directly.

Remark 3. Let us remark that this choice of the form of the control is important. Indeed, even the stabilizability by means of a simpler feedback law of the form $\partial_t \mathcal{C} = f(\mathcal{S})$ does not necessarily imply the stabilizability by means of a feedback law of the form $\mathcal{C} = f(\mathcal{S})$. See, for instance, [5] in the context of finite-dimensional systems.

We can now be more specific about the problem under study.

Definition 1.1. *Given a feedback law*

$$(\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3) = (\mathcal{C}_1(\mathcal{S}), \mathcal{C}_2(\mathcal{S}), \mathcal{C}_3(\mathcal{S})),$$

1. *we shall call "the closed-loop system," with $\mathcal{S}$ as the unknown, the system* (1.5), (1.7), (1.8), *with boundary conditions given by* (1.12)–(1.15);
2. *we shall call $\mathcal{S} = (\omega, \lambda_1, \dots, \lambda_g)$ a solution of the closed-loop system if*
   - *$\mathcal{S} \in C^0([0,T^*) \times \overline{\Omega}; \mathbb{R}) \times C^0([0,T^*), \mathbb{R})^g$ (for some $T^* > 0$);*
   - *$v(t,\cdot)$ being for each $t \in [0,T^*)$ the unique solution (in the sense of distributions in Ω) of* (1.5), *the functions $(\lambda_i)_{i=1,\dots,g}$ satisfy* (1.8) *for all $t \in [0,T^*)$, and ω satisfies* (1.7) *in the sense of distributions in $(0,T^*) \times \Omega$ and* (1.13)–(1.15) *in the sense of distributions on the open manifold*

$$\Big\{(t,x) \in (0,T^*) \times \Sigma, \ \mathcal{C}_1[\mathcal{S}(t)](x) < 0\Big\};$$

3. *we call "maximal" any solution that cannot be extended over its maximal time T^*.*

The purpose of this paper is to establish the following result.

THEOREM 1.2. *If Σ meets every connected component of the boundary, one can find three continuous functions $\mathcal{C}_1$, $\mathcal{C}_2$, and $\mathcal{C}_3$ defined on $C^0(\overline{\Omega};\mathbb{R}) \times \mathbb{R}^g$ and with values in $C^0(\Sigma;\mathbb{R})$, $C^0(\Sigma^-;\mathbb{R})$, and $C^0(\Sigma^-;\mathbb{R})$, respectively, such that the following properties are fulfilled:*

P1. *For any $(\omega_0, \lambda_1^0, \dots, \lambda_g^0) \in C^0(\overline{\Omega};\mathbb{R}) \times \mathbb{R}^g$, the closed-loop system with initial condition*

$$\mathcal{S}(0) = (\omega_0, \lambda_1^0, \dots \lambda_g^0) \tag{1.16}$$

has a global in time solution, and any local in time solution can be extended to $T^* = +\infty$ (in other words, any maximal solution is global).

P2. *For any $\varepsilon > 0$, there exists $\eta > 0$ such that if*

$$\max(\|\omega_0\|_{L^\infty(\Omega)}, |\lambda_1^0|, \dots, |\lambda_g^0|) < \eta,$$

then one has

$$\max(\|\omega(t,\cdot)\|_{L^\infty(\Omega)}, |\lambda_1(t)|, \dots, |\lambda_g(t)|) \le \varepsilon,$$

for all $t \ge 0$ and any global in time solution of the closed-loop system satisfying (1.16).

P3. *Any global in time solution of the closed-loop system satisfies*

$$\max(\|\omega(t,\cdot)\|_{L^\infty(\Omega)}, |\lambda_1(t)|, \dots, |\lambda_g(t)|) \to 0 \text{ as } t \to +\infty.$$

We will describe in section 2 the feedback law $(\mathcal{C}_1(\mathcal{S}), \mathcal{C}_2(\mathcal{S}), \mathcal{C}_3(\mathcal{S}))$ which involves Theorem 1.2. The precise result expressed in terms of this feedback law is given in section 2.5.

1.3. A related problem concerning the stationary equation. As pointed out in [3], the problem of asymptotic stabilizability by stationary feedback is connected with a problem concerning the stationary equation. Indeed, Brockett established a necessary condition for a finite-dimensional control system to be stabilizable; see [1].

THEOREM 1.3 (see Brockett [1]). *A necessary condition for the control system $\dot{x} = f(x,u)$ to be locally asymptotically stabilizable at the equilibrium point x_0 (satisfying $f(x_0,0) = 0$) by a stationary feedback is that the image by f of any neighborhood of $(x_0,0)$ is a neighborhood of 0.*

The corresponding statement of this necessary condition in the infinite-dimensional system considered here is precisely what is proved in [6], that is, the existence of solutions for the stationary problem with a small force term (and by scaling arguments, with any force term). Hence, the study in [6] can be viewed as a preliminary step before this one. As we will see in section 2, some tools developed in [6] are essential in the construction here. For more details, see [3], [4], and [6].

1.4. Structure of the paper. In the next section, we begin by giving the main ideas concerning the construction of the feedback law that yields Theorem 1.2, then we detail this construction (which is rather involved), and finally we state our precise result (Theorem 2.4), which takes the stated form of the feedback law into account and clearly involves Theorem 1.2.

In section 3, we fix the notation and give preliminary elementary statements, which are classical for the construction of global in time solutions to the Euler system

in two dimensions. At the end of this section, a proposition that is central in the proof is stated.

Sections 4 and 5 prove the existence of local in time solutions of the closed-loop system defined with the control law introduced in section 2. This is done in two steps. In section 4, we construct a certain operator $\mathcal{F}$. Then section 5 proves by Schauder's fixed point theorem that the operator $\mathcal{F}$ has fixed points, which actually give local in time solutions to the closed-loop system.

Section 6 finishes the proof of Theorem 2.4, by proving that maximal solutions of the closed-loop system are global and satisfy the asymptotic stability properties P2 and P3 described in Theorem 1.2.

Finally, we put in the appendix the proofs of the most technical lemmas.

2. Description of the feedback.

2.1. Basic ideas. The most important feature of the 2-D Euler equation is a straightforward consequence of (1.6) (or (1.7)), precisely the following:

(2.1) the vorticity ω follows the flow of the velocity v.

A direct consequence of this fact is that, to perform P3 in Theorem 1.2, a global solution of the closed-loop has to satisfy

(2.2) the flow of the velocity v makes any point in $\overline{\Omega}$ go out of the domain

(except perhaps for configurations with important zones of null vorticity from the beginning, but this situation is essentially nongeneric).

To get (2.2), we examine the Hodge decomposition of the velocity, which in non-simply connected domains takes the following form (as usual, $\nabla^\perp := (-\partial_{x^2}, \partial_{x^1})$):

$$v(t,x) = \nabla\phi(t,x) + \nabla^\perp\psi(t,x) + \sum_{k=1}^{g} \mu_i(t)\nabla^\perp\tau_i, \tag{2.3}$$

where $\tau_i \in C^\infty(\overline{\Omega};\mathbb{R})$, $i \in \{1,\dots,g\}$, is the solution of the system

$$\begin{cases} \Delta\tau_i = 0 & \text{in } \Omega, \\ \tau_i = 1 & \text{on } \Gamma_i, \\ \tau_i = 0 & \text{on } \partial\Omega\backslash\Gamma_i, \end{cases} \tag{2.4}$$

and where the different terms satisfy, for each t,

$$\begin{cases} \Delta\phi(t,x) = 0 & \text{in } \Omega, \\ \partial_n\phi(t,x) = v(t,x)\cdot n(x) & \text{on } \partial\Omega, \end{cases} \tag{2.5}$$

$$\begin{cases} \Delta\psi(t,x) = \operatorname{curl} v(t,x) & \text{in } \Omega, \\ \psi(t,x) = 0 & \text{on } \partial\Omega, \end{cases} \tag{2.6}$$

$$\lambda_i(t) = -\sum_{j=1}^{g} \mu_j(t)\left(\int_\Omega \nabla\tau_i\cdot\nabla\tau_j\right) - \int_\Omega \omega(t,x)\tau_i(x)dx \quad \forall\, i \in \{1,\dots,g\}. \tag{2.7}$$

The family $\nabla^\perp\tau_i$ being clearly linearly independent, the matrix $(\int_\Omega \nabla\tau_i.\nabla\tau_j)_{i,j=1,\dots,n}$ is invertible.

Now to obtain (2.2), it seems rather arduous to rely on the last two terms in (2.3), because μ_j and $\nabla^\perp\psi$ are fixed at the beginning by the state and then slowly evolve

according to the flow of v itself. On the contrary, the $\nabla\phi$ part is directly obtained from the control. Hence a natural idea, which is also present in [2], [4], and [6], is to fix the $v \cdot n$ part of the control so that the $\nabla\phi$ part in (2.3) should prevail over the other two in such a way that (2.2) is satisfied.

As in [4], this program is fulfilled by finding a function $\theta : \overline{\Omega} \to \mathbb{R}$ such that

$$\begin{cases} \Delta\theta = 0 \quad \text{in } \Omega, \\ \partial_n\theta = 0 \quad \text{on } \partial\Omega \setminus \Sigma, \\ |\nabla\theta(x)| > 0 \quad \text{in } \overline{\Omega}. \end{cases} \tag{2.8}$$

Indeed, given such a function θ, one can hope that the control

$$\mathcal{C}_1 = f(\omega, \lambda_1, \dots, \lambda_g)\partial_n\theta \tag{2.9}$$

with $f(\omega, \lambda_1, \dots, \lambda_g)$ an adequate nonnegative function, which should be large when $(\omega, \lambda_1, \dots, \lambda_g)$ is large, will satisfy the requirements.

Once this part of the control is imposed, the idea is to choose the vorticity part of the control in the form

$$\partial_t\omega = -K(\omega, \lambda_1, \dots, \lambda_g)\omega \quad \text{in } \Sigma^-,$$

where $K(\omega, \lambda_1, \dots, \lambda_g)$ is an appropriate positive function. In this way, one can hope that the vorticity inside the domain will gradually be replaced by a smaller one.

However, there remain two issues:

- This might not be sufficient to get rid of the velocity circulations. The natural idea to diminish these circulations is to inject additional vorticity through $\Sigma^- \cap \Gamma_i$, as motivated by (1.8). This can raise a problem, because this injected vorticity could influence the other λ_j. In order to avoid this, we make this vorticity leave the domain through Γ_0.
- Because of (2.1), at a point of $\partial\Sigma^-$ where v is *pointing inside* Σ^-, there must be compatibility conditions on the control in vorticity so that the solution will have proper regularity. The reason for this is that, on one side of this point in $\partial\Omega$, the vorticity is determined by the control, whereas on the other side, it is determined by the incoming flow along the uncontrolled part of the boundary (see the points A in figures below). This is the main reason we must consider a control in the form (1.11)–(1.15): the continuity of the entering vorticity at this point is ensured by the $\mathcal{C}_3$-part of the control.
  Moreover, it will be technically simpler if these points in $\partial\Sigma^-$ where v is pointing inside Σ^- do not depend on the state. In fact, it can be expected that by choosing f in (2.9) properly, these points will be exactly those for which $\nabla\theta$ is pointing inside Σ^-.

Remark 4. It would seem natural to require the function θ to satisfy, besides (2.8),

(2.10)
at any point of $\partial\gamma^-$, $\nabla\theta$ is pointing outside γ^-, where $\gamma^- = \{x \in \Sigma \ / \ \partial_n\theta(x) < 0\}$.

In the case of a simply connected domain, this is possible; see [4]. But this is no longer possible in the case of a nonsimply connected domain, since this would result in a null index of the vector field $\nabla\theta$ around the outer component of the boundary

and in positive indices of $\nabla\theta$ around inner components, which would be inconsistent with (2.8).

Now that we have sketched the main ideas, we can construct a function θ satisfying (2.8); other conditions are required, either for technical reasons or to address the above-mentioned issues. The feedback law, which relies on θ, is constructed subsequently.

2.2. The function θ. The function θ that we introduce here has been used to prove the existence of solutions for the stationary problem; see section 1.3 and reference [6]. Precisely, we have the following result.

PROPOSITION 2.1 (see [6, Prop. 1]). *Consider Ω a nonempty, bounded, connected and regular domain in $\mathbb{R}^2$, assumed to be not simply connected. Denote $\Gamma_0, \dots, \Gamma_g$ the connected components of its boundary. Let n be the unit outward normal on $\partial\Omega$. Consider Σ an open part of $\partial\Omega$, which meets each connected component of $\partial\Omega$. Then there exists a function $\tilde{\theta} \in C^\infty(\overline{\Omega};\mathbb{R})$ that satisfies the following conditions:*

$$\Delta\tilde{\theta} = 0 \quad \text{in } \Omega, \tag{2.11}$$

$$\partial_n\tilde{\theta} = 0 \quad \text{on } \partial\Omega\backslash\Sigma, \tag{2.12}$$

$$|\nabla\tilde{\theta}(x)| > 0 \quad \text{for any } x \text{ in } \overline{\Omega}, \tag{2.13}$$

$$\text{for } \gamma^+(\tilde{\theta}) := \{x \in \partial\Omega \ / \ \partial_n\tilde{\theta} > 0\} \quad \text{and} \quad \gamma^-(\tilde{\theta}) := \{x \in \partial\Omega \ / \ \partial_n\tilde{\theta} < 0\}, \tag{2.14}$$
$$\text{one has: } \overline{\gamma^+(\tilde{\theta})} \cap \overline{\gamma^-(\tilde{\theta})} = \emptyset,$$

$$\gamma^+(\tilde{\theta}) \text{ and } \gamma^-(\tilde{\theta}) \text{ are unions of a finite number of intervals of } \partial\Omega \text{ with disjoint closures,} \tag{2.15}$$

$$\text{there exist } g \text{ points } \tilde{M}_1, \dots, \tilde{M}_g \text{ in } \gamma^-(\tilde{\theta}) \cap \Gamma_0, \text{ sent respectively on } \gamma^+(\tilde{\theta}) \cap \Gamma_1, \dots, \gamma^+(\tilde{\theta}) \cap \Gamma_g \text{ by the flow of } \nabla\tilde{\theta}, \tag{2.16}$$
$$\text{with the trajectories not touching } \partial\Omega\backslash[\gamma^+(\tilde{\theta}) \cup \gamma^-(\tilde{\theta})].$$

To describe properties of the flow, it is more convenient to work in a domain that is invariant by the flow. To that end, we consider $R > 0$ such that $\overline{\Omega} \subset B_R$ and introduce an operator π that extends continuous (resp., C^1) vector fields defined on Ω to continuous (resp., C^1) and compactly supported vector fields on B_R; see a more precise definition of π in section 3.1.

We have the following technical refinement of Proposition 2.1.

COROLLARY 2.2. *One can add the following requirement on $\tilde{\theta}$ (call $\tilde{\Phi}$ the flow of $\pi(\nabla\tilde{\theta})$):*

$$\text{given any point } E \text{ in } \partial\gamma^+(\tilde{\theta}) \text{ such that } \nabla\tilde{\theta}(E) \text{ is pointing outside } \gamma^+(\tilde{\theta}), \tag{2.17}$$
$$\text{then for } t > 0,\ \tilde{\Phi}(t,0,E) \text{ does not meet another point in } \partial\gamma^+(\tilde{\theta}) \text{ pointing outside } \gamma^+(\tilde{\theta}) \text{ before leaving } \overline{\Omega}.$$

The proof of this corollary is postponed to the appendix.

In fact, the function θ used in this paper is given by $-\tilde{\theta}$; hence θ satisfies (2.11), (2.12), (2.13), (2.14), and (2.15). However, (2.16) must be replaced by

$$\text{there exist } g \text{ points } M_1, \dots, M_g \text{ in } \gamma^-(\theta) \cap \Gamma_1, \dots, \gamma^-(\theta) \cap \Gamma_g, \text{ sent respectively on } \gamma^+(\theta) \cap \Gamma_0 \text{ by } \Phi, \text{ with the trajectories not touching } \partial\Omega\backslash[\gamma^+(\theta) \cup \gamma^-(\theta)] \tag{2.18}$$

(here Φ is the flow of $\pi(\nabla\theta)$, and we define $\gamma^+(\theta) := \gamma^-(\tilde{\theta})$ and $\gamma^-(\theta) := \gamma^+(\tilde{\theta})$), and (2.17) must be replaced by

(2.19) given any point A in $\partial\gamma^-(\theta)$ such that $\nabla\theta(A)$ is pointing inside $\gamma^-(\theta)$, $\Phi(t,0,A)$ has not met another point in $\partial\gamma^-(\theta)$ at which $\nabla\theta$ is pointing inside $\gamma^-(\theta)$ for $t<0$ such that $\Phi([t,0],0,E)\subset\overline{\Omega}$.

A representation of what $\nabla\theta$ may look like is given in Figure 2.1 below. The dotted lines represent some flow lines of $\nabla\theta$.

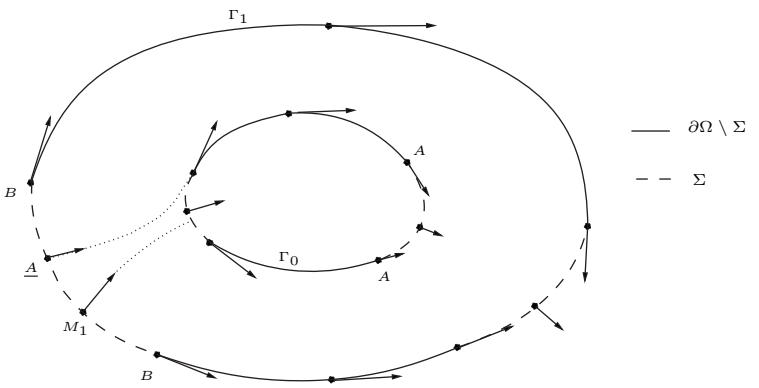


FIG. 2.1. *A representation of* $\nabla\theta$.

2.3. Some constructions relying on θ. We denote

$$\gamma^+ = \gamma^+(\theta) = \left\{x\in\partial\Omega \;\Big/\; \partial_n\theta(x) > 0\right\} \text{ and } \gamma^- = \gamma^-(\theta) = \left\{x\in\partial\Omega \;\Big/\; \partial_n\theta(x) < 0\right\}.$$

We also introduce

$$V(\theta) = \max_{\overline{\Omega}}\theta - \min_{\overline{\Omega}}\theta. \tag{2.20}$$

As in [6] we call A the points in $\partial\gamma^-$ on which $\nabla\theta$ is pointing *inside* γ^- and B the points in $\partial\gamma^-$ on which $\nabla\theta$ is pointing *outside* γ^-. In what follows, we denote by $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{M}$ the sets of A, B, and M_i points, respectively.

For each $A\in\mathcal{A}$, we introduce γ_A as the component of $\partial\Omega\setminus(\overline{\gamma^-}\cup\overline{\gamma^+})$ whose closure contains A (and the same for B). We also consider the points $\underline{A}$ defined as

$$\underline{A} := \Phi(t'_A,0,A), \text{ where } t'_A := \min\left\{t\le 0 \;/\; \Phi([t,0],0,A)\subset\overline{\Omega}\right\}. \tag{2.21}$$

Using (2.19), one sees that $\underline{A}\in\gamma^-(\theta)\cup\mathcal{B}$.

Given θ, we shall introduce some functions on $\overline{\gamma^-}$, called Γ_A and Λ_i, defined for each $A\in\mathcal{A}$ and each $M_i\in\mathcal{M}$, respectively, and supported in a neighborhood of this point in $\overline{\gamma^-}$. Precisely, given $A\in\mathcal{A}$, call $\mathcal{V}_A$ a closed neighborhood of A in $\overline{\gamma^-}$, small enough that it contains neither any M_i point, nor any other point of $\mathcal{A}$, nor any $\underline{A}$ or

B point. Then define a function $\Gamma_A \in C^\infty(\overline{\gamma^-};\mathbb{R})$ satisfying

$$\begin{cases} -1 \leq \Gamma_A \leq 1, \\ \mathrm{Supp}(\Gamma_A) \subset \mathcal{V}_A, \\ \Gamma_A \equiv 1 \quad \text{in a neighborhood of } A, \\ \int_{\mathcal{V}_A} \Gamma_A(x) \nabla\theta(x) \cdot n(x) dx = 0. \end{cases} \tag{2.22}$$

Now, given an $M_i \in \mathcal{M}$, call $\mathcal{V}_{M_i}$ a closed neighborhood of M_i in $\gamma^- \cap \Gamma_i$, small enough that it contains neither any $\underline{A}$ point nor points of $\partial\gamma^-$, and such that all $\mathcal{V}_{M_i}$ are sent by Φ to $\gamma^+(\theta) \cap \Gamma_0$, with the trajectories not touching $\partial\Omega \backslash (\overline{\gamma^+} \cup \overline{\gamma^-})$ (as made possible by (2.18) and Gronwall's lemma; see Lemma 3.4 in section 3.3 below). Then define $\Lambda_i \in C^\infty(\overline{\gamma^-};\mathbb{R})$ satisfying

$$\begin{cases} \Lambda_i \leq 0, \\ \mathrm{Supp}(\Lambda_i) \subset \mathcal{V}_{M_i}, \\ \int_{\mathcal{V}_{M_i}} \Lambda_i(x) \nabla\theta(x) \cdot n(x) dx = 1. \end{cases} \tag{2.23}$$

Note that the last condition can be easily obtained since $\nabla\theta(x) \cdot n(x)$ is negative on $\mathcal{V}_{M_i}$.

We reduce if necessary the supports of Γ_A and Λ_i in order to obtain

$$\begin{array}{l} \mathrm{Supp}(\Gamma_A) \cap \mathrm{Supp}(\Lambda_i) = \emptyset \quad \text{for any } A \in \mathcal{A} \text{ and any } i = 1, \dots, g, \\ \mathrm{Supp}(\Gamma_A) \cap \mathrm{Supp}(\Gamma_{A'}) = \emptyset \quad \text{for any } A, A' \in \mathcal{A} \text{ such that } A \neq A'. \end{array} \tag{2.24}$$

To make the notation lighter, we write

$$\mathrm{Supp}(\Lambda) := \bigcup_{i=1}^{g} \mathrm{Supp}(\Lambda_i),$$
$$\mathrm{Supp}(\Gamma) := \bigcup_{A \in \mathcal{A}} \mathrm{Supp}(\Gamma_A).$$

Also, we introduce

$$\begin{aligned} &\|\Lambda\|_\infty := \max_{i=1}^{g} \|\Lambda_i\|_\infty, \\ &\mathcal{T}(\Gamma) = \sum_{A \in \mathcal{A}} \int_{\gamma^-} |\Gamma_A(x) \nabla\theta(x) \cdot n(x)| dx. \end{aligned} \tag{2.25}$$

We denote by ℓ a strict minimizer of the distance between the connected components of $\gamma^+ \cup \gamma^-$ and of the distances between the various $\mathrm{Supp}(\Gamma_A)$ with $A \in \mathcal{A}$, $\mathrm{Supp}(\Lambda_i)$ with $i \in \{1, \dots, g\}$, and points $B \in \mathcal{B}$.

The requirements on the supports are summarized in Figure 2.2 (where the arrows represent $\nabla\theta$).

2.4. The feedback law. Let us now describe the feedback law that we use. It is given by the following rule:

- If $(\omega(t), \lambda_1(t), \dots, \lambda_g(t)) = 0$, then fix

$$v \cdot n = \mathcal{C}_1 := 0 \text{ on } \Sigma. \tag{2.26}$$

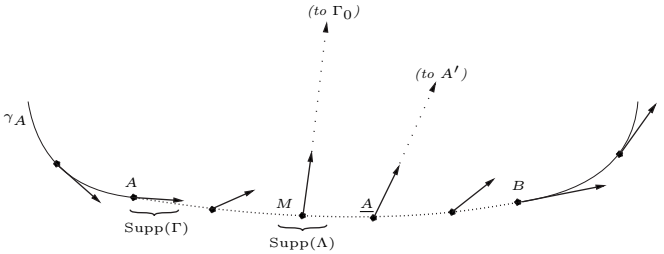


FIG. 2.2. *A representation of* $\Sigma^- \cap \Gamma_i$.

- If $(\omega(t), \lambda_1(t), \dots, \lambda_g(t)) \neq 0$, then fix

(2.27)
$$\begin{cases} v \cdot n = \mathcal{C}_1 := K \max(\|\omega(t,\cdot)\|_{L^\infty(\Omega)}, |\lambda_1(t)|, \dots, |\lambda_g(t)|) \nabla\theta(x) \cdot n(x) \text{ on } \Sigma, \\ \omega = \omega_1 + \omega_2 \text{ on } \gamma^-, \end{cases}$$

where ω_1 and ω_2 are given by

(2.28)
$$\begin{cases} \partial_t \omega_1 &= \mathcal{C}_2 := -M \max(\|\omega(t,\cdot)\|_{L^\infty(\Omega)}, |\lambda_1(t)|, \dots, |\lambda_g(t)|)\omega_1 \text{ on } \gamma^-, \\ \omega_2 &= \mathcal{C}_3 := \sum_{A\in\mathcal{A}} \omega(t,A)\Gamma_A(x) - \sum_{i=1}^g \lambda_i(t)\Lambda_i(x) \text{ on } \gamma^-. \end{cases}$$

Consequently, we will have $\Sigma^- = \gamma^-$ except in the case $(\omega(t), \lambda_1(t), \dots, \lambda_g(t)) = 0$.

Remark that ω_1 is a function of the state since

$$\omega_1(t,\cdot) = \omega(t,\cdot) - \sum_{A\in\mathcal{A}} \omega(t,A)\Gamma_A(\cdot) + \sum_{i=1}^g \lambda_i(t)\Lambda_i(\cdot) \text{ on } \gamma^-.$$

The constants K and M are to be chosen large enough, as will be seen more precisely later.

Remark 5. Let us remark that, as the vorticity functions ω considered here are in the class $C^0([0,T)\times\overline{\Omega})$, the functions $t \mapsto \omega(t,A)$ are well-defined and continuous. Consequently, the feedback law is equivalent (in a distributional sense) to

$$\begin{aligned} \partial_t\omega(t,x) = &- M \max(\|\omega(t,\cdot)\|_{L^\infty(\Omega)}, |\lambda_1(t)|, \dots, |\lambda_g(t)|) \\ &\times \left[\omega(t,x) - \sum_{A\in\mathcal{A}} \omega(t,A)\Gamma_A(x) + \sum_{i=1}^g \lambda_i(t)\Lambda_i(x)\right] \\ &+ \sum_{A\in\mathcal{A}} \partial_t\omega(t,A)\Gamma_A(x) - \sum_{i=1}^g \left(\frac{d}{dt}\lambda_i(t)\right)\Lambda_i(x), \end{aligned} \tag{2.29}$$

where $\partial_t\omega(t,A)$ and $\frac{d}{dt}\lambda_i(t)$ can be recovered, in a formal sense for the first one, from the state thanks to (1.7)–(1.8).

2.5. The result. We rewrite the definition of the solutions of the system with the above described feedback.

DEFINITION 2.3. *A function* $(\omega, \lambda_1, \dots, \lambda_g)$ *in* $C^0([0,T^*)\times\overline{\Omega};\mathbb{R})\times C^0([0,T^*);\mathbb{R})^g$ *is a solution of the closed-loop system with the feedback law of section* 2.4 *if and only if it satisfies*

- *the relation* (1.8) *for all* $t \in [0,T^*)$ *and the equation* (1.7) *in the sense of distributions, where* v *is defined for each* t *by* (1.5)*, with* $\mathcal{C}_1$ *fixed by* (2.27)*,*
- *that on the domain* $\{t \in [0,T^*) \ / \ (\omega(t,\cdot),\lambda_1(t),\dots,\lambda_g(t)) \not\equiv 0\} \times \gamma^-$ *(which is an open bidimensional manifold), the function*

$$\omega_1(t,x) = \omega(t,x) - \sum_{A\in\mathcal{A}} \omega(t,A)\Gamma_A(x) + \sum_{i=1}^{g} \lambda_i(t)\Lambda_i(x)$$

satisfies (2.28) *in a distributional sense.*

The following theorem is the main result of the paper; it clearly involves Theorem 1.2.

THEOREM 2.4. *If the constant* K *is large enough, and* M *is large enough depending on* K*, then for any initial condition* $(\omega_0,\lambda_1^0,\dots,\lambda_g^0)$ *in* $C^0(\overline{\Omega};\mathbb{R}) \times \mathbb{R}^g$*, there are solutions in* $C^0([0,T^*)\times\overline{\Omega};\mathbb{R}) \times C^0([0,T^*);\mathbb{R})^g$ *of the closed-loop system (for some* $T^*>0$*) satisfying*

$$(\omega,\lambda_1,\dots,\lambda_g)_{|t=0} = (\omega_0,\lambda_1^0,\dots,\lambda_g^0). \tag{2.30}$$

Moreover, any maximal solution is global and satisfies, for some $\mathcal{K}>0$ *depending only on* Ω *and* Σ *(and on the functions* θ*,* Γ_A*, and* Λ_i *constructed on* (Ω,Σ)*),*

$$\max(\|\omega(t,\cdot)\|_{L^\infty(\Omega)},|\lambda_1(t)|,\dots,|\lambda_g(t)|) \leq \mathcal{K}\max(\|\omega_0\|_{L^\infty(\Omega)},|\lambda_1^0|,\dots,|\lambda_g^0|) \quad \forall\ t\geq 0, \tag{2.31}$$

$$\max(\|\omega(t,\cdot)\|_{L^\infty(\Omega)},|\lambda_1(t)|,\dots,|\lambda_g(t)|) \to 0 \quad \text{as } t\to+\infty. \tag{2.32}$$

3. Notation and prerequisites.

3.1. Notation. We essentially keep the notation of [4]. The velocity field will now be designated by y. We write $\Omega_T := [0,T]\times\overline{\Omega}$ and $\Sigma_T := [0,T]\times\partial\Omega$. In that context we write $pr_1(t,x)=t$ and $pr_2(t,x)=x$.

For X a nonempty compact subset of $\mathbb{R}^n$ and f a continuous function $X\to\mathbb{R}$, we introduce $\Xi_X[f]$ as the following function $\mathbb{R}^{+*}\to\mathbb{R}^{+*}\cup\{+\infty\}$:

$$\Xi_X[f](\varepsilon) := \sup\left\{\eta>0 \ / \ \forall x,x'\in X,\ |x-x'|\leq\eta \Rightarrow |f(x)-f(x')|\leq\varepsilon\right\}; \tag{3.1}$$

for $x\in X$ we introduce $\Xi_X^x[f]$ as

$$\Xi_X^x[f](\varepsilon) := \sup\left\{\eta>0 \ / \ \forall x'\in X,\ |x-x'|\leq\eta \Rightarrow |f(x)-f(x')|\leq\varepsilon\right\}. \tag{3.2}$$

These two functions are clearly related to the modulus of continuity of f.

Given K a compact set in $\mathbb{R}^2$ and f a continuous function $K\to\mathbb{R}^2$, we introduce the *log-Lipschitz* norm

$$q_K(f) := \|f\|_\infty + \sup\left\{\frac{|f(x)-f(x')|}{r(|x-x'|)},\ (x,x')\in K^2,\ x\neq x'\right\}, \tag{3.3}$$

where $r(s)=s-s\ln(s)$ in $(0,1)$ and $r(s)=s$ in $[1,+\infty)$.

We call log-Lipschitz the functions for which $q_K(f)<+\infty$, and denote by $\mathcal{LL}(K)$ their space, which we endow with the norm described in (3.3). We denote by $\mathcal{L}ip(K)$ the space of Lipschitz functions on K.

In the notation of a functional space, an index 0 refers to functions with compact support.

Let R be a positive real number, large enough so that $\overline{\Omega}$ is included in the open 0-centered ball with radius R, denoted by B_R.

We consider a linear continuous extension operator $\pi : C^0(\overline{\Omega}) \to C^0_0(B_R)$, which maps any $\mathcal{LL}(\overline{\Omega})$ function to an $\mathcal{LL}_0(B_R)$ function, and any $C^1(\overline{\Omega})$ function to a $C^1_0(B_R)$ function. We fix c_π a constant such that

$$\|\pi(f)\|_{C^0(\overline{B_R})} \le c_\pi \|f\|_{C^0(\overline{\Omega})}, \qquad \|\pi(f)\|_{\mathcal{LL}(\overline{B_R})} \le c_\pi \|f\|_{\mathcal{LL}(\overline{\Omega})}, \quad \text{and} \quad \|\pi(f)\|_{C^1(\overline{B_R})} \le c_\pi \|f\|_{C^1(\overline{\Omega})}. \tag{3.4}$$

Frequently, we write (ω, λ_i) for $(\omega, \lambda_1, \dots, \lambda_g)$. Also, using the canonical isomorphism $C^0([0,T^*) \times \overline{\Omega}; \mathbb{R}) \cong C^0([0,T^*); C^0(\overline{\Omega}; \mathbb{R}))$, we often write $\omega(t)$ for $\omega(t,\cdot)$ and $y(t)$ for $y(t,\cdot)$. In the same way, $\|\omega(\cdot)\|_{L^\infty(\Omega)}$ denotes the function

$$t \mapsto \operatorname*{ess\,sup}_{x\in\Omega} |\omega(t,x)|.$$

3.2. The Wolibner–Yudovich theorem. In this section, we introduce a classical tool to deal with flows of vector fields which do not satisfy the Lipschitz condition (in fact, the existence is the Peano theorem, and the uniqueness is the Osgood theorem). One has the following theorem.

THEOREM 3.1 (Wolibner–Yudovich theorem). *Consider $T > 0$ and a vector field $y \in L^\infty([0,T]; C^0_0(B_R; \mathbb{R}^2))$ such that for some constant C*

$$q_{\overline{B_R}}(y(t)) \le C \ \text{ a.e. in } [0,T]. \tag{3.5}$$

Then there exists a unique map $\Phi^y \in C^0([0,T]\times[0,T]\times B_R; B_R)$, $(t,s,x) \mapsto \Phi^y(t,s,x)$, which is a flow of y, i.e., a function that satisfies

$$\Phi^y(t,s,x) = x + \int_s^t y(\tau, \Phi^y(\tau,s,x))d\tau \quad \forall (t,s,x) \in [0,T]\times[0,T]\times B_R. \tag{3.6}$$

Moreover, there are two positive constants C_{WY} and δ_{WY} depending only on (R,T,C) such that for any $(s,s',t,t',x,x') \in [0,T]^4 \times {B_R}^2$, one has

$$|\Phi^y(t,s,x) - \Phi^y(t',s',x')| \le C_{WY}(|s-s'|^{\delta_{WY}} + |t-t'|^{\delta_{WY}} + |x-x'|^{\delta_{WY}}). \tag{3.7}$$

For a proof of this theorem, we refer to Wolibner [9], Yudovich [11, Lemma 6.3], or Kato [7].

Estimates such as (3.5) can easily be established by using the following theorem.

THEOREM 3.2 (Wolibner). *Consider $\omega \in C^0(\overline{\Omega}; \mathbb{R})$, $\lambda_1, \dots, \lambda_g \in \mathbb{R}$. Then the function y defined in $C^0(\overline{\Omega}; \mathbb{R}^2)$ by*

$$\begin{cases} \operatorname{curl} y(x) = \omega(x) \quad \text{in } \Omega, \\ \operatorname{div} y(x) = 0 \quad \text{in } \Omega, \\ y(x)\cdot n(x) = 0 \quad \text{on } \partial\Omega, \\ \displaystyle\int_{\Gamma_i} y(x)\cdot \vec{\tau}(x)dx = \lambda_i \quad \text{for } i = 1, \dots g, \end{cases} \tag{3.8}$$

satisfies the estimate

$$\|y\|_{\mathcal{LL}(\overline{\Omega})} \le C_{\mathcal{LL}} \max\left(\|\omega\|_{C^0}, |\lambda_1|, \dots, |\lambda_g|\right). \tag{3.9}$$

We refer, for instance, to [7, Lemma 1.4] or [9].

3.3. Elementary tools. Here we recall two elementary Gronwall inequalities.

LEMMA 3.3. *Consider two vector fields $y_1 \in L^\infty([0,T];\mathcal{L}ip_0(B_R;\mathbb{R}^2))$ and $y_2 \in L^\infty([0,T];\mathcal{L}\mathcal{L}_0(B_R;\mathbb{R}^2))$ and their corresponding flows Φ_1 and Φ_2 (obtained with the Cauchy–Lipschitz theorem for the first and with the Wolibner–Yudovich theorem for the latter). Then one has, for all $t \in [0,T]$,*

(3.10)
$$\|\Phi_1(t,0,\cdot) - \Phi_2(t,0,\cdot)\|_{L^\infty(B_R)} \leq \exp\left(\int_0^t \|y_1(\tau)\|_{\mathcal{L}ip(B_R)} d\tau\right) \|y_1 - y_2\|_{L^1([0,t],L^\infty(B_R))}.$$

LEMMA 3.4. *Consider $y \in L^\infty([0,T],\mathcal{L}ip_0(B_R;\mathbb{R}^2))$ and its flow Φ^y. One has, for all $(t,x,x') \in [0,T]\times B_R^2$,*

$$|\Phi^y(t,0,x) - \Phi^y(t,0,x')| \leq \exp\left(\int_0^t \|y(\tau)\|_{\mathcal{L}ip(B_R)} d\tau\right) |x - x'|. \tag{3.11}$$

These are very classical and elementary statements.

3.4. A proposition concerning flows under the feedback law. We begin with a remark.

Remark 6. The flow Φ of $\pi(\nabla\theta)$ satisfies the following properties:

(i) for x in $\mathcal{B}\cup\gamma^-(\theta)$ and for any $t\in(0,T]$, $\exists \nu>0$ s.t. $\Phi([t-\nu,t],t,x)\subset \mathbb{R}^2\backslash\overline{\Omega}$;
(ii) for x in $\gamma^-(\theta)$ and for any $t\in[0,T)$, $\exists\nu>0$ s.t. $\Phi([t,t+\nu],t,x)\subset\Omega$;
(iii) for any $B\in\mathcal{B}$ and for any $t\in[0,T)$, $\exists\nu>0$ s.t. $\Phi([t,t+\nu],t,B)\subset\gamma_B$;
(iv) for any $A\in\mathcal{A}$, for any $t\in[0,T)$, and for $\tau<t$ s.t. $(t-\tau)\|\nabla\theta\|_\infty\leq \ell/2$, one has $\Phi(\tau,t,A)\in\gamma_A$;
(v) for any $A\in\mathcal{A}$, for any $t\in(0,T]$, and for $\tau>t$ s.t. $(\tau-t)\|\nabla\theta\|_\infty\leq\ell/2$, one has $\Phi(\tau,t,A)\in\Omega$;
(vi) for all $M_i\in\mathcal{M}$, one has $\Phi(t,0,M_i)\notin\partial\Omega\backslash[\overline{\gamma^+}\cup\overline{\gamma^-}]$ for $t>0$ s.t. $\Phi([0,t],0,M_i)\subset\overline{\Omega}$; that is, the trajectories of M_i do not touch the set $\partial\Omega\backslash[\overline{\gamma^+}\cup\overline{\gamma^-}]$ before leaving the domain.

These properties are easy to prove using the form of $\nabla\theta$, the uniqueness of the flow, and the definition of ℓ.

The idea of the following proposition is to prove that if one imposes a control of the form (2.27) with K large enough, some of the properties in Proposition 2.1 and Remark 6 are also true for the flow of the resulting velocity y.

PROPOSITION 3.5. *There exist $\kappa>0$ and $\overline{K}:=\overline{K}(\theta)>0$ such that, for any $K\geq\overline{K}$, any $T>0$, any $(\omega,\lambda_i)\in C^0(\Omega_T;\mathbb{R})\times C^0([0,T];\mathbb{R}^g)$, and any $\alpha\in C^0([0,T],\mathbb{R}^+)$ positive satisfying*

$$\alpha(t)\geq\max\left(|\lambda_1(t)|,\dots,|\lambda_g(t)|,\|\omega(t)\|_\infty\right), \tag{3.12}$$

the solution $y\in C^0(\Omega_T;\mathbb{R}^2)$ of

$$\begin{cases} \operatorname{curl} y(t,x)=\omega(t,x) \quad \text{for } (t,x)\in\Omega_T,\\ \operatorname{div} y(t,x)=0 \quad \text{for } (t,x)\in\Omega_T,\\ y(t,x)\cdot n(x)=K\alpha(t)\nabla\theta(x)\cdot n(x) \quad \text{for } (t,x)\in\Sigma_T,\\ \displaystyle\int_{\Gamma_i} y(t,x)\cdot\vec{\tau}(x)dx=\lambda_i(t) \quad \text{for } t\in[0,T] \text{ and } i\in\{1,\dots,g\} \end{cases} \tag{3.13}$$

satisfies

$$y(t,x)\cdot\nabla\theta(x)\geq\kappa K\alpha(t) \text{ in } \Omega_T, \tag{3.14}$$

and, Φ^y *being the flow of* $\pi(y)$ *in* B_R,

for any point x *in* $\mathcal{B} \cup \gamma^-(\theta)$ *and for any* $t \in (0,T)$,

$$\exists \nu > 0 \text{ such that } \Phi^y([t-\nu,t),t,x) \subset \mathbb{R}^2 \backslash \overline{\Omega}, \tag{3.15}$$

$$\exists \nu > 0 \text{ such that } \Phi^y((t,t+\nu],t,x) \subset \Omega \cup [\partial\Omega \backslash (\overline{\gamma^-} \cup \overline{\gamma^+})], \tag{3.16}$$

$$\text{for any } x \in \mathrm{Supp}(\Lambda) \text{ and for any } t \in [0,T], \text{ one has } \Phi^y(\tau,t,x) \notin \bigcup_{i=1}^{g} (\Gamma_i \cap \overline{\gamma^+}) \text{ for } \tau \in (t,T] \text{ such that } \Phi^y([t,\tau],t,x) \subset \overline{\Omega}, \tag{3.17}$$

(3.18)
$$\text{for any } A \in \mathcal{A} \text{ and for any } t \in (0,T], \text{ one has } \Phi^y(\tau,t,A) \in \gamma_A \text{ for } \tau \in [0,t) \text{ such that } c_\pi(K\|\nabla\theta\|_\infty + C_{\mathcal{LL}}) \left(\int_\tau^t \alpha(s)ds \right) \leq \ell/2,$$

(3.19)
$$\text{for any } A \in \mathcal{A} \text{ and for any } t \in [0,T), \text{ one has } \Phi^y(\tau,t,A) \in \Omega \text{ for } \tau \in (t,T] \text{ such that } c_\pi(K\|\nabla\theta\|_\infty + C_{\mathcal{LL}}) \left(\int_t^\tau \alpha(s)ds \right) \leq \ell/2,$$

$$\text{for any } A \in \mathcal{A} \text{ and for any } t \in [0,T], \text{ one has } \Phi^y(\tau,t,A) \notin \mathrm{Supp}(\Gamma) \cup \mathrm{Supp}(\Lambda) \text{ for } \tau \in [0,t) \text{ such that } \Phi^y([\tau,t],t,A) \subset \overline{\Omega}. \tag{3.20}$$

Of course, the previous flow has to be understood in the Wolibner–Yudovich sense.

The proof of Proposition 3.5 is delayed to the appendix.

Remark 7. Let us remark that, as a consequence of (3.14), the points A and B defined for $\nabla\theta$ are still valid for the velocity y described in (3.13); that is, for all t in $[0,T]$, $y(t,A)$ (resp., $y(t,B)$) is tangent to $\partial\Omega$ and pointing inside (resp., outside) γ^- (for $K \geq \overline{K}$, provided $\alpha(t) > 0$).

In what follows, we will systematically suppose $K \geq \overline{K}$.

4. Construction of the operator $\mathcal{F}$. In this section, we construct an operator $\mathcal{F} = (F, G_1, \ldots, G_g)$, whose fixed points give local in time solutions to the closed-loop system. Roughly speaking, $F[\omega,\lambda_i]$ is the solution of an initial-boundary problem, which is approximately the closed-loop system described above, where (1.7) is replaced by the following *linear* equation:

$$\partial_t F[\omega,\lambda_i] + \mathrm{div}(y_{\omega,\lambda_i} F[\omega,\lambda_i]) = 0 \quad \text{in } (0,T^*) \times \Omega,$$

where y_{ω,λ_i} is the solution of (1.5) corresponding to (ω,λ_i). Then G_i corresponds approximately to the solution of (1.8).

4.1. The domain X. First let us define the space X on which $\mathcal{F}$ is to be defined. The operator $\mathcal{F}$ is split into $\mathcal{F} = (F, G_i)$, with $F : X \to C^0([0,T] \times \overline{\Omega}; \mathbb{R})$

and $G_i : X \to C^0([0,T];\mathbb{R})$ for $i \in \{1,\dots,g\}$. We introduce

$$
\begin{aligned}
(4.1)\quad X := \Big\{ & (\omega,\lambda_i) \in C^0([0,T]\times\overline{\Omega};\mathbb{R}) \times \left[C^0([0,T];\mathbb{R})\right]^g \ \Big/ \\
& \text{(a)} \quad \omega(0,\cdot) = \omega_0, \\
& \text{(b)} \quad \|\omega(t,\cdot)\|_\infty \leq \mathcal{N}_{\omega_0,\lambda_i^0} \quad \text{for } t\in[0,T], \\
& \text{(c)} \quad \lambda_i(0) = \lambda_i^0 \quad \text{for } i = 1\dots g, \\
& \text{(d)} \quad \left\| \frac{\partial\omega}{\partial t} \right\|_{L_t^\infty(H_x^{-1})} \leq \kappa_1 \max\left(\|\omega_0\|_\infty, |\lambda_1^0|,\dots,|\lambda_g^0|\right)^2, \\
& \text{(e)} \quad |\lambda_i(t)| \leq \mathcal{M}_{\omega_0,\lambda_i^0} \quad \text{for } t\in[0,T], \\
& \text{(f)} \quad \left|\frac{d\lambda_i}{dt^-}\right|(t) \leq \kappa_2 \max\left(\|\omega_0\|_\infty, |\lambda_1^0|,\dots,|\lambda_g^0|\right)^2, \\
& \text{(g)} \quad \forall A\in\mathcal{A},\ \forall\varepsilon\in(0,1),\ \Xi_{[0,T]}[\omega(A,\cdot)](\varepsilon) \geq \tfrac{1}{c}\left(\Xi_{\gamma_A}[\omega_0](\varepsilon)\right)^{\frac{1}{\delta}}, \\
& \text{(h)} \quad \forall\varepsilon\in(0,1),\ \Xi_{[0,T]}[\|\omega(\cdot)\|_{L^\infty(\Omega)}](\varepsilon) \geq \tfrac{1}{c}\min\left[\left(\Xi_{\overline{\Omega}}[\omega_0](\varepsilon)\right)^{\frac{1}{\delta}}, \varepsilon\right]\Big\},
\end{aligned}
$$

where the constant c depends on Ω, θ, T, and (ω_0,λ_i^0) and will be chosen large enough later. The other constants are fixed as follows:

$$
(4.2)\qquad \mathcal{N}_{\omega_0,\lambda_i^0} := \left[3 + |\Omega| + \frac{V(\theta)}{\kappa}\mathcal{T}(\Gamma)\right](1+\|\Lambda\|_\infty)\max(|\lambda_1^0|,\dots,|\lambda_g^0|,\|\omega_0\|_\infty),
$$

$$
(4.3)\qquad \mathcal{M}_{\omega_0,\lambda_i^0} := \left[2 + |\Omega| + \frac{V(\theta)}{\kappa}\mathcal{T}(\Gamma)\right]\max(|\lambda_1^0|,\dots,|\lambda_g^0|,\|\omega_0\|_\infty),
$$

and δ is defined with reference to section 3.2 as

$$
(4.4)\qquad \delta := \delta_{WY}(R,T,c_\pi(C_{\mathcal{LL}} + K\|\nabla\theta\|_{\mathcal{LL}(\overline{\Omega})})\mathcal{N}_{\omega_0,\lambda_i^0}).
$$

The constants κ_1 and κ_2 depend on the domain and on the choice of θ, Λ, and Γ but not on T:

$$
(4.5)\qquad \kappa_1 := 2|\Omega|^{\frac{1}{2}}(C_{\mathcal{LL}} + K\|\nabla\theta\|_{L^\infty(\Omega)})\left[3 + |\Omega| + \frac{V(\theta)}{\kappa}\mathcal{T}(\Gamma)\right]^2(1+\|\Lambda\|_\infty)^2,
$$

$$
(4.6)\qquad \kappa_2 := |\Sigma|K\|\nabla\theta\|_{L^\infty(\Omega)}\left[3 + |\Omega| + \frac{V(\theta)}{\kappa}\mathcal{T}(\Gamma)\right]^2(1+\|\Lambda\|_\infty)^2.
$$

In (4.2)–(4.6), $|\Omega|$ stands for the Lebesgue measure of Ω and $|\Sigma|$ for the length of Σ, κ is the constant in (3.14), $V(\theta)$ is defined in (2.20), $\mathcal{T}(\Gamma)$ and $\|\Lambda\|_\infty$ are defined in (2.25), c_π is defined in (3.4), and $C_{\mathcal{LL}}$ is introduced in (3.9).

The time T is chosen in the following way:

- if $\max(|\lambda_1^0|,\dots,|\lambda_g^0|,\|\omega_0\|_\infty) = 0$, then 0 is a clear solution of the system and we pass (throughout sections 4 and 5 we will suppose $(\omega_0,\lambda_i^0) \neq (0,0)$);
- if $\|\omega_0\|_\infty = 0$, but $|\lambda_k^0| > 0$ for some $k\in\{1,\dots,g\}$, we fix

$$
(4.7)\qquad \underline{T} := \frac{|\lambda_k^0|}{2\kappa_2\max\left(\|\omega_0\|_\infty,|\lambda_1^0|,\dots,|\lambda_g^0|\right)^2};
$$

- if $\|\omega_0\|_\infty \neq 0$, then we fix

$$
(4.8)\qquad \underline{T} := \frac{\|\omega_0\|_\infty}{2\kappa_1\max\left(\|\omega_0\|_\infty,|\lambda_1^0|,\dots,|\lambda_g^0|\right)^2}.
$$

Finally, we define T as

$$T := \min\left(\frac{\ell}{4c_\pi(C_{\mathcal{LL}} + K\|\nabla\theta\|_\infty)\mathcal{N}_{\omega_0,\lambda_i^0}}, \underline{T}\right). \tag{4.9}$$

It is quite clear that X is convex, closed, and nonempty (since, for example, it contains the constant function $t \mapsto (\omega_0, \lambda_i^0)$).

Let us finish this paragraph with a remark concerning the choice of T.

Remark 8. Let us remark that T allows us to have the following properties:

- for any $A \in \mathcal{A}$ and any $(\omega, \lambda_i) \in X$, if one puts y as in (3.13) (with α satisfying (3.12)), then for any $t \in [0,T]$ one has $\Phi^y([0,t],t,A) \subset \overline{\gamma_A}$ (this is a consequence of the first part in the minimum in (4.9), using (3.18));
- for any $(\omega, \lambda_i) \in X$, one has as a consequence of the definition of $\underline{T}$ and of points (d) and (f) in (4.1), that
  - if $\underline{T}$ is defined by (4.7), then for all $t \in [0,T]$,

$$|\lambda_k(t)| \geq \frac{|\lambda_k^0|}{2} > 0, \tag{4.10}$$

  - if $\underline{T}$ is defined by (4.8), then for all $t \in [0,T]$,

$$\|\omega(t,\cdot)\|_\infty \geq \kappa_3^{-1}\frac{\|\omega_0\|_{H^{-1}(\Omega)}}{2} > 0, \tag{4.11}$$

    where κ_3 is some constant such that $\|\cdot\|_{H^{-1}(\Omega)} \leq \kappa_3\|\cdot\|_{L^\infty(\Omega)}$.

4.2. The operator F. Let us now describe the operator F. Consider $(\omega, \lambda_i) \in X$. First, we associate with (ω, λ_i) the function $\alpha_{\omega,\lambda_i} \in C^0([0,T], \mathbb{R}^{+*})$ by

$$\alpha_{\omega,\lambda_i}(t) := \max\left(|\lambda_1(t)|, \ldots, |\lambda_g(t)|, \|\omega(t)\|_\infty\right). \tag{4.12}$$

Then, we can associate the following vector field $y_{\omega,\lambda_i} \in C^0(\Omega_T, \mathbb{R}^2)$ as the solution of

$$\begin{cases} \operatorname{curl} y_{\omega,\lambda_i}(t,x) = \omega(t,x) \quad \text{for } (t,x) \in \Omega_T, \\ \operatorname{div} y_{\omega,\lambda_i}(t,x) = 0 \quad \text{for } (t,x) \in \Omega_T, \\ y_{\omega,\lambda_i}(t,x)\cdot n(x) = K\alpha_{\omega,\lambda_i}(t)\nabla\theta(x)\cdot n(x) \quad \text{for } (t,x) \in \Sigma_T, \\ \displaystyle\int_{\Gamma_i} y_{\omega,\lambda_i}(t,x)\cdot\vec{\tau}(x)dx = \lambda_i(t) \quad \text{for } t \in [0,T], \quad i = 1,\ldots,g, \end{cases} \tag{4.13}$$

where $K \geq \overline{K}(\theta)$. Then we extend this vector field to $[0,T]\times B_R$ by

$$\tilde{y}_{\omega,\lambda_i}(t,\cdot) = \pi[y_{\omega,\lambda_i}(t,\cdot)]. \tag{4.14}$$

By the Wolibner–Yudovich theorem (see section 3.2), this vector field yields a flow $\Phi^{\omega,\lambda_i} : [0,T]\times[0,T]\times B_R \to B_R$, i.e., a solution of (3.6). Now, given this flow, we can introduce the following two functions on $[0,T]\times\overline{\Omega}$ (which, roughly speaking, represent, respectively, the time and location of entrance in the domain of the point located at x at time t, when following the flow):

$$s_{\omega,\lambda_i}(t,x) := \max\left\{\tau \in [0,t],\ \Phi^{\omega,\lambda_i}(\tau,t,x) \in \overline{\gamma^-}\right\}, \tag{4.15}$$

$$a_{\omega,\lambda_i}(t,x) := \Phi^{\omega,\lambda_i}(s_{\omega,\lambda_i}(t,x),t,x), \tag{4.16}$$

with the convention that when $\{\tau \in [0,t],\ \Phi^{\omega,\lambda_i}(\tau,t,x) \in \overline{\gamma^-}\} = \emptyset$, then

$$s_{\omega,\lambda_i}(t,x) := 0,$$

and correspondingly

$$a_{\omega,\lambda_i}(t,x) := \Phi^{\omega,\lambda_i}(0,t,x).$$

Note that in all cases, one has $a_{\omega,\lambda_i}(t,x) \in \overline{\Omega}$. Note that the function s_{ω,λ_i} is not necessarily continuous (contrary to what happened in the simply connected case; see [4, equations (3.36)–(3.37)]).

Now we can define $F[\omega,\lambda_i](t,x)$ for $(t,x) \in \Omega_T$. In that order, we distinguish four cases, corresponding to different situations for $a_{\omega,\lambda_i}(t,x)$. In what follows, the constant M (which appears in (2.28)) is to be chosen large enough later.

Case α: $a_{\omega,\lambda_i}(t,x) \in \Omega \cup (\partial\Omega \backslash \overline{\gamma^-})$. This case is possible only if $s_{\omega,\lambda_i}(t,x) = 0$. In that case, we fix

$$F[\omega,\lambda_i](t,x) := \omega_0(a_{\omega,\lambda_i}(t,x)). \tag{4.17}$$

Case β: $a_{\omega,\lambda_i}(t,x) \in \overline{\gamma^-} \backslash [\mathrm{Supp}(\Gamma) \cup \mathrm{Supp}(\Lambda)]$. In that case, we fix

$$F[\omega,\lambda_i](t,x) := \omega_0(a_{\omega,\lambda_i}(t,x)) \exp\left(-M \int_0^{s_{\omega,\lambda_i}(t,x)} \alpha_{\omega,\lambda_i}(\tau) d\tau\right). \tag{4.18}$$

Case γ: $a_{\omega,\lambda_i}(t,x) \in \mathrm{Supp}(\Gamma_A)$ for some $A \in \mathcal{A}$. In that case, we fix

$$\begin{aligned} F[\omega,\lambda_i](t,x) := {} & \Big[\omega_0(a_{\omega,\lambda_i}(t,x)) - \omega_0(A)\Gamma_A(a_{\omega,\lambda_i}(t,x))\Big] \exp\left(-M \int_0^{s_{\omega,\lambda_i}(t,x)} \alpha_{\omega,\lambda_i}(\tau) d\tau\right) \\ & + \omega_0(\Phi^{\omega,\lambda_i}(0, s_{\omega,\lambda_i}(t,x), A))\Gamma_A(a_{\omega,\lambda_i}(t,x)). \end{aligned} \tag{4.19}$$

Case δ: $a_{\omega,\lambda_i}(t,x) \in \mathrm{Supp}(\Lambda_k)$ for some $k \in \{1,\dots,g\}$. In that case, we fix

$$\begin{aligned} F[\omega,\lambda_i](t,x) := {} & \Big[\omega_0(a_{\omega,\lambda_i}(t,x)) + \lambda_k^0 \Lambda_k(a_{\omega,\lambda_i}(t,x))\Big] \exp\left(-M \int_0^{s_{\omega,\lambda_i}(t,x)} \alpha_{\omega,\lambda_i}(\tau) d\tau\right) \\ & - \lambda_k(s_{\omega,\lambda_i}(t,x))\Lambda_k(a_{\omega,\lambda_i}(t,x)). \end{aligned} \tag{4.20}$$

Another way to express this is that $F[\omega,\lambda_i]$ is given by

$$\begin{aligned} F[\omega,\lambda_i](t,x) = {} & \left[\omega_0(a_{\omega,\lambda_i}(t,x)) - \sum_{A\in\mathcal{A}} \omega_0(A)\Gamma_A(a_{\omega,\lambda_i}(t,x)) + \sum_{k=1}^{g} \lambda_k^0 \Lambda_k(a_{\omega,\lambda_i}(t,x))\right] \\ & \times \exp\left(-M \int_0^{s_{\omega,\lambda_i}(t,x)} \alpha_{\omega,\lambda_i}(\tau) d\tau\right) \\ & + \sum_{A\in\mathcal{A}} \omega_0(\Phi^{\omega,\lambda_i}(0, s_{\omega,\lambda_i}(t,x), A))\Gamma_A(a_{\omega,\lambda_i}(t,x)) \\ & - \sum_{k=1}^{g} \lambda_k(s_{\omega,\lambda_i}(t,x))\Lambda_k(a_{\omega,\lambda_i}(t,x)) \end{aligned} \tag{4.21}$$

with at most one nonnull term in each summation.

We also define on $[0,T]\times\overline{\gamma^-}$

(4.22)
$$\begin{cases} \omega^\sharp(t,x) := \left[\omega_0(x) - \displaystyle\sum_{A\in\mathcal{A}} \omega_0(A)\Gamma_A(x) + \sum_{k=1}^{g} \lambda_k^0\Lambda_k(x)\right] \exp\left(-M\displaystyle\int_0^t \alpha_{\omega,\lambda_i}(\tau)d\tau\right), \\ \omega^\flat(t,x) := \displaystyle\sum_{A\in\mathcal{A}} \omega_0(\Phi^{\omega,\lambda_i}(0,t,A))\Gamma_A(x), \\ \omega^\natural(t,x) := -\displaystyle\sum_{k=1}^{g} \lambda_k(t)\Lambda_k(x). \end{cases}$$

Finally, let us call $\tilde{F}$ the same operator as F, with Λ_k replaced by 0 for each $k\in\{1,\dots,g\}$; that is, $\tilde{F}$ is constant along the flow of Φ^{ω,λ_i}, and on $\overline{\gamma^-}$, one has

$$\tilde{F}[\omega,\lambda_i](t,x) := \left[\omega_0(x) - \sum_{A\in\mathcal{A}} \omega_0(A)\Gamma_A(x)\right] \exp\left(-M\int_0^t \alpha_{\omega,\lambda_i}(\tau)d\tau\right) + \sum_{A\in\mathcal{A}} \omega_0(\Phi^{\omega,\lambda_i}(0,t,A))\Gamma_A(x). \tag{4.23}$$

4.3. The operators G_i. Define the function $\mathcal{T}_{\omega_0,\lambda_i^0}:\mathbb{R}\to\mathbb{R}$ as

$$\begin{cases} \mathcal{T}_{\omega_0,\lambda_i^0}(x) = x \quad \text{in } [-\mathcal{M}_{\omega_0,\lambda_i^0}, \mathcal{M}_{\omega_0,\lambda_i^0}], \\ \mathcal{T}_{\omega_0,\lambda_i^0}(x) = \mathcal{M}_{\omega_0,\lambda_i^0} \quad \text{in } [\mathcal{M}_{\omega_0,\lambda_i^0}, +\infty), \\ \mathcal{T}_{\omega_0,\lambda_i^0}(x) = -\mathcal{M}_{\omega_0,\lambda_i^0} \quad \text{in } (-\infty, -\mathcal{M}_{\omega_0,\lambda_i^0}]. \end{cases} \tag{4.24}$$

Let us now introduce the operators G_k, $k=1,\dots,g$. We define $G_k(\omega,\lambda_1,\dots,\lambda_g)\in C^0([0,T],\mathbb{R})$ by

(4.25)
$$G_k(\omega,\lambda_1,\dots,\lambda_g)(t) := \mathcal{T}_{\omega_0,\lambda_i^0}\left[\lambda_k^0 + \int_0^t\int_{\Gamma_k} y_{\omega,\lambda_i}(s,x)\cdot n(x)F[\omega,\lambda_i](s,x)dsdx\right].$$

5. Proof that $\mathcal{F}$ admits a fixed point. In this section, we prove that the operator $\mathcal{F} := (F, G_1,\dots,G_g)$ that we have just constructed admits a fixed point. This is done by using the Leray–Schauder fixed point theorem. Accordingly, we have to prove three properties:

- $\mathcal{F}(X)\subset X$;
- $\mathcal{F}(X)$ is compact in X for the C^0 topology;
- $\mathcal{F}$ is continuous for the C^0 topology.

We prove this in three distinct subsections.

5.1. $\mathcal{F}(X)\subset X$. The first point to prove is that, for $(\omega,\lambda_i)\in X$, $F[\omega,\lambda_i]$ is a continuous function of (t,x). Fixing $(t,x)\in[0,T]\times\overline{\Omega}$, let us prove that $F[\omega,\lambda_i]$ is continuous at the point (t,x). Again, we distinguish the four cases α, β, γ, and δ, corresponding respectively to the case when $a_{\omega,\lambda_i}(t,x)\in\Omega\cup(\partial\Omega\backslash\overline{\gamma^-})$, $a_{\omega,\lambda_i}(t,x)\in\overline{\gamma^-}\backslash[\mathrm{Supp}(\Gamma)\cup\mathrm{Supp}(\Lambda)]$, $a_{\omega,\lambda_i}(t,x)\in\mathrm{Supp}(\Gamma)$, and $a_{\omega,\lambda_i}(t,x)\in\mathrm{Supp}(\Lambda)$.

Case α: $a_{\omega,\lambda_i}(t,x)\in\Omega\cup(\partial\Omega\backslash\overline{\gamma^-})$. Therefore, $s_{\omega,\lambda_i}(t,x)=0$. By the continuity of the flow Φ^{ω,λ_i}, there exists a neighborhood of (t,x) on which $\Phi^{\omega,\lambda_i}(0,t',x')\in$

$\Omega \cup (\partial\Omega \backslash \overline{\gamma^-})$. Then the continuity of $F[\omega, \lambda_i]$ at the point (t, x) comes directly from the continuities of the flow and of ω_0, and from (4.17).

Case β: $a_{\omega,\lambda_i}(t,x) \in \overline{\gamma^-} \backslash [\mathrm{Supp}(\Gamma) \cup \mathrm{Supp}(\Lambda)]$. In this case, using (3.15) and the continuity of the flow, one sees that s_{ω,λ_i} is continuous at the neighborhood of (t, x). To be more precise the following hold:

– s_{ω,λ_i} is always upper semicontinuous, as follows from the continuity of the flow. Indeed, consider $s > s_{\omega,\lambda_i}(t,x)$; the trajectory

$$\Phi^{\omega,\lambda_i}(\tau, t, x) \quad \text{for } \tau \in [s, t]$$

does not touch $\overline{\gamma^-}$. Consequently, for (t', x') close enough to (t, x), the corresponding trajectory

$$\Phi^{\omega,\lambda_i}(\tau, t', x') \quad \text{for } \tau \in [s, t']$$

does not touch $\overline{\gamma^-}$ either. This leads to

$$\overline{\lim_{(t',x')\to(t,x)}} s_{\omega,\lambda_i}(t', x') \leq s_{\omega,\lambda_i}(t, x). \tag{5.1}$$

– s_{ω,λ_i} is lower semicontinuous in this neighborhood, as a consequence of (3.15). Indeed, for $s \in [s_{\omega,\lambda_i}(t,x) - \nu, s_{\omega,\lambda_i}(t,x))$, one has $\Phi^{\omega,\lambda_i}(s, t, x) \in B_R \backslash \overline{\Omega}$. Now using the continuity of the flow, this gives

$$\underline{\lim_{(t',x')\to(t,x)}} s_{\omega,\lambda_i}(t', x') \geq s_{\omega,\lambda_i}(t, x). \tag{5.2}$$

Then again, once we have obtained the continuity of s_{ω,λ_i}, the continuity of $F[\omega, \lambda_i]$ at the point (t, x) comes from the continuities of the flow and of ω_0, and from (4.21).

(Cases α and β are the only ones that arise in the simply connected case; see [4, Lemma 3.3].)

Case γ: $a_{\omega,\lambda_i}(t,x) \in \mathrm{Supp}(\Gamma_A)$ for some $A \in \mathcal{A}$. In this case, $s_{\omega,\lambda_i}(t,x)$ can be discontinuous, but only in the case where $a_{\omega,\lambda_i}(t,x) = A$, for the same reason as in case β. Indeed, when $a_{\omega,\lambda_i}(t,x) \neq A$, (3.15) is still valid, so the same argument stands true. So we suppose from now on that $a_{\omega,\lambda_i}(t,x) = A$. Consider (t', x') in a neighborhood of (t, x). We distinguish some subcases according to the locus of $a_{\omega,\lambda_i}(t', x')$.

- *Cases β' and δ':* $a_{\omega,\lambda_i}(t', x') \in \overline{\gamma^-} \setminus \mathrm{Supp}(\Gamma_A)$ (including $a_{\omega,\lambda_i}(t', x') \in \mathrm{Supp}(\Gamma_{A'})$ for some $A' \in \mathcal{A} \setminus \{A\}$). These cases cannot happen if the neighborhood around (t, x) is chosen small enough (this is a clear consequence of the continuity of the flow).
- *Case γ':* $a_{\omega,\lambda_i}(t', x') \in \mathrm{Supp}(\Gamma_A)$ (including A). Let us prove that in this case $a_{\omega,\lambda_i}(t', x')$ is close to $a_{\omega,\lambda_i}(t, x) = A$ and that $s_{\omega,\lambda_i}(t', x')$ is close to $s_{\omega,\lambda_i}(t, x)$ in the following sense: take a sequence (t'_n, x'_n) in the region of points in Case γ', converging to (t, x); then one has the corresponding convergences

$$a_{\omega,\lambda_i}(t'_n, x'_n) \to a_{\omega,\lambda_i}(t, x) \quad \text{and} \quad s_{\omega,\lambda_i}(t'_n, x'_n) \to s_{\omega,\lambda_i}(t, x) \quad \text{as } n \to +\infty.$$

  Indeed,
  * given $\varepsilon > 0$, one has for n large enough $s_{\omega,\lambda_i}(t'_n, x'_n) \geq s_{\omega,\lambda_i}(t, x) - \varepsilon$. If not, for some subsequence of (t'_n, x'_n) (that we still call (t'_n, x'_n)), one has $s_{\omega,\lambda_i}(t'_n, x'_n) \to \overline{s}$, with

$$\overline{s} \leq s_{\omega,\lambda_i}(t, x) - \varepsilon.$$

By continuity of the flow, $\Phi^{\omega,\lambda_i}(s_{\omega,\lambda_i}(t'_n,x'_n),t'_n,x'_n)$ converges to $\Phi^{\omega,\lambda_i}(\overline{s},t,x)$ as $n\to+\infty$. But as $\overline{s}\le s_{\omega,\lambda_i}(t,x)-\varepsilon$ and using (3.18), one must have $\Phi^{\omega,\lambda_i}(\overline{s},t,x)\in\gamma_A$, which contradicts the fact that it is a limit point of a sequence in $\overline{\gamma^-}$.

* we argue in the same way to get $s_{\omega,\lambda_i}(t'_n,x'_n)\le s_{\omega,\lambda_i}(t,x)+\varepsilon$. If this does not happen, one finds a subsequence of (t'_n,x'_n) for which $s_{\omega,\lambda_i}(t'_n,x'_n)$ converges to $\overline{s}\ge s_{\omega,\lambda_i}(t,x)+\varepsilon$. This yields a contradiction with the continuity of the flow and (3.19).

Now, using the continuity of the flow and the convergence of $s_{\omega,\lambda_i}(t'_n,x'_n)$, one gets the continuity of $a_{\omega,\lambda_i}(t'_n,x'_n)$.

It follows from the choice of T—see in particular Remark 8—that

$$F^{\omega,\lambda_i}(t,A)=\omega_0(\Phi^{\omega,\lambda_i}(0,t,A))$$

and hence that $t\mapsto F^{\omega,\lambda_i}(t,A)$ is continuous. Now, using the continuity of Γ_A, we get a neighborhood of (t,x) in which the points in the Case γ' satisfy

$$|F^{\omega,\lambda_i}(t',x')-F^{\omega,\lambda_i}(t,x)|<\epsilon.$$

- *Case* α': $a_{\omega,\lambda_i}(t',x')\in\Omega\cup(\partial\Omega\backslash\overline{\gamma^-})$. Then one has

$$F[\omega,\lambda_i](t',x')=\omega_0(\Phi^{\omega,\lambda_i}[0,t',x']).$$

But by (4.19) we also have

$$F[\omega,\lambda_i](t,x)=\omega_0\left(\Phi^{\omega,\lambda_i}[0,s_{\omega,\lambda_i}(t,x),a_{\omega,\lambda_i}(t,x)]\right)=\omega_0(\Phi^{\omega,\lambda_i}[0,t,x]).$$

(Remember $a_{\omega,\lambda_i}(t,x)=A$.) Then again, using only the continuity of the flow and the continuity of ω_0, we get that $F[\omega,\lambda_i](t,x')$ can be made arbitrarily close to $F[\omega,\lambda_i](t,x)$ if we restrict x' to a small neighborhood of x of points in Case α'.

Case δ: $a_{\omega,\lambda_i}(t,x)\in\mathrm{Supp}(\Lambda_k)$ for some $k\in\{1,\dots,g\}$. This is again, as in Case β, a situation where s_{ω,λ_i} is continuous at the neighborhood of (t,x). Then the continuity in this case is a consequence of the continuities of the flow, of ω_0 and λ_k, and of (4.21).

From the continuity of $F(\omega,\lambda_i)$ and (4.25), we get that the functions $G_k(\omega,\lambda_1,\dots,\lambda_g)$ are time continuous.

Once this is proved, we have to check that the points (a) to (h) in the definition of X are satisfied by $\mathcal{F}(\omega,\lambda_i)$ for $(\omega,\lambda_i)\in X$.

(a) That $F(\omega,\lambda_i)(0,\cdot)=\omega_0$ is a clear consequence of the construction of F.

(c) We have also $G_i(0)=\lambda_i^0$ for $i=1,\dots,g$, as a direct consequence of (4.3), (4.24), and (4.25).

(b) Let us check that for all $(t,x)\in[0,T]\times\overline{\Omega}$ one has $|F[\omega,\lambda_i](t,x)|\le\mathcal{N}_{\omega_0,\lambda_i^0}$ by separating the four cases. Let us therefore consider (t,x) which achieves the maximum of $|F[\omega,\lambda_i](t,x)|$.

Case α: Suppose $a_{\omega,\lambda_i}(t,x)\in\Omega\cup(\partial\Omega\backslash\overline{\gamma^-})$. Then one has

$$|F[\omega,\lambda_i](t,x)|=|\omega_0(a_{\omega,\lambda_i}(t,x))|\le\|\omega_0\|_\infty.$$

Case β: Suppose $a_{\omega,\lambda_i}(t,x)\in\overline{\gamma^-}\backslash[\mathrm{Supp}(\Gamma)\cup\mathrm{Supp}(\Lambda)]$. Then one has

$$\begin{aligned}|F[\omega,\lambda_i](t,x)|&=|\omega_0(a_{\omega,\lambda_i}(t,x))|\exp\left(-M\int_0^{s_{\omega,\lambda_i}(t,x)}\alpha_{\omega,\lambda_i}(\tau)d\tau\right)\\&\le\|\omega_0\|_\infty.\end{aligned}$$

Case γ: Suppose $a_{\omega,\lambda_i}(t,x) \in \mathrm{Supp}(\Gamma_A)$ for some $A \in \mathcal{A}$. It is a consequence of (3.20) (or Remark 8) that

$$|F[\omega,\lambda_i](t,A)| \leq \|\omega_0\|_\infty.$$

Then one has, using (2.22) and (4.19),

$$\begin{aligned}|F[\omega,\lambda_i](t,x)| &= |\omega^\sharp(s_{\omega,\lambda_i}(t,x),a_{\omega,\lambda_i}(t,x)) + \omega^\flat(s_{\omega,\lambda_i}(t,x),a_{\omega,\lambda_i}(t,x))| \\ &\leq 3\|\omega_0\|_\infty \leq \mathcal{N}_{\omega_0,\lambda_i^0}.\end{aligned}$$

Case δ: Suppose $a_{\omega,\lambda_i}(t,x) \in \mathrm{Supp}(\Lambda_k)$ for some $k \in \{1,\dots,g\}$. As $F[\omega,\lambda_i](t,x)$ is transported by the flow inside Ω, it suffices to prove that for $(t,x) \in [0,T] \times \mathrm{Supp}(\Lambda_k)$ one has

$$\begin{aligned}&|\omega^\sharp(t,x) + \omega^\natural(t,x)| \\ &\qquad \leq \left[3 + |\Omega| + \frac{V(\theta)}{\kappa}\mathcal{T}(\Gamma)\right](1+\|\Lambda\|_\infty)\max(|\lambda_1^0|,\dots,|\lambda_g^0|,\|\omega_0\|_\infty).\end{aligned}$$

Of course, one has

$$|\omega^\sharp(t,x)| \leq \|\omega_0\|_\infty + \|\Lambda\|_\infty|\lambda_k^0| \quad \text{on } [0,T]\times\mathrm{Supp}(\Lambda_k).$$

Now, using point (e), one gets, for $(t,x) \in [0,T]\times\mathrm{Supp}(\Lambda_k)$,

$$\begin{aligned}|\omega^\natural(t,x)| &\leq |\lambda_k(t)|\|\Lambda\|_\infty \\ &\leq \left[2 + |\Omega| + \frac{V(\theta)}{\kappa}\mathcal{T}(\Gamma)\right]\|\Lambda\|_\infty\max(|\lambda_1^0|,\dots,|\lambda_g^0|,\|\omega_0\|_\infty).\end{aligned}$$

Hence, one still gets the estimate (b) for $F(\omega,\lambda_i)$.

(e) That the functions G_i satisfy the constraint (e) is a direct consequence of (4.3) and (4.24).

(f) Point (f) is obtained as a consequence of (4.25). Consider $k \in \{1,\dots,g\}$ and $t \in (0,T]$. Then either t is a left accumulation of points where

$$\left|\lambda_k^0 + \int_0^t \int_{\Gamma_k} y_{\omega,\lambda_i}(t,x)\cdot n(x) F[\omega,\lambda_i](t,x)dx\right| \geq \mathcal{M}_{\omega_0,\lambda_i^0},$$

and in that case

$$\frac{dG_k(\omega,\lambda_i)}{dt^-} = 0,$$

or it is not, and one can write

$$\begin{aligned}\frac{dG_k(\omega,\lambda_i)}{dt^-} &= \int_{\Gamma_k} y_{\omega,\lambda_i}(t,x)\cdot n(x)F[\omega,\lambda_i](t,x)dx \\ &= K\int_{\Gamma_k}\alpha_{\omega,\lambda_i}(t)\nabla\theta(x)\cdot n(x)F[\omega,\lambda_i](t,x)dx.\end{aligned}$$

Using the fact that $(\omega,\lambda_i) \in X$ and consequently satisfies points (b) and (e), we get that

$$\|\alpha_{\omega,\lambda_i}(t)\nabla\theta(x)\cdot n(x)\|_{C^0([0,T]\times\partial\Omega)} \leq \mathcal{N}_{\omega_0,\lambda_i^0}\|\nabla\theta\|_\infty.$$

Using the fact that $F[\omega,\lambda_i]$ satisfies the estimate (b), this leads to

$$\left|\frac{dG_k(\omega,\lambda_i)}{dt^-}\right|(t) \le \kappa_2 \max\left(\|\omega_0\|_\infty, |\lambda_1^0|,\dots,|\lambda_g^0|\right)^2.$$

(d) Define

$$\check{y}_{\omega,\lambda_i} = y_{\omega,\lambda_i} - K\alpha_{\omega,\lambda_i}\nabla\theta.$$

Using (3.9) and (b) and (e) in (4.1), one gets

$$\|\check{y}_{\omega,\lambda_i}\|_{L^\infty([0,T];\mathcal{LL}(\Omega))} \le C_{\mathcal{LL}} \max(\mathcal{M}_{\omega_0,\lambda_i^0},\mathcal{N}_{\omega_0,\lambda_i^0}) = C_{\mathcal{LL}}\mathcal{N}_{\omega_0,\lambda_i^0}.$$

It follows that one has

$$\text{(5.3)}\qquad \begin{cases} \|y_{\omega,\lambda_i}\|_{L^\infty([0,T]\times\Omega)} \le C_{\mathcal{LL}}\mathcal{N}_{\omega_0,\lambda_i^0} + K\mathcal{N}_{\omega_0,\lambda_i^0}\|\nabla\theta\|_\infty, \\ \|y_{\omega,\lambda_i}\|_{L^\infty([0,T];\mathcal{LL}(\overline{\Omega}))} \le C_{\mathcal{LL}}\mathcal{N}_{\omega_0,\lambda_i^0} + K\mathcal{N}_{\omega_0,\lambda_i^0}\|\nabla\theta\|_{\mathcal{LL}}. \end{cases}$$

Moreover, we have from point (b) that

$$\max_{t\in[0,T]} \|F[\omega,\lambda_i](t,\cdot)\|_{L^\infty(\Omega)} \le \mathcal{N}_{\omega_0,\lambda_i^0}.$$

Consequently one gets

$$\|y_{\omega,\lambda_i}F[\omega,\lambda_i]\|_{L^\infty([0,T],L^\infty(\Omega))} \le (C_{\mathcal{LL}} + K\|\nabla\theta\|_\infty)\mathcal{N}^2_{\omega_0,\lambda_i^0}.$$

But it follows from the construction that $F[\omega,\lambda_i]$ satisfies

$$\partial_t F[\omega,\lambda_i] + \operatorname{div}(y_{\omega,\lambda_i}F[\omega,\lambda_i]) = 0 \text{ in } \mathcal{D}'((0,T)\times\Omega).$$

This leads to the fact that $F[\omega,\lambda_i]$ satisfies constraint (d).

(g) This point follows from (3.7), (3.18), and (4.19). Indeed, one has, for any $A\in\mathcal{A}$ and any $(t,t')\in[0,T]^2$,

$$F[\omega,\lambda_i](t,A) - F[\omega,\lambda_i](t',A) = \omega_0(\Phi^{\omega,\lambda_i}(0,t,A)) - \omega_0(\Phi^{\omega,\lambda_i}(0,t',A)),$$

with $\Phi^{\omega,\lambda_i}(0,t',A)$ and $\Phi^{\omega,\lambda_i}(0,t,A)$ in $\overline{\gamma_A}$. Hence, so that

$$|F[\omega,\lambda_i](t,A) - F[\omega,\lambda_i](t',A)| \le \varepsilon,$$

it is sufficient that $|\Phi^{\omega,\lambda_i}(0,t,A) - \Phi^{\omega,\lambda_i}(0,t',A)| \le \Xi_{\overline{\gamma_A}}[\omega_0](\varepsilon)$. Using (3.7) and (5.3), one sees that it is sufficient that

$$|t-t'| \le \left[\frac{\Xi_{\overline{\gamma_A}}[\omega_0](\varepsilon)}{C_{WY}(R,T,c_\pi\mathcal{N}_{\omega_0,\lambda_i^0}[C_{\mathcal{LL}}+K\|\nabla\theta\|_{\mathcal{LL}}])}\right]^{\frac{1}{\delta}}.$$

(h) We write $\hat{\omega} := F(\omega,\lambda_i)$. We divide the proof that $\hat{\omega}$ satisfies point (h) into two steps. First we estimate $\hat{t}-t$ so that

$$\text{(5.4)}\qquad \|\hat{\omega}(t,\cdot)\|_\infty - \|\hat{\omega}(\hat{t},\cdot)\|_\infty \le \varepsilon \quad \text{for } \hat{t} > t,$$

and then we estimate $\hat{t}-t$ so that

$$\text{(5.5)}\qquad \|\hat{\omega}(t,\cdot)\|_\infty - \|\hat{\omega}(\hat{t},\cdot)\|_\infty \ge -\varepsilon \quad \text{for } \hat{t} > t.$$

In what follows, we suppose $\hat{t} > t$.

- *Sufficient condition for* (5.4). First, we state a lemma.
LEMMA 5.1. *There exist* $\rho > 0$ *and* $\overline{\eta} > 0$ *such that for any* $\eta \in (0, \overline{\eta})$ *and any* $\overline{x} \in \overline{\Omega}$, *there is some* $\tilde{x} \in \Omega$ *such that*

$$d(\tilde{x}, \overline{x}) \leq \eta \quad \textit{and} \quad d(\tilde{x}, \partial\Omega) \geq \rho\eta.$$

Proof of Lemma 5.1. We introduce $\mathcal{V}$, a tubular neighborhood of $\partial\Omega$ in $\mathbb{R}^2$. It is easy to see that the following procedure, which associates a point $\tilde{x}$ to any $\overline{x}$, allows us to find relevant ρ and $\overline{\eta}$:
  - for $\overline{x}$ in $\mathcal{V} \cap \overline{\Omega}$, we pick a point $\tilde{x}$ in the direction of the inner normal to $\partial\Omega$,
  - for $\overline{x}$ in $\overline{\Omega} \backslash \mathcal{V}$, we pick $\tilde{x} = \overline{x}$.

The details are left to the reader.
We go back to the sufficient condition for (5.4). Let us consider $t \in [0, T)$. We introduce $\overline{x} \in \overline{\Omega}$ such that

$$|\hat{\omega}(t, \overline{x})| = \|\hat{\omega}(t, \cdot)\|_{\infty}.$$

We have two possible situations.
  - *First situation:* $d(\overline{x}, \overline{\gamma^+}) \geq \ell/4$. Then, considering (4.9), (5.3), the fact that $\hat{\omega}$ is constant along the flow of y_{ω,λ_i}, and the fact that a point following the flow of y_{ω,λ_i} can leave the domain only through $\overline{\gamma^+}$, one deduces that for $\hat{t} \in (t, T]$

$$\|\hat{\omega}(\hat{t})\|_{C^0(\overline{\Omega})} \geq |\hat{\omega}(\hat{t}, \Phi^{\omega,\lambda_i}(\hat{t}, t, \overline{x}))| = |\hat{\omega}(t, \overline{x})| = \|\hat{\omega}(t)\|_{C^0(\overline{\Omega})},$$

    which is stronger than (5.4).
  - *Second situation:* $d(\overline{x}, \overline{\gamma^+}) \leq \ell/4$, and hence, considering the definition of ℓ, one has

$$d(\overline{x}, \overline{\gamma^-}) \geq 3\ell/4.$$

    Considering (4.9) and (5.3), one deduces that

$$\Phi^{\omega,\lambda_i}(s, t, x) \in \overline{\Omega} \backslash \overline{\gamma^-} \quad \forall s \in [0, T] \text{ and any } x \in \overline{\Omega} \text{ such that } d(x, \overline{x}) < \ell/2. \tag{5.6}$$

    Moreover, using the fact that $\hat{\omega}$ is constant along the flow of y_{ω,λ_i} and the fact that a point following the flow of y_{ω,λ_i} cannot leave the domain except through $\overline{\gamma^+}$, we see that in order to have (5.4), it is sufficient that

$$|t - \hat{t}| \leq d(\tilde{x}, \overline{\gamma^+}) / \|y_{\omega,\lambda_i}\|_{L^\infty(\Omega_T)},$$

    where $\tilde{x} \in \Omega$ is some point satisfying

$$|\hat{\omega}(t, \tilde{x}) - \hat{\omega}(t, \overline{x})| \leq \varepsilon. \tag{5.7}$$

    Using (5.6), one sees that to get (5.7), it is sufficient to have

$$d(x, \overline{x}) < \ell/2 \text{ and } |\Phi^{\omega,\lambda_i}(0, t, \tilde{x}) - \Phi^{\omega,\lambda_i}(0, t, \overline{x})| \leq \Xi_{\overline{\Omega}}[\omega_0](\varepsilon),$$

and hence, using (3.7), that

$$|\tilde{x}-\overline{x}| \le \min\left\{ \left(\frac{\Xi_{\overline{\Omega}}[\omega_0](\varepsilon)}{C_{WY}(R,T,c_\pi(C_{\mathcal{LL}}+K\|\nabla\theta\|_{\mathcal{LL}})\mathcal{N}_{\omega_0,\lambda_i^0})} \right)^{\frac{1}{\delta}}, \ell/4 \right\}.$$

It follows from Lemma 5.1 that one can find such a point $\tilde{x}$ satisfying (5.7) such that

$$d(\tilde{x},\overline{\gamma^+}) \ge m \min\Big(\big(\Xi_{\overline{\Omega}}[\omega_0](\varepsilon)\big)^{\frac{1}{\delta}}, 1\Big)$$

for some $m>0$ independent from ε. So, using (5.3), one deduces that in order to have $\|\hat{\omega}(t,\cdot)\|_\infty - \|\hat{\omega}(\hat{t},\cdot)\|_\infty \ge -\varepsilon$ (given $\varepsilon\in(0,1)$), it is sufficient that

$$|t-\hat{t}| \le \frac{1}{c}\min\Big\{ \big(\Xi_{\overline{\Omega}}[\omega_0](\varepsilon)\big)^{\frac{1}{\delta}}, \varepsilon\Big\}$$

for some $c>0$ large enough depending on (ω_0,λ_i^0), on the domain, and on K and on θ, but not on (ω,λ_i) or on ε.

- *Sufficient condition for* (5.5). Let us prove that in order for (5.5) to happen, it is sufficient that

$$\|\hat{\omega}_{|\gamma^-}(s)\|_\infty - \|\hat{\omega}_{|\gamma^-}(t)\|_\infty \le \varepsilon \quad \forall\ s\in[t,\hat{t}]. \tag{5.8}$$

Indeed, let us consider $\hat{x}\in\overline{\Omega}$ such that

$$|\hat{\omega}(\hat{t},\hat{x})| = \|\hat{\omega}(\hat{t},\cdot)\|_\infty.$$

– If $\Phi^{\omega,\lambda_i}([t,\hat{t}],\hat{t},\hat{x})$ meets $\overline{\gamma^-}$, then clearly, using again the fact that $\hat{\omega}$ is constant along the flow of y_{ω,λ_i}, one deduces that

$$\|\hat{\omega}(\hat{t})\|_\infty \le \sup_{s\in[t,\hat{t}]} \|\hat{\omega}_{|\gamma^-}(s)\|_\infty,$$

and hence

$$\begin{aligned}\|\hat{\omega}(\hat{t})\|_\infty - \|\hat{\omega}(t)\|_\infty &\le \sup_{s\in[t,\hat{t}]} \|\hat{\omega}_{|\gamma^-}(s)\|_\infty - \|\hat{\omega}(t)\|_\infty \\ &\le \sup_{s\in[t,\hat{t}]} \|\hat{\omega}_{|\gamma^-}(s)\|_\infty - \|\hat{\omega}_{|\gamma^-}(t)\|_\infty.\end{aligned}$$

– Otherwise, it is quite clear that

$$\|\hat{\omega}(\hat{t})\|_\infty = |\hat{\omega}(\hat{t},\hat{x})| = |\hat{\omega}(t,\Phi^{\omega,\lambda_i}(t,\hat{t},\hat{x}))| \le \|\hat{\omega}(t)\|_{C^0(\overline{\Omega})},$$

which is stronger than (5.5).

Now, using (2.22) and the decomposition (4.22) of $F[\omega,\lambda_i]$, one gets on $(0,T)\times\gamma^-$,

$$\|\omega^\flat(t)-\omega^\flat(\hat{t})\|_\infty \le \max_{A\in\mathcal{A}} |\omega_0(\Phi^{\omega,\lambda_i}(0,t,A)) - \omega_0(\Phi^{\omega,\lambda_i}(0,\hat{t},A))|, \tag{5.9}$$

and

$$\frac{\partial}{\partial t}\left(\omega^\natural+\omega^\sharp\right)=-\sum_{k=1}^{g}\Lambda_k(x)\frac{d}{dt}\lambda_k(t)-M\alpha_{\omega,\lambda_i}(t) \tag{5.10}$$
$$\times\left[\omega_0(x)-\sum_{A\in\mathcal{A}}\omega_0(A)\Gamma_A(x)+\sum_{k=1}^{g}\lambda_k^0\Lambda_k(x)\right]\exp\left(-M\int_0^t\alpha_{\omega,\lambda_i}(\tau)d\tau\right).$$

In order to have (5.8), it is sufficient that

$$\|\omega^\flat(t)-\omega^\flat(\hat{t})\|_\infty\leq\varepsilon/2\quad\text{and}\quad\|\omega^\natural(t)+\omega^\sharp(t)-\omega^\natural(\hat{t})+\omega^\sharp(\hat{t})\|_\infty\leq\varepsilon/2$$

and hence, using points (b), (e), and (f) in the definition of X and (3.18), (5.9), and (5.10), that

$$|t-\hat{t}|\leq\frac{1}{c}\min(\varepsilon,\Xi_{\overline{\Omega}}[\omega_0](\varepsilon))$$

for some c large enough depending on ω_0 and λ_i^0, but not on ε, $\hat{t}$, or (ω,λ_i). So in all cases, in order to get (5.4) and (5.5), it is sufficient to have

$$|t-\hat{t}|\leq\frac{1}{c}\min\left(\varepsilon,\Xi_{\overline{\Omega}}[\omega_0](\varepsilon)\right)$$

for proper c, which allows us to conclude.

This ends the proof that $\mathcal{F}(X)\subset X$. □

5.2. $\mathcal{F}(X)$ is compact in $C^0([0,T]\times\overline{\Omega};\mathbb{R})\times[C^0([0,T];\mathbb{R})]^g$. Consider a sequence $(\omega_n,\lambda_i^n)_{n\geq1}$ in X. Let us prove that one can extract a converging subsequence from $\mathcal{F}(\omega_n,\lambda_i^n)$ in $C^0([0,T]\times\overline{\Omega};\mathbb{R})\times\left[C^0([0,T];\mathbb{R})\right]^g$. We will have to extract subsequences from $(\omega_n,\lambda_i^n)_{n\geq1}$ several times to get the convergence and, in order to avoid too heavy notation, we will continue to write those subsequences (ω_n,λ_i^n) (instead of $(\omega_{\varphi(n)},\lambda_i^{\varphi(n)})$, for instance). Moreover, we put an index n to objects constructed in section 4.2, corresponding to (ω_n,λ_i^n): each (ω_n,λ_i^n) yields a function α_n by (4.12) and then a vector field y_{ω_n,λ_i^n} on $[0,T]\times\overline{\Omega}$ by (4.13), which in turn yields a vector field $\tilde{y}_{\omega_n,\lambda_i^n}$ by (4.14). To these $\tilde{y}_{\omega_n,\lambda_i^n}$ one can associate a flow Φ_n by (3.6).

Using (3.9), (4.1), and (4.13), one easily gets that for some $C>0$,

$$q_{\overline{\Omega}}(y_{\omega_n,\lambda_i^n}(t,\cdot)-K\alpha_n(t)\nabla\theta(\cdot))\leq C\quad\forall t\in[0,T],\ \forall n\geq1,$$

and hence, using (4.1) again, the regularity of the function $\nabla\theta$, and (3.4), that for some $C'>0$,

$$q_{\overline{B_R}}(\tilde{y}_{\omega_n,\lambda_i^n}(t,\cdot))\leq C'\quad\forall t\in[0,T],\ \forall n\geq1.$$

Therefore, it follows from the Wolibner–Yudovich (see (3.7)) and Ascoli–Arzela theorems that Φ_n is relatively compact in $C^0([0,T]\times[0,T]\times B_R;B_R)$, say

$$\Phi_n\longrightarrow\overline{\Phi}\quad\text{in } C^0([0,T]\times[0,T]\times B_R;B_R). \tag{5.11}$$

We now have to prove

$$F(\omega_n,\lambda_i^n)\longrightarrow\overline{F}\quad\text{in } C^0([0,T]\times\overline{\Omega})\text{ as } n\to+\infty \tag{5.12}$$

for a certain $\overline{F}$ in $C^0([0,T]\times\overline{\Omega};\mathbb{R})$. To that end, let us first prove that one can get some compactness on the sequence $F(\omega_n,\lambda_i^n)$ on the boundary, precisely

$$F(\omega_n,\lambda_i^n)_{|[0,T]\times\overline{\gamma^-}} \longrightarrow \overline{\omega} \text{ uniformly on } [0,T]\times\overline{\gamma^-} \text{ as } n\to+\infty, \tag{5.13}$$

for some $\overline{\omega}$ in $C^0([0,T]\times\overline{\gamma^-};\mathbb{R})$.

For this, let us decompose $F(\omega_n,\lambda_i^n)_{|[0,T]\times\overline{\gamma^-}}$ into

$$F(\omega_n,\lambda_i^n)=\omega_n^\flat+\omega_n^\sharp+\omega_n^\natural \quad \text{on } [0,T]\times\overline{\gamma^-},$$

as described in (4.22). Then we prove (5.13) in several steps.

- First, it follows from points (b) and (g) in the definition of X and the Ascoli–Arzela theorem that one can extract subsequences s.t.

$$\omega_n^\flat \longrightarrow \overline{\omega}^\flat:=\sum_{A\in\mathcal{A}}\overline{\omega}(\cdot,A)\Gamma_A(\cdot) \quad \text{uniformly on } [0,T]\times\overline{\gamma^-} \text{ as } n\to+\infty. \tag{5.14}$$

- From the Ascoli–Arzela theorem and points (b) and (e) in the definition of X, one deduces that the sequence of functions $\Upsilon_n : t\mapsto M\int_0^t\alpha_n(\tau)d\tau$ is relatively compact in $C^0([0,T];\mathbb{R}^+)$ and hence, up to a subsequence, one has

$$\omega_n^\sharp \longrightarrow \overline{\omega}^\sharp:=\left[\omega_0(\cdot)-\sum_{A\in\mathcal{A}}\omega_0(A)\Gamma_A(\cdot)+\sum_{i=1}^g\lambda_i^0\Lambda_i(\cdot)\right]\exp(-\overline{\Upsilon}), \tag{5.15}$$
$$\text{uniformly on } [0,T]\times\overline{\gamma^-} \text{ as } n\to+\infty.$$

- Extracting again a subsequence if necessary, one can get from points (e) and (f) in the definition of X that $\lambda_i^n\to\overline{\lambda_i}$ in $C^0([0,T],\mathbb{R})$. Consequently one gets

$$\omega_n^\natural \longrightarrow \overline{\omega}^\natural:=\sum_{i=1}^g\Lambda_i(\cdot)\Big[-\overline{\lambda_i}\Big] \quad \text{uniformly on } [0,T]\times\overline{\gamma^-} \text{ as } n\to+\infty. \tag{5.16}$$

We get (5.13) from (5.14), (5.15), and (5.16).

Furthermore, using point (h) and the Ascoli–Arzela theorem one can extract a converging subsequence from $\|\omega_n(\cdot)\|_{L^\infty(\Omega)}$:

$$\|\omega_n(t)\|_{L^\infty(\Omega)}\longrightarrow\overline{N}(t) \text{ uniformly on } [0,T].$$

This yields, as $n\to+\infty$,

$$\begin{aligned}\alpha_n(t):=\max(|\lambda_1^n(t)|,\dots,|\lambda_g^n(t)|,\|\omega_n(t)\|_\infty)\\ \longrightarrow\overline{\alpha}(t):=\max(|\overline{\lambda}_1(t)|,\dots,|\overline{\lambda}_g(t)|,\overline{N}(t)) \text{ in } C^0([0,T];\mathbb{R}^+).\end{aligned} \tag{5.17}$$

Let us prove that this implies that $\overline{\Phi}$ satisfies the conclusions of Proposition 3.5. We proceed exactly as for [4, equation (3.57)ff]. Let us recall the argument. Define

$$\check{y}_{\omega_n,\lambda_i^n}:=y_{\omega_n,\lambda_i^n}-K\alpha_n(t)\nabla\theta.$$

Hence $\check{y}_{\omega_n,\lambda_i^n}$ satisfies (3.8) for (ω_n,λ_i^n). From (4.1) and usual elliptic estimates concerning (3.8), one deduces that, for any $r\in(2,+\infty)$,

$$\check{y}_{\omega_n,\lambda_i^n} \text{ is bounded in } C^0([0,T],W^{1,r}(\Omega;\mathbb{R}^2)),$$
$$\frac{\partial}{\partial t}\check{y}_{\omega_n,\lambda_i^n} \text{ is bounded in } L^\infty([0,T],H^{-1}(\Omega;\mathbb{R}^2)).$$

Using [8, Appendix C, Lemma C1] with $X=W^{1,r}(\Omega;\mathbb{R}^2)$ and $Y=H^{-1}(\Omega;\mathbb{R}^2)$, one deduces that $\check{y}_{\omega_n,\lambda_i^n}$ is relatively compact in $C^0([0,T],W^{1,r}(\Omega;\mathbb{R}^2)-w)$ and hence, using the Rellich–Kondrakov theorem, relatively compact in $C(\Omega_T)$. Hence using (5.17), the sequence y_{ω_n,λ_i^n} is itself relatively compact in $C(\Omega_T)$.

Hence, up to a subsequence, one has

$$y_{\omega_n,\lambda_i^n}\longrightarrow \overline{Y} \quad \text{in } C^0(\Omega_T;\mathbb{R}^2).$$

As in [4, equation (3.60)], we get that $\overline{\Phi}=\Phi^{\pi(\overline{Y})}$: this is a consequence of the definition of the flow and of the dominated convergence theorem.

Now let us observe that $\overline{Y}$ satisfies the assumptions of Proposition 3.5. This is a consequence of the fact that the sequence (ω_n,λ_i^n) satisfies them and of the fact that

$$\|\operatorname{curl}\overline{Y}\|_\infty \le \liminf_{n\to+\infty}\|\omega_n\|_\infty.$$

Note that by (5.17) and by (4.10)–(4.11), one has $\overline{\alpha}>0$.

Let us now show that, together with (5.11), this yields a convergence for $F(\omega_n,\lambda_i^n)$ in $[0,T]\times\overline{\Omega}$. The flow $\overline{\Phi}$ yields functions $\overline{s}$ and $\overline{a}$ as for (4.16) with Φ^{ω,λ_i} replaced by $\overline{\Phi}$. Then one can define $\overline{F}$ by

$$\overline{F}(t,x):=\overline{\omega}(\overline{s},\overline{a}), \tag{5.18}$$

where we extend the definition of $\overline{\omega}$ on $\{0\}\times\overline{\Omega}$ by ω_0. Note that in this setting, the function $\overline{\omega}$ is well-defined and continuous in $(\{0\}\times\overline{\Omega})\cup([0,T]\times\overline{\gamma^-})$.

We have determined the potential limit $\overline{F}$; it remains to prove (5.12). Toward this end, let us prove the following equivalent assertion (by using a compactness argument):

(5.19)
$$\forall\varepsilon>0 \text{ and } \forall(t,x)\in[0,T]\times\overline{\Omega},\exists N\in\mathbb{N} \text{ and } \exists\mathcal{V} \text{ a vicinity of } (t,x) \text{ in } [0,T]\times\overline{\Omega}$$
$$\text{such that } \forall n\ge N, \text{ one has } \|F(\omega_n,\lambda_i^n)-\overline{F}\|_{C^0(\mathcal{V})}\le\varepsilon.$$

To prove (5.19), we fix $\varepsilon>0$ and $(t,x)\in[0,T]\times\overline{\Omega}$, and discuss them relative to the location of $\overline{a}(t,x)$.

Case α: $\overline{a}(t,x)$ in $\Omega\cup(\partial\Omega\backslash\overline{\gamma^-})$. Then, by continuity of $\overline{\Phi}$, one has, for any (t',x') in a neighborhood $\mathcal{V}_1$ of (t,x) in $[0,T]\times\overline{\Omega}$, that $\overline{a}(t',x')\in\Omega\cup(\partial\Omega\backslash\overline{\gamma^-})$. Then by (5.11) we get that for n large enough, $a_n(t',x')\in\Omega\cup(\partial\Omega\backslash\overline{\gamma^-})$ for $(t',x')\in\mathcal{V}_1$. So on $\mathcal{V}_1$, for such n, the expression of $F(\omega_n,\lambda_i^n)$ is $\omega_0(\Phi_n(t,0,x))$. So enlarging N and reducing $\mathcal{V}_1$ if necessary, using (5.18), we get the conclusion of (5.19) in this case.

Cases β and δ: $\overline{a}(t,x)$ in $\overline{\gamma^-}\backslash\operatorname{Supp}(\Gamma)$. In this case—remember that $\overline{\Phi}$ satisfies the conclusions of Proposition 3.5—one can show that (5.1)–(5.2) is true for $\overline{s}$, exactly as in Case β in section 5.1. Hence, $(\overline{s},\overline{a})$ is continuous in a neighborhood of (t,x). Let us distinguish two subcases.

– *Subcase* (i): Suppose $\overline{s}(t,x) > 0$. In some neighborhood $\mathcal{V}_1$ of (t,x) in $[0,T] \times \overline{\Omega}$, one has $\overline{s}(t',x') > 0$. We introduce a neighborhood $\mathcal{W}$ of $(\overline{s}(t,x), \overline{a}(t,x))$ in $[0,T] \times (\overline{\gamma^-} \setminus \mathrm{Supp}(\Gamma))$, small enough that on it,

$$|\overline{\omega}(\tau,y) - \overline{\omega}(\overline{s}(t,x), \overline{a}(t,x))| \leq \varepsilon/2$$

(and $\tilde{\mathcal{W}}$ containing $\mathcal{W}$ in which this is valid with ε). Reducing $\mathcal{V}_1$ if necessary, we have

$$(\overline{s}(t',x'), \overline{a}(t',x')) \in \mathcal{W}$$

for (t',x') in $\mathcal{V}_1$. Using (3.15)–(3.16) and (5.11), one gets that for N large enough, one has

$$(s_n(t',x'), a_n(t',x')) \in \tilde{\mathcal{W}}$$

for (t',x') in $\mathcal{V}_1$, which yields the conclusion of (5.19).

– *Subcase* (ii): Suppose $\overline{s}(t,x) = 0$. Let us call $\mathcal{W}$ a vicinity of $(0, \overline{a}(t,x))$ in $(\{0\} \times \overline{\Omega}) \cup ([0,T] \times \overline{\gamma^-})$ such that for (τ,y) in $\mathcal{W}$ one has $|\overline{\omega}(\tau,y) - \overline{\omega}(0, \overline{a}(t,x))| \leq \varepsilon/2$. For (t',x') in a certain neighborhood $\mathcal{V}$ of (t,x) in $[0,T] \times \overline{\Omega}$ and N large enough, we have, for all $n \geq N$, either $s_n(t',x') \in pr_1(\mathcal{W})$ or $a_n(t',x') \in pr_2(\mathcal{W})$, because otherwise, we could find a subsequence for which $\Phi_n(\cdot,t',x')$ meets $\overline{\gamma^-} \setminus pr_2(\mathcal{W})$ for any n, which would be in contradiction with (5.11). With (5.13), this yields again the conclusion of (5.19).

Case γ: $\overline{a}(t,x)$ in $\mathrm{Supp}(\Gamma_A)$ for some $A \in \mathcal{A}$. We divide again into subcases:

– *Subcase* (i): Suppose $\overline{a}(t,x) \neq A$. Then one can reproduce the proof of the previous Cases β and δ if we take care that $\mathcal{W}$ stays at positive distance from $[0,T] \times \{A\}$.

– *Subcase* (ii): Suppose $\overline{a}(t,x) = A$ and $\overline{s}(t,x) > 0$. Note that in this case, $\overline{F}(t,x) = \omega_0(\overline{\Phi}(0,t,x))$ (thanks to (4.19) and (5.14)–(5.16)).
We fix $\mathcal{W}_1$ as an open vicinity of $(\overline{s}(t,x), \overline{a}(t,x))$ in $[0,T] \times \overline{\gamma^-}$ and $\mathcal{W}_2$ as an open vicinity of $(0, \overline{\Phi}(0,t,x))$ in $\{0\} \times \overline{\Omega}$, small enough such that, on both $\mathcal{W}_1$ and $\mathcal{W}_2$, we have $|\overline{\omega}(t',x') - \overline{\omega}(0, \overline{a}(t,x))| \leq \varepsilon/2$. Reduce them so that they are disjoint (this is possible thanks to (3.18)). Let us prove that for (t',x') in some neighborhood of (t,x) and for n large enough, we have

$$(s_n(t',x'), a_n(t',x')) \in \mathcal{W}_1 \cup \mathcal{W}_2. \tag{5.20}$$

If not, we would have an increasing sequence of integers $\varphi(n)$ and a sequence of points (t'_n, x'_n) converging to (t,x), for which

$$(s_{\varphi(n)}(t'_n,x'_n), a_{\varphi(n)}(t'_n,x'_n)) \notin \mathcal{W}_1 \cup \mathcal{W}_2.$$

By compactness of $[0,T] \times \overline{\Omega}$, one would have, up to a subsequence,

$$(s_{\varphi(n)}(t'_n,x'_n), a_{\varphi(n)}(t'_n,x'_n)) \to (\hat{s},\hat{a}) \notin \mathcal{W}_1 \cup \mathcal{W}_2.$$

This would be in contradiction with $(t'_n,x'_n) \to (t,x)$ and (5.11):

– Suppose indeed that $\hat{s} > 0$. By continuity of $\overline{\Phi}$ one has

$$\overline{\Phi}(s_{\varphi(n)}(t'_n, x'_n), t'_n, x'_n) \to \overline{\Phi}(\hat{s}, t, x) \quad \text{as } n \to +\infty.$$

This involves $\hat{s} = \overline{s}$, because for $\tau \neq s$, we have $\overline{\Phi}(\tau, t, x) \notin \overline{\gamma}$. Consequently, we have $\hat{a} \in \overline{\gamma^-} \backslash pr_2(\mathcal{W}_1)$. But the trajectory from x to $\overline{\Phi}(0, t, x)$ has no such point, so this is impossible.

– Suppose now that $\hat{s} = 0$. Then by (5.11) one should have

$$a_{\varphi(n)}(t'_n, x'_n) \sim \overline{\Phi}(0, t'_n, x'_n) \to \overline{\Phi}(0, t, x)$$

and hence $(\hat{s}, \hat{a}) \in \mathcal{W}_2$.

Now (5.20) gives again the conclusion in (5.19).

– *Subcase* (iii): Suppose $\overline{a}(t, x) = A$ and $\overline{s}(t, x) = 0$. Again, we have $\overline{F}(t, x) = \omega_0(\overline{\Phi}(0, t, x))$. We fix $\mathcal{W}$ as a vicinity of $(0, A)$ in $(\{0\} \times \overline{\Omega}) \cup ([0, T] \times \overline{\gamma^-})$ on which again $|\overline{\omega}(t', x') - \overline{\omega}(0, \overline{a}(t, x))| \leq \varepsilon/2$ occurs. Then again, as in Subcase (ii) one can see that, for (t', x') in a small neighborhood of (t, x) and n large enough, one has

$$(s_n(t', x'), a_n(t', x')) \in \mathcal{W},$$

which leads to the conclusion.

So in all cases (5.19) is obtained; thus we get (5.12), and then the relative compactness of the sequence $G_i(\omega_n, \lambda_1^n, \ldots, \lambda_g^n)$ follows. This concludes this section.

5.3. $\mathcal{F}$ is continuous for the $C^0([0,T] \times \overline{\Omega}; \mathbb{R}) \times [C^0([0,T]; \mathbb{R})]^g$ topology. Using the previous section; we see that it is enough to prove that if $(\omega_n, \lambda_i^n) \to (\overline{\omega}, \overline{\lambda}_i)$ for the C^0 topology, then $F[\omega_n, \lambda_i^n] \to F(\overline{\omega}, \overline{\lambda})$ pointwise. This is essentially the same argument as in the previous section; we do not repeat it. Now, as the convergence of $F(\omega_n, \lambda_i^n)$ is established, obtaining the convergence of $G_k(\omega_n, \lambda_i^n)$, $k = 1, \ldots, g$, is straightforward.

5.4. Conclusion. Hence we get by the Leray–Schauder fixed point theorem a fixed point $(\omega^*, \lambda_i^*) \in X$ of the operator $\mathcal{F}$ described in section 4.2.

It follows from the construction that on $[0, T] \times \Omega$, one has

$$\partial_t F(\omega^*, \lambda_i^*) + \operatorname{div}(y_{\omega^*, \lambda_i^*} F(\omega^*, \lambda_i^*)) = 0 \tag{5.21}$$

and

$$F(\omega^*, \lambda_i^*)_{|t=0} = \omega_0. \tag{5.22}$$

Let us assume for the moment that the following lemma is proven.

LEMMA 5.2. *If M has been chosen large enough (depending on Ω, Σ, θ, and K), then for any $k \in \{1, \ldots, g\}$ and all $t \in [0, T]$,*

$$\left| \lambda_k^0 + \int_0^t \int_{\Gamma_k} y_{\omega^*, \lambda_i^*}(\sigma, x).n(x) \omega^*(\sigma, x) dx d\sigma \right| \leq \mathcal{M}_{\omega_0, \lambda_i^0}. \tag{5.23}$$

If (5.23) is true, then by (1.8) one has

$$\lambda_k^* = G_k(\omega^*, \lambda_1^*, \ldots, \lambda_g^*)(t) = \lambda_k^0 + \int_0^t \int_{\Gamma_k} y_{\omega^*, \lambda_i^*}(\sigma, x).n(x) \omega^*(\sigma, x) dx d\sigma. \tag{5.24}$$

Hence $(\omega^*, \lambda_1^*, \dots, \lambda_g^*)$ satisfies (1.6)–(1.8). Moreover, this fixed point satisfies the initial conditions (2.30). That $F(\omega^*, \lambda_i^*)$ satisfies the boundary condition (2.27)–(2.28) is a clear consequence of the construction of $\mathcal{F}$. So it remains only to prove Lemma 5.2.

Proof of Lemma 5.2. In the proof of Lemma 5.2, we will not use the specific form of T (in particular, Remark 8). This will be useful in section 6. Denote $y^* := y_{\omega^*,\lambda_i^*}$, $\alpha^* := \alpha_{\omega^*,\lambda_i^*}$, and $\Phi^* := \Phi^{\omega^*,\lambda_i^*}$. For any $k \in \{1, \dots, g\}$, we have (using (2.22), (2.23), and (4.22))

$$\begin{aligned}
&\int_0^t \int_{\Gamma_k} y^*(\sigma,x).n(x)F[\omega^*,\lambda_i^*](\sigma,x)dxd\sigma \\
&\quad = K\int_0^t \int_{\Gamma_k} \alpha^*(\sigma)\nabla\theta(x).n(x)\omega^*(\sigma,x)dxd\sigma \\
&\quad = K\int_0^t \alpha^*(\sigma)\left[-\lambda_k^*(\sigma) + \int_{\Gamma_k\cap\gamma^-} \nabla\theta(x).n(x)\omega^{*\sharp}(x)\right. \\
&\quad \left. + \int_{\Gamma_k\cap\gamma^+} \nabla\theta(x).n(x)\omega^*(\sigma,x)\right]dxd\sigma \\
&\quad = -K\int_0^t \alpha^*(\sigma)\lambda_k^*(\sigma)dxd\sigma \\
&\quad +K\int_0^t \alpha^*(\sigma)\int_{\Gamma_k\cap\gamma^-} \nabla\theta(x).n(x)\left[\omega_0(x) + \lambda_k^0\Lambda_k(x)\right] \\
&\quad \cdot \exp\left(-M\int_0^\sigma \alpha^*(\tau)d\tau\right)dxd\sigma \\
&\quad +K\int_0^t \int_{\Gamma_k\cap\gamma^+} \alpha^*(\sigma)\nabla\theta(x).n(x)\omega^*(\sigma,x)dxd\sigma. \qquad (5.25)
\end{aligned}$$

Put

$$\mathcal{C}_0 = -\int_{\gamma^-} \nabla\theta(x).n(x)\left|\omega_0(x) + \sum_{i=1}^g \lambda_i^0\Lambda_i(x)\right|dx \qquad (5.26)$$

and

$$\mathcal{C}_1 = \int_{\gamma^-} |\nabla\theta(x).n(x)|dx. \qquad (5.27)$$

We will show that (5.23) is valid provided M is large enough (in terms of Ω, Σ, θ, and K) to satisfy

$$\frac{K\mathcal{C}_0}{M} < \frac{\max(|\lambda_1^0|, \dots, |\lambda_g^0|, \|\omega_0\|_\infty)}{4} \quad \text{and} \quad \frac{K(2\mathcal{C}_1 + \mathcal{T}(\Gamma))}{M} < \frac{1}{4}, \qquad (5.28)$$

which we suppose from now on. In fact, the most problematic term in (5.25) is the last one. To estimate λ_k^*, we thus introduce the following function:

$$h(t) = K\int_0^t \int_{\bigcup_{k=1}^g (\Gamma_k\cap\gamma^+)} \alpha^*(\sigma)\nabla\theta(x)\cdot n(x)|\omega^*(\sigma,x)|dxd\sigma.$$

In order to estimate $h(t)$, let us consider $\tilde{\omega} := \tilde{F}[\omega^*, \lambda_i^*]$. As y^* satisfies the assumptions of Proposition 3.5, one easily sees (using (3.17)) that $\tilde{\omega} = \omega^*$ on $[0,T] \times \cup_{k=1}^{g}(\Gamma_k \cap \gamma^+)$. Also, by (4.23), one has

$$\tilde{\omega}(t,x) = \left(\omega_0(x) - \sum_{A\in\mathcal{A}} \omega_0(A)\Gamma_A(x)\right) \exp\left(-M\int_0^t \alpha^*(\tau)d\tau\right) + \sum_{A\in\mathcal{A}} \omega^*(t,A)\Gamma_A(x) \quad \text{on } [0,T]\times\overline{\gamma^-}. \tag{5.29}$$

But as $\tilde{\omega}$ satisfies

$$\partial_t\tilde{\omega} + \operatorname{div}(y^*\tilde{\omega}) = 0,$$

one deduces that

$$\frac{d}{dt}\left(\int_\Omega |\tilde{\omega}|\right)(t) = -\int_{\partial\Omega}(y^*\cdot n)|\tilde{\omega}|(t) = -K\int_{\partial\Omega}\alpha^*(t)\nabla\theta(x)\cdot n(x)|\tilde{\omega}|(t,x)dx.$$

Consequently, one gets

$$\begin{aligned} h(t) &\le K\int_0^t\int_{\gamma^+}\alpha^*(\sigma)\nabla\theta(x)\cdot n(x)|\tilde{\omega}(\sigma,x)|d\sigma dx \\ &\le \int_\Omega|\tilde{\omega}(0,\cdot)| - K\int_0^t\int_{\gamma^-}\alpha^*(\sigma)\nabla\theta(x)\cdot n(x)|\tilde{\omega}(\sigma,x)|d\sigma dx. \end{aligned}$$

Using (5.29), one gets

$$\begin{aligned} h(t) \le \int_\Omega|\omega(0,\cdot)| + 2K\mathcal{C}_1\|\omega_0\|_\infty\int_0^t\alpha^*(\sigma)\exp\left(-M\int_0^\sigma\alpha^*(\tau)d\tau\right)d\sigma \\ - K\sum_{A\in\mathcal{A}}\int_0^t\int_{\gamma^-}\alpha^*(\sigma)\nabla\theta(x)\cdot n(x)|\omega^*(\sigma,A)\Gamma_A(x)|d\sigma dx, \end{aligned}$$

and hence

$$h(t) \le \int_\Omega|\omega(0,\cdot)| + 2K\frac{\mathcal{C}_1\|\omega_0\|_\infty}{M} + K\mathcal{T}(\Gamma)\max_{A\in\mathcal{A}}\int_0^t\alpha^*(\sigma)|\omega^*(\sigma,A)|d\sigma. \tag{5.30}$$

Let us now concentrate on the last term. For $A\in\mathcal{A}$ and $\sigma\in[0,T]$, let us define

$$\begin{aligned} &\mathfrak{s}(\sigma,A) := \min\left\{t\in[0,\sigma] \;\middle/\; \Phi^*([t,\sigma],\sigma,A)\subset\overline{\Omega}\right\}, \\ &\mathfrak{a}(\sigma,A) := \Phi^*(\mathfrak{s}(\sigma,A),\sigma,A). \end{aligned} \tag{5.31}$$

Using (3.20), (2.28), and the fact that ω^* is constant along the flow, one deduces that for any $A\in\mathcal{A}$ and any σ, one has

$$\omega(\sigma,A) = \omega_0(\mathfrak{a}(\sigma,A))\exp\left(-M\int_0^{\mathfrak{s}(\sigma,A)}\alpha^*(\tau)d\tau\right). \tag{5.32}$$

To estimate the last term in (5.30), we fix $A\in\mathcal{A}$.

- If t is such that

$$\int_0^t \alpha^*(\sigma)d\sigma \leq \frac{V(\theta)}{\kappa K}, \tag{5.33}$$

then using (5.32), one easily gets

$$\int_0^t \alpha^*(\sigma)|\omega(\sigma, A)|d\sigma \leq \frac{V(\theta)}{\kappa K}\|\omega_0\|_\infty.$$

(When considering a time T with the specific form (4.9), one could prove that on such a time interval we always have (5.33).)

- If not, call ξ the (unique) time for which

$$\int_0^\xi \alpha^*(\sigma)d\sigma = \frac{V(\theta)}{\kappa K}.$$

Let us prove that for $\sigma > \xi$, one has

$$\int_{\mathfrak{s}(\sigma,A)}^\sigma \alpha^*(\tau)d\tau \leq \frac{V(\theta)}{\kappa K}\left(= \int_0^\xi \alpha^*(\tau)d\tau\right). \tag{5.34}$$

This results from the fact that if one defines

$$\mu(t) = \theta(\Phi^*[t, \mathfrak{s}(\sigma, A), A]),$$

then one has (in the classical sense) for any $t \in [\mathfrak{s}(\sigma, A), \sigma]$

$$\begin{aligned}\frac{d\mu}{dt} &= y^*(t, \Phi^*[t, \mathfrak{s}(t, A), A]) \cdot \nabla\theta(\Phi^*[t, \mathfrak{s}(t, A), A]) \\ &\geq \kappa K \alpha^*(t).\end{aligned}$$

Integrating this inequality between $\mathfrak{s}(\sigma, A)$ and σ yields (5.34). In particular, as a consequence of (5.34), one gets that for $\sigma > \xi$, one has $\mathfrak{s}(\sigma, A) > 0$. Consequently, using (5.32),

$$\begin{aligned}\int_0^t \alpha^*(\sigma)|\omega(\sigma, A)|d\sigma &= \int_0^\xi \alpha^*(\sigma)|\omega(\sigma, A)|d\sigma + \int_\xi^t \alpha^*(\sigma)|\omega(\sigma, A)|d\sigma \\ &\leq \frac{V(\theta)}{\kappa K}\|\omega_0\|_\infty \\ &\quad + \int_\xi^t \alpha^*(\sigma)\|\omega_0\|_\infty \exp\left(-M\int_0^{\mathfrak{s}(\sigma,A)} \alpha^*(\tau)d\tau\right)d\sigma.\end{aligned}$$

Now using (5.34), one gets

$$\begin{aligned}\int_0^{\mathfrak{s}(\sigma,A)} \alpha^*(\tau)d\tau &= \int_0^\sigma \alpha^*(\tau)d\tau - \int_{\mathfrak{s}(\sigma,A)}^\sigma \alpha^*(\tau)d\tau \\ &\geq \int_0^\sigma \alpha^*(\tau)d\tau - \int_0^\xi \alpha^*(\tau)d\tau.\end{aligned}$$

Hence,

$$\int_\xi^t \alpha^*(\sigma)\|\omega_0\|_\infty \exp\left(-M\int_0^{\mathfrak{s}(\sigma,A)}\alpha^*(\tau)d\tau\right)d\sigma \\ \leq \int_\xi^t \alpha^*(\sigma)\|\omega_0\|_\infty \exp\left(-M\int_\xi^{\sigma}\alpha^*(\tau)d\tau\right)d\sigma \leq \frac{\|\omega_0\|_\infty}{M}.$$

Finally, in all cases we get

$$h(t) \leq \int_\Omega |\omega_0| + \left(\frac{2K\mathcal{C}_1}{M} + \mathcal{T}(\Gamma)\left[\frac{V(\theta)}{\kappa}+\frac{K}{M}\right]\right)\|\omega_0\|_\infty.$$

Let us go back to λ_k^*. At times t for which (5.23) is valid in $[0,t]$ (this is at least the case for times in a neighborhood of 0), one has (5.24) and consequently, one gets

$$|\lambda_k^*(t)| \leq |\lambda_k^0| + h(t) + K\int_0^t\left[-\alpha^*(s)\lambda_k^*(s) + \mathcal{C}_0\alpha^*(s)\exp\left(-M\int_0^s\alpha^*(\tau)d\tau\right)\right]ds \\ \leq |\lambda_k^0| + \int_\Omega|\omega(0,\cdot)| + \frac{V(\theta)\mathcal{T}(\Gamma)}{\kappa}\|\omega_0\|_\infty + K\frac{(2\mathcal{C}_1+\mathcal{T}(\Gamma))\|\omega_0\|_\infty+\mathcal{C}_0}{M} \\ -K\int_0^t\alpha^*(s)\lambda_k^*(s)ds.$$

Hence, with $\alpha^*(t)\geq 0$, we get

$$|\lambda_k^*(t)| \leq |\lambda_k^0| + \int_\Omega|\omega(0,\cdot)| + \frac{V(\theta)\mathcal{T}(\Gamma)}{\kappa}\|\omega_0\|_\infty \\ + K\frac{(2\mathcal{C}_1+\mathcal{T}(\Gamma))\|\omega_0\|_\infty+\mathcal{C}_0}{M}. \tag{5.35}$$

Using (5.28), one gets

$$|\lambda_k(t)| < \left(\frac{3}{2}+|\Omega|+\frac{V(\theta)\mathcal{T}(\Gamma)}{\kappa}\right)\max(|\lambda_1^0|,\dots,|\lambda_g^0|,\|\omega_0\|_\infty) \\ < \mathcal{M}_{\omega_0,\lambda_i^0}. \tag{5.36}$$

Hence (5.23) propagates during the whole time interval $[0,T]$.

So at this point, we have proven that for any (ω_0,λ_i^0), there exists a local solution of the closed-loop system.

6. End of the proof. To finish the proof, we still have to establish two propositions:

- any maximal solution of the closed-loop system is global,
- for any global solution of the closed-loop system, 0 is asymptotically stable.

6.1. Maximal solutions are global solutions. Consider a maximal solution (ω,λ_i) of the closed-loop system, say it is defined on $[0,T^*)$, with T^* maximal. Let us prove that $T^*=+\infty$. Toward this end, let us suppose by contradiction that $T^*<+\infty$, and prove that

$$(\omega(t),\lambda_i(t)) \longrightarrow (\omega(T^*),\lambda_i(T^*)) \text{ as } t\to T^{*-} \tag{6.1}$$

in $C^0(\overline{\Omega};\mathbb{R})\times\mathbb{R}^g$. Using again the local existence result, this yields a contradiction. This is done as in [4, Proposition 3.4]. We first establish the following lemma.

LEMMA 6.1. *Let $T>0$ and let $(\omega,\lambda_i)\in C^0([0,T]\times\overline{\Omega})\times C^0([0,T])^g$ be a solution of the closed-loop system. Then one has for $(t,x)\in\Omega_T$ and any $s\in[0,t]$*

$$\omega(t,x)=\omega(s_{\omega,\lambda_i}(t,x),a_{\omega,\lambda_i}(t,x)), \tag{6.2}$$

and for any t in $[0,T]$, one has

$$\begin{aligned}\max(\|\omega(t)\|_\infty,|\lambda_1|(t),\dots,|\lambda_g|(t))\\ \le\left(3+|\Omega|+\frac{V(\theta)\mathcal{T}(\Gamma)}{\kappa}\right)(1+\|\Lambda\|_\infty)\max(\|\omega_0\|_\infty,|\lambda_1^0|,\dots,|\lambda_g^0|).\end{aligned} \tag{6.3}$$

Proof of Lemma 6.1. First, such a solution satisfies

$$\partial_t\omega+\operatorname{div}(y_{\omega,\lambda_i}\omega)=0.$$

A classical regularization argument shows that ω is constant along the flow of y_{ω,λ_i}, which yields (6.2).

We suppose that $(\omega(t),\lambda_1(t),\dots,\lambda_g(t))$ does not vanish. If it does, then using the definition of the feedback and the fact that the vorticity is constant along the flow, $(\omega(t),\lambda_1(t),\dots,\lambda_g(t))$ stays null. From now on, we work on the initial interval where $(\omega(t),\lambda_1(t),\dots,\lambda_g(t))$ is not zero.

Now, the "λ_i" part in (6.3) can be reproduced from what was already done in (5.36), because we did not use the particular form of T but only (2.27)–(2.28), the fact that the vorticity follows the flow, and the fact that the velocity satisfies Proposition 3.5.

It remains to prove the "ω" part of (6.3). Having proved the estimate on the λ_i, this is done as for point (b) in the proof of $\mathcal{F}(X)\subset X$ (see section 5.1), except that now the estimate

$$|\omega(\cdot,A)|\le\|\omega_0\|_\infty\quad\text{for any } A\in\mathcal{A}$$

comes now from (5.32) (and not from the choice of T).

Having proved Lemma 6.1, we get Hölder estimates on the flow from (3.7), which can consequently be extended on $[0,T^*]$, and then we get (6.1) approximately as for the continuity of $F(\omega,\lambda_i)$ in section 5.1 (we omit the details). Hence, using again the local existence theorem, we find a contradiction to $T^*<+\infty$.

6.2. 0 is asymptotically stable. Now that we have proved (6.3), it remains to prove (2.32). We consider again a global solution (ω,λ_i) of the closed-loop system; let us show that $\|(\omega,\lambda_i)(t)\|_\infty\to0$ as $t\to+\infty$.

We suppose that $(\omega(t,\cdot),\lambda_i(t))$ never vanishes. If it does vanish for some $T>0$, it follows from (2.26) and from the fact that the vorticity is constant along the flow of y_{ω,λ_i} that (ω,λ_i) is null in the neighborhood in time of $+\infty$; hence the result is valid.

This is done in several steps. First, we prove that $\omega(t,\cdot)\to0$ on the entering zone γ^- and in a second step that this convergence holds in the rest of the domain. The convergence to zero of $\lambda_1,\dots,\lambda_g$ is proved in the same step.

Again, we denote

$$\alpha(t):=\max\left(|\lambda_1(t)|,\dots,|\lambda_g(t)|,\|\omega(t)\|_\infty\right).$$

We first prove the following lemma.

LEMMA 6.2. *Let $(\omega, \lambda_i) \in C^0([0,+\infty) \times \overline{\Omega}) \times C^0([0,+\infty))^g$ be a global solution of the closed-loop system. Then it satisfies*

$$\|\omega(t)\|_{C^0(\gamma^- \setminus [\mathrm{Supp}(\Lambda) \cup \mathrm{Supp}(\Gamma)])} \longrightarrow 0 \text{ as } t \to +\infty. \tag{6.4}$$

Proof of Lemma 6.2. Consider $x \in \gamma^- \setminus (\mathrm{Supp}(\Gamma) \cup \mathrm{Supp}(\Lambda))$. If $\omega(t_0, x) = 0$ for some time t_0, then it follows from (2.28) that $\omega(t, x) = 0$ for all t. Let us suppose that $\omega_0(x) \neq 0$. Then it follows from (2.28) that on $\gamma^- \setminus [\mathrm{Supp}(\Lambda) \cup \mathrm{Supp}(\Gamma)]$,

$$\partial_t |\omega(t,x)| \leq -M |\omega(t,x)|^2;$$

hence

$$|\omega(t,x)| \leq \frac{|\omega_0(x)|}{1 + M|\omega_0(x)|t},$$

and hence $\omega(t,x) \to 0$ as $t \to +\infty$. One sees that the estimate is uniform and hence that (6.4) holds.

Now, we have the following lemma.

LEMMA 6.3. *Let $(\omega, \lambda_i) \in C^0([0,+\infty) \times \overline{\Omega}) \times C^0([0,+\infty))^g$ be a global solution of the closed-loop system. Then it satisfies*

$$\|\omega(t)\|_{C^0(\mathrm{Supp}(\Gamma))} \longrightarrow 0 \text{ as } t \to +\infty. \tag{6.5}$$

Proof of Lemma 6.3. Let us first prove that for any $A \in \mathcal{A}$, one has

$$\omega(t, A) \longrightarrow 0 \text{ as } t \to +\infty. \tag{6.6}$$

It follows from (3.20) that

$$\Phi^{\omega, \lambda_i}(\mathfrak{s}(t,A), t, A) \notin \mathrm{Supp}(\Gamma) \cup \mathrm{Supp}(\Lambda),$$

with $\mathfrak{s}(t,A)$ given by (5.31). Now, we fix $\varepsilon > 0$ and let t_0 be a time such that for $t \geq t_0$, one has

$$\|\omega(t)\|_{C^0(\overline{\gamma^- \setminus [\mathrm{Supp}(\Lambda) \cup \mathrm{Supp}(\Gamma)]})} \leq \varepsilon.$$

Then if t_1 is such that $|\omega(t_1, A)| > \varepsilon$, one deduces that $\mathfrak{s}(t_1, A) \leq t_0$. Hence, using $\Phi^{\omega,\lambda_i}(s, t_1, A) \in \overline{\Omega}$ for $t_0 \leq s \leq t_1$ and the fact that the vorticity is constant along the flow, one gets

$$\|\omega(s)\|_{C^0(\overline{\Omega})} \geq \varepsilon \text{ for } t_0 \leq s \leq t_1.$$

But (3.14) implies that, in the classical sense,

$$\frac{d}{ds}\left[\theta(\Phi^{\omega,\lambda_i}(s, t_0, x))\right] = y_{\omega,\lambda_i}(s, \Phi^{\omega,\lambda_i}(s,t_0,x)) \cdot \nabla\theta(\Phi^{\omega,\lambda_i}(s,t_0,x)) \geq \kappa K \alpha(s) \geq \kappa K \varepsilon. \tag{6.7}$$

With the boundedness of θ in $\overline{\Omega}$, one sees that $|t_1 - t_0|$ must be bounded, which gives (6.6).

Now, consider $x \in \mathrm{Supp}(\Gamma_A)$ for a certain $A \in \mathcal{A}$ and t large enough. Then, (6.5) follows from the fact, due to (2.28), that for $t \in [0,T]$ and $x \in \mathrm{Supp}(\Gamma_A)$, one has

$$\omega(t,x) = (\omega_0(x) - \omega_0(A)\Gamma_A(x)) \exp\left(-M\int_0^t \alpha(\tau)d\tau\right) + \omega(t,A)\Gamma_A(x). \tag{6.8}$$

Indeed, given any $\varepsilon > 0$, for t_0 large enough, one has $|\omega_2(s,x)| \leq \varepsilon$ on $\mathrm{Supp}(\Gamma_A)$ for any $s \geq t_0$ (with the notation of ω_1 and ω_2 in (2.27)). One gets for any $s \geq t_0$ that

$$|\alpha(s)| \geq K \max(0, |\omega_1(s,x)| - |\omega_2(s,x)|) \geq K \max(0, |\omega_1(s,x)| - \varepsilon).$$

Then

- if $|\omega_1(t,x)| \leq 2\varepsilon$, then, because ω_1 has the form

$$\omega_1(s,x) = \left(\omega_0(x) - \sum_{A\in\mathcal{A}} \omega_0(A)\Gamma_A(x)\right) \exp\left(-M\int_0^s \alpha(\tau)d\tau\right),$$

  this inequality stays valid for $s \geq t$, or
- if $|\omega_1(t,x)| \geq 2\varepsilon$, then for $s \geq t$ such that this is still valid, one has $|\alpha(s)| \geq \varepsilon$; then with (6.8), one sees that $|\omega_1(s,x)|$ decreases until it reaches the previous situation.

Then, we establish the following lemma.

LEMMA 6.4. *Let $(\omega, \lambda_i) \in C^0([0,+\infty) \times \overline{\Omega}) \times C^0([0,+\infty))^g$ be a global solution of the closed-loop system. Then it satisfies*

$$\|\omega(t)\|_{C^0(\gamma^+\cap\Gamma_k)} \longrightarrow 0 \text{ as } t \to +\infty \quad \forall k = 1, \ldots, g. \tag{6.9}$$

Proof of Lemma 6.4. The limit (6.9) follows from (3.17), (6.2), and Lemmas 6.2 and 6.3. Indeed, we introduce t_0 such that for $t \geq t_0$, one has $|\omega(t,\cdot)| \leq \varepsilon$ on $\overline{\gamma^-} \setminus \mathrm{Supp}(\Lambda)$. Suppose that we could find, for times arbitrarily large, some points in $\cup_{i=1}^g (\overline{\gamma}^+ \cap \Gamma_k)$ for which $|\omega(t,x)| > \varepsilon$ and hence, by (3.17), such that $s_{\omega,\lambda_i}(t,x) \leq t_0$. This contradicts (6.7) and the boundedness of θ in $\overline{\Omega}$.

Now, we have the following lemma.

LEMMA 6.5. *Let $(\omega, \lambda_i) \in C^0([0,+\infty) \times \overline{\Omega}) \times C^0([0,+\infty))^g$ be a global solution of the closed-loop system. Then it satisfies*

$$\|\omega(t)\|_{C^0(\mathrm{Supp}(\Lambda))} \longrightarrow 0 \text{ as } t \to +\infty. \tag{6.10}$$

Proof of Lemma 6.5. Fix $k \in \{1, \ldots, g\}$. It follows from (1.8), (2.22), (2.23), (2.27), and (2.28) that λ_k satisfies

$$\begin{aligned}\frac{d}{dt}\lambda_k(t) = -K\alpha(t)\lambda_k(t) + K\alpha(t)\int_{\Gamma_k\cap\gamma^-} \nabla\theta(x)\cdot n(x)\omega_1(t,x)dx \\ + K\alpha(t)\int_{\Gamma_k\cap\gamma^+} \nabla\theta(x)\cdot n(x)\omega(t,x)dx.\end{aligned} \tag{6.11}$$

But ω_1 converges uniformly to 0 (this is proved exactly as Lemma 6.2), and by Lemma 6.4, the second integral in (6.11) converges to 0 (remember that $\alpha(t)$ is bounded thanks to (6.3)). Hence, given $\varepsilon > 0$, there exists t_0 such that for $t \geq t_0$,

$$\left|\int_{\Gamma_k\cap\gamma^-} \nabla\theta(x)\cdot n(x)\omega_1(t,x)dx\right| + \left|\int_{\Gamma_k\cap\gamma^+} \nabla\theta(x)\cdot n(x)\omega(t,x)dx\right| \leq \varepsilon.$$

Consequently, for $t \geq t_0$, if $\lambda_k(t) \geq 2\varepsilon$, using $|\lambda_k(t)| \leq \alpha(t)$, one gets

$$\frac{d}{dt}\lambda_k(t) \leq -\varepsilon K \lambda_k(t),$$

and if $\lambda_k(t) \leq -2\varepsilon$, one gets

$$\frac{d}{dt}\lambda_k(t) \geq -\varepsilon K \lambda_k(t).$$

This yields

$$\lambda_i(t) \longrightarrow 0 \text{ as } t \to +\infty \text{ for } i = 1, \ldots, g. \tag{6.12}$$

Then having proved (6.12), (6.10) follows from the same procedure as the one at the end of Lemma 6.3.

These lemmas allow us to establish the following proposition.

PROPOSITION 6.6. *Let $(\omega, \lambda_i) \in C^0([0,+\infty) \times \overline{\Omega}) \times C^0([0,+\infty))^g$ be a global solution of the closed-loop system. Then it satisfies*

$$\max(\|\omega(t)\|_{C^0(\overline{\Omega})}, |\lambda_1(t)|, \ldots, |\lambda_g(t)|) \longrightarrow 0 \text{ as } t \to +\infty. \tag{6.13}$$

Proof of Proposition 6.6. The $\lambda_i(t)$-part is precisely (6.12). For the ω-part, consider $\tau(\varepsilon)$ such that for $t \geq \tau(\varepsilon)$, one has

$$\|\omega_{|(\gamma^+ \cup \gamma^-)}(t,\cdot)\|_{L^\infty} \leq \varepsilon \quad \text{and} \quad |\lambda_i(t)| \leq \varepsilon \quad \forall i = 1, \ldots, g.$$

Suppose that for any $\tilde{\tau}$, one can find $t \geq \tilde{\tau}$ for which $\|\omega(t,\cdot)\|_{C^0(\overline{\Omega})} > \varepsilon$. Then there is some $x \in \overline{\Omega}$ for which $|\omega(t,x)| > \varepsilon$, and hence by (6.2) one has $s_{\omega,\lambda_i}(t,x) \leq \tau(\varepsilon)$ and hence $\alpha(t) \geq \varepsilon$ on $[\tau(\varepsilon), t]$. Consequently, there exists $x_0 \in \overline{\Omega}$ such that $|\omega(\tau(\varepsilon), x_0)| > \varepsilon$ and for which

$$\Phi^{\omega,\lambda_i}([\tau(\varepsilon), t], \tau(\varepsilon), x_0) \subset \overline{\Omega}.$$

With (6.7), this contradicts the fact that θ is bounded in $\overline{\Omega}$.

7. Appendix.

7.1. Proof of Corollary 2.2. We reduce Σ a little in order to keep some kind of margin. Introduce $\tilde{\theta}$ as in Proposition 2.1. We describe a procedure that allows us to slightly modify $\tilde{\theta}$ to get rid of problematic points "E," while preserving (2.11)–(2.16). The idea is the following: consider an E point as in (2.17); by $\tilde{\Phi}$ it is first transported along $\partial\Omega \backslash (\overline{\gamma^+(\tilde{\theta})} \cup \overline{\gamma^-(\tilde{\theta})})$ (remember (2.12) and (2.13)); call γ_E the corresponding connected component of $\partial\Omega \backslash (\overline{\gamma^+(\tilde{\theta})} \cup \overline{\gamma^-(\tilde{\theta})})$. Consider t_E the biggest positive time for which $\tilde{\Phi}((0,t),0,E) \subset \gamma_E$. There are two cases:

- If $\tilde{\Phi}(t_E, 0, E) \in \overline{\gamma^+(\tilde{\theta})}$, it is clear from (2.12)–(2.13) that this point is in $\partial\gamma^+(\tilde{\theta})$, pointing inside $\gamma^+(\tilde{\theta})$. And consequently for t just after t_E, one has $\tilde{\Phi}(t,0,E) \in B_R \backslash \overline{\Omega}$, so the E point under consideration satisfies (2.17).
- If $\tilde{\Phi}(t_E, 0, E) \in \overline{\gamma^-(\tilde{\theta})}$, we consider the following time t'_E:

$$t'_E = \sup\left\{ t \in (t_E, +\infty),\ \tilde{\Phi}((t_E, t), 0, E) \in \Omega \right\}.$$

It is indeed quite clear that for times $t > t_E$ with $t - t_E$ small, $\tilde{\Phi}(t, 0, E) \in \Omega$, and it follows from $|\nabla\tilde{\theta}|(x) > 0$ in $\overline{\Omega}$ that the preceding set is bounded from above. Then $\tilde{\Phi}(t'_E, 0, E) \in \overline{\gamma^+(\tilde{\theta})}$ (for it cannot be in $\partial\Omega\backslash(\overline{\gamma^-(\tilde{\theta})} \cup \overline{\gamma^+(\tilde{\theta})})$ because of the uniqueness of the flow and it cannot be in $\overline{\gamma^-(\tilde{\theta})}$ because points in $\overline{\gamma^-(\tilde{\theta})}$ come from $(B_R\backslash\overline{\Omega}) \cup \partial\Omega$ by the flow of $\nabla\theta$).
If $\tilde{\Phi}(t'_E, 0, E) \in \gamma^+(\tilde{\theta})$, then (2.17) is valid for this E; we now suppose that $E_2 := \tilde{\Phi}(t'_E, 0, E) \in \partial\gamma^+(\tilde{\theta})$.
We consider a small connected neighborhood $\mathcal{U}$ of E_2 in $\partial\Omega$; thanks to the margin we kept on Σ, one can require $\mathcal{U} \subset \Sigma$. We also consider a point $F \in \gamma^+(\tilde{\theta})$ and $\mathcal{V}_F$ a small connected neighborhood of F in $\gamma^+(\tilde{\theta})$, not touching $\partial\gamma^+(\tilde{\theta})$ or $\mathcal{U}$.
We introduce a function on $\partial\Omega$, say ψ, supported in $\mathcal{U} \cup \mathcal{V}_F$, nonnegative in $\mathcal{U}$, nonpositive in $\mathcal{V}_F$, and such that

$$\int_{\partial\Omega} \psi = 0 \quad \text{and} \quad \psi(E_2) = 1.$$

Then we define $\hat{\theta} \in C^\infty(\Omega; \mathbb{R})$ by

$$\left\{ \begin{array}{l} \Delta\hat{\theta} = 0 \quad \text{in } \Omega, \\ \partial_n\hat{\theta} = \psi \quad \text{on } \partial\Omega, \\ \int_\Omega \hat{\theta} = 0. \end{array} \right. \tag{7.1}$$

Using elliptic estimates and Lemma 3.3, one sees that $\tilde{\theta} + \varepsilon\hat{\theta}$ still satisfies (2.11)–(2.16) for $\varepsilon > 0$ small enough. Let us particularly emphasize that, for $\varepsilon > 0$ small enough, one has $\partial_n(\tilde{\theta} + \varepsilon\hat{\theta}) > 0$ on $\mathcal{V}_F$. The E considered now satisfies (2.17). The procedure has not added an E point, but it has slightly moved the frontier of $\gamma^+(\tilde{\theta})$: introduce

$$\gamma^+(\tilde{\theta} + \varepsilon\hat{\theta}) := \{x \in \partial\Omega \ / \ \partial_n(\tilde{\theta} + \varepsilon\hat{\theta}) > 0\}.$$

Then $E_2 \in \gamma^+(\tilde{\theta} + \varepsilon\hat{\theta})$, and the new frontier of γ^+ is now given by

$$\partial\gamma^+(\tilde{\theta} + \varepsilon\hat{\theta}) = \partial\gamma^+(\tilde{\theta}) \cup \{E_3\} \setminus \{E_2\},$$

where E_3 is the point in $\partial\mathcal{U}$ that does not belong to $\gamma^+(\tilde{\theta})$.
Now if E_2 satisfied (2.17), then E_3 also does for ε small enough: we have two cases:

- $\nabla\tilde{\theta}$ is pointing inside $\gamma^+(\tilde{\theta})$ at E_2. This case is in fact not possible because of the definition of E_2 and t'_E: points in $\partial\gamma^+(\tilde{\theta})$ at which $\nabla\tilde{\theta}$ is pointing inside $\gamma^+(\tilde{\theta})$ come from $\partial\Omega$ when following the flow.
- $\nabla\tilde{\theta}$ is pointing outside $\gamma^+(\tilde{\theta})$ at E_2. Then using (2.13), one sees that the trajectory of E_2 under the flow of $\nabla\tilde{\theta}$ follows the connected component of E_2 in $\partial\Omega\backslash[\gamma^+(\tilde{\theta}) \cup \gamma^-(\tilde{\theta})]$. In particular, this trajectory meets E_3. But for ε small enough, the trajectories under the flow of $\nabla(\tilde{\theta} + \varepsilon\hat{\theta})$ are almost the same as the ones in the flow of $\nabla\tilde{\theta}$ (as seen by Lemma 3.3 and elliptic estimates). This yields the conclusion.

So one can get rid of problematic points one after another.

7.2. Proof of Proposition 3.5.

- *Proof of* (3.14). Introduce $\hat{y}$ as the solution of

$$\begin{cases} \operatorname{curl} \hat{y}(t,x) = \omega(t,x) \quad \text{for } (t,x) \in \Omega_T, \\ \operatorname{div} \hat{y}(t,x) = 0 \quad \text{for } (t,x) \in \Omega_T, \\ \hat{y}(t,x) \cdot n(x) = 0 \quad \text{for } (t,x) \in \Sigma_T, \\ \displaystyle\int_{\Gamma_i} \hat{y}(t,x) \cdot \vec{\tau}(x) dx = \lambda_i(t) \quad \text{for } t \in [0,T], \text{ for } i = 1, \dots, g. \end{cases}$$

Of course, one has $y = \hat{y} + K\alpha(t)\nabla\theta(x)$. Now (3.9) involves

$$\|\hat{y}(t)\|_{L^\infty(\Omega)} \leq C_{\mathcal{LL}} \max\left(|\lambda_1(t)|, \dots, |\lambda_g(t)|, \|\omega(t)\|_\infty\right) \ \forall t \in [0,T].$$

(The constant $C_{\mathcal{LL}}$ does not depend on t.) Hence

$$y(t,x) \cdot \nabla\theta(x) \geq K\alpha(t)|\nabla\theta(x)|^2 - C_{\mathcal{LL}}\|\nabla\theta\|_\infty \alpha(t). \tag{7.2}$$

Equation (2.13) and the compactness of $\overline{\Omega}$ allow us to introduce

$$\underline{m} := \min_{x \in \overline{\Omega}} |\nabla\theta(x)| > 0.$$

One easily deduces from (7.2) that (3.14) holds if $\overline{K} \geq 2C_{\mathcal{LL}}\|\nabla\theta\|_\infty / \underline{m}^2$ and $\kappa = \underline{m}^2/2$ (which we suppose in what follows).

- *Proof of* (3.17). Property (3.17) will essentially follow from Gronwall's inequality (3.11), from (2.13), and from (2.16).
We extend the definition of α and (ω, λ^i) for times $t \geq T$ by $\alpha(T)$ and $(\omega, \lambda^i)(T)$, respectively.
We write $\Theta^\alpha(t,x) := K\alpha(t)\theta(x)$. We consider Φ, Φ^y, and Φ^α the respective flows of $\pi(\nabla\theta)$, $\pi(y(t,x))$, and $\pi(\nabla\Theta^\alpha(t,x))$.
First, by a compactness argument and using (2.13), one sees that there exist $T_\theta > 0$ and $d_\theta > 0$ such that

$$\forall x \in \overline{\Omega}, \ \exists t \in [0, T_\theta] \text{ such that } \operatorname{dist}(\Phi(t,0,x), \overline{\Omega}) \geq d_\theta.$$

It suffices, for instance, to observe that

$$\frac{d}{dt}\theta(\Phi(t,0,x)) = |\nabla\theta(\Phi(t,0,x))|^2$$

if x and t are such that $\Phi(t,0,x) \in \overline{\Omega}$, and to use (2.13) and the boundedness of θ on $\overline{\Omega}$.
Hence

$$\forall x \in \overline{\Omega}, \ \exists \mathcal{T}(x) \text{ such that } \int_0^{\mathcal{T}(x)} K\alpha(\tau)d\tau \leq T_\theta \text{ and such that } \operatorname{dist}(\Phi^\alpha(\mathcal{T}(x),0,x), \overline{\Omega}) \geq d_\theta.$$

Now by Lemma 3.3 one has

$$|\Phi^y(t,0,x) - \Phi^\alpha(t,0,x)| \leq \exp\left(K\|\pi(\nabla\theta)\|_{\mathcal{L}ip} \int_0^t \alpha(\tau)d\tau\right) \|\pi(\hat{y})\|_{L^1([0,T],L^\infty(B_R))}.$$

Consequently, one sees that for $x \in \overline{\Omega}$ and t such that $0 \le t \le \mathcal{T}(x)$,

$$\begin{aligned}|\Phi^y(t,0,x) - \Phi^\alpha(t,0,x)| &\le \exp(T_\theta \|\pi(\nabla\theta)\|_{\mathcal{L}ip(B_R)}) \|\hat{y}\|_{L^1([0,t],L^\infty_x)} \\ &\le C_{\mathcal{L}\mathcal{L}} \exp(T_\theta \|\pi(\nabla\theta)\|_{\mathcal{L}ip(B_R)}) \|(\omega,\lambda_i)\|_{L^1([0,t],L^\infty_x)}.\end{aligned}$$

Now one deduces from (3.12) that for $0 \le t \le \mathcal{T}(x)$,

$$\|(\omega,\lambda_i)\|_{L^1([0,t],L^\infty_x)} \le \int_0^t \alpha(\tau)d\tau \le \frac{T_\theta}{K}.$$

This yields

$$|\Phi^y(t,0,x) - \Phi^\alpha(t,0,x)| \le \frac{C_{\mathcal{L}\mathcal{L}} \exp(T_\theta \|\pi(\nabla\theta)\|_{\mathcal{L}ip(B_R)}) T_\theta}{K}. \tag{7.3}$$

Consequently, if K is large enough (in terms of only θ), one has

$$\begin{aligned}\forall x \in \overline{\Omega},\ \exists \mathcal{T}(x) \text{ such that } &\int_0^{\mathcal{T}(x)} K\alpha(\tau)d\tau \le T_\theta \\ &\text{and such that } \operatorname{dist}(\Phi^y(\mathcal{T}(x),0,x),\overline{\Omega}) \ge d_\theta/2,\end{aligned}$$

with (7.3) valid between times 0 and $\mathcal{T}(x)$. With (2.18), this gives (3.17) for K large enough.

- *Proof of* (3.18). This is due to the uniqueness of the flow: on $\overline{\gamma_A}$, $y(s,x)$ is of the form $\lambda(s,x)\vec{\tau}(x)$, the sign of $\lambda(s,x)$ being constant in such a way that the direction of $y(s,A)$ is pointing inside γ^- (indeed, thanks to (3.13) and (3.14), $y(s,x)$ has the same direction as $\nabla\theta(x)$ on $\partial\Omega$). So one finds a local in time backward solution of (3.6) inside γ_A. This solution does not go outside γ_A for times $\tau \in [0,t)$ if $t-\tau$ is small enough so that

$$c_\pi(K\|\nabla\theta\|_\infty + C_{\mathcal{L}\mathcal{L}}) \left(\int_\tau^t \alpha(s)ds\right) \le \ell/2, \tag{7.4}$$

because the velocity is estimated by

$$\begin{aligned}|\pi[y](s,x)| &\le \|\pi[\hat{y}](s,\cdot)\|_{L^\infty(B_R)} + K\alpha(s)\|\pi[\nabla\theta]\|_{L^\infty(B_R)} \\ &\le c_\pi(C_{\mathcal{L}\mathcal{L}} + K\|\nabla\theta\|_\infty)\alpha(s)\end{aligned} \tag{7.5}$$

and because of the definition of ℓ.

- *Proof of* (3.15)–(3.16). This is mutatis mutandis [4, Lemma 3.3]. Let us treat separately the points in γ^- and the B points.
  - Let us consider (3.16) for a point $B \in \mathcal{B}$. Using again (3.13) and (3.14), one sees that, as for $s \in [0,T]$, $y(s,B)$ is tangent to $\partial\Omega$ and by (3.14) pointing outside γ^-, there is a solution for the flow starting from B and that stays inside $\partial\Omega\backslash(\overline{\gamma^-} \cup \overline{\gamma^-})$ at least for small times. So the uniqueness of the flow gives (3.16).
  - Consider $x \in \gamma^-$. Let us use the coordinates in the reference frame given by $(\vec{\tau}(x), n(x))$. Then using (3.9) and (3.13), we see that for any $\varepsilon > 0$, one finds some neighborhood $\mathcal{U}$ of (t,x) in $(0,T] \times B_R$ for which one has for any $(\tilde{t},\tilde{x})$ in $\mathcal{U}$,

$$|\{\pi[y](\tilde{t},\tilde{x}) - K\alpha(\tilde{t})\pi[\nabla\theta](\tilde{x})\} \cdot n(x)| \le \varepsilon.$$

(And the left-hand side is null when $\tilde{x}$ is on γ^-.) Hence for suitable ε and $\mathcal{U}$ the second coordinate of $y(\tilde{t},\tilde{x})$ is positive in a neighborhood of (t,x) in $[0,T]\times B_R$. Using (3.6), one deduces (3.15) and (3.16) except for B points.

Before dealing with (3.15) for points $B\in\mathcal{B}$, let us make (3.15) more precise for points in γ^-. We consider $\tau<t$ sufficiently close to t for (7.4) to hold. Consider $\underline{s}$ the smallest time $s\in[\tau,t]$ such that $\Phi^y((s,t),t,E)\subset B_R\backslash\overline{\Omega}$, and let us suppose $\underline{s}>\tau$. One can estimate the velocity by (7.5) and consequently in our case, one has $\Phi^y(\underline{s},t,x)\notin\overline{\gamma^+}$. Certainly, one also has that $\Phi^y(\underline{s},t,x)\notin\Omega\cup\gamma^-\cup\mathcal{B}$ because of (3.16) that we just proved for points in $\gamma^-\cup\mathcal{B}$. This point can thus only be in $\partial\Omega\backslash(\overline{\gamma^+}\cup\gamma^-\cup\mathcal{B})$ (unless it is in $B_R\backslash\overline{\Omega}$). Hence it must lie in γ_B for the other components are too far because of (7.4). But by uniqueness of the flow, this is not possible. Consequently one has, for any $s\in[\tau,t]$ with τ satisfying (7.4),

$$\Phi^y(s,t,x)\in B_R\backslash\overline{\Omega}. \tag{7.6}$$

– Now, let us deal with (3.15) for B points. We see, using Remark 6(i) and the same procedure as for the proof of (3.17), that for some ν, one has $\Phi^y(t-\nu,t,B)\in B_R\backslash\overline{\Omega}$. Now we claim that, at least if ν has been chosen small enough, $\Phi^y(s,t,B)\in B_R\backslash\overline{\Omega}$ for any s in $[t-\nu,t)$. Indeed, when considering a sequence of points x_n in γ^- converging to B, by continuity of the flow we have that $\Phi^y(s,t,x_n)$ is converging to $\Phi^y(s,t,B)$. Using (7.6), one gets that $\Phi^y(s,t,B)\notin\Omega$ for $s\in[t-\nu,t)$. But $\Phi^y(s,t,B)$ cannot belong to γ^- by (3.15), which is already established for points in γ^-; nor can it belong to γ_B by uniqueness of the flow (and the other components of $\partial\Omega\backslash(\gamma^+\cup\gamma^-)$ are too far by the choice of ν). Hence, one must have $\Phi^y(s,t,B)\in B_R\backslash\overline{\Omega}$.

- *Proof of* (3.19). This is a consequence of the continuity of the flow. Suppose indeed by contradiction that for some ω, $(\lambda_i)_{i=1,\dots,g}$, and α, one has (3.19) not satisfied by an A point. Then for some $\tau>t$, $\Phi^y(\tau,t,A)\in B_R\backslash\overline{\Omega}$ (note indeed that this point cannot be in $\overline{\gamma_A}$ by uniqueness of the flow or in other components of $\partial\Omega\backslash(\gamma^+\cup\gamma^-)$ by the choice of τ (which is not too far from t), and if this point is in γ^-, we conclude by (3.15) that there indeed exists a point $\Phi^y(\tau-\nu,t,A)$ in $B_R\backslash\overline{\Omega}$). We look at the trajectories starting from x in γ^- close to A. By (3.16) they are inside Ω for small time and, in fact, by the same argument as previously, in Ω as long as $\tau-t$ is small enough so that

$$c_\pi(K\|\nabla\theta\|_\infty+C_{\mathcal{LL}})\left(\int_t^\tau\alpha(s)ds\right)\leq\ell/2.$$

  So our assumption would be in contradiction with the continuity of the flow.
- *Proof of* (3.20). This follows again from the procedure of the proof of (3.17) and the choices of $\mathrm{Supp}(\Gamma_A)$ and $\mathrm{Supp}(\Lambda_i)$ (which are at positive distance from $\underline{A}$).

Acknowledgment. The author thanks the anonymous referees for their useful comments on a first version of this paper.

REFERENCES

[1] R. W. Brockett, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory (Houghton, MI, 1982), Progr. Math. 27, Birkhäuser Boston, Boston, 1983, pp. 181–191.

[2] J.-M. Coron, *On the controllability of* 2*-D incompressible perfect fluids*, J. Math. Pures Appl. (9), 75 (1996), pp. 155–188.

[3] J.-M. Coron, *Sur la stabilisation des fluides parfaits incompressibles bidimensionnels*, in Séminaire: Équations aux Dérivées Partielles (1998–1999), École Polytechnique, Centre de Mathématiques, Palaiseau, France, 1999, pp. VII-1–VII-26 (in French).

[4] J.-M. Coron, *On the null asymptotic stabilization of two-dimensional incompressible Euler equations in a simply connected domain*, SIAM J. Control Optim., 37 (1999), pp. 1874–1896.

[5] J.-M. Coron and L. Praly, *Adding an integrator for the stabilization problem*, Systems Control Lett., 17 (1991), pp. 89–104.

[6] O. Glass, *Existence of solutions for the two-dimensional stationary Euler system for ideal fluids with arbitrary force*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 20 (2003), pp. 921–946.

[7] T. Kato, *On classical solutions of the two-dimensional nonstationary Euler equation*, Arch. Rational Mech. Anal., 25 (1967), pp. 188–200.

[8] P.-L. Lions, *Mathematical Topics in Fluid Mechanics. Vol.* 1. *Incompressible Models*, Oxford Lecture Ser. Math. Appl. 3, The Clarendon Press, Oxford University Press, New York, 1996.

[9] W. Wolibner, *Un théorème sur l'existence du mouvement plan d'un fluide parfait, homogène, incompressible, pendant un temps infiniment long*, Math. Z., 37 (1933), pp. 698-726 (in French).

[10] V. I. Yudovich, *The flow of a perfect, incompressible liquid through a given region*, Dokl. Akad. Nauk SSSR, 146 (1962), pp. 561–564 (in Russian); Soviet Physics Dokl., 7 (1962), pp. 789–791 (in English).

[11] V. I. Yudovich, *Non-stationary flows of an ideal incompressible fluid*, Ž. Vyčisl. Mat. i Mat. Fiz., 3 (1963), pp. 1032–1066 (in Russian); U.S.S.R. Comput. Math. Math. Phys., 3 (1963), pp. 1407–1456 (in English).

SIAM J. CONTROL OPTIM.
Vol. 44, No. 3, pp. 1148–1163

CHARACTERISTIC FREQUENCIES, POLYNOMIAL-EXPONENTIAL TRAJECTORIES, AND LINEAR EXACT MODELING WITH MULTIDIMENSIONAL BEHAVIORS*

EVA ZERZ†

Abstract. The characteristic frequencies of a linear, shift-invariant multidimensional behavior correspond to its nonzero exponential trajectories. The set of polynomial-exponential trajectories belonging to a fixed characteristic frequency of a behavior is investigated: A test is derived for determining whether this space is finite-dimensional, and if so, a basis is constructed. If it is infinite-dimensional, one considers only the polynomial-exponential trajectories up to a certain degree of the polynomial part, and a characterization is given of the asymptotic growth of the dimensions of these spaces as the degree bound tends to infinity. A dual problem is concerned with linear exact modeling, that is, the construction of the so-called most powerful unfalsified model (MPUM): Given a finite set of polynomial-exponential trajectories, the goal is to construct a behavior that contains the data and as little else as possible.

Key words. multidimensional systems, linear PDE with constant coefficients, behavioral approach, polynomial ideals and modules

AMS subject classifications. 93B25, 93C05, 93C20, 93A30, 13C99, 35E20

DOI. 10.1137/S0363012904441738

Introduction. The study of multidimensional, linear, shift-invariant behaviors (i.e., the solution spaces of linear systems of partial differential or difference equations with constant coefficients) using commutative algebra originated some 15 years ago in the seminal paper [9]. Since then, the theory has been advanced by several authors. In [13], the notion of a characteristic point (or frequency) of a multidimensional behavior was introduced. The present paper continues this line of research. In particular, the subset of a behavior consisting of all polynomial-exponential trajectories belonging to a fixed characteristic frequency will be studied. After some preliminaries, the question whether this set is finite-dimensional as a vector space will be addressed in section 1.1. If it is, a vector space basis can be constructed as described in section 1.2. In section 1.3, we determine the asymptotic behavior of the dimensions of the spaces of polynomial-exponential solutions belonging to a fixed characteristic frequency as the degree of the polynomial part grows.

In the second part of the paper, we study the problem of linear exact modeling within the multidimensional setting. Suppose that one observes a finite number of polynomial-exponential signals. The goal is to find a model for these data, that is, a behavior that contains the given trajectories, or in other words, a model that is unfalsified by the observations. Of course, the problem can always be solved by making this behavior large enough. However, the more solutions a model admits, the less it explains. This leads to the requirement that the model should be as powerful as possible, that is, it should not admit more solutions than necessary. In other words, we are looking for the most powerful unfalsified model (MPUM) that has been studied for one-dimensional behaviors in [1]. A preliminary version of the present paper can

*Received by the editors March 3, 2004; accepted for publication (in revised form) April 3, 2005; published electronically October 3, 2005.
http://www.siam.org/journals/sicon/44-3/44173.html
†Department of Mathematics, University of Kaiserslautern, 67663 Kaiserslautern, Germany (zerz@mathematik.uni-kl.de).

be found in [15], where the computer algebraic implementation is described in more detail, and some geometric interpretations are given.

1. Characteristic frequencies and polynomial-exponential trajectories. We deal exclusively with linear and shift-invariant behaviors. For the sake of brevity, we focus on behaviors given by differential (rather than difference) equations, and we study only their smooth solutions, although the case of distributional solutions, and the discrete counterpart, can be treated analogously.

Let $\mathcal{A} := \mathcal{C}^\infty(\mathbb{R}^n)$ denote the space of complex-valued smooth functions defined on $\mathbb{R}^n$, and let $\mathcal{D} := \mathbb{C}[s_1, \dots, s_n]$ be the polynomial ring, in n variables, with complex coefficients. Let $R \in \mathcal{D}^{g\times q}$ be a polynomial matrix. Then

$$\mathcal{B} := \{w \in \mathcal{A}^q \mid R(\partial)w := R(\partial_1, \dots, \partial_n)w = 0\}$$

is the smooth solution space of a linear system of partial differential equations with constant coefficients. Such a set $\mathcal{B}$ will be called a *behavior* throughout this paper. If R has full column rank (by the rank of a polynomial matrix R, its generic rank is meant, i.e., its rank as a matrix over the field of rational functions $\mathbb{C}(s_1, \dots, s_n)$), then $\mathcal{B}$ is called *autonomous*. This means that the system does not contain free variables, that is, none of the components w_j of w is unconstrained by the system law $R(\partial)w = 0$. More precisely, none of the projections $\pi_j : \mathcal{B} \to \mathcal{A}$, $w \mapsto w_j$ is surjective. Algebraically speaking, autonomy is also equivalent to the fact that $P := \mathcal{D}^{1\times q}/\mathcal{D}^{1\times g}R$ is a torsion module, that is, for all $p \in P$ there exists $0 \neq d \in \mathcal{D}$ such that $dp = 0$ (indeed, we even have $dP = 0$ for some $0 \neq d \in \mathcal{D}$, that is, P has a nonzero annihilator).

Let a behavior $\mathcal{B}$ be given. An n-tuple of complex numbers, $\lambda \in \mathbb{C}^n$, is called a *characteristic frequency* (or characteristic point or pole point [13]) of $\mathcal{B}$ if there exists $0 \neq c \in \mathbb{C}^q$ such that

$$w := c\exp_\lambda \tag{1.1}$$

belongs to $\mathcal{B}$, where the scalar function $\exp_\lambda$ is defined by

$$\exp_\lambda(t) := \exp(t_1\lambda_1 + \cdots + t_n\lambda_n) \qquad \text{for } t \in \mathbb{R}^n.$$

The function w from (1.1) is called an *exponential trajectory with frequency vector* λ. Thus λ is a characteristic frequency of $\mathcal{B}$ if and only if the system contains a nonzero exponential trajectory with frequency vector λ.

The formula

$$R(\partial)c\exp_\lambda = R(\lambda)c\exp_\lambda$$

is easy to verify. It shows that w from (1.1) belongs to $\mathcal{B}$ if and only if $R(\lambda)c = 0$. This observation leads to the following lemma [13].

LEMMA 1. *The vector $\lambda \in \mathbb{C}^n$ is a characteristic frequency of $\mathcal{B}$ if and only if $R(\lambda) \in \mathbb{C}^{g\times q}$ does not have full column rank.*

Thus, λ is a characteristic frequency of $\mathcal{B}$ if and only if $\operatorname{rank}(R(\lambda)) < q$, that is, if and only if λ is common zero of all the $q \times q$ minors of $R \in \mathcal{D}^{g\times q}$. Therefore, the set V of all characteristic frequencies of $\mathcal{B}$ is an algebraic variety in $\mathbb{C}^n$, called the characteristic variety of $\mathcal{B}$ in [13]. If the rank of R is less than q, then all its minors of order q are (identically) zero, and then $V = \mathbb{C}^n$, that is, every λ is a characteristic frequency of $\mathcal{B}$. On the other hand, if the rank of R equals q, that is, if $\mathcal{B}$ is autonomous, then V is a proper subset of $\mathbb{C}^n$.

As is well known from the one-dimensional case ($n = 1$), i.e., ordinary differential equations, it is not sufficient to consider exponential trajectories alone. We have to admit polynomial-exponential solutions.

For this, we define the space of *polynomial-exponential functions* $\mathcal{A}_{pe}$ as the subspace of $\mathcal{A} = \mathcal{C}^\infty(\mathbb{R}^n)$ containing all finite sums of functions of the form

$$a := b \exp_\lambda,$$

where $b \in \mathbb{C}[t_1, \ldots, t_n]$ is a polynomial and $\lambda \in \mathbb{C}^n$. Just like $\mathcal{A}$ itself, the space $\mathcal{A}_{pe}$ is a module over $\mathbb{C}[\partial_1, \ldots, \partial_n]$. Moreover, we have [13]

$$\mathfrak{M}(\mathcal{B} \cap \mathcal{A}_{pe}^q) = \mathfrak{M}(\mathcal{B}),$$

where

$$\mathfrak{M}(B) := \{r \in \mathcal{D}^{1\times q} \mid r(\partial)w = 0 \text{ for all } w \in B\}$$

for an arbitrary set $B \subseteq \mathcal{A}^q$. In particular, $\mathcal{B}$ is uniquely determined by $\mathcal{B}_{pe} := \mathcal{B} \cap \mathcal{A}_{pe}^q$. Indeed, $\mathcal{B}_{pe}$ is dense in $\mathcal{B}$ [7], with respect to the standard $\mathcal{C}^\infty$-topology, that is, the topology of uniform convergence of all derivatives on compact sets. This fact is also referred to as the Malgrange approximation theorem, e.g., in [11]. If $\mathcal{B}$ is finite-dimensional (as a complex vector space) [2, 8, 10, 11, 12], then

$$\mathcal{B}_{pe} = \mathcal{B}.$$

This is true, for example, for every autonomous one-dimensional ($n = 1$) behavior, because the solution space of a homogeneous system of linear constant-coefficient ordinary differential equations is finite-dimensional. In dimensions $n \geq 2$, however, an autonomous system is not necessarily finite-dimensional (take for instance $n = 2$, and $R = s_1$; then $\mathcal{B}$ consists of all smooth functions of t_2). On the other hand, a finite-dimensional behavior is always autonomous.

One calls $\mathcal{B}_{pe}$ the set of *polynomial-exponential trajectories of* $\mathcal{B}$. Below, we will investigate the polynomial-exponential trajectories of $\mathcal{B}$ with respect to a fixed frequency vector.

For this, let $\lambda \in \mathbb{C}^n$. Via the affine change of variables

$$\tilde{R}(s) := R(s + \lambda)$$

we obtain $\tilde{\mathcal{B}} := \{w \in \mathcal{A}^q \mid \tilde{R}(\partial)w = 0\}$. There is a bijective relation between the polynomial-exponential solutions with frequency vector λ to $R(\partial)w = 0$ and the polynomial solutions to $\tilde{R}(\partial)w = 0$, which is given by

$$p \in \tilde{\mathcal{B}} \Leftrightarrow p \exp_\lambda \in \mathcal{B} \tag{1.2}$$

for $p \in \mathbb{C}[t_1, \ldots, t_n]^q$. This can be seen by comparing the formulas

$$(\partial + \lambda)^\mu t^\nu = \sum_\rho \binom{\mu}{\rho} (\partial^\rho t^\nu) \lambda^{\mu-\rho}$$

and

$$\partial^\mu t^\nu \exp_\lambda(t) = \sum_\rho \binom{\mu}{\rho} (\partial^\rho t^\nu)(\partial^{\mu-\rho} \exp_\lambda(t)) = \exp_\lambda(t) \sum_\rho \binom{\mu}{\rho} (\partial^\rho t^\nu) \lambda^{\mu-\rho},$$

which hold for all $\mu, \nu \in \mathbb{N}^n$, where the summation runs over all $\rho \in \mathbb{N}^n$ with $\rho_i \leq \mu_i$ for all $1 \leq i \leq n$, and the usual multi-index notation is used, e.g., $t^\nu := t_1^{\nu_1} \cdots t_n^{\nu_n}$, $\binom{\mu}{\rho} := \binom{\mu_1}{\rho_1} \cdots \binom{\mu_n}{\rho_n}$, etc. Thus

$$\exp_\lambda(t)(\partial + \lambda)^\mu t^\nu = \partial^\mu t^\nu \exp_\lambda(t).$$

By taking linear combinations, this shows that

$$\exp_\lambda(t) R(\partial + \lambda) p(t) = R(\partial) p(t) \exp_\lambda(t)$$

for any $R \in \mathcal{D}^{g \times q}$ and any $p \in \mathbb{C}[t_1, \ldots, t_n]^q$. Therefore, omitting the argument t,

$$R(\partial + \lambda)p = 0 \Leftrightarrow R(\partial) p \exp_\lambda = 0,$$

which is precisely the equivalence from (1.2).

Therefore we may assume, without loss of generality, that the characteristic frequency under consideration lies at the origin, and we write again R and $\mathcal{B}$ instead of $\tilde{R}$ and $\tilde{\mathcal{B}}$, respectively. For the remainder of this section, let $R \in \mathcal{D}^{g \times q}$ be a given polynomial matrix, and let $\mathcal{B} = \{w \in \mathcal{A}^q \mid R(\partial)w = 0\}$ be the corresponding behavior. We will investigate the polynomial-exponential trajectories of $\mathcal{B}$ with frequency vector zero. In other words, we are interested in the polynomial solutions to $R(\partial)w = 0$. Define

$$\mathfrak{P} := \{p \in \mathbb{C}[t_1, \ldots, t_n]^q \mid R(\partial)p = 0\}$$

and for any integer $d \geq 0$,

$$\mathfrak{P}_d := \{p \in \mathbb{C}[t_1, \ldots, t_n]^q \mid R(\partial)p = 0 \text{ and } \deg(p) \leq d - 1\}.$$

Here $\deg(p) := \max_{1 \leq j \leq q} \deg(p_j)$, where $\deg(\cdot)$ denotes the total degree of a nonzero polynomial, and $\deg(0) := -1$. Clearly,

$$\{0\} = \mathfrak{P}_0 \subseteq \mathfrak{P}_1 \subseteq \mathfrak{P}_2 \subseteq \cdots \subseteq \mathfrak{P}. \tag{1.3}$$

We have

$$\mathfrak{P} = \mathcal{B} \cap \mathbb{C}[t_1, \ldots, t_n]^q = \bigcup_{d=0}^{\infty} \mathfrak{P}_d$$

and

$$\mathfrak{P}_d = \{w \in \mathcal{A}^q \mid R(\partial)w = 0 \text{ and } \partial^\mu w = 0 \text{ for all } \mu \in \mathbb{N}^n \text{ with } |\mu| = d\},$$

where $|\mu| := \mu_1 + \cdots + \mu_n$. Here we use that w is a polynomial of degree less than d if and only if all its derivatives of order d vanish.

Let $M := \mathcal{D}^{1 \times g} R \subseteq \mathcal{D}^{1 \times q}$ be the polynomial module generated by the rows of R. According to [9], we have

$$M = \mathfrak{M}(\mathcal{B}) = \{r \in \mathcal{D}^{1 \times q} \mid r(\partial)w = 0 \text{ for all } w \in \mathcal{B}\}.$$

Conversely, we write

$$\mathcal{B} = \mathfrak{B}(M) := \{w \in \mathcal{A}^q \mid r(\partial)w = 0 \text{ for all } r \in M\}.$$

Due to the injective cogenerator property of $\mathcal{A}$ [9], the maps $\mathfrak{M}$ and $\mathfrak{B}$ are inclusion-reversing bijections between the set of behaviors in $\mathcal{A}^q$ and the set of submodules of $\mathcal{D}^{1\times q}$, and they are inverse to each other. Moreover, we note for later use that

$$\mathfrak{B}(M_1 \cap M_2) = \mathfrak{B}(M_1) + \mathfrak{B}(M_2). \tag{1.4}$$

Using these maps, we have

$$\mathfrak{M}(\mathfrak{P}_d) = M + I^d\mathcal{D}^{1\times q} \quad \text{and} \quad \mathfrak{P}_d = \mathfrak{B}(M + I^d\mathcal{D}^{1\times q}),$$

where $I := \langle s_1, \dots, s_n\rangle$ is the ideal in $\mathcal{D}$ generated by $s_1, \dots, s_n$, and the ideal I^d consists of all products of d elements of I, that is, it is generated by the monomials s^μ, where $\mu \in \mathbb{N}^n$ satisfies $|\mu| = d$. Finally,

$$I^d\mathcal{D}^{1\times q} := I^d \times \cdots \times I^d \subseteq \mathcal{D}^{1\times q}$$

is the module of all polynomial row vectors whose entries belong to I^d.

Each $\mathfrak{P}_d$ is a finite-dimensional complex vector space, and thus autonomous. We have [8, 10, 12]

$$\dim_{\mathbb{C}} \mathfrak{P}_d = \dim_{\mathbb{C}} \mathcal{D}^{1\times q}/\mathfrak{M}(\mathfrak{P}_d) = \dim_{\mathbb{C}} \mathcal{D}^{1\times q}/(M + I^d\mathcal{D}^{1\times q}). \tag{1.5}$$

Let

$$h_d := \dim_{\mathbb{C}} \mathfrak{P}_{d+1}/\mathfrak{P}_d. \tag{1.6}$$

In what follows, we will

1. decide whether $\mathfrak{P}$ is finite-dimensional as a complex vector space (this is true if and only if the sequence (1.3) becomes stationary, or, equivalently, $h_d = 0$ for large enough d);
2. construct a basis of $\mathfrak{P}_d$ (in the case where $\mathfrak{P}$ is finite-dimensional, this yields a basis of $\mathfrak{P}$ itself);
3. determine the behavior of h_d as d tends to infinity.

1.1. Deciding whether $\mathfrak{P}$ is finite-dimensional. Let $M = \mathcal{D}^{1\times g}R \subseteq \mathcal{D}^{1\times q}$, $I = \langle s_1, \dots, s_n\rangle$, and

$$(M : I) = \{r \in \mathcal{D}^{1\times q} \mid rI \subseteq M\}.$$

Since $\mathcal{D} = I^0 \supseteq I = I^1 \supseteq I^2 \supseteq \cdots$, we have

$$M \subseteq (M : I) \subseteq (M : I^2) \subseteq (M : I^3) \subseteq \cdots. \tag{1.7}$$

The *saturation* of M by I is defined by [4, 6, 12]

$$N := (M : I^\infty) := \bigcup_{d=0}^{\infty} (M : I^d).$$

Due to the Noetherian property of $\mathcal{D}$, the sequence (1.7) must become stationary, that is, there exists an integer $l \geq 0$ such that

$$(M : I^\infty) = (M : I^l).$$

The module N, being a submodule of $\mathcal{D}^{1\times q}$, possesses a representation $N = \mathcal{D}^{1\times h}S$ for some $S \in \mathcal{D}^{h\times q}$. The calculation of such a matrix S is implemented, e.g., in the computer algebra system SINGULAR [5]. Since $M \subseteq N$, we have $R = XS$ for some $\mathcal{D}$-matrix X. This implies that $\operatorname{rank}(R) \le \operatorname{rank}(S)$. Indeed, the rows of R and the rows of S generate the same $\mathbb{C}(s_1, \dots, s_n)$-vector space, and thus, the ranks of R and S coincide. Still, it may happen that $\operatorname{rank}(R(0)) < \operatorname{rank}(S(0))$. The inclusion $M \subseteq N$ also implies that $\mathfrak{B}(N) = \{w \in \mathcal{A}^q \mid S(\partial)w = 0\} \subseteq \mathfrak{B}(M) = \mathcal{B}$. The following theorem can be found in [12], for the special case of M being an ideal ($q = 1$). However, the version given below holds for arbitrary q.

THEOREM 1. *The following are equivalent:*

1. $\mathfrak{P}$ *is finite-dimensional over* $\mathbb{C}$*;*
2. $\mathfrak{B}(N)$ *contains no nonzero polynomial trajectories;*
3. *zero is not a characteristic frequency of* $\mathfrak{B}(N)$*;*
4. $\operatorname{rank}(S(0)) = q$.

Moreover, if these equivalent assertions are true, then $\mathcal{B}$ *is autonomous.*

Proof. The equivalence of assertions 3 and 4 is a consequence of Lemma 1. The implication "2 $\Rightarrow$ 3" follows trivially from the definition of a characteristic frequency. Conversely, suppose that $\mathfrak{B}(N)$ contains a nonzero polynomial vector, say $S(\partial)p = 0$, where $p(t) = \sum_{|\nu|<d} \frac{1}{\nu!} p_\nu t^\nu$ for some $d \in \mathbb{N}$, $p_\nu \in \mathbb{C}^q$. (The factors $\nu! := \nu_1! \cdots \nu_n!$ are extracted from the coefficient vectors for notational convenience.) Let d be chosen as small as possible, that is, there exists μ such that $|\mu| = d - 1$ and $p_\mu \neq 0$. Then $c := \partial^\mu p = p_\mu$ and hence $0 \neq c \in \mathbb{C}^q$. On the other hand, $S(\partial)c = S(\partial)\partial^\mu p = \partial^\mu S(\partial)p = 0$, that is, $c \in \mathfrak{B}(N)$, showing that zero is a characteristic frequency of $\mathfrak{B}(N)$.

Thus, assertions 2–4 are equivalent, and it remains to show their equivalence with the first assertion. For this, let $\mu \in \mathbb{N}^n$. Using the injective cogenerator property of $\mathcal{A}$ [9], one can show that $\partial^\mu\mathcal{B} := \{\partial^\mu w \mid w \in \mathcal{B}\}$ is again a behavior, and that

$$\mathfrak{M}(\partial^\mu\mathcal{B}) = (\mathfrak{M}(\mathcal{B}) : s^\mu) := \{r \in \mathcal{D}^{1\times q} \mid rs^\mu \in \mathfrak{M}(\mathcal{B})\}.$$

Hence, with $M = \mathfrak{M}(\mathcal{B})$, we have $\partial^\mu\mathcal{B} = \mathfrak{B}(M : s^\mu)$ and thus

$$\mathfrak{B}(M : I^d) = \mathfrak{B}\Big(\bigcap_{|\mu|=d} (M : s^\mu)\Big) = \sum_{|\mu|=d} \mathfrak{B}(M : s^\mu) = \sum_{|\mu|=d} \partial^\mu\mathcal{B},$$

where the relation (1.4) was used for the second equality. The set $\mathfrak{P}$ is infinite-dimensional if and only if there exist polynomial elements of $\mathcal{B}$ of arbitrarily high degree. This means that for all d, the behavior $\sum_{|\mu|=d} \partial^\mu\mathcal{B}$ possesses nonzero polynomial elements, and hence, similarly as shown in the first part of the proof, it also possesses nonzero constant solutions, that is, zero is a characteristic frequency of $\mathfrak{B}(M : I^d)$ for all d. Still equivalently, zero is a characteristic frequency of $\mathfrak{B}(M : I^\infty) = \mathfrak{B}(N)$.

The final statement follows from the fact that if $\operatorname{rank}(S(0)) = q$, then the rank of S, and hence of R, must be q. ∎

The set $\mathfrak{P}$ is finite-dimensional if and only if there exists an upper bound for the degree of any polynomial trajectory. This means that $h_d = 0$ for large enough d. Let d^* be the smallest integer such that $h_{d^*} = 0$, that is,

$$\mathfrak{P}_{d^*} = \mathfrak{P}_{d^*+1}.$$

Then $d^* - 1$ is the maximal degree of any polynomial solution to $R(\partial)w = 0$. This is due to the fact that the first equality in (1.3) will already yield stationarity, as shown in the subsequent lemma.

LEMMA 2. *If* $\mathfrak{P}_d = \mathfrak{P}_{d+1}$, *then*

$$\mathfrak{P}_d = \mathfrak{P}_{d+l} \quad \textit{for all } l \geq 0.$$

Thus $\mathfrak{P} = \mathfrak{P}_{d^*}$, *where* d^* *is the smallest integer with* $\mathfrak{P}_{d^*} = \mathfrak{P}_{d^*+1}$.

Proof. It suffices to show that $\mathfrak{P}_{d+2} \subseteq \mathfrak{P}_{d+1}$. We have

$$\mathfrak{P}_d = \mathfrak{B}(M + I^d\mathcal{D}^{1\times q}).$$

Hence the assumption $\mathfrak{P}_{d+1} \subseteq \mathfrak{P}_d$ is equivalent to

$$M + I^d\mathcal{D}^{1\times q} \subseteq M + I^{d+1}\mathcal{D}^{1\times q}$$

or

$$I^d\mathcal{D}^{1\times q} \subseteq M + I^{d+1}\mathcal{D}^{1\times q}.$$

From the last inclusion, it is easy to conclude that

$$I^{d+1}\mathcal{D}^{1\times q} \subseteq M + I^{d+2}\mathcal{D}^{1\times q}$$

and hence one obtains the desired result, by adding M, and by applying the inclusion-reversing map $\mathfrak{B}$. □

Example. Consider

$$R = \begin{bmatrix} s_1(s_2+1) & 0 \\ 0 & s_3 \\ (s_2+1)(s_3+1) & s_1 \end{bmatrix},$$

whose behavior has zero as a characteristic frequency, because $\operatorname{rank}(R(0)) = 1$. Using SINGULAR, it turns out that $N = M$. Hence we may choose $S = R$ and we may conclude that $\mathfrak{P}$ is not finite-dimensional. However, if we add an additional row and consider

$$\hat{R} = \begin{bmatrix} s_1(s_2+1) & 0 \\ 0 & s_3 \\ (s_2+1)(s_3+1) & s_1 \\ 0 & s_2^2 \end{bmatrix},$$

then the resulting behavior still has the characteristic frequency zero, but SINGULAR returns

$$\hat{S} = \begin{bmatrix} s_1(s_2+1) & 0 \\ 0 & 1 \\ (s_2+1)(s_3+1) & 0 \end{bmatrix},$$

and hence $\operatorname{rank}(\hat{S}(0)) = 2$, that is, $\hat{\mathfrak{P}}$ is finite-dimensional. In fact, $\dim_{\mathbb{C}} \hat{\mathfrak{P}} = 4$, and $\hat{\mathfrak{P}}_2 \neq \hat{\mathfrak{P}}_3 = \hat{\mathfrak{P}}_4$, that is, $d^* = 3$ and thus the largest degree of a polynomial solution to $\hat{R}(\partial)w = 0$ equals 2.

1.2. Computing a basis of $\mathfrak{P}_d$. A basis of $\mathfrak{P}_d$ can be computed in various ways, e.g., see [10, 12]. Here, we use the method described in [8]. For this, let $\delta := \dim_{\mathbb{C}} \mathfrak{P}_d$. Then $\mathfrak{P}_d$ possesses a representation

$$\mathfrak{P}_d = \{w \mid \exists x_0 \in \mathbb{C}^\delta : w(t) = C \exp(t_1 A_1 + \cdots + t_n A_n) x_0 \text{ for all } t \in \mathbb{R}^n\},$$

where $A_i \in \mathbb{C}^{\delta\times\delta}$ are nilpotent, pairwise commuting matrices, and $C \in \mathbb{C}^{q\times\delta}$. Thus, the columns of $C \exp(t_1 A_1 + \cdots + t_n A_n)$ are a basis of $\mathfrak{P}_d$.

The matrices A_i and C can be constructed as follows [8]:

Let

$$M^{(d)} := \mathfrak{M}(\mathfrak{P}_d) = M + I^d \mathcal{D}^{1\times q}.$$

According to (1.5), $P^{(d)} := \mathcal{D}^{1\times q}/M^{(d)}$ has dimension δ as a $\mathbb{C}$-vector space, and a concrete $\mathbb{C}$-basis $\{b_1, \ldots, b_\delta\}$ of $P^{(d)}$ can be found, e.g., using Gröbner bases. Thus there is a $\mathbb{C}$-vector space isomorphism $\mathbb{C}^\delta \cong P^{(d)}$, $e_k \leftrightarrow b_k$, where e_k is the kth natural basis vector of $\mathbb{C}^\delta$. Using this isomorphism, the multiplication by s_i in $P^{(d)}$ yields matrices $F_i \in \mathbb{C}^{\delta\times\delta}$, which are pairwise commuting (because $s_i s_j = s_j s_i$ in $P^{(d)}$) and nilpotent (because $s_i^d P^{(d)} = 0$ for all i). Let $A_i := F_i^T$ and let $C \in \mathbb{C}^{q\times\delta}$ be the matrix obtained from expressing each $[e_j] \in P^{(d)}$, where e_j is the jth natural basis vector of $\mathcal{D}^{1\times q}$, in terms of the basis elements of $P^{(d)}$ according to

$$[e_j] = \sum_{k=1}^{\delta} C_{jk} b_k.$$

This construction method works for any finite-dimensional behavior $\mathcal{B}$, not only for the behaviors of the form $\mathfrak{P}_d$. For instance, consider a one-dimensional and scalar system ($n = q = 1$), say $r(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ for $r = s^\delta + \alpha_{\delta-1}s^{\delta-1} + \cdots + \alpha_1 s + \alpha_0$. The procedure from above yields (choosing $b_1 = [1], \ldots, b_\delta = [s^{\delta-1}]$)

$$A = \begin{bmatrix} 0 & 1 & & \\ \vdots & & \ddots & \\ 0 & & & 1 \\ -\alpha_0 & \cdots & \cdots & -\alpha_{\delta-1} \end{bmatrix} \quad \text{and} \quad C = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix},$$

which is the so-called observability form. Similarly, one obtains the observer form by a suitable choice of the basis elements b_k. Note that the matrix A in this example is not in general nilpotent. This is because $\mathcal{B} = \{w \in \mathcal{A} \mid r(\frac{\mathrm{d}}{\mathrm{d}t})w = 0\}$ is finite-dimensional, but does not have the form $\mathfrak{P}_d$. For this $\mathcal{B}$, we have

$$\mathfrak{P}_d = \left\{ w \in \mathcal{A} \middle| r\left(\frac{\mathrm{d}}{\mathrm{d}t}\right) w = 0, \frac{\mathrm{d}^d}{\mathrm{d}t^d} w = 0 \right\} = \left\{ w \in \mathcal{A} \middle| \frac{\mathrm{d}^e}{\mathrm{d}t^e} w = 0 \right\},$$

where $e = \min\{d, \min\{k \mid \alpha_k \neq 0\}\}$, that is, s^e is the greatest common divisor of r and s^d. Thus, the only one-dimensional and scalar case that fits into the situation outlined above is $r = s^e$ and then all $\alpha_k = 0$ and hence we obtain a nilpotent matrix A.

In section 2, we will discuss another realization of a finite-dimensional behavior that is dual to the given one in a certain sense, and that corresponds to the controllability/controller forms in the one-dimensional scalar case.

Example. Let us return to $\hat{R}$ from the previous example. We have $d^* = 3$ and $\hat{\mathfrak{P}} = \hat{\mathfrak{P}}_3 = \{w \in \mathcal{A}^2 \mid Q(\partial)w = 0\}$, where

$$Q := \begin{bmatrix} s_3 & 0 \\ 0 & s_3 \\ s_2^2 & 0 \\ 0 & s_2^2 \\ s_1 & 0 \\ s_2 + 1 & s_1 \end{bmatrix}. \tag{1.8}$$

This matrix has been obtained by computing a Gröbner basis of $M^{(3)} = \mathcal{D}^{1\times 4}\hat{R} + I^3\mathcal{D}^{1\times 2} = \mathcal{D}^{1\times 6}Q$. Here, $\delta = 4$. As a basis of $P^{(3)} = \mathcal{D}^{1\times 2}/M^{(3)}$, we choose $b_1 = [e_1]$, $b_2 = [e_2]$, $b_3 = [s_2e_1]$, $b_4 = [s_2e_2]$. Then

$$A_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad A_3 = 0$$

and

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Thus all polynomial solutions to $\hat{R}(\partial)w = 0$ have the form

$$w(t) = C\exp(t_1A_1 + t_2A_2 + t_3A_3)x_0 = \begin{bmatrix} c_1 + c_3t_2 \\ c_2 - (c_1 + c_3)t_1 + c_4t_2 - c_3t_1t_2 \end{bmatrix}, \tag{1.9}$$

where $x_0 = [c_1, c_2, c_3, c_4]^T \in \mathbb{C}^4$.

1.3. Asymptotic behavior of the numbers h_d. Consider the function $h : \mathbb{N} \to \mathbb{N}$, $d \mapsto h_d$, where h_d is defined in (1.6). As above, let $I = \langle s_1, \dots, s_n\rangle$. The set $\mathcal{D} \setminus I$ consists of all polynomials with a nonvanishing constant term, and thus, it is multiplicatively closed. The *localization* [4, 6] of $\mathcal{D}$ at I is defined by

$$\mathcal{D}_0 := \left\{ \frac{f}{g} \mid f \in \mathcal{D}, g \in \mathcal{D} \setminus I \right\}.$$

We identify $\mathcal{D}$ with a subring of $\mathcal{D}_0$ via $f = \frac{f}{1}$. Similarly, for $M = \mathcal{D}^{1\times g}R \subseteq \mathcal{D}^{1\times q}$, we define the localization of M at I as

$$M_0 := \left\{ \frac{f}{g} \mid f \in M, g \in \mathcal{D} \setminus I \right\}$$

and we identify M with a submodule of M_0. The (Krull) *dimension* [4, 6] of M_0 is an integer between -1 and n. The case $\dim(M_0) = -1$ corresponds to $\operatorname{rank}(R(0)) = q$, that is, to zero not being a characteristic frequency of $\mathcal{B}$. Then we have $\mathfrak{P}_d = \{0\}$ for all d, and the function h is identically zero. If we exclude this special case in the following theorem, we have $0 \le \dim(M_0) \le n$. Note that $\dim(M_0) = n$ holds if and only if $\mathcal{B}$ is not autonomous. The dimension can be computed by means of computer algebra systems such as SINGULAR.

THEOREM 2. *If* $\operatorname{rank}(R(0)) = q$, *then the function* h *is identically zero. Let* $\operatorname{rank}(R(0)) < q$. *Then the function* h *agrees, for large* d, *with a polynomial of degree* $\dim(M_0) - 1$, *where* M_0 *is the localization of* M *at* I, *and* $\dim(\cdot)$ *denotes the Krull dimension.*

Proof. Let $\mathcal{D}_0$ be the localization of $\mathcal{D}$ at I. Then $M_0 = \mathcal{D}_0^{1\times g}R \subseteq \mathcal{D}_0^{1\times q}$ and $\dim(M_0) = \dim(P_0)$, where $P_0 := \mathcal{D}_0^{1\times q}/M_0$. Since $\mathcal{D}_0$ is a local ring with maximal ideal $I_0 := \mathcal{D}_0 I$, it follows from the theory of the Hilbert–Samuel function [4, 6] that

$$\dim_{\mathbb{C}} P_0/I_0^d P_0$$

coincides, for large d, with a polynomial of degree $\dim(M_0)$. We have

$$P_0/I_0^d P_0 \cong \mathcal{D}_0^{1\times q}/M_0 + I_0^d \mathcal{D}_0^{1\times q} \cong \mathcal{D}^{1\times q}/M + I^d\mathcal{D}^{1\times q}. \tag{1.10}$$

The first isomorphism is straightforward, and can be directly verified by checking that the mapping

$$\begin{aligned} \mathcal{D}_0^{1\times q}/M_0 + I_0^d\mathcal{D}_0^{1\times q} &\to \mathcal{D}_0^{1\times q}/M_0/I_0^d\big(\mathcal{D}_0^{1\times q}/M_0\big), \\ r_0 + M_0 + I_0^d\mathcal{D}_0^{1\times q} &\mapsto \big(r_0 + M_0\big) + I_0^d\big(\mathcal{D}_0^{1\times q}/M_0\big) \end{aligned}$$

has the desired properties. The second isomorphism is more interesting. In [3], it is proven for zero-dimensional ideals. The generalization to the module case is easy, but it is given here for the sake of completeness. Consider the homomorphism

$$\varphi : \mathcal{D}^{1\times q} \to \mathcal{D}_0^{1\times q}/M_0 + I_0^d\mathcal{D}_0^{1\times q}, \qquad r \mapsto \varphi(r) := \tfrac{r}{1} + M_0 + I_0^d\mathcal{D}_0^{1\times q}.$$

It is clear that $r \in M + I^d\mathcal{D}^{1\times q}$ implies $\varphi(r) = 0$, and hence we have

$$M + I^d\mathcal{D}^{1\times q} \subseteq \ker(\varphi).$$

In order to establish the second isomorphism in (1.10), it suffices, in view of the homomorphism theorem, to show the converse inclusion, and the surjectivity of φ. Both facts follow from the following observation: For each $g \in \mathcal{D}\setminus I$ there exists $h \in \mathcal{D}$ such that $(1-hg)e_j \in I^d\mathcal{D}^{1\times q}$ for all $1 \le j \le q$, where e_j denotes the jth natural basis vector of $\mathcal{D}^{1\times q}$. To see this, note that $g_0 := 1 - \frac{g}{g(0)} \in I$, where $g(0) \in \mathbb{C}$ is the nonzero constant term of $g \notin I$. Then $h := \frac{1}{g(0)}(1 + g_0 + \cdots + g_0^{d-1})$ yields $hg = \frac{1}{g(0)}(1 + g_0 + \cdots + g_0^{d-1})g(0)(1-g_0) = 1 - g_0^d$. Thus $1 - hg = g_0^d \in I^d$, and therefore $(1-hg)e_j \in I^d\mathcal{D}^{1\times q}$ for all j. In other words, $e_j \equiv hge_j$ modulo $I^d\mathcal{D}^{1\times q}$.

Now let $r \in \ker(\varphi)$. Then $\frac{r}{1} = \frac{f}{g}$ for some $g \in \mathcal{D}\setminus I$ and some $f = gr \in M + I^d\mathcal{D}^{1\times q}$. Let $h \in \mathcal{D}$ be as constructed above. Then we have $hgr = hf \in M + I^d\mathcal{D}^{1\times q}$, and on the other hand,

$$hgr = \sum r_j hge_j \equiv \sum r_j e_j = r \text{ modulo } M + I^d\mathcal{D}^{1\times q}.$$

This implies that $r \in M + I^d\mathcal{D}^{1\times q}$. The surjectivity of φ is proven similarly.

Combining (1.10) with the dimension formula (1.5), one obtains that

$$\dim_{\mathbb{C}} \mathfrak{P}_d = \dim_{\mathbb{C}} \mathcal{D}^{1\times q}/M + I^d\mathcal{D}^{1\times q} = \dim_{\mathbb{C}} P_0/I_0^d P_0$$

is equal, for large d, to a polynomial of degree $\delta := \dim(M_0)$, say,

$$\dim_{\mathbb{C}} \mathfrak{P}_d = \alpha_\delta d^\delta + \sum_{k=0}^{\delta-1} \alpha_k d^k$$

for suitable coefficients α_k, and large enough d. Then

$$h_d = \dim_{\mathbb{C}} \mathfrak{P}_{d+1} - \dim_{\mathbb{C}} \mathfrak{P}_d = \alpha_\delta\big((d+1)^\delta - d^\delta\big) + \sum_{k=0}^{\delta-1} \alpha_k\big((d+1)^k - d^k\big)$$
$$= \alpha_\delta \delta d^{\delta-1} + \sum_{k=0}^{\delta-2} \beta_k d^k$$

for some coefficients β_k. It follows that h_d is a polynomial of degree $\dim(M_0) - 1$, for large d. □

If M is homogeneous, then the polynomial from Theorem 2 coincides with the *Hilbert polynomial* [4, 6] of M. For an interpretation of the Hilbert polynomial in the context of multidimensional behaviors, see [14].

As a consequence of the proof of Theorem 2, the function $\dim_{\mathbb{C}} \mathfrak{P}_d$ itself agrees, for large d, with a polynomial of degree $\dim(M_0)$. The connection between Theorems 1 and 2 is given by the fact that $\operatorname{rank}(S(0)) = q$ is equivalent to $\dim(M_0) \leq 0$, and $\operatorname{rank}(R(0)) < \operatorname{rank}(S(0)) = q$ is equivalent to $\dim(M_0) = 0$.

Example. Let us return to the matrix R from the previous example. We have already seen that $\mathfrak{P}$ is not finite-dimensional. Indeed, the numbers $\dim_{\mathbb{C}} \mathfrak{P}_d$ satisfy

$$\dim_{\mathbb{C}} \mathfrak{P}_d = 2d - 1$$

for all $d \geq 1$ and thus $h_d = 2$ for large enough d, a polynomial of degree 0. Note that the module $M = \mathcal{D}^{1\times 3}R$ has Krull dimension 2, but M_0, its localization at zero, has Krull dimension 1.

2. Polynomial-exponential trajectories and the MPUM. Let

$$p \in \mathbb{C}[t_1, \ldots, t_n]^q$$

be a vector of polynomials. In this section, we construct the smallest behavior $\mathcal{B}$ that contains p. This behavior will be called the *most powerful unfalsified model* (MPUM) of p [1]. For this, we write

$$p(t) = \sum_{\nu \in \mathbb{N}^n, |\nu| < d} \frac{1}{\nu!} p_\nu t^\nu, \qquad p_\nu \in \mathbb{C}^q,$$

where $\nu! = \nu_1! \cdots \nu_n!$, and we assume, without loss of generality, that d is chosen as small as possible. We have

$$\partial^\mu p = 0 \quad \text{for all } \mu \in \mathbb{N}^n \text{ with } |\mu| = d.$$

Consider the module $P := \mathcal{D}/I^d$. As a $\mathbb{C}$-vector space, P is generated by $[s^\nu]$, where $\nu \in \mathbb{N}^n$ is such that $|\nu| < d$. Thus the underlying vector space of P isomorphic to $\mathbb{C}^\delta$, where $\delta := |\{\nu \in \mathbb{N}^n \mid |\nu| < d\}|$. A simple combinatorial argument shows that

$$\delta = \binom{n+d-1}{n}.$$

The multiplication by s_i in P yields a linear transformation in $\mathbb{C}^\delta$ which we may identify with a matrix F_i after fixing a basis of $\mathbb{C}^\delta$. The easiest way to do this is to enumerate the elements of $\{\nu \in \mathbb{N}^n \mid |\nu| < d\} =: \{\nu_1, \ldots, \nu_\delta\}$, e.g., in lexicographical order, and to identify the element $[s^{\nu_k}]$ of P with the kth natural basis vector

of $\mathbb{C}^\delta$. The matrices obtained this way are pairwise commuting and nilpotent. This construction is a special case of the procedure described in section 1.2, but note that we put $A_i := F_i$ here (without transposition).

THEOREM 3. *The MPUM of p is given by*

$$\begin{aligned}\mathcal{B} &= \{w \in \mathcal{A}^q \mid \exists x \in \mathcal{A}^\delta : \partial_i x = A_i x \text{ for } 1 \le i \le n \text{ and } w = Cx\} \\ &= \{w \in \mathcal{A}^q \mid \exists x_0 \in \mathbb{C}^\delta : w(t) = C\exp(t_1A_1 + \cdots + t_nA_n)x_0 \text{ for all } t \in \mathbb{R}^n\},\end{aligned}$$

where C is the $q \times \delta$ matrix obtained from putting the coefficient vector p_{ν_k} into the kth column of C. In particular, the MPUM is finite-dimensional over $\mathbb{C}$, and hence autonomous.

This construction of the MPUM is similar to the one given in [1] for the case $n = 1$. The second equality in Theorem 3 is well known [2, 8, 10, 11, 12].

Proof. We show that $\mathcal{B}$ from above is really the MPUM of p. Note that

$$\xi(t) := \begin{bmatrix} \frac{1}{\nu_1!}t^{\nu_1} \\ \vdots \\ \frac{1}{\nu_\delta!}t^{\nu_\delta} \end{bmatrix}$$

satisfies $\partial_i\xi = A_i\xi$ for all i, and thus $p = C\xi \in \mathcal{B}$. Moreover, $\xi_\nu := \partial^\nu\xi$ also satisfies $\partial_i\xi_\nu = A_i\xi_\nu$ and we get $\partial^\nu p = C\xi_\nu \in \mathcal{B}$ for all $\nu \in \mathbb{N}^n$. Thus the given polynomial vector p and all its derivatives belong to $\mathcal{B}$. On the other hand, we have

$$\mathcal{B} = \{w \in \mathcal{A}^q \mid \exists x_0 \in \mathbb{C}^\delta : w(t) = C\exp(t_1A_1 + \cdots + t_nA_n)x_0 \text{ for all } t \in \mathbb{R}^n\}.$$

Since $\{\xi_\nu(0) \mid \nu \in \mathbb{N}^n, |\nu| < d\}$, where $\xi_0 = \xi$, is the set of all natural basis vectors of $\mathbb{C}^\delta$, we conclude that $\{C\xi_\nu \mid \nu \in \mathbb{N}^n, |\nu| < d\}$ is a generating set of $\mathcal{B}$. Therefore, $\mathcal{B}$ is the linear span of the given polynomial vector p and its derivatives, and thus it is the smallest behavior containing p. ∎

In the one-dimensional and scalar case, say, $p(t) = \alpha_{\delta-1}t^{\delta-1} + \cdots + \alpha_1 t + \alpha_0$, this yields

$$A = \begin{bmatrix} 0 & \cdots & 0 & 0 \\ 1 & & & \vdots \\ & \ddots & & \vdots \\ & & 1 & 0 \end{bmatrix} \quad \text{and} \quad C = \begin{bmatrix} \alpha_0 & \cdots & k!\alpha_k & \cdots & (\delta-1)!\alpha_{\delta-1} \end{bmatrix}$$

and then $C\exp(tA) = (p(t), p'(t), \ldots, p^{(\delta-1)}(t))$, in particular, $C\exp(tA)e_1 = p(t)$. This form is reminiscent of the controllability form. The controller form analogue is obtained similarly (by choosing another enumeration of the basis).

Example. Consider the polynomial vector from (1.9) as an element of $\mathbb{C}[t_1, t_2]^2$. Then $d = 3$ and $\delta = 6$. We have

$$A_1 = \begin{bmatrix} 0&0&0&0&0&0 \\ 0&0&0&0&0&0 \\ 0&0&0&0&0&0 \\ 1&0&0&0&0&0 \\ 0&1&0&0&0&0 \\ 0&0&0&1&0&0 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} 0&0&0&0&0&0 \\ 1&0&0&0&0&0 \\ 0&1&0&0&0&0 \\ 0&0&0&0&0&0 \\ 0&0&0&1&0&0 \\ 0&0&0&0&0&0 \end{bmatrix}$$

and

$$C = \begin{bmatrix} c_1 & c_3 & 0 & 0 & 0 & 0 \\ c_2 & c_4 & 0 & -(c_1+c_3) & -c_3 & 0 \end{bmatrix}.$$

Hence

$$w(t) = \begin{bmatrix} c_1 + c_3 t_2 & c_3 & 0 & 0 & 0 & 0 \\ c_2 + c_4 t_2 - (c_1+c_3)t_1 - c_3 t_1 t_2 & c_4 - c_3 t_1 & 0 & -c_1 - c_3 - c_3 t_2 & -c_3 & 0 \end{bmatrix} x_0.$$

A kernel representation of the MPUM, that is, a matrix $\hat{Q}$ such that $w \in \mathcal{B}$ if and only if $\hat{Q}(\partial)w = 0$, is given by

$$\hat{Q} := \begin{bmatrix} s_2^2 & 0 \\ 0 & s_2^2 \\ s_1 & 0 \\ s_2 + 1 & s_1 \end{bmatrix}.$$

This corresponds to the last four rows of Q from (1.8). The first two rows of Q say that the solution does not depend on t_3; they would also be present in $\hat{Q}$ if we had worked over $\mathbb{C}[t_1, t_2, t_3]$. Note that the vector space dimension of the MPUM equals 4, and thus it is strictly smaller than $\delta = 6$. This shows that the "realization" of the MPUM constructed in Theorem 3 is not minimal, in general.

The minimality of MPUM realizations, and the reduction of a given realization to minimality, are discussed below, where we derive a generalization of the Kalman observability decomposition. Note that this is the key ingredient to keeping the computational effort moderate: The construction of A_i, C is quite cheap, since it consists in putting ones and zeros into the correct entries of each A_i, and in putting the given coefficient vectors of p into the correct positions of C, respectively, and thus, it is more or less a matter of bookkeeping. The computational cost becomes relevant, however, when we want to calculate $C \exp(t_1 A_1 + \cdots + t_n A_n)$ for the explicit description of the MPUM trajectories, or if we wish to construct a kernel representation of the MPUM (using the fundamental principle [9] for eliminating the latent variable x). Clearly, these operations are sensitive with respect to the size of the matrices A_i.

For this, let $A_i \in \mathbb{C}^{\delta \times \delta}$, $1 \le i \le n$, be pairwise commuting matrices, and let $C \in \mathbb{C}^{q \times \delta}$. Consider $\mathcal{B} = \{w \in \mathcal{A}^q \mid \exists x \in \mathcal{A}^\delta : \partial_i x = A_i x \text{ for } 1 \le i \le n \text{ and } w = Cx\}$. We call $(A_1, \ldots, A_n, C)$ a *realization* of $\mathcal{B}$ of size δ, and a realization is said to be *minimal* if there exists no realization of $\mathcal{B}$ of strictly smaller size. A realization $(A_1, \ldots, A_n, C)$ is called *observable* if

$$\mathcal{O}(A_1, \ldots, A_n, C) := \bigcap_{\mu \in \mathbb{N}^n} \ker(C A_1^{\mu_1} \cdots A_n^{\mu_n}) = \{0\}.$$

Observability means that x is uniquely determined by w with $w = Cx$ and $\partial_i x = A_i x$ for all i. This follows from $(\partial^\mu w)(0) = C A^\mu x_0$, where $A^\mu := A_1^{\mu_1} \cdots A_n^{\mu_n}$. Thus x_0, and hence x, is uniquely determined by w if and only if the realization is observable.

THEOREM 4. *A realization $(A_1, \ldots, A_n, C)$ of $\mathcal{B}$ is minimal if and only if it is observable. Given $(A_1, \ldots, A_n, C)$, there exists a nonsingular matrix $T \in \mathbb{C}^{\delta \times \delta}$ such that*

$$T A_i T^{-1} = \begin{bmatrix} A_i^1 & 0 \\ * & * \end{bmatrix} \quad \text{and} \quad C T^{-1} = \begin{bmatrix} C^1 & 0 \end{bmatrix},$$

where $(A_1^1, \ldots, A_n^1, C^1)$ is observable, and thus a minimal realization of $\mathcal{B}$.

Proof. Let $(A_1, \dots, A_n, C)$ be given. Let $\mathcal{K} := \mathcal{K}(A_1, \dots, A_n, C)$ be the subspace of $\mathbb{C}^{1\times\delta}$ spanned by the rows of the matrices CA^μ, where $\mu \in \mathbb{N}^n$ and $A^\mu = A_1^{\mu_1} \cdots A_n^{\mu_n}$. Note that it suffices to consider $0 \le \mu_i \le \delta - 1$. Let $r := \dim(\mathcal{K})$ and let $T_1 \in \mathbb{C}^{r\times\delta}$ be a matrix whose rows are a basis of $\mathcal{K}$. Choose T_2 such that $T^T = \left[T_1^T, T_2^T\right]$ is square and nonsingular. Since $\mathcal{K}$ is invariant under right multiplication by any A_i, that is, $x \in \mathcal{K}$ implies $xA_i \in \mathcal{K}$, we have $T_1 A_i = A_i^1 T_1$ for some $A_i^1 \in \mathbb{C}^{r\times r}$, and thus

$$TA_iT^{-1} = \begin{bmatrix} A_i^1 & 0 \\ * & * \end{bmatrix}.$$

Similarly, since the rows of C are contained in $\mathcal{K}$, we have $C = C^1 T_1$ for some $C^1 \in \mathbb{C}^{q\times r}$ and thus $CT^{-1} = [C^1, 0]$. The matrices A_i^1 are pairwise commuting, and $(A_1^1, \dots, A_n^1, C^1)$ is a realization of $\mathcal{B}$ of size $r = \delta - \dim(\mathcal{O}(A_1, \dots, A_n, C))$. Thus, a nonobservable realization can be reduced in size. Moreover, we have $\dim(\mathcal{K}) = \dim(\mathcal{K}(A_1^1, \dots, A_n^1, C^1))$ and thus $\mathcal{O}(A_1^1, \dots, A_n^1, C^1) = \{0\}$.

It remains to be shown that if $(A_1, \dots, A_n, C)$ is not minimal, then it does not satisfy $\mathcal{O}(A_1, \dots, A_n, C) = \{0\}$. Let $(\tilde{A}_1, \dots, \tilde{A}_n, \tilde{C})$ be a realization of $\mathcal{B}$ of strictly smaller size $\tilde{\delta} < \delta$. By considering $(\partial^\mu w)(0)$ for $w \in \mathcal{B}$ and $\mu \in \mathbb{N}^n$, we obtain that for every $x_0 \in \mathbb{C}^\delta$, there exists $\tilde{x}_0 \in \mathbb{C}^{\tilde{\delta}}$ such that $CA^\mu x_0 = \tilde{C}\tilde{A}^\mu \tilde{x}_0$ for all $\mu \in \mathbb{N}^n$. This implies that $\dim(\mathcal{K}) \le \dim(\tilde{\mathcal{K}})$, where $\mathcal{K}$ is as above and $\tilde{\mathcal{K}} = \mathcal{K}(\tilde{A}_1, \dots, \tilde{A}_n, \tilde{C})$, and thus $\dim(\mathcal{O}(A_1, \dots, A_n, C)) = \delta - \dim(\mathcal{K}) \ge \delta - \tilde{\delta} > 0$. ∎

Example. Returning to the example from above, a suitable transformation matrix is given by

$$T = \left[\begin{array}{cccccc} c_1 & c_3 & 0 & 0 & 0 & 0 \\ c_2 & c_4 & 0 & -(c_1+c_3) & -c_3 & 0 \\ c_3 & 0 & 0 & 0 & 0 & 0 \\ c_4 & 0 & 0 & -c_3 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right],$$

where the horizontal line denotes the partition between T_1 and T_2, and where we assume $c_3 \neq 0$. Then the new realization of the MPUM takes the form

$$A_1^1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \end{bmatrix}, \qquad A_2^1 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad C^1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

which is minimal, and happens to coincide with the system that generated (1.9) in the first place.

Finally, it remains to generalize the result of Theorem 3 to the case where we have more than one given polynomial trajectory, and to the case where the data trajectories are polynomial-exponential rather than purely polynomial. We do this in three steps.

First, let $p_1, \dots, p_N \in \mathbb{C}[t_1, \dots, t_n]^q$ be given. Let $\mathcal{B}_l$ be the MPUM of p_l for $l = 1, \dots, N$. Then the MPUM of $D = \{p_1, \dots, p_N\}$ is given by

$$\mathcal{B} = \mathcal{B}_1 + \cdots + \mathcal{B}_N.$$

Moreover, if

$$\mathcal{B}_l = \left\{w \in \mathcal{A}^q \mid \exists x_l \in \mathcal{A}^{\delta_l} : \partial_i x_l = A_{il} x_l \text{ for } 1 \le i \le n \text{ and } w = C_l x_l\right\},$$

then

$$\mathcal{B} = \{w \in \mathcal{A}^q \mid \exists x \in \mathcal{A}^\delta : \partial_i x = A_i x \text{ for } 1 \le i \le n \text{ and } w = Cx\},$$

where $\delta := \delta_1 + \cdots + \delta_N$ and

$$A_i := \operatorname{diag}(A_{i1}, \ldots, A_{iN}) \quad \text{and} \quad C := \begin{bmatrix} C_1 & \ldots & C_N \end{bmatrix}. \tag{2.1}$$

Second, the equivalence (1.2) enables translating these results to the case where $p_1, \ldots, p_N$ are polynomial-exponential rather than purely polynomial. Let

$$\tilde{\mathcal{B}} = \{w \in \mathcal{A}^q \mid \exists x \in \mathcal{A}^\delta : \partial_i x = A_i x \text{ for } 1 \le i \le n \text{ and } w = Cx\}$$

be the MPUM of p. Then

$$\mathcal{B} = \{w \in \mathcal{A}^q \mid \exists x \in \mathcal{A}^\delta : \partial_i x = \lambda_i x + A_i x \text{ for } 1 \le i \le n \text{ and } w = Cx\}$$

is the MPUM of $p \exp_\lambda$.

Finally, let $D = \{p_1 \exp_{\lambda^{(1)}}, \ldots, p_N \exp_{\lambda^{(N)}}\}$ be given. Let

$$\mathcal{B}_l = \{w \in \mathcal{A}^q \mid \exists x_l \in \mathcal{A}^{\delta_l} : \partial_i x_l = \lambda_i^{(l)} x_l + A_{il} x_l \text{ for } 1 \le i \le n \text{ and } w = C_l x_l\}$$

be the MPUM of $p_l \exp_{\lambda^{(l)}}$. Then

$$\mathcal{B} = \{w \in \mathcal{A}^q \mid \exists x \in \mathcal{A}^\delta : \partial_i x = \Lambda_i x + A_i x \text{ for } 1 \le i \le n \text{ and } w = Cx\}$$

is the MPUM of D, where $\delta = \delta_1 + \cdots + \delta_N$,

$$\Lambda_i := \operatorname{diag}\left(\lambda_i^{(1)} I_{\delta_1}, \ldots, \lambda_i^{(N)} I_{\delta_N}\right),$$

and A_i and C are as in (2.1). Using Theorem 4, the resulting realization of the MPUM can be reduced to a minimal one.

Conclusion. The characteristic frequencies of a behavior correspond to its nonzero exponential trajectories. Similarly as with ordinary differential equations, it does not suffice to consider purely exponential solutions, one has to take polynomial-exponential functions into account. Since a fixed characteristic frequency can be shifted to the origin without loss of generality, the problem can be reduced to finding polynomial solutions. We have seen how to decide whether the polynomial solution set of a given system of PDE is finite-dimensional. If yes, we can construct a basis. In the general case, we can only construct a basis of the space of polynomial solutions up to a certain total degree d, and we can make a statement about the dimensions of these spaces as d tends to infinity. Finally, we have constructed the MPUM for the case of a single observed trajectory of polynomial type. This result has been generalized to the case where we have any finite number of polynomial-exponential trajectories. We have shown that the resulting realization of the MPUM is minimal if and only if it is observable, and we have described a method for reducing a given realization to a minimal one, which is analogous to the Kalman observability decomposition.

Acknowledgments. The author would like to thank Prof. Wilhelm Plesken and Prof. Gerhard Pfister for useful suggestions.

REFERENCES

[1] A. C. Antoulas and J. C. Willems, *A behavioral approach to linear exact modeling*, IEEE Trans. Automat. Control, 38 (1993), pp. 1776–1802.

[2] J.-E. Björk, *Rings of Differential Operators*, North-Holland, Amsterdam, 1979.

[3] D. Cox, J. Little, and D. O'Shea, *Using Algebraic Geometry*, Springer, New York, 1998.

[4] D. Eisenbud, *Commutative Algebra with a View Toward Algebraic Geometry*, Springer, New York, 1995.

[5] G.-M. Greuel, G. Pfister, and H. Schönemann, *Singular* 2.0. *A Computer Algebra System for Polynomial Computations*, Centre for Computer Algebra, University of Kaiserslautern, 2001, www.singular.uni-kl.de/.

[6] G.-M. Greuel and G. Pfister, *A Singular Introduction to Commutative Algebra,* Springer, New York, 2002.

[7] L. Hörmander, *Linear Partial Differential Operators*, Springer, New York, 1976.

[8] V. Lomadze and E. Zerz, *Partial differential equations of Krull dimension zero*, in Proceedings of MTNS 2000, Perpignan, France, 2000, www.univ-perp.fr/mtns2000/.

[9] U. Oberst, *Multidimensional constant linear systems*, Acta Appl. Math., 20 (1990), pp. 1–175.

[10] U. Oberst, *Finite dimensional systems of partial differential or difference equations*, Adv. Appl. Math., 17 (1996), pp. 337–356.

[11] H. K. Pillai and S. Shankar, *A behavioral approach to control of distributed systems*, SIAM J. Control Optim., 37 (1999), pp. 388–408.

[12] B. Sturmfels, *Solving Systems of Polynomial Equations,* AMS, Providence, RI, 2002.

[13] J. Wood, U. Oberst, E. Rogers, and D. H. Owens, *A behavioral approach to the pole structure of one-dimensional and multidimensional linear systems*, SIAM J. Control Optim., 38 (2000), pp. 627–661.

[14] J. Wood, E. Rogers, P. Rocha, and D. H. Owens, *Structure indices for multidimensional systems*, IMA J. Math. Control Inform., 17 (2000), pp. 227–256.

[15] E. Zerz, *Poles, polynomial-exponential trajectories, and the MPUM*, in Proceedings of MTNS 2004, Leuven, Belgium, 2004.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 4, pp. 1165–1191

A UNIFIED APPROACH FOR STOCHASTIC AND MEAN SQUARE STABILITY OF CONTINUOUS-TIME LINEAR SYSTEMS WITH MARKOVIAN JUMPING PARAMETERS AND ADDITIVE DISTURBANCES*

MARCELO D. FRAGOSO† AND OSWALDO L. V. COSTA‡

Abstract. Necessary and sufficient conditions for stochastic stability (SS) and mean square stability (MSS) of continuous-time linear systems subject to Markovian jumps in the parameters and additive disturbances are established. We consider two scenarios regarding the additive disturbances: one in which the system is driven by a Wiener process, and one characterized by functions in $L_2^m(\Omega, \mathcal{F}, \mathbb{P})$, which is the usual scenario for the H_∞ approach. The Markov process is assumed to take values in an infinite countable set $\mathcal{S}$. It is shown that SS is equivalent to the spectrum of an augmented matrix lying in the open left half plane, to the existence of a solution for a certain Lyapunov equation, and implies (is equivalent for $\mathcal{S}$ finite) asymptotic wide sense stationarity (AWSS). It is also shown that SS is equivalent to the state $x(t)$ belonging to $L_2^n(\Omega, \mathcal{F}, \mathbb{P})$ whenever the disturbances are in $L_2^m(\Omega, \mathcal{F}, \mathbb{P})$. For the case in which $\mathcal{S}$ is finite, SS and MSS are equivalent, and the Lyapunov equation can be written down in two equivalent forms with each one providing an easier-to-check sufficient condition.

Key words. stochastic stability, mean square stability, jump parameter, continuous-time linear systems, Markov chain

AMS subject classifications. 93E15, 93C05, 93C60, 60J75, 60J27

DOI. 10.1137/S0363012903434753

1. Introduction.

In recent years, there has been a steadily rising level of activity with linear systems which are subject to abrupt changes in their structures. Most of the literature considers the case in which the abrupt changes are modeled by a Markov chain, namely, linear systems with Markovian jump parameters (LSMJP). These changes arise quite often in practice and may be due, for instance, to component and/or interconnection failures, *inter alia*. This is to be found, for instance, in robotic manipulator systems, aircraft control systems, large scale flexible structures for space stations (such as antenna, solar arrays, etc.), and flexible manufacturing systems, on which an actuator or a sensor failure is a quite common occurrence. Without any intention of being exhaustive here, we mention [4], [5], [6], [7], [8], [11], [12], [13], [14], [15], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [34], [35], [37], [38], [39], [40], [41], [42], [43], [46], [49], and the references therein, as a small sample of works dealing with different aspects of LSMJP problems.

Although mean square and almost sure stability for linear stochastic systems with random parameter perturbation have a long and successful history, the interest in

*Received by the editors September 15, 2003; accepted for publication (in revised form) March 24, 2005; published electronically October 7, 2005. This work was supported in part by the Brazilian National Research Council-CNPq under grants 472920/03-0, 302587/2004-7, and 304866/03-2, by the Research Council of the State of São Paulo-FAPESP under grant 03/06736-7, by PRONEX under grant 015/98, and by IM-AGIMB. A preliminary version of this manuscript was presented as a regular paper at the 39th IEEE Conference on Decision and Control (CDC00).

http://www.siam.org/journals/sicon/44-4/43475.html

†National Laboratory for Scientific Computing - LNCC/CNPq, Av. Getulio Vargas 333, Petrópolis, Rio de Janeiro, CEP 25651-070, Brazil (frag@lncc.br).

‡Departamento de Engenharia de Telecomunicações e Controle - Escola Politécnica da Universidade de São Paulo, CEP 05508, São Paulo, S.P., Brazil (oswaldo@lac.usp.br).

getting easy-to-check conditions for special structures continues unabated. Among the techniques used, we mention the one which relies on martingale methods and the idea of projecting onto the unit sphere or the projective space to compute the asymptotic exponential growth rate (the Lyapunov exponent) and the one using the Lyapunov function approach. For the class $\dot{x}(t) = (A + P(t))x(t)$, where $P(t)$ is a matrix whose elements belong to a particular class of stochastic processes (either ergodic processes or a Markov jump, stationary, etc.), or $\dot{x}(t) = A(p(t))x(t)$, with $p(t)$ a white noise or an ergodic processes, the readers are referred, for instance, to [3], [16], and [31] (and the many references therein). Stability questions for LSMJP (the case in which $p(t)$ is a Markov chain) have given rise to a significant number of papers dealing with this subject (see, for instance, [4], [8], [11], [12], [13], [18], [19], [22], [23], [24], [26], [27], [34], [35], [37], [38], [39], [40], [43], [46], [49], and the references therein). See also [14] for a historical account on the discrete-time case. It is important to stress that most of these papers consider the case in which the Markov process takes values in a finite state space. This assumption has important consequences and, in particular, it is a well-known fact that mean square stability is equivalent to stochastic stability. For the case in which the state space of the Markov chain is *infinite countable*, these concepts are no longer equivalent (for examples, see [12] or [14] for the discrete-time case and [26] for the continuous-time scenario). In addition, they consider the real case.

In [37] and [40], necessary and sufficient conditions for mean square stability (MSS) were obtained for the homogeneous continuous-time noise-free real case. A common feature in these papers is that the MSS criteria is expressed as the maximal real part of a certain matrix being less than zero, and in fact this number was shown to be the mean square Lyapunov exponent of the system. In [37], the sample path Lyapunov exponent λ and the p-moment Lyapunov exponent $g(p)$ for a real homogeneous LSMJP is studied for the case of a stationary, ergodic Markov chain. Using extensions of results in [1], in conjunction with ideas developed in [2], relations between λ and $g(p)$ are derived. In addition, under a certain assumption which guarantees that the operation mode has the so-called property S, an exact expression for the mean square Lyapunov exponent is obtained. The result for $g(2)$ is given in terms of the maximal real part of a matrix representation of a certain linear operator (the matrix is not exhibited explicitly). The result closest to ours is that obtained in [40], using a technique completely different from that used here.

In [35] necessary and sufficient conditions for MSS of the discrete-time noise-free case were obtained in terms of the existence of a solution of a Lyapunov equation. The continuous-time counterpart of this result is derived in [24] (see also [43]), including a study on the relationship among various moment and sample path stability. In the time varying case with state dependent additive Wiener disturbance, exponential stability in mean square (ESMS) for the zero solution of the system are studied in [19] and [39]. In [19], after deriving a certain Itô-type formula, the authors show equivalence results for ESMS, including a Lyapunov-type equation. After discussing some properties of the solution for the nonlinear general case, it is given in [39] a sufficient condition for the pth moment exponential stability based on the Lyapunov-type result and sufficient condition for pth moment exponential stability to imply almost sure exponential stability, using essentially the idea of Lyapunov exponent. The author uses also the so-called M-matrices theory (in fact, nonsingular M-matrices) to derive sufficient conditions for the pth moment and almost sure exponential stability. In fact it has to be checked if, for $0 < p < 2$, a certain matrix $\mathcal{A}(p)$ is a nonsingular M-matrix. All these papers consider the real case with the Markov process taking values in a finite set.

Almost sure stability for Markovian jump linear systems (MJLS) is examined, for instance, in [11], [18], [22], [23], [38], [40], and [41] (see also [39] for almost sure exponential stability). A historical account on earlier works can be found in [22]. Some important issues regarding MSS for the case in which the state space of the Markov chain is infinite countable can be found in [12], [25], [26], and [27].

Asymptotic stability in distribution for LSMJP has been studied in [4] and [49]. This seems an adequate notion to adopt when, for instance, degeneracy of the noise is allowed. In [4], asymptotic stability in distribution for the semilinear stochastic differential equation with Markov jump parameters is discussed. Using the interesting concept of asymptotic flatness in the pth mean, the authors derive sufficient conditions for stability in distribution by assuming MSS of the autonomous systems and the rather interesting condition $dk_0^2 < 1/\Omega_p$, with d denoting the system dimension, k_0 being a certain Lipschitz condition for the diffusion matrix σ, and $\Omega_p = \max\{\Omega_{p_1}, \ldots, \Omega_{p_N}\}$, with Ω_{p_i} denoting the largest eigenvalue of the matrix P_i, which satisfies a certain Lyapunov equation. In [49], asymptotic stability in the distribution for a class of nonlinear diffusion with a Markov jump is derived. Sufficient conditions are established, provided certain properties of the solution are satisfied (called properties (P1) and (P2)). In addition, sufficient conditions for the mentioned properties in terms of Lyapunov functions are also exhibited. In order to make their result more applicable, the authors also derive sufficient conditions in terms of the M-matrix.

This paper deals with stochastic and mean square stability for continuous-time LSMJP with the Markov process taking values in an *infinite countable set* $\mathcal{S}$ in a unified way. This makes an important difference with respect to the previous papers and, in particular, we have that, unlike the finite-dimensional case, stochastic stability (SS) and MSS are not equivalent (see [26]). In this case, it is SS, as defined in the paper, that is equivalent to the spectrum of a certain matrix lying in the open left half plane. This equivalence does not hold anymore for MSS. In the *finite case* our paper provides, based on *an easily computable criterion* (the spectrum of a matrix), a unified result for SS, MSS, and asymptotic wide sense stationarity (AWSS) in Theorem 5.6 (necessary and sufficient conditions for the three scenarios considered in the paper). It extends and encompasses the results in [28].

A well-known feature of the class of LSMJP is that the state $x(t)$ is not Markov. The primary hindrance here then is related to the kind of analytical tools for handling the problem of getting easy-to-check stability criteria. A distinctive feature of our approach is that it can be used, also in a unified way, for continuous and discrete-time case, since it does not rely on a specific Itô type of result. It relies rather on the fact that the augmented state $(x(t), \theta(t))$ is Markov, and, therefore, an operator theoretical approach is possible, provided we can connect, in a suitable way, $x(t)$ with measurable functions of $(x(t), \theta(t))$. Briefly, the idea behind our approach is as follows: we know that $x(t)$ in (3.2) (also (3.3) and (3.4)) is not Markov, but $(x(t), \theta(t))$ is Markov. Therefore, if we work with the measurable function of $(x(t), \theta(t))$, i.e., $f(x(t), \theta(t))$, we can then think in tools such as infinitesimal generator, which allows us to think about the dynamics. The following question then arises: is it possible to bypass the non-Markovian property of $x(t)$ using $f(x(t), \theta(t))$? It comes up, bearing in mind that $x(t) = \sum_{i \in \mathcal{S}} x(t) 1_{\{\theta_t = i\}}$ and $x(t)x(t)^* = \sum_{i \in \mathcal{S}} x(t)x(t)^* 1_{\{\theta_t = i\}}$, that a natural answer to the above question is to work with $x(t)1_{\{\theta_t = i\}}$ and $x(t)x(t)^* 1_{\{\theta_t = i\}}$, since in this case we can think about differential equations for the moments (first and second moment) in terms of the operators defined in section 4. This approach was used in [11], which is, to some extent, the discrete-time version of this paper, and we believe that it may be useful elsewhere. In addition, since operator theory is the technical

underpinning of the paper, it has the advantage of being suitable for treating the general complex and infinite-dimensional (due to the fact that $\mathcal{S}$ is infinite countable) setting.

We deal with two scenarios regarding additive disturbances: the one in which the disturbances are characterized via a Wiener process, and the one characterized by any function in $L_2^m(\Omega, \mathcal{F}, \mathbb{P})$. It is shown that SS is equivalent to the real part of the spectrum of an augmented matrix being less than zero, or to the existence of a solution of a Lyapunov equation, or that the state $x(t) \in L_2^n(\Omega, \mathcal{F}, \mathbb{P})$ for $L_2^m(\Omega, \mathcal{F}, \mathbb{P})$-disturbances, and implies (is equivalent for $\mathcal{S}$ finite) AWSS stability. It is also shown that these equivalence relations hold even for the case in which only real initial conditions are considered. The first criterion (based on the eigenvalues) translates clearly the intuitive idea that unstable modes of operation do not necessarily compromise the global stability of the system. In fact it can be shown that the stability of all modes of operation is neither necessary nor sufficient for the global stability of the system (see, e.g., [34]). The eigenvalue criterion shows clearly the connection between SS and the probability of visits to the unstable modes. A cursory examination of the augmented matrix reveals that a balance between the modes and the transition probability matrix is essential for SS. For the case of one mode operation (no jumps in the parameters) our criteria reconcile to a well-known stability results for continuous-time linear systems. When the state space of the Markov process $\mathcal{S}$ is finite, MSS and SS are equivalent, and the Lyapunov equation can be written in two equivalent forms with each of these forms providing easier-to-check sufficient conditions. These results, we believe, provide a flexible theory which gives a rather complete and unified picture of SS and MSS for LSMJP.

Finally, it is noteworthy here that, besides the interest in its own right, matrices with complex coefficients are also interesting from an application point of view. As pointed out in [32] and [7], in several engineering applications, as in communication application of signals systems, whirling shafts, vibrational systems, etc., complex coefficients come into play. In [32] some systems with complex coefficients that naturally arise in mechanics and signal processing are presented, motivating the study of equations with complex coefficients and illustrating more general situations such as satellite and cosmic vehicles control. We refer to these papers and the references therein for further examples and results on systems with complex coefficients.

An outline of the content of this paper is as follows. In section 2 we provide the bare essentials of notational conventions and some preliminaries. The model and problem statement are described in section 3. Stochastic and mean square stability for the homogeneous case (including Lyapunov equations) is treated in section 4. For the case in which the $\mathcal{S}$ is finite it is proved that the Lyapunov equation can be written down in two equivalent forms with each one providing an easier-to-check sufficient condition. Section 5 accounts the case with additive disturbances. It embodies two kinds of analysis: one for the $L_2^m(\Omega, \mathcal{F}, \mathbb{P})$-disturbances and one for the linear jump-diffusion case. For all the cases, it is shown that SS is equivalent to the spectrum of an augmented matrix lying in the open left half plane, and to the existence of a solution for a certain Lyapunov equation, and implies (is equivalent for $\mathcal{S}$ finite) AWSS. In the appendix we present the proof of some auxiliary results.

2. Notation and preliminaries. For $\mathbb{X}$ and $\mathbb{Y}$ complex Banach spaces we set $\mathbb{B}(\mathbb{X}, \mathbb{Y})$ for the Banach space of all bounded linear operators of $\mathbb{X}$ into $\mathbb{Y}$, with the uniform induced norm represented by $\|.\|$. For simplicity we shall set $\mathbb{B}(\mathbb{X}) := \mathbb{B}(\mathbb{X}, \mathbb{X})$. For $T \in \mathbb{B}(\mathbb{X})$ we denote by $\sigma(T)$ the spectrum of T. If $\mathbb{X}$ is a Hilbert space, then

$\langle .;. \rangle$ will stand for the inner product, and for $\mathcal{T} \in \mathbb{B}(\mathbb{X})$, $\mathcal{T}^\star$ will indicate the adjoint operator of $\mathcal{T}$. As usual, $\mathcal{T} \geq 0$ ($\mathcal{T} > 0$) will mean that the operator $\mathcal{T} \in \mathbb{B}(\mathbb{X})$ is positive semidefinite (positive definite), respectively. In particular, we shall denote by $\mathbb{C}^n$ the n-dimensional complex Euclidean spaces and by $\mathbb{B}(\mathbb{C}^n, \mathbb{C}^m)$ the normed bounded linear space of all $m \times n$ complex matrices with $\mathbb{B}(\mathbb{C}^n) := \mathbb{B}(\mathbb{C}^n, \mathbb{C}^n)$ and $\mathbb{B}(\mathbb{C}^n)^+ := \{L \in \mathbb{B}(\mathbb{C}^n); L = L^* \geq 0\}$. In this case, the superscripts $-$, $\prime$, and $*$ will denote complex conjugate, transpose, and conjugate transpose, respectively. Either the uniform induced norm in $\mathbb{B}(\mathbb{C}^n)$ or the standard Euclidean norm in $\mathbb{C}^n$ is represented by $\|.\|$. We also use $\mathbb{R}^+$ to denote the interval $[0, \infty)$. Unless otherwise stated, we define $\mathcal{S} := \{1, 2, \dots\}$. For $D_i \in \mathbb{B}(\mathbb{C}^n)$, $i \in \mathcal{S}$, $\mathrm{diag}(D_i)$ is an infinity square matrix where the matrices D_i are put together corner-to-corner diagonally with all other entries being zero. We refer to I_ℓ as the $\ell \times \ell$ identity matrix and to $L_2^n(\mathbb{R}^+)$ as the space of functions $f : [0, \infty) \to \mathbb{C}^n$ such that each component $f^i(.)$ is in the standard L_2 $(\mathbb{R}^+)$ space of Lebesgue square integrable functions. Similarly, $L_2^n(\Omega, \mathcal{F}, \mathbb{P})$ is the space of square integrable stochastic process. $E[.]$ denotes the mathematical expectation, and we set for a second order random variable $x(t)$, $\|x(t)\|_2^2 := E[\|x(t)\|^2]$, and for $x = \{x(t); t \in \mathbb{R}^+\} \in L_2^n(\Omega, \mathcal{F}, \mathbb{P})$, $\|x\|_2^2 := \int_0^\infty E[\|x(t)\|^2]dt$. We denote by $\mathbb{R}_e\{\lambda\}$ the real part of a complex number λ. We recall that (see Theorem 1.1.2 of [48]) a linear operator $\mathcal{L}$ on a Banach space $\mathbb{X}$ is the infinitesimal generator of a uniformly continuous semigroup $\phi_{\mathcal{L}}(t)$ if and only if $\mathcal{L} \in \mathbb{B}(\mathbb{X})$. Furthermore, in this case $\phi_{\mathcal{L}}(t) = e^{\mathcal{L}t}(.)$, which is defined as

$$e^{\mathcal{L}t}(.) := \sum_{n=0}^{\infty} \frac{1}{n!} \mathcal{L}^n(.) t^n \in \mathbb{B}(\mathbb{X}). \tag{2.1}$$

We set

$$\mathbb{R}_e\{\lambda(\mathcal{L})\} := \sup\{\mathbb{R}_e\{\lambda\}; \lambda \in \sigma(\mathcal{L})\}.$$

We recall that the trace operator $\mathrm{tr}(.) : \mathbb{B}(\mathbb{C}^n) \to \mathbb{C}$ is a linear functional with the following properties: (i) $\mathrm{tr}(KL) = \mathrm{tr}(LK)$; (ii) for any $M, P \in \mathbb{B}(\mathbb{C}^n)^+$ with $P > 0$, we have

$$\Big(\min_{i=1,\dots,n} \lambda_i(P)\Big)\mathrm{tr}(M) \leq \mathrm{tr}(MP) \leq \Big(\max_{i=1,\dots,n} \lambda_i(P)\Big)\mathrm{tr}(M). \tag{2.2}$$

We shall denote by ℓ_1^n, ℓ_2^n, $\ell_{\sup}^n$, respectively, the sets made up of all infinite sequences of complex vectors $x = (x_1, x_2, \dots)$, $x_i \in \mathbb{C}^n$, such that $\sum_{i=1}^\infty \|x_i\| < \infty$, $\sum_{i=1}^\infty \|x_i\|^2 < \infty$, and $\sup\{\|x_i\|, i = 1, 2, \dots\} < \infty$. We denote the usual norms $\|.\|_1$, $\|.\|_2$, $\|.\|_{\sup}$ in ℓ_1^n, ℓ_2^n, and $\ell_{\sup}^n$, respectively, by $\|x\|_1 = \sum_{i=1}^\infty \|x_i\|$, $\|x\|_2 = (\sum_{i=1}^\infty \|x_i\|^2)^{\frac{1}{2}}$, and $\|x\|_{\sup} = \sup\{\|x_i\|, i = 1, 2, \dots\}$. We have that $(\ell_1^n, \|.\|_1)$, $(\ell_2^n, \|.\|_2)$, $(\ell_{\sup}^n, \|.\|_{\sup})$ are Banach spaces and, in fact, $(\ell_2^n, \|.\|_2)$ is a Hilbert space equipped with the usual inner product $\langle x; y \rangle := \sum_{i=1}^\infty x_i^* y_i$ (see [47]).

Set $\mathbb{H}_1^{n,m}$ (respectively, $\mathbb{H}_2^{n,m}$, $\mathbb{H}_{\sup}^{n,m}$), the linear space made up of all sequence of complex matrices $V = (V_1, V_2, \dots)$ with $V_i \in \mathbb{B}(\mathbb{C}^n, \mathbb{C}^m)$ such that $\sum_{i=1}^\infty \|V_i\| < \infty$ ($\sum_{i=1}^\infty \mathrm{tr}(V_i^* V_i) < \infty$, $\sup\{\|V_i\|; i = 1, 2, \dots\} < \infty$). For simplicity, set $\mathbb{H}_\iota^n := \mathbb{H}_\iota^{n,n}$, $\iota = 1, 2, \sup$. For $V = (V_1, \dots) \in \mathbb{H}_\iota^{n,m}$, we consider the following norms $\|.\|_\iota$ in $\mathbb{H}_\iota^{n,m}$,

$\iota = 1, 2, \sup$:

$$\|V\|_1 := \sum_{i=1}^{\infty} \|V_i\|, \tag{2.3}$$

$$\|V\|_2 := \left(\sum_{i=1}^{\infty} \operatorname{tr}(V_i^* V_i) \right)^{1/2}, \tag{2.4}$$

$$\|V\|_{\sup} := \sup\{\|V_i\|; i = 1, 2, \dots\}. \tag{2.5}$$

It is easy to verify that $(\mathbb{H}_\iota^{n,m}, \|.\|_\iota)$ and $(\ell_\iota^{nm}, \|.\|_\iota)$, $\iota = 1, 2, \sup$, are uniformly homeomorphic. Therefore, $(\mathbb{H}_\iota^{n,m}, \|.\|_\iota)$ are Banach spaces and, in fact, $(\mathbb{H}^{n,m}, \|.\|_2)$ is a Hilbert space, with the inner product given, for $S = (S_1, \dots)$ and $V = (V_1, \dots)$ in $\mathbb{H}_2^{n,m}$, by

$$\langle V; S \rangle = \sum_{i \in \mathcal{S}} \operatorname{tr}(V_i^* S_i). \tag{2.6}$$

For $V = (V_1, \dots) \in \mathbb{H}_\iota^{n,m}$ we shall write $V^* = (V_1^*, \dots) \in \mathbb{H}_\iota^{m,n}$ and say that $V \in \mathbb{H}_\iota^n$ is Hermitian if $V = V^*$. We define $\mathbb{H}_\iota^{n*} := \{V = (V_1, \dots) \in \mathbb{H}_\iota^n; V_i = V_i^*, i = 1, \dots\}$ and $\mathbb{H}_\iota^{n+} := \{V = (V_1, \dots) \in \mathbb{H}_\iota^{n*}; V_i \geq 0, i = 1, \dots\}$ and shall write, for $V = (V_1, \dots) \in \mathbb{H}_\iota^n$ and $S = (S_1, \dots) \in \mathbb{H}_\iota^n$, that $V \geq S$ if $V - S = (V_1 - S_1, \dots) \in \mathbb{H}_\iota^{n+}$, and that $V > S$ if $V_i - S_i > 0$ for $i \in \mathcal{S}$. Finally we say that $V = (V_1, \dots) \in \widetilde{\mathbb{H}}_{\sup}^{n+}$ if $V \in \mathbb{H}_{\sup}^{n+}$ and for some $\alpha > 0$, $V_i \geq \alpha I_n$ for each $i \in \mathcal{S}$.

For the case in which $\mathcal{S} = \{1, \dots, N\}$ norms (2.3), (2.4), (2.5) are equivalent, since the underline spaces are finite-dimensional. For notational simplicity we just write in this case $\mathbb{H}^n$ instead of $\mathbb{H}_\iota^n$.

Define now the operators φ and $\hat{\varphi}$ in the following way: for $\iota = 1, 2, \sup$, $V = (V_1, \dots) \in \mathbb{H}_\iota^{n,m}$, with $V_i = (v_{i1} \dots v_{in}) \in \mathbb{B}(\mathbb{C}^n, \mathbb{C}^m)$, $v_{ij} \in \mathbb{C}^m$,

$$\varphi(V_i) := \begin{bmatrix} v_{i1} \\ \cdot \\ \cdot \\ \cdot \\ v_{in} \end{bmatrix} \in \mathbb{C}^{mn} \qquad \text{and} \qquad \hat{\varphi}(V) := \begin{bmatrix} \varphi(V_1) \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} \in \ell_\iota^{mn}.$$

Furthermore, for $v = (v_1, \dots) \in \ell_\iota^{nm}$, $v_i \in \mathbb{C}^{mn}$, we define $V_i \in \mathbb{B}(\mathbb{C}^n, \mathbb{C}^m)$ such that $V_i := \varphi^{-1}(v_i)$, and $V = (V_1, \dots) \in \mathbb{H}_\iota^{n,m}$, as

$$V := \hat{\varphi}^{-1} \left(\begin{bmatrix} v_1 \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} \right) := \left[\hat{\varphi}_1^{-1}(v), \dots \right] = \left[\varphi^{-1}(v_1), \dots \right].$$

REMARK 2.1. *Notice that the mapping φ stacks up the column of the matrix from left to right and makes a long vector out of the matrix. Furthermore, it can be shown, through the mapping $\hat{\varphi}$, that $(\mathbb{H}_2^{n,m}, \|.\|_2)$ and $(\ell_2^{mn}, \|.\|_2)$ are isometrically isomorphic spaces (if $V \in \mathbb{H}_2^{n,m}$, then $\|V\|_2 = \|\hat{\varphi}(V)\|_2$).*

With the Kronecker product $L \otimes K \in \mathbb{B}(\mathbb{C}^{sn}, \mathbb{C}^{rm})$ defined in the usual way for any $L \in \mathbb{B}(\mathbb{C}^s, \mathbb{C}^r)$ and $K \in \mathbb{B}(\mathbb{C}^n, \mathbb{C}^m)$, the following properties hold (see, e.g., [9]):

$$\text{(i)} \quad (L \otimes K)^* = L^* \otimes K^* \qquad \text{and} \qquad \text{(ii)} \quad \varphi(LKH) = (H' \otimes L)\varphi(K). \tag{2.7}$$

Recall also that for $L \in \mathbb{B}(\mathbb{C}^n)$ and $K \in \mathbb{B}(\mathbb{C}^m)$ the Kronecker sum is defined as

$$L \oplus K := L \otimes I_m + I_n \otimes K \in \mathbb{B}(\mathbb{C}^{nm}).$$

The proof of the next result can be found in [13].

LEMMA 2.1. *For any $H \in \mathbb{H}_\iota^n$ there exist $H^i \in \mathbb{H}_\iota^{n+}$, $i = 1, 2, 3, 4$, such that*

$$H = (H^1 - H^2) + \sqrt{-1}(H^3 - H^4). \tag{2.8}$$

The next result follows immediately from section 4.2 of [48].

PROPOSITION 2.2. *Let $\mathbb{X}$ be a Banach space. Let $\mathcal{L} \in \mathbb{B}(\mathbb{X})$ be the infinitesimal generator of the uniformly continuous semigroup $\phi_{\mathcal{L}}(t) : \mathbb{X} \to \mathbb{X}$. The following assertions are equivalent:*

(i) $\mathbb{R}_e\{\lambda(\mathcal{L})\} < 0$.

(ii) *There are constants $k > 0, b > 0$ such that*

$$\|\phi_{\mathcal{L}}(t)\| \leq k e^{-bt} \quad \text{for all} \quad t \geq 0.$$

(iii) $\int_0^\infty \|\phi_{\mathcal{L}}(t)x\| dt < \infty$ *for every $x \in \mathbb{X}$.*

If X is finite-dimensional, then (iii) *can be replaced by*

(iii′) $\|\phi_{\mathcal{L}}(t)x\| \to 0$ *as $t \to \infty$ for every $x \in \mathbb{X}$.*

We recall from (2.1) that under the conditions of Proposition 2.2, $\phi_{\mathcal{L}}(t) = e^{\mathcal{L}t}$. From the decomposition of square matrices into positive semidefinite matrices as in Lemma 2.1 and Proposition 2.2 we have the following result.

PROPOSITION 2.3. *Let $\mathcal{L} \in \mathbb{B}(\mathbb{H}_1^n)$ be the infinitesimal generator of the uniformly continuous semigroup $e^{\mathcal{L}t} : \mathbb{H}_1^n \to \mathbb{H}_1^n$. The following assertions are equivalent.*

(i) $\mathbb{R}_e\{\lambda(\mathcal{L})\} < 0$.

(ii) *There are constants $k > 0, b > 0$ such that*

$$\|e^{\mathcal{L}t}\| \leq k e^{-bt} \quad \text{for all} \quad t \geq 0.$$

(iii) $\int_0^\infty \|e^{\mathcal{L}t}(V)\| dt < \infty$ *for every $V \in \mathbb{H}_1^{n+}$.*

For the finite-dimensional case (iii) *can be replaced by*

(iii′) $\|e^{\mathcal{L}t}(V)\| \to 0$ *as $t \to \infty$ for every $V \in \mathbb{H}^{n+}$.*

Proof. See the appendix for the proof. □

The next proposition, adapted from [48], will be useful in deriving some stability results.

PROPOSITION 2.4. *Let $\mathcal{A} \in \mathbb{B}(\ell_1^{n^2})$ and $\{f(t); t \in \mathbb{R}^+\}$ be a continuous function in $\ell_1^{n^2}$ such that $\lim_{t\to\infty} f(t) = f_0$. Consider*

$$\dot{y}(t) = \mathcal{A}y(t) + f(t). \tag{2.9}$$

If $\mathbb{R}_e\{\lambda(\mathcal{A})\} < 0$, then for any initial condition $y(0) = y_0 \in \ell_1^{n^2}$,

$$\lim_{t\to\infty} y(t) = -\mathcal{A}^{-1} f_0.$$

3. The models and problem statement. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space equipped with a filtration $\{\mathcal{F}_t, t \in \mathbb{R}^+\}$ satisfying the usual hypothesis, that is, a right continuous filtration augmented by all null sets in the $\mathbb{P}$-completion of $\mathcal{F}$, and carrying the following statistically mutually independent objects:

(0.1) An m-dimensional Wiener process $W = \{(w(t), \mathcal{F}_t), t \in \mathbb{R}^+\}$ with an incremental covariance operator $R dt$.

(0.2) A homogeneous Markov process $\theta = \{(\theta_t, \mathcal{F}_t), t \in \mathbb{R}^+\}$ with right continuous trajectories and taking values on the set $\mathcal{S}$. We assume also that

$$P(\theta_{t+h} = j | \theta_t = i) = \begin{cases} \lambda_{ij} h + o(h), & i \neq j, \\ 1 + \lambda_{ii} h + o(h), & i = j, \end{cases} \tag{3.1}$$

where $[(\lambda_{ij})]$ is the stationary infinite-dimensional transition rate matrix of $\{\theta\}$ with $0 \leq \lambda_{ij}$, $i \neq j$, and $0 \leq \lambda_i := -\lambda_{ii} = \sum_{\{j:\ j \neq i\}} \lambda_{ij} \leq \varrho$ for all $i \in \mathcal{S}$, i.e., the process is supposed to be conservative (see, e.g., [36]). The notation $o(h)$ denotes an infinitesimal of higher order than h, i.e., $\lim_{h \downarrow 0} \frac{o(h)}{h} = 0$. We define $p_{ij}(t) := \mathbb{P}(\theta_{t+s} = j | \theta_s = i)$, $i, j \in \mathcal{S}$, and denote $p_i(t) := \mathbb{P}(\theta_t = i)$ for any $i \in \mathcal{S}$. Notice that, in this setting, $P_t := (p_1(t), \dots)'$ satisfies the Kolmogorov forward differential equation $dP_t/dt = \Lambda P_t$; $P_0 = P, t \in \mathbb{R}^+$, where $\Lambda := [(\lambda_{ij})]'$. In addition, we assume that $\{(\theta_t, \mathcal{F}_t), t \in \mathbb{R}^+\}$ has initial distribution $\{v(i); i \in \mathcal{S}\}$.

(0.3) A random variable $x_0 \colon \Omega \to \mathbb{C}^n$ with $E[\|x_0\|^2] < \infty$.

We assume that the filtration $\mathcal{F}_t$ contains the filtration generated by (0.1)–(0.3), i.e., the natural filtration of $\{(x_0, w(s), \theta(s))\}$.

We deal with three types of linear systems with Markovian jump parameters. First, in order to bring to bear some basic results in its more general form and put SS in a unified basis, we consider the homogeneous system

$$\dot{x}(t) = A(\theta_t) x(t), \quad x(0) = x_0, \quad \theta_0 = v, \quad t \in \mathbb{R}^+. \tag{3.2}$$

We next consider the class of dynamical systems modeled by the following stochastic differential equation:

$$\dot{x}(t) = A(\theta_t) x(t) + B(\theta_t) w(t), \quad x(0) = x_0, \quad \theta_0 = v, \quad t \in \mathbb{R}^+, \tag{3.3}$$

where the additive disturbance, $w(.)$, is modeled by

(0.1b) $\{w(t); t \in \mathbb{R}^+\}$ is any $L_2^m(\Omega, \mathcal{F}, \mathbb{P})$-function,

which is the usual scenario for the H_∞ approach. In addition, in order to reflect the types of additive disturbances encountered in the specialized literature in detail, we shall consider also the class of dynamical systems modeled by the following Itô stochastic differential equation:

$$dx(t) = A(\theta_t) x(t) dt + B(\theta_t) dw(t), \quad x(0) = x_0, \quad \theta_0 = v, \quad t \in \mathbb{R}^+, \tag{3.4}$$

where we require, in addition to (0.1)–(0.3), that

(0.4) $\theta = \{(\theta_t, \mathcal{F}_t), t \in \mathbb{R}^+\}$ is an irreducible positive Harris recurrent Markov process with initial distribution $\{v(i); i \in \mathcal{S}\}$. We recall that, in this setting, it is a standard result of the Markov chain theory (see [42]) that there exist limiting probabilities $\{\pi_i;\ i \in \mathcal{S}\}$ which do not depend on the initial distribution with $\{\sum_{i=1}^{\infty} \pi_i = 1\}$ and satisfy

$$\lim_{t \to \infty} \sup_{j \in \mathcal{S}} \{ \mid p_j(t) - \pi_j \mid \} = 0. \tag{3.5}$$

Furthermore, $A(.)$ and $B(.)$ are such that $A(\theta_t) = A_j$ and $B(\theta_t) = B_j$ for $\theta_t = j$, $j \in \mathcal{S}$, with A_j, B_j, $j \in \mathcal{S}$, being constant matrices in $\mathbb{B}(\mathbb{C}^n)$ and $\mathbb{B}(\mathbb{C}^m, \mathbb{C}^n)$, respectively. It

is assumed that $A := (A_1, \dots) \in \mathbb{H}_{\sup}^n$ and $B := (B_1, \dots) \in \mathbb{H}_{\sup}^{m,n}$ for the case (0.1b), and $B \in \mathbb{H}_2^{m,n}$ for the case (0.1). In addition, define for $t \in \mathbb{R}^+$

$$q(t) := E(x(t)) \ \in \ \mathbb{C}^n, \tag{3.6}$$

$$\mathcal{Q}(\tau, t) := E(x(t+\tau)x(t)^*) \ \in \ \mathbb{B}(\mathbb{C}^n), \tag{3.7}$$

$$Q(t) := \mathcal{Q}(0, t) \ \in \ \mathbb{B}(\mathbb{C}^n)^+ \tag{3.8}$$

and

$$q_i(t) := E(x(t)1_{\{\theta_t = i\}}) \ \in \ \mathbb{C}^n, \tag{3.9}$$

$$Q_i(t) := E(x(t)x(t)^* 1_{\{\theta_t = i\}}) \ \in \ \mathbb{B}(\mathbb{C}^n)^+, \tag{3.10}$$

$$\mathcal{Q}_i(s, t) := E(x(t+s)x(t)^* 1_{\{\theta_{t+s} = i\}}) \ \in \ \mathbb{B}(\mathbb{C}^n), \tag{3.11}$$

where $1_{\{.\}}$ stands for the Dirac measure. Set also

$$\hat{q}(t) := \begin{bmatrix} q_1(t) \\ \vdots \end{bmatrix},$$

$$\hat{Q}(t) := (Q_1(t), \dots),$$

$$\hat{\mathcal{Q}}(s, t) := (\mathcal{Q}_1(s, t), \dots).$$

Since

$$\|\hat{Q}(t)\|_1 = \sum_{i=1}^{\infty} \|Q_i(t)\| \le \sum_{i=1}^{\infty} E[\|x(t)\|^2 1_{\{\theta_t = i\}}] = E[\|x(t)\|^2], \tag{3.12}$$

it follows that $\hat{Q}(t) \in \mathbb{H}_1^{n+}$. Similarly we have that $\hat{q}(t) \in \ell_1^n$ and $\hat{\mathcal{Q}}(s, t) \in \mathbb{H}_1^n$.

Preserving the terminology, often used in the literature for MJLS, we define the following.

DEFINITION 3.1. *A linear system with a Markovian jump parameter is stochastically stable if for arbitrary initial conditions x_0 and an arbitrary initial distribution v we have*

$$\int_0^{\infty} \|x(t)\|^2 dt < \infty.$$

DEFINITION 3.2. *A linear system with a Markovian jump parameter is mean square stable if there exist $q \in \mathbb{C}^n$ and $Q \in \mathbb{B}(\mathbb{C}^n)^+$ such that for arbitrary initial conditions x_0 and arbitrary initial distribution v we have*

(a) $\|q(t) - q\| \to 0$ *as* $t \to \infty$,

(b) $\|Q(t) - Q\| \to 0$ *as* $t \to \infty$.

DEFINITION 3.3. *A linear system with a Markovian jump parameter is asymptotically wide sense stationary if there exist $q \in \mathbb{C}^n$ and $\mathcal{Q}(\tau) \in \mathbb{B}(\mathbb{C}^n)^+$ such that for arbitrary initial conditions x_0 and an arbitrary initial distribution v we have*

(a) $\|q(t) - q\| \to 0$ *as* $t \to \infty$,

(b) $\|\mathcal{Q}(\tau, t) - \mathcal{Q}(\tau)\| \to 0$ *as* $t \to \infty$.

REMARK 3.1. *In the case of systems* (3.3), (3.4), *we assume also that q, Q, and $\mathcal{Q}(\tau)$ are independent of $w(t)$.*

4. SS for the homogeneous case. In this section necessary and sufficient conditions for SS of the homogeneous case are established. It is required that either the sum of the real part of all the elements in the spectrum of an augmented infinite dimensional matrix is less than zero or that there exists a unique solution of a Lyapunov equation. Moreover, it is shown that for real matrices $A(i)$, the system is stochastically stable for the complex state space if and only if the system is stochastically stable for the real state space. For the case in which $\mathcal{S}$ is finite, more specific results can be obtained and, in particular, it is proved that the Lyapunov equation can be written down in two equivalent forms with each one providing an easier-to-check sufficient condition. In this case it is well known that SS and MSS are equivalent concepts (see, for instance, [24]).

We begin by providing part of the basic common notational machinery, definitions, and some important basic results, necessary for the analysis of the situations described in section 3 which resort here, in part, to the methods of operator theory.

4.1. Main operators and auxiliary results. We consider first the homogeneous equation (3.2), restated here for convenience:

$$dx(t) = A(\theta_t)x(t)dt, \quad x(t_0) = x_0, \quad \theta_{t_0} = v, \quad t \in \mathbb{R}^+. \tag{4.1}$$

In addition, let T_n denote the nth jump time of the Markov process $\{\theta_t; t \geq 0\}$ and define $\Upsilon := \{\omega \in \Omega; T_n(\omega) \to \infty\}$. For each realization of the Markov process $\{\theta_t; t \geq 0\}$ in Υ we have that $\{A(\theta_t); t \geq 0\}$ are matrix-valued functions on $\mathbb{R}^+$ of the class PC (piecewise continuous, see [10, p. 411]) and, therefore, according to [10, p. 11], there exists aunique continuous solution $\Phi(., t_0)$ from $\mathbb{R}^+$ to $\mathbb{B}(\mathbb{C}^n)$ of the homogeneous linear matrix differential equation

$$\frac{\partial \Phi(t, t_0)}{\partial t} = A(\theta_t)\Phi(t, t_0), \qquad \Phi(t_0, t_0) = I$$

for almost all $t \in \mathbb{R}^+$. Moreover, the solution of (4.1) is given by

$$x(t) = \Phi(t, t_0)x_0. \tag{4.2}$$

In what follows we shall be using the following notation:

$$\begin{array}{l} F := \Lambda' \otimes I_n + \operatorname{diag}(A_i); \quad V := \Lambda' \otimes I_{n^2}; \quad G := \operatorname{diag}(I_n \oplus A_i); \\ H := \operatorname{diag}(\bar{A}_i \oplus A_i); \quad \mathcal{A} := V + H; \quad \mathcal{B} := V + G. \end{array} \tag{4.3}$$

We define also the following linear operators:

$$\begin{array}{ll} \mathcal{E}(.) = (\mathcal{E}_1(.), \dots); & \mathcal{F}(.) = (\mathcal{F}_1(.), \dots); \\ \mathcal{L}(.) = (\mathcal{L}_1(.), \dots); & \mathcal{T}(.) = (\mathcal{T}_1(.), \dots); \end{array} \tag{4.4}$$

where for $P = (P_1, \dots) \in \mathbb{H}_1^n$, $V = (V_1, \dots) \in \mathbb{H}_{\sup}^n$, and $i \in \mathcal{S}$,

$$\begin{aligned} \mathcal{E}_i(P) &:= \sum_{j=1}^{\infty} \lambda_{ji} P_j, \\ \mathcal{F}_i(P) &:= A_i P_i + \sum_{j=1}^{\infty} \lambda_{ji} P_j, \\ \mathcal{L}_i(P) &:= A_i P_i + P_i A_i^* + \sum_{j=1}^{\infty} \lambda_{ji} P_j, \\ \mathcal{T}_i(V) &:= A_i^* V_i + V_i A_i + \sum_{j=1}^{\infty} \lambda_{ji} V_j. \end{aligned} \tag{4.5}$$

We now have the following auxiliary results.

LEMMA 4.1. *For $\mathcal{E}$, $\mathcal{L}$, $\mathcal{T}$, and $\mathcal{F}$ defined as in* (4.4) *and* (4.5), *we have that*

(a) $\mathcal{E} \in \mathbb{B}(\mathbb{H}_1^n)$, $\mathcal{L} \in \mathbb{B}(\mathbb{H}_1^n)$, *and* $\mathcal{F} \in \mathbb{B}(\mathbb{H}_1^n)$;

(b) $\mathcal{T} \in \mathbb{B}(\mathbb{H}_{\sup}^n)$;

(c) *for any* $Q = (Q_1, \dots) \in \mathbb{H}_1^n$, $\mathcal{L}(Q)^* = \mathcal{L}(Q^*)$;

(d) $Q \in \mathbb{H}_1^{n+}$ *implies* $e^{\mathcal{L}t}(Q) \in \mathbb{H}_1^{n+}$ *for any* $t \in \mathbb{R}^+$.

Items (c) *and* (d) *also hold replacing* $\mathbb{H}_1^n$ *by* $\mathbb{H}_{\sup}^n$ *and* $\mathcal{L}$ *by* $\mathcal{T}$.

Proof. See the appendix for the proof. □

LEMMA 4.2. *Let f(t) be $\mathcal{F}_t$-measurable and assume that* $E(f(t)1_{\{\theta_t=i\}}) := f_i(t)$ *exists. Then*

$$E(f(t)d(1_{\{\theta_t=i\}})) = \sum_{j=1}^{\infty} \lambda_{ji} f_j(t)dt + o(dt).$$

Proof. See the appendix for the proof. □

PROPOSITION 4.3. *There exist constants $c_1 > 0$ and $c_2 > 0$ such that for any $H = (H_1, \dots) \in \mathbb{H}_1^n$, $c_1\|\hat{\varphi}(H)\|_1 \le \|H\|_1 \le c_2\|\hat{\varphi}(H)\|_1$.*

Proof. See the appendix for the proof. □

The next proposition provides differential equations to compute the first and second moments of the state variable of (3.2).

PROPOSITION 4.4. *For $t \in \mathbb{R}^+$, we have for* (3.2) *that*

(a) $\dot{\hat{q}}(t) = F\hat{q}(t)$,

(b) $\dot{\hat{Q}}(t) = \mathcal{L}(\hat{Q}(t))$.

Proof. The proof relies essentially on (3.2), Lemma 4.2, and (3.9)–(3.10). See [28] for details. □

The following result is germane to our approach.

PROPOSITION 4.5. *For F, $\mathcal{A}$, $\mathcal{B}$ defined as in* (4.3) *we have that* $F \in \mathbb{B}(\ell_1^n)$, $\mathcal{A} \in \mathbb{B}(\ell_1^{n^2})$, $\mathcal{B} \in \mathbb{B}(\ell_1^{n^2})$, *and* $\mathcal{A}^* \in \mathbb{B}(\ell_{\sup}^{n^2})$. *Moreover, for any* $Q \in \mathbb{H}_1^n$, $\mathcal{Q} \in \mathbb{H}_1^n$, *and* $V \in \mathbb{H}_{\sup}^n$ *we have*

(a) $\hat{\varphi}(\mathcal{L}(Q)) = \mathcal{A}\hat{\varphi}(Q)$;

(b) $\hat{\varphi}(\mathcal{T}(V)) = \mathcal{A}^*\hat{\varphi}(V)$;

(c) $\hat{\varphi}(\mathcal{F}(\mathcal{Q})) = \mathcal{B}\hat{\varphi}(\mathcal{Q})$.

Proof. See the appendix for the proof. □

We need also the following auxiliary results.

LEMMA 4.6. *Let $\mathcal{A} \in \mathbb{B}(\ell_1^{n^2})$, as defined in* (4.3), *and consider the homogeneous system $\dot{y}(t) = \mathcal{A}y(t)$, $t \in \mathbb{R}^+$, with the initial condition $y(0) = \hat{\varphi}(Q)$, $Q \in \mathbb{H}_1^{n+}$. Then*

(a) $y(t) = \hat{\varphi}(e^{\mathcal{L}t}(Q))$,

(b) $\hat{\varphi}^{-1}(y(t)) \in \mathbb{H}_1^{n+}$ *and consequently* ${\hat{\varphi}_j}^{-1}(y(t)) \in \mathbb{B}(\mathbb{C}^n)^+, j \in \mathcal{S}$.

The result also holds replacing $\ell_1^{n^2}$, $\mathbb{H}_1^{n+}$, $\mathcal{A}$, and $\mathcal{L}$ by $\ell_{\sup}^{n^2}$, $\mathbb{H}_{\sup}^{n+}$, $\mathcal{A}^*$, and $\mathcal{T}$, respectively.

Proof. See the appendix for the proof. □

LEMMA 4.7. *For any second order random variable z taking values in $\mathbb{C}^n$ and $t, \tau \in \mathbb{R}^+$, $t \ge \tau$,*

$$\|\Phi(t,\tau)z\|_2^2 \le n\|e^{\mathcal{L}(t-\tau)}\|\|z\|_2^2. \tag{4.6}$$

Proof. See the appendix for the proof. □

PROPOSITION 4.8. *If $\mathbb{R}_e\{\lambda(\mathcal{A})\} < 0$, then $\mathbb{R}_e\{\lambda(F)\} < 0$.*

Proof. See the appendix for the proof. □

REMARK 4.1. *It is not difficult to see that* $\mathbb{R}_e\{\lambda(F)\} < 0$ *does not imply* $\mathbb{R}_e\{\lambda(\mathcal{A})\} < 0$. *Indeed, consider, for instance,* $n = 1, \mathcal{S} = \{1, 2\}$, $\lambda_{11} = \lambda_{22} = -1$, $A_1 = \frac{1}{2}$, *and* $A_2 = -5$. *It is then a straightforward exercise to show that*

$$\lambda_1(F) = \frac{-6.5 - \sqrt{34.25}}{2} < 0, \quad \lambda_2(F) = \frac{-6.5 + \sqrt{34.25}}{2} < 0$$

and that

$$\lambda_1(\mathcal{A}) = \frac{-11 + \sqrt{125}}{2} > 0, \quad \lambda_2(\mathcal{A}) = \frac{-11 - \sqrt{125}}{2} < 0.$$

4.2. SS results. Our main SS results will follow from the next propositions. The next result shows a SS result in the spirit of the classical linear case. It shows that SS is equivalent to the spectrum of an augmented matrix, $\mathcal{A}$, lying in the open left half plane.

PROPOSITION 4.9. *The following affirmatives are equivalent:*

(i) *System* (3.2) *is stochastically stable according to Definition* 3.1.
(ii) $\mathbb{R}_e\{\lambda(\mathcal{L})\} < 0$.
(iii) $\mathbb{R}_e\{\lambda(\mathcal{A})\} < 0$.

Proof. From Proposition 4.4(b) we have that $\dot{\hat{Q}}(t) = \mathcal{L}(\hat{Q}(t))$. Therefore, from (3.12),

$$\int_0^\infty \|e^{\mathcal{L}t}(\hat{Q}(0))\|_1 dt = \int_0^\infty \|\hat{Q}(t)\|_1 dt \le \int_0^\infty E(\|x(t)\|^2) dt \tag{4.7}$$

for any initial condition x_0 and initial distribution v. Suppose (i) holds, so that $\int_0^\infty E(\|x(t)\|^2)dt < \infty$ for any initial condition x_0 and initial distribution v. Consider any $H = (H_1, \dots) \in \mathbb{H}_1^{n+}$. By taking an initial condition x_0 and v such that $Q_i(0) = H_i$, it follows from (4.7) that for any $H = (H_1, \dots) \in \mathbb{H}_1^{n+}$, $\int_0^\infty \|e^{\mathcal{L}t}(H)\|_1 dt < \infty$, which implies, from Proposition 2.3, that (ii) holds. From Propositions 2.2 and 4.4(b) it is immediate that if (ii) holds, then (i) holds. The equivalence between (ii) and (iii) follows from Propositions 4.3, 4.5, and 2.2, Lemma 4.6, and the fact that for any $y \in \ell_1^{n^2}$,

$$\int_0^\infty \|e^{\mathcal{A}t}y\|_1 dt = \int_0^\infty \|\hat{\varphi}(e^{\mathcal{L}t}(\hat{\varphi}^{-1}(y)))\|_1 dt \le \frac{1}{c_1}\int_0^\infty \|e^{\mathcal{L}t}(\hat{\varphi}^{-1}(y)\|_1 dt,$$

and similarly, for any $H \in \mathbb{H}_1^n$, $\int_0^\infty \|e^{\mathcal{L}t}(H)\|_1 dt \le c_2 \int_0^\infty \|e^{\mathcal{A}t}(\hat{\varphi}(H))\|_1 dt$. □

PROPOSITION 4.10. *If system* (3.2) *is stochastically stable according to Definition* 3.1, *then for every* $S = (S_1, \dots) \in \widetilde{\mathbb{H}}_{\sup}^{n+}$, *there exists a unique* $G = (G_1, \dots) \in \widetilde{\mathbb{H}}_{\sup}^{n+}$ *such that*

$$\mathcal{T}(G) + S = 0. \tag{4.8}$$

Proof. See Theorem 8 in [27]. □

PROPOSITION 4.11. *If there exists* $G = (G_1, \dots) \in \widetilde{\mathbb{H}}_{\sup}^{n+}$ *such that* (4.8) *is satisfied for some* $S = (S_1, \dots) \in \widetilde{\mathbb{H}}_{\sup}^{n+}$, *then system* (3.2) *is stochastically stable according to Definition* 3.1.

Proof. See Theorem 7 in [27]. □

PROPOSITION 4.12. *If* $\mathbb{R}_e\{\lambda(\mathcal{A})\} < 0$, *then system* (3.2) *is mean square stable according to Definition* 3.2 *with* $q = 0$ *and* $Q = 0$.

Proof. First notice that since $\mathbb{R}_e\{\lambda(\mathcal{A})\} < 0$, we have from Proposition 4.8 that $\mathbb{R}_e\{\lambda(F)\} < 0$. Thus, from Propositions 4.4(a) and 2.2 it follows that $\hat{q}(t) \to 0$ as $t \to \infty$, and since $q(t) = \sum_{i=1}^{\infty} q_i(t)$ we have that $q(t) \to 0$ as $t \to \infty$. Now, from Propositions 4.4(b) and 4.5(b) we have $\hat{\varphi}(\hat{Q}(t)) = \mathcal{A}\hat{\varphi}(\hat{Q}(t))$. If we now define $y(t) := \hat{\varphi}(\hat{Q}(t))$, we get $\dot{y}(t) = \mathcal{A}y(t)$. Now by Proposition 2.2, it follows that $\hat{\varphi}(\hat{Q}(t)) \to 0$ as $t \to \infty$. By noting that $Q(t) = \sum_{i=1}^{\infty} Q_i(t)$ it follows that $Q(t) \to 0$ as $t \to \infty$. ∎

In order to show that our setup encompasses the real case, we consider now the situation in which the initial condition x_0 in (4.1) is real, as in Definition 2.1 of [24] stated below.

DEFINITION 4.13 (see [24]). *System* (3.2) *is stochastically stable for the real state space case if for any* $x_0 \in \mathbb{R}^n$ *and any probability distribution* v *for* θ_0,

$$\int_0^{\infty} E(\|x(t)\|^2)dt < \infty. \tag{4.9}$$

Notice that in Definition 4.13, we consider only real initial conditions for x_0. We next show that in fact Definitions 3.1 and 4.13 are equivalent.

PROPOSITION 4.14. *System* (3.2) *is stochastically stable for the real state space case as in Definition* 4.13 *if and only if it is stochastically stable according to Definition* 3.1.

Proof. Clearly if system (3.2) is stochastically stable according to Definition 3.1, it is stochastically stable for the real state space case as in Definition 4.13. Suppose now that system (3.2) is stochastically stable for the real state space case as in Definition 4.13. For any $H \in \mathbb{H}^{n+}$, define $\mathbb{I}(H) \in \mathbb{H}_1^{n+}$ as follows:

$$\mathbb{I}(H) := (\|H_1\|I_n, \|H_2\|I_n, \dots).$$

Clearly $H \le \mathbb{I}(H)$. We consider real initial conditions x_0 and v such that $Q_i(0) = \|H_i\|I_n$ so that $\hat{Q}(0) = \mathbb{I}(H)$ (for instance, x_0 and θ_0 are independent of $E(x_0x_0^*) = \|H\|_1 I_n$ and $v(i) = \frac{\|H_i\|}{\|H\|_1}$). Let $x_H(t)$ denote the trajectory for this initial condition. From Lemma 4.1, we have that for all $t \in \mathbb{R}^+$,

$$\mathbf{0} \le e^{\mathcal{L}t}(H) \le e^{\mathcal{L}t}(\mathbb{I}(H)). \tag{4.10}$$

From (3.12) and (4.10),

$$\int_0^{\infty} \|e^{\mathcal{L}t}(H)\|_1 dt \le \int_0^{\infty} \|e^{\mathcal{L}t}(\mathbb{I}(H))\|_1 dt \le \int_0^{\infty} E(\|x_H(t)\|^2)dt < \infty$$

for all $H \in \mathbb{H}_1^{n+}$, and thus from Propositions 2.3 and 4.9, system (3.2) is stochastically stable according to Definition 3.1. ∎

4.3. The case in which $\mathcal{S}$ is finite. We consider in this subsection the case in which $\mathcal{S} = \{1, \dots, N\}$ so that more specific results can be obtained. The first result reads as follows.

LEMMA 4.15. *$\mathcal{T}^\star = \mathcal{L}$, i.e., $\mathcal{T}$ is the adjoint operator of $\mathcal{L}$ in the Hilbert space* $(\mathbb{H}^n, \|.\|_2)$.

Proof. This can be proved easily. ∎

REMARK 4.2. *From Proposition* 4.5 *and Lemma* 4.15 *it is easy to verify that the following assertions are equivalent:*

(i) $\mathbb{R}_e\{\lambda(\mathcal{A})\} < 0$.

(ii) $\mathbb{R}_e\{\lambda(\mathcal{L})\} < 0$.
(iii) $\mathbb{R}_e\{\lambda(\mathcal{A}^*)\} < 0$.
(iv) $\mathbb{R}_e\{\lambda(\mathcal{T})\} < 0$.

The next result shows that MSS implies that $\mathbb{R}_e\{\lambda(\mathcal{A})\} < 0$.

PROPOSITION 4.16. *If system* (3.2) *is mean square stable according to Definition* 3.2(b) *with* $Q = 0$, *then* $\mathbb{R}_e\{\lambda(\mathcal{A})\} < 0$.

Proof. We have that $Q(t) = \sum_{i=1}^{N} Q_i(t)$ and from Proposition 4.4(b) and Proposition 4.5(a) $\hat{\varphi}(\dot{\hat{Q}}(t)) = \mathcal{A}\hat{\varphi}(\hat{Q}(t))$. Therefore, $\hat{\varphi}(\hat{Q}(t)) = e^{\mathcal{A}t}\hat{\varphi}(\hat{Q}(0))$ and

$$Q(t) = \sum_{i=1}^{N} \hat{\varphi}_i^{-1}(e^{\mathcal{A}t}\hat{\varphi}(\hat{Q}(0))).$$

Now, by hypothesis, $Q(t) \to 0$ as $t \to \infty$ for any $Q(0) = E(x_0 x_0^*)$ and the initial distribution $v(\imath) = 1$. Furthermore, from Lemma 4.6(b) we have that $\hat{\varphi}_j^{-1}(e^{\mathcal{A}t}\hat{\varphi}(\hat{Q}(0))) \geq 0$ for all $j = 1, \ldots, N$, which shows that $e^{\mathcal{A}t}\hat{\varphi}(\hat{Q}(0)) \to 0$ as $t \to \infty$. Moreover, bearing in mind that for any $H_i \in \mathbb{H}^{n+}$, $i = 1, \ldots, N$, we can show by straightforward arguments that there are second order random variables x_i such that $E(x_i x_i^*) = H_i$, we have that $e^{\mathcal{A}t}\hat{\varphi}(H_1, \ldots, H_N) \to 0$ as $t \to \infty$ and from Proposition 2.3 (for finite-dimensional spaces), $\mathbb{R}_e\{\lambda(\mathcal{A})\} < 0$. □

The next result captures an MSS result in the spirit of the classical linear case. It shows that MSS stability is equivalent to the spectrum of an augmented matrix, $\mathcal{A}$, lying in the open left half plane.

THEOREM 4.17. *System* (3.2) *is mean square stable according to Definition* 3.2 *if and only if* $\mathbb{R}_e\{\lambda(\mathcal{A})\} < 0$.

Proof. The proof follows from Propositions 4.12 and 4.16. □

The next examples, borrowed from [41], illustrate how sometimes the switching between operation modes can play tricks with our intuition. They show that system (3.2) carries a great deal of subtlety which distinguishes it from the linear case and provides us with a very rich structure. The need to combine the transition probability of the Markov chain with the eigenvalues of the matrices A_i in order to get an adequate criterion for MSS is portrayed by matrix $\mathcal{A}$.

EXAMPLE 4.1 (each mode is unstable but overall system is stable). *Consider the model*

$$A_1 = \begin{bmatrix} \frac{1}{2} & -1 \\ 0 & -2 \end{bmatrix}, \quad A_2 = \begin{bmatrix} -2 & -1 \\ 0 & \frac{1}{2} \end{bmatrix}, \quad \Lambda = \begin{bmatrix} -\beta & \beta \\ \beta & -\beta \end{bmatrix}.$$

Therefore, each mode is unstable. We have that depending on the value of β *the overall system will be mean square stable. For* $0 < \beta \leq 1.33$ *we have that for some* i, $\mathbb{R}_e\{\lambda_i(\mathcal{A})\} > 0$, *and thus the system is not mean square stable. But for* $\beta > 1.34$, *we get that* $\mathbb{R}_e\{\lambda(\mathcal{A})\} < 0$, *and the system is mean square stable. This shows that as the number of jumps per unit time increases, the effect of switching between the unstable modes makes the overall system mean square stable.*

EXAMPLE 4.2 (each mode is stable but overall system is unstable). *Consider now the model*

$$A_1 = \begin{bmatrix} -1 & 10 \\ 0 & -1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} -1 & 0 \\ 10 & -1 \end{bmatrix}, \quad \Lambda = \begin{bmatrix} -\beta & \beta \\ \beta & -\beta \end{bmatrix}.$$

Therefore, each mode is stable. We have that depending on the value of β *the overall system will be mean square unstable in the mean square sense. For* $0 < \beta \leq 0.04$ *we*

have that $\mathbb{R}_e\{\lambda(\mathcal{A})\} < 0$*, and the system is mean square stable. But for* $\beta > 0.05$*, we get that for some* i*,* $\mathbb{R}_e\{\lambda_i(\mathcal{A})\} > 0$*, and thus the system is not mean square stable. This shows that as the number of jumps per unit time increases, the effect of switching between the stable modes makes the overall system mean square unstable.*

More specific results can also be obtained from Proposition 4.10, as follows.

PROPOSITION 4.18. *If* $\mathbb{R}_e\{\lambda(\mathcal{A})\} < 0$*, then for every* $S = (S_1, \ldots, S_N) \in \mathbb{H}^n$*, there exists a unique* $G = (G_1, \ldots, G_N) \in \mathbb{H}^n$ *such that*

$$\mathcal{L}(G) + S = 0. \tag{4.11}$$

Moreover,

(a) $G_i = -\hat{\varphi}_i^{-1}(\mathcal{A}^{-1}\hat{\varphi}(S))$*;*
(b) $\hat{\varphi}(G) = \int_0^\infty e^{\mathcal{A}t}\hat{\varphi}(S)dt$*;*
(c) $S \in \mathbb{H}^*$ *if and only if* $G \in \mathbb{H}^*$*;*
(d) $S \in \mathbb{H}^+$ *implies* $G \in \mathbb{H}^+$*;*
(e) $S \in \tilde{\mathbb{H}}^+$ *implies* $G \in \tilde{\mathbb{H}}^+$*.*

These results also hold replacing $\mathcal{L}$ *by* $\mathcal{T}$*.*

Proof. (a) From Proposition 4.5(a), we have that $\hat{\varphi}(\mathcal{L}(G_1, \ldots, G_N)) = \mathcal{A}\hat{\varphi}(G_1, \ldots, G_N)$, and, therefore, (4.11) is equivalent to

$$\mathcal{A}\hat{\varphi}(G_1, \ldots, G_N) = -\hat{\varphi}(S_1, \ldots, S_N). \tag{4.12}$$

The expression for G_i follows immediately from the assumption on $\mathcal{A}$ and the definition of $\hat{\varphi}$. Assume now that there exists $\bar{G} = (\bar{G}_1, \ldots, \bar{G}_N) \in \mathbb{H}^n$ such that $\mathcal{L}_i(\bar{G}) + S_i = 0$. Then, bearing in mind (4.12), we have $\mathcal{A}\hat{\varphi}(\bar{G} - G) = 0$, or $\hat{\varphi}(\bar{G} - G) = 0$, which implies that $\bar{G} - G = 0$ and the uniqueness follows.

(b) Item (b) follows from (4.12) bearing in mind that $\mathbb{R}_e\{\lambda(\mathcal{A})\} < 0$, Proposition 2.2, and that $\mathcal{A}\phi_{\mathcal{A}}(t) = d/dt(\phi_{\mathcal{A}}(t))$.

(c) From (4.8) and Lemma 4.1(c) we have that $\mathcal{L}(G^*) + S^* = 0$, and the result follows from the fact that $\mathcal{L}(G^* - G) + (S^* - S) = 0$.

(d,e) Items (d) and (e) are a consequence of $e^{\mathcal{A}t}\hat{\varphi}(Q) = \hat{\varphi}(e^{\mathcal{L}t}(Q))$, bearing in mind Lemma 4.1. □

We next present equivalent forms of the Lyapunov equation and the Lyapunov inequality. The next result shows that the dual representation ($\mathcal{L}$ instead of $\mathcal{T}$) of (4.8) also implies MSS.

THEOREM 4.19. *The following assertions are equivalent to MSS of system* (3.2)*:*

(a) $\mathbb{R}_e\{\lambda(\mathcal{A})\} < 0$*.*
(b) *For some* $G_j > 0 \in \mathbb{B}(\mathbb{C}^n)$*,* $j \in \mathcal{S}$*, we have* $\mathcal{L}_i(G) < 0,\ i \in \mathcal{S}$*.*
(c) *For any* $S_i > 0 \in \mathbb{B}(\mathbb{C}^n)$*,* $i \in \mathcal{S}$*, there is a unique* $G = (G_1, \ldots, G_N)$*,* $G_i > 0 \in \mathbb{B}(\mathbb{C}^n)$*,* $i \in \mathcal{S}$*, such that* $\mathcal{L}_i(G) + S_i = 0,\ i \in \mathcal{S}$*. Moreover,*

$$G_i = \hat{\varphi}_i^{-1}(-\mathcal{A}^{-1}\hat{\varphi}(S_1, \ldots, S_N)), \quad i \in \mathcal{S}.$$

Proof. Clearly (c) implies (b). Suppose that (b) holds. We consider the homogeneous system

$$\dot{y}(t) = \mathcal{A}^* y(t), \quad t \in \mathbb{R}^+, \qquad y(0) \in \hat{\varphi}(\mathbb{H}^{n+}), \tag{4.13}$$

where

$$\hat{\varphi}(\mathbb{H}^{n+}) = \{y \in \mathbb{C}^{Nn^2}; y = \hat{\varphi}(Q), Q \in \mathbb{H}^{n+}\}.$$

From Proposition 4.5, we have that

$$\hat{\varphi}_j^{-1}(\dot{y}(t)) = \mathcal{T}_j(\hat{\varphi}_1^{-1}(y(t)), \ldots, \hat{\varphi}_N^{-1}(y(t))) \ , \quad j \in \mathcal{S} \tag{4.14}$$

with $\hat{\varphi}_j^{-1}(y(0)) \in \mathbb{B}(\mathbb{C}^n)^+$. It follows from Lemma 4.6 that $\hat{\varphi}_j^{-1}(y(t)) \in \mathbb{B}(\mathbb{C}^n)^+$ for all $j \in \mathcal{S}$ and all $t \in \mathbb{R}^+$, and thus $y(t) \in \hat{\varphi}(\mathbb{H}^{n+}),\ t \in \mathbb{R}^+$. Define now the function $\phi : \hat{\varphi}(\mathbb{H}^{n+}) \to \mathbb{R}$ as

$$\phi_j(y) := \mathrm{tr}(\hat{\varphi}_j^{-1}(y)G_j) = \mathrm{tr}(G_j^{1/2}\hat{\varphi}_j^{-1}(y)G_j^{1/2}) \geq 0, \quad j \in \mathcal{S},$$
$$\phi(y) := \sum_{j=1}^{N} \phi_j(y) \geq 0.$$

In order to prove that ϕ is a Lyapunov function for the system (4.13) we do need to show that

(i) $\phi(y) \to \infty$ whenever $\|y\| \to \infty$ and $y \in \hat{\varphi}(\mathbb{H}^{n+})$;
(ii) $\phi(0) = 0$;
(iii) $\phi(y) > 0$ for all $y \neq 0 \in \hat{\varphi}(\mathbb{H}^{n+})$;
(iv) ϕ is continuous; and
(v) $\dot{\phi}(y(t)) < 0$ whenever $y(t) \neq 0 \in \hat{\varphi}(\mathbb{H}^{n+})$.

Now, for $y \in \hat{\varphi}(\mathbb{H}^{n+})$ let $\lambda_{ij}(y) \geq 0$ denote the ith eigenvalue of $\hat{\varphi}_j^{-1}(y)$ and $\lambda_i(G_j) > 0$ the ith eigenvalue of G_j. Define

$$c_o := \min_{1 \leq i \leq n \ 1 \leq j \leq N} \ \lambda_i(G_j) > 0$$

(since $G_j > 0$) and

$$c_1 := \max_{1 \leq i \leq n \ 1 \leq j \leq N} \lambda_i(G_j) > 0.$$

From (2.2(ii)) we have that

$$c_0 \left(\sum_{j=1}^{N} \sum_{i=1}^{n} \lambda_{ji}(y) \right) \leq \phi(y) \leq c_1 \left(\sum_{j=1}^{N} \sum_{i=1}^{n} \lambda_{ji}(y) \right). \tag{4.15}$$

Note that

$$\|y\|^2 = \mathrm{tr}(yy^*) = \sum_{j=1}^{N} \mathrm{tr}((\hat{\varphi}_j^{-1}(y))^2) = \sum_{j=1}^{N} \sum_{i=1}^{n} (\lambda_{ji}(y))^2,$$

and bearing in mind the positiveness of $\lambda_{ji}(y)$ we get that $\|y\| \to \infty$ if and only if

$$\sum_{j=1}^{N} \sum_{i=1}^{n} (\lambda_{ji}(y)) \to \infty,$$

and $y = 0$ if and only if $\lambda_{ji} = 0$, $i = 1, \ldots, n$, $j \in \mathcal{S}$. Thus, from these results and (4.15) we get (i)–(iii). Since continuity of ϕ is easily verified, it only remains to show (v). Now, from the definition of ϕ, (4.14), and Lemmas 4.1 and 4.15, we have

$$\begin{aligned}
\dot{\phi}(y(t)) &= \sum_{j=1}^{N} \dot{\phi}_j(y(t)) = \sum_{j=1}^{N} \mathrm{tr}\left(\hat{\varphi}_j^{-1}(\dot{y}(t))G_j\right) \\
&= \sum_{j=1}^{N} \mathrm{tr}\big(\mathcal{T}_j(\hat{\varphi}_1^{-1}(y(t)), \ldots, \hat{\varphi}_N^{-1}(y(t)))G_j\big) \\
&= \langle \mathcal{T}(\hat{\varphi}_1^{-1}(y(t)), \ldots, \hat{\varphi}_N^{-1}(y(t)))^*; G\rangle = \langle \mathcal{T}\big((\hat{\varphi}^{-1}(y(t)))^*\big); G\rangle \\
&= \langle \hat{\varphi}^{-1}(y(t))^*; \mathcal{L}(G)\rangle < 0
\end{aligned}$$

whenever $y(t) \neq 0 \in \hat{\varphi}(\mathbb{H}^{n+})$. Therefore, we have shown that (4.13) is asymptotically stable (cf. [33]) and thus $\|\exp(\mathcal{A}^* t)y\| \to 0$ as $t \to \infty$ for all $y \in \hat{\varphi}(\mathbb{H}^{n+})$ which yields from Proposition 2.3 that $\mathbb{R}_e(\lambda(\mathcal{A}^*)) = \mathbb{R}_e(\lambda(\mathcal{A})) < 0$.

Finally, from Proposition 4.18 we have that (a) implies (c) and from Theorem 4.17 that (a) is equivalent to MSS. □

REMARK 4.3. *The above theorem also holds if we replace $\mathcal{L}$ by $\mathcal{T}$.*

REMARK 4.4. *Note that for the case in which the coefficients are all real, the Lyapunov operator $\mathcal{L}$ works on a Hilbert space of dimension $\frac{Nn(n+1)}{2}$ rather than Nn^2, the dimension of the matrix $\mathcal{A}$. This information can then be used to write the Lyapunov operator $\mathcal{L}$ as a square matrix of dimension $\frac{Nn(n+1)}{2}$. Once this is done, it would be more advantageous to check if $\mathbb{R}_e\{\lambda(\mathcal{L})\} < 0$ by looking at the eigenvalues of this reduced order matrix.*

We show now that from Theorem 4.19 we can derive some *easier-to-check conditions* for MSS of (3.2).

COROLLARY 4.20. *Conditions* (i) *and* (ii) *below are equivalent:*

(i) $\exists\ \alpha_j > 0$, $j \in \mathcal{S}$, *such that* $\alpha_i \lambda_{max}(A_i + A_i^*) + \sum_{i=1}^N \lambda_{ij}\alpha_j < 0$;
(ii) $\exists\ \alpha_j > 0$, $j \in \mathcal{S}$, *such that* $\alpha_i \lambda_{max}(A_i + A_i^*) + \sum_{i=1}^N \lambda_{ji}\alpha_j < 0$;

where $\lambda_{max}(\mathcal{T}) := \max\{\lambda : \lambda$ is an eigenvalue of the operator $\mathcal{T}\}$. Moreover, if one of the above conditions is satisfied, then system (3.2) *is mean square stable.*

Proof. Consider the homogeneous scalar system

$$\dot{\tilde{x}}(t) = \tilde{a}(\theta_t)\tilde{x}(t), \quad t \in \mathbb{R}^+, \tag{4.16}$$

where $\tilde{a}_i := \frac{1}{2}\lambda_{max}(A_i + A_i^*)$, $i \in \mathcal{S}$. Then by applying Theorem 4.19 (bearing in mind Remark 4.3) to system (4.16) we obtain that conditions (i) and (ii) are equivalent. Suppose now that condition (i) is satisfied and set $G_j = \alpha_j I_n > 0$, $j \in \mathcal{S}$. Since

$$\begin{aligned} A_i^* G_i + G_i A_i + \sum_{i=1}^N \lambda_{ij} G_j &= \alpha_i (A_i + A_i^*) + \sum_{i=1}^N \lambda_{ij}\alpha_j I_n \\ &\leq \left(\alpha_i \lambda_{max}(A_i + A_i^*) + \sum_{i=1}^N \lambda_{ij}\alpha_j \right) I_n \ < \ 0, \end{aligned}$$

we get from Theorem 4.19(a) that system (3.2) is mean square stable. □

COROLLARY 4.21. *If for some real number $\delta_i > 0$, $i \in \mathcal{S}$, one of the following conditions is satisfied:*

(1) $\lambda_{max}[A_i + A_i^* + \frac{1}{\delta_i}(\sum_{\{j:\ j \neq i\}} \lambda_{ij}\delta_j) I_n] < -\lambda_{ii}$,
(2) $\lambda_{max}[A_i + A_i^* + \frac{1}{\delta_i}(\sum_{\{j:\ j \neq i\}} \lambda_{ji}\delta_j) I_n] < -\lambda_{ii}$,

then system (3.2) *is mean square stable. Moreover, these conditions are weaker than those in Corollary* 4.20.

Proof. Suppose that condition (1) is satisfied and set $G_i = \delta_i I_n > 0$, $j \in \mathcal{S}$. Then,

following the arguments of the previous corollary we have

$$A_i^*G_i + G_iA_i + \sum_{i=1}^{N}\lambda_{ij}G_j = \delta_i(A_i + A_i^*) + \sum_{i=1}^{N}\lambda_{ij}\delta_j I_n$$
$$= \delta_i\left[A_i + A_i^* + \frac{1}{\delta_i}\sum_{\{j:\ j\neq i\}}\lambda_{ij}\delta_j I_n + \lambda_{ii}I_n\right]$$
$$\leq \delta_i\left[\lambda_{max}\left(A_i + A_i^* + \frac{1}{\delta_i}\sum_{\{j:\ j\neq i\}}\lambda_{ij}\delta_j I_n\right)I_n + \lambda_{ii}I_n\right] < 0$$

and, therefore, system (3.2) is mean square stable. The proof for condition (2) is similar. Now note that if the conditions of Corollary 4.20 are satisfied and defining $|\lambda|_{max}(\mathcal{T}) := \max\{|\lambda|{:}\lambda \text{ is an eigenvalue of the operator } \mathcal{T}\}$, then from condition (i) we get

$$\lambda_{max}\left[A_i + A_i^* + \frac{1}{\delta_i}\left(\sum_{\{j:\ j\neq i\}}\lambda_{ij}\delta_j\right)I_n\right] \leq |\lambda|_{max}\left[A_i + A_i^* + \frac{1}{\delta_i}\left(\sum_{\{j:\ j\neq i\}}\lambda_{ij}\delta_j\right)I_n\right]$$
$$= \left\|A_i + A_i^* + \frac{1}{\delta_i}\left(\sum_{\{j:\ j\neq i\}}\lambda_{ij}\delta_j\right)I_n\right\|$$
$$\leq \|A_i + A_i^*\| + \frac{1}{\delta_i}\sum_{\{j:\ j\neq i\}}\lambda_{ij}\delta_j$$
$$= \lambda_{max}(A_i + A_i^*) + \frac{1}{\delta_i}\sum_{\{j:\ j\neq i\}}\lambda_{ij}\delta_j < -\lambda_{ii},$$

which implies that condition (1) above is satisfied. Similarly we can show that condition (ii) of Corollary 4.20 implies condition (2) above. □

5. The $L_2^m(\Omega, \mathcal{F}, \mathbb{P})$ and jump diffusion case. We consider in this section two scenarios regarding the additive disturbance: the one as in (3.3), characterized by functions in $L_2^m(\Omega, \mathcal{F}, \mathbb{P})$, and that in (3.4), a jump diffusion, where the noise is characterized via a Wiener process. For both cases, under suitable conditions, it is shown that SS implies (is equivalent for $\mathcal{S}$ finite) to AWSS. In addition, it is shown that the state $x(t) \in L_2^n(\Omega, \mathcal{F}, \mathbb{P})$ for $L_2^m(\Omega, \mathcal{F}, \mathbb{P})$-disturbances.

5.1. The $L_2^m(\Omega, \mathcal{F}, \mathbb{P})$ disturbance case. We consider in this subsection the class of dynamical systems modeled by the stochastic equation

$$\dot{x}(t) = A(\theta_t)x(t) + B(\theta_t)w(t), \quad x(0) = x, \quad \theta_0 = v, \quad t \in \mathbb{R}^+, \tag{5.1}$$

where the additive disturbance $\{w(t); t \in \mathbb{R}^+\}$ is any $L_2^m(\Omega, \mathcal{F}, \mathbb{P})$-function. The equation above is restated here just for the sake of convenience. We need the following result.

LEMMA 5.1. *Let $\{w(t); t \in \mathbb{R}^+\} \in L_2^m(\Omega, \mathcal{F}, \mathbb{P})$ and define, for $\lambda > 0$,*

$$\chi(t) = \int_0^t e^{-\lambda(t-\tau)}\|w(\tau)\|_2 d\tau. \tag{5.2}$$

Then $\chi(t) \to 0$ *as* $t \to \infty$.

Proof. See the appendix for the proof. ∎

We shall prove the following result.

THEOREM 5.2. *The following assertions are equivalent:*

1. $\mathbb{R}_e\{\lambda(\mathcal{A})\} < 0$.
2. $\{x(t); t \in \mathbb{R}^+\} \in L_2^n(\Omega, \mathcal{F}, \mathbb{P})$ *for every* $\{w(t); t \in \mathbb{R}^+\} \in L_2^m(\Omega, \mathcal{F}, \mathbb{P})$ *and initial conditions* x_0 *and* v.

Moreover, if (1) *or* (2) *is satisfied, then the following condition is satisfied.*

3. $\lim_{t\to\infty} E(x(t+s)^* x(t)) = 0$ *for every* $s \geq 0$, $\{w(t); t \in \mathbb{R}^+\} \in L_2^m(\Omega, \mathcal{F}, \mathbb{P})$ *and initial conditions* x_0 *and* v.

Proof. (1 ⇒ 2): First notice that over the set Υ, we have from (5.1) and Theorem 2.1.70 of [10, p. 17] that

$$x(t) = \Phi(t,0)x(0) + \int_0^t \Phi(t,\tau)B(\theta_\tau)w(\tau)d\tau. \tag{5.3}$$

By the triangular inequality and recalling that $\mathbb{P}(\Upsilon) = 1$, it follows from (5.3) that

$$\|x(t)\|_2 = \|\Phi(t,0)x(0)\|_2 + \int_0^t \|\Phi(t,\tau)B(\theta_\tau)w(\tau)\|_2 d\tau. \tag{5.4}$$

From (4.6),

$$\|\Phi(t,0)x(0)\|_2^2 \leq n\|e^{\mathcal{L}t}\|\|x(0)\|_2^2. \tag{5.5}$$

From (4.6) again, with $z = B(\theta_\tau)w(\tau)$, we have

$$\begin{aligned} \|\Phi(t,\tau)B(\theta_\tau)w(\tau)\|_2^2 &\leq n\|e^{\mathcal{L}(t-\tau)}\|\|B(\theta_\tau)w(\tau)\|_2^2 \\ &\leq n\|B\|_{\sup}^2\|e^{\mathcal{L}(t-\tau)}\|\|w(\tau)\|_2^2. \end{aligned} \tag{5.6}$$

From Proposition 4.9 and the hypothesis that $\mathbb{R}_e\{\lambda(\mathcal{A})\} < 0$, it follows that $\mathbb{R}_e\{\lambda(\mathcal{L})\} < 0$, and thus there exist constants $\lambda > 0$ and $\mu > 0$ such that

$$\|e^{\mathcal{L}t}\| \leq \mu e^{-2\lambda t}. \tag{5.7}$$

From (5.4), (5.5), (5.6), and (5.7) we have, for some $a > 0$,

$$\|x(t)\|_2 \leq a\left(e^{-\lambda t}\|x(0)\|_2 + \int_0^t e^{-\lambda(t-\tau)}\|w(\tau)\|_2 d\tau\right). \tag{5.8}$$

Consider $\chi(t)$ as in (5.2). If we define

$$f(t) = \begin{cases} e^{-\lambda t}, & t \geq 0, \\ 0, & t < 0, \end{cases}$$

$$g(t) = \begin{cases} \|w(t)\|_2, & t \geq 0, \\ 0, & t < 0, \end{cases}$$

we have that χ can be written as the convolution (denoted here by $*$) of f and g, that is, $\chi(t) = (f * g)(t)$. Since $f \in L_1(\mathbb{R}^+)$ and $g \in L_2(\mathbb{R}^+)$, it follows that the convolution $f * g \in L_2(\mathbb{R}^+)$ and, moreover, for some $b > 0$,

$$\int_0^\infty \chi(t)^2 dt \leq b^2 \int_0^\infty f(t)dt \int_0^\infty g(t)^2 dt = \frac{b^2}{\lambda}\|w\|_2^2 \tag{5.9}$$

(cf. [45]). Taking (5.8) into square, we get

$$\|x(t)\|_2^2 \leq 2a^2(e^{-2\lambda t}\,\|x(0)\|_2^2 + \chi(t)^2) \tag{5.10}$$

and from (5.9) and (5.10) it follows that for some $c > 0$,

$$\|x\|_2^2 = \int_0^\infty \|x(t)\|_2^2 dt \leq c(\|x(0)\|_2^2 + \|w\|_2^2)$$

showing the desired result.

$(2 \Rightarrow 1)$: Take $w(t) = 0$ for all $t \in \mathbb{R}^+$. It follows from the hypothesis that for any initial condition x_0 and initial distribution v, $\|x\|_2^2 < \infty$, that is,

$$\|x\|_2^2 = \int_0^\infty \|x(t)\|_2^2 dt < \infty.$$

Thus system (3.2) is stochastically stable as in Definition 3.1, and the result follows from Proposition 4.9.

$(1 \Rightarrow 3)$: From Lemma 5.1 and (5.10) it follows that $E(\|x(t)\|^2) \to 0$ as $t \to \infty$. The result follows since

$$|\,E(x(t+s)^* x(t))\,| \leq E(\|x(t+s)\|\|x(t)\|) \leq (E(\|x(t+s)\|^2)E(\|x(t)\|^2))^{1/2} \to 0$$

as $t \to \infty$. □

5.2. The jump diffusion case. In this subsection we deal with MSS issues for the class of systems described in section 3 by (3.4), i.e.,

$$dx(t) = A(\theta_t)x(t)dt + B(\theta_t)dw(t), \quad x(0) = x_0, \quad \theta_0 \text{ with distribution } v, \quad t \in \mathbb{R}^+,$$

under assumptions (0.1)–(0.4). We recall that in this case we assume that $B \in \mathbb{H}_2^{m,n}$.

Let $\hat{R}(t) := (R_1(t), \dots) \in \mathbb{H}_1^{n+}$ with $R_i(t) := B_i R B_i^* p_i(t) \in \mathbb{B}(\mathbb{C}^n)$, where $p_i(t) = \mathbb{P}(\theta_t = i)$. The next proposition provides differential equations to compute the first and second moments of the state variable and can be easily proved using a suitable version of Itô's rule in conjunction with Lemma 4.2.

PROPOSITION 5.3. *For $t \in \mathbb{R}^+$ we have*

(a) $\dot{\hat{q}}(t) = F\hat{q}(t)$,

(b) $\dot{\hat{Q}}(t) = \mathcal{L}(\hat{Q}(t)) + \hat{R}(t)$,

(c) $\dot{\hat{\mathcal{Q}}}(s,t) = \mathcal{F}(\hat{\mathcal{Q}}(s,t))$.

PROPOSITION 5.4. *If $\mathbb{R}_e\{\lambda(\mathcal{A})\} < 0$, then system (3.4) is mean square stable according to Definition 3.2 and asymptotically wide sense stationary according to Definition 3.3, with $q = 0$ and*

$$Q = \sum_{i=1}^{\infty} \hat{\varphi}_i^{-1}(-\mathcal{A}^{-1}\hat{\varphi}(\mathcal{R})), \tag{5.11}$$

$$\mathcal{Q}(s) = \sum_{i=1}^{\infty} \hat{\varphi}_i^{-1}(e^{\mathcal{B}s}\mathcal{A}^{-1}\hat{\varphi}(\mathcal{R})), \tag{5.12}$$

where $\mathcal{R} = (\mathcal{R}_1, \dots) \in \mathbb{H}_1^{n+}$, $\mathcal{R}_i := \bar{R}_i \pi_i$, and $\bar{R}_i := B_i R B_i^*$; $i \in \mathcal{S}$.

Proof. First notice that since $\mathbb{R}_e\{\lambda(\mathcal{A})\} < 0$, we have from Proposition 4.8 that $\mathbb{R}_e\{\lambda(F)\} < 0$. Thus, from Propositions 4.4(a) and 2.3 it follows that $\hat{q}(t) \to 0$ as

$t \to \infty$, and since $q(t) = \sum_{i=1}^{\infty} q_i(t)$ we have that $q(t) \to 0$ as $t \to \infty$. Now, from Propositions 5.3(b) and 4.5(a) we have

$$\hat{\varphi}(\dot{\hat{Q}}(t)) = \mathcal{A}\hat{\varphi}(\hat{Q}(t)) + \hat{\varphi}(\hat{R}(t)).$$

If we now define $y(t) := \hat{\varphi}(\hat{Q}(t))$ and $f(t) := \hat{\varphi}(\hat{R}(t))$, we get

$$\dot{y}(t) = \mathcal{A}y(t) + f(t).$$

From the forward differential equation to the Markov chain (section 3), bearing in mind the definition of $\hat{R}(t)$, it follows that $f(t)$ is continuous. Furthermore, recalling that θ has limiting probabilities $\{\pi_i;\ i \in \mathcal{S}\}$ satisfying expression (3.5), we get

$$\|\hat{R}(t) - \mathcal{R}\|_1 = \sum_{i=1}^{\infty} \|R_i(t) - \mathcal{R}_i\| \le \|R\| \|B\|_2^2 \sup_{j \in \mathcal{S}} \{ \mid p_j(t) - \pi_j \mid \}$$

and thus $\lim_{t\to\infty} \hat{R}(t) = \mathcal{R}$ in $\mathbb{H}_1^n$. Now by Proposition 2.4, it follows that $\hat{\varphi}(\hat{Q}(t)) \to -\mathcal{A}^{-1}\hat{\varphi}(\mathcal{R})$ as $t \to \infty$. By noting that $Q(t) = \sum_{i=1}^{\infty} Q_i(t)$ it follows that $Q(t) \to Q$ as $t \to \infty$, with Q as in (5.11). From Propositions 4.5(c) and 5.3(c) we have that $\dot{\hat{\varphi}}(\hat{\mathcal{Q}}(s,t)) = \mathcal{B}\hat{\varphi}(\hat{\mathcal{Q}}(s,t))$ and, therefore, $\hat{\varphi}(\hat{\mathcal{Q}}(s,t)) = e^{\mathcal{B}s}(\hat{\mathcal{Q}}(0,t))$. Moreover, as seen above, we have $\hat{\varphi}(\hat{Q}(t)) \to \mathcal{A}^{-1}\hat{\varphi}(\mathcal{R})$ as $t \to \infty$. It follows that $\hat{\varphi}(\hat{\mathcal{Q}}(s,t)) \to e^{\mathcal{B}s}\mathcal{A}^{-1}\hat{\varphi}(\mathcal{R})$ as $t \to \infty$ and since $\mathcal{Q}(s,t) = \sum_{i=1}^{\infty} \mathcal{Q}_i(s,t)$, we have $\mathcal{Q}(t,s) \to \sum_{i=1}^{\infty} \hat{\varphi}_i^{-1}(e^{\mathcal{B}s}\mathcal{A}^{-1}\hat{\varphi}(\mathcal{R})) = \mathcal{Q}(s)$. □

REMARK 5.1. *It is a consequence of Proposition* 5.4 *that the* L_2^m*-result of Theorem* 5.2 *does not apply for the Wiener disturbance setting.*

For the case in which $\mathcal{S} = \{1, \ldots, N\}$ we have the following result.

PROPOSITION 5.5. *The following affirmatives are equivalent:*

(a) $\mathbb{R}_e\{\lambda(\mathcal{A})\} < 0$.

(b) *System* (3.4) *is mean square stable according to Definition* 3.2(b).

(c) *System* (3.4) *is asymptotically wide sense stationary according to Definition* 3.3.

Proof. From Proposition 5.4 we have that (a) implies (b) and (c). In addition, it is obvious that AWSS implies MSS. It remains to prove that (b) implies (a). First, we have $Q(t) = \sum_{i=1}^{N} Q_i(t)$ and from Propositions 4.4(b) and 4.5(a) $\hat{\varphi}(\dot{\hat{Q}}(t)) = \mathcal{A}\hat{\varphi}(\hat{Q}(t)) + \hat{\varphi}(\hat{R}(t))$. Therefore, $\hat{\varphi}(\hat{Q}(t)) = e^{\mathcal{A}t}\hat{\varphi}(\hat{Q}(0)) + \int_0^t e^{\mathcal{A}(t-s)}\hat{\varphi}(\hat{R}(s))ds$ and

$$Q(t) = \sum_{i=1}^{N} \hat{\varphi}_i^{-1}(e^{\mathcal{A}t}\hat{\varphi}(\hat{Q}(0))) + \sum_{i=1}^{N} \hat{\varphi}_i^{-1}\left(\int_0^t e^{\mathcal{A}(t-s)}\hat{\varphi}(\hat{R}(s))ds\right). \tag{5.13}$$

Now, by hypothesis, there exists $Q \in \mathbb{H}^{n+}$ (which depends only on R) such that $Q(t) \to Q$ as $t \to \infty$ for any $Q(0) = E(x_0 x_0^*)$ and Q does not depend on x_0. Furthermore, notice that for $x_0 = 0$ we have that the second term on the right-hand side of (5.13) converges to Q as $t \to \infty$ and thus the first term goes to zero, for any x_0 and v. The remaining of the proof follows the proof of Proposition 4.16. □

Finally, we conclude with the following unifying result.

THEOREM 5.6. *Consider the following assertions:*

(a) $\mathbb{R}_e\{\lambda(\mathcal{A})\} < 0$.

(b) *There exists* $G \in \widetilde{\mathbb{H}}_{\sup}^{n+}$ *such that* (4.8) *is satisfied for some* $S \in \widetilde{\mathbb{H}}_{\sup}^{n+}$.

(c) *System* (3.2) *is stochastically stable according to Definition* 3.1.

(d) *System* (3.2) *is stochastically stable for the real state space case according to Definition* 4.13.
(e) *System* (3.3) *is stochastically stable according to Definition* 3.1.
(f) *System* (3.2) *is mean square stable according to Definition* 3.2.
(g) *System* (3.2) *is mean square stable according to Definition* 3.2(b).
(h) *System* (3.4) *is mean square stable according to Definition* 3.2.
(i) *System* (3.4) *is mean square stable according to Definition* 3.2(b).
(j) *System* (3.4) *is asymptotically wide sense stationary according to Definition* 3.3.
(k) *System* (3.3) *is mean square stable according to Definition* 3.2.
(l) *System* (3.3) *is mean square stable according to Definition* 3.2(b).
(m) *System* (3.3) *is asymptotically wide sense stationary according to Definition* 3.3.

The affirmatives (a), (b), (c), (d), (e) *are equivalent and any of them implies affirmatives* (f), (g), (h), (i), (j), (k), (l), (m). *Moreover, if* $\mathcal{S}$ *is finite, then all affirmatives are equivalent.*

Proof. It follows, essentially, from Propositions 4.9, 4.10, 4.11, 4.12, 4.14, 4.16, 5.4, 5.5 and Theorem 5.2. □

Appendix. In this section we present the proof of some auxiliary results.

Proof of Proposition 2.3. Suppose $\int_0^\infty \|e^{\mathcal{L}t}(V)\|_1 dt < \infty$ for every $V \in \mathbb{H}_1^{n+}$. From the decomposition (2.8), for any $H \in \mathbb{H}_1^n$ we can find $H^i \in \mathbb{H}_1^{n+}$, $i = 1, 2, 3, 4$, such that $H = (H^1 - H^2) + \sqrt{-1}(H^3 - H^4)$. From linearity of the semigroup $e^{\mathcal{L}t}$, we have

$$\int_0^\infty \|e^{\mathcal{L}t}(H)\|_1 dt \le \sum_{i=1}^4 \int_0^\infty \|e^{\mathcal{L}t}(H^i)\|_1 dt < \infty,$$

showing the desired result. □

Proof of Lemma 4.1. For (a), consider $P = (P_1, \dots) \in \mathcal{H}_1^n$. Then

$$\begin{aligned}\|\mathcal{L}_i(P)\| &= \left\| A_i P_i + P_i A_i^* + \sum_{j=1}^\infty \lambda_{ji} P_j \right\| \\ &\le 2\|A\|_{\sup}\|P_i\| + \sum_{j=1}^\infty \mid \lambda_{ji} \mid \|P_j\|.\end{aligned}$$

Taking the sum over i and recalling that $0 \le \lambda_j \le \varrho$, we get

$$\begin{aligned}\sum_{i=1}^\infty \|\mathcal{L}_i(P)\| &\le 2\|A\|_{\sup} \sum_{i=1}^\infty \|P_i\| + \sum_{j=1}^\infty \sum_{i=1}^\infty \mid \lambda_{ji} \mid \|P_j\| \\ &\le 2\|A\|_{\sup}\|P\|_1 + 2\sum_{j=1}^\infty \lambda_j \|P_j\| \\ &\le 2\|A\|_{\sup}\|P\|_1 + 2\varrho \sum_{j=1}^\infty \|P_j\| \\ &= 2\|P\|_1(\|A\|_{\sup} + \varrho),\end{aligned}$$

showing the desired result. The proof for $\mathcal{E}$ and $\mathcal{F}$ follows the same steps and will be omitted.

For (b), it follows that for any $P = (P_1, \dots) \in \mathbb{H}^n_{\sup}$,

$$\begin{aligned}\|\mathcal{T}_i(P)\| &= \left\| A_i^* P_i + P_i A_i + \sum_{j=1}^{\infty} \lambda_{ij} P_j \right\| \\ &\le 2\|A\|_{\sup}\|P\|_{\sup} + \|P\|_{\sup} \sum_{j=1}^{\infty} |\lambda_{ji}| \\ &= 2\|P\|_{\sup}(\|A\|_{\sup} + \lambda_i) \le 2\|P\|_{\sup}(\|A\|_{\sup} + \varrho),\end{aligned}$$

showing the desired result.

Part (c) follows from the definition of $\mathcal{L}(.)$. To prove (d) it suffices to prove that $Y_i(t) \in \mathbb{B}(\mathbb{C}^n)^+$ for $i \in \mathcal{S}$ and any $t \in \mathbb{R}^+$, where $Y(t) := (Y_1(t), \dots)$, with $Y(0) = Q$ and $t \in \mathbb{R}^+$, satisfies $\dot{Y}(t) = \mathcal{L}(Y(t))$ or $\dot{Y}_i(t) = \mathcal{L}_i(Y(t))$. From (4.5) and defining $\tilde{A}_i = A_i + \frac{1}{2}\lambda_{ii} I$, we have

$$\dot{Y}_i(t) = \tilde{A}_i Y_i(t) + Y_i(t)\tilde{A}_i^* + \sum_{\{j \ne i\}} \lambda_{ji} Y_j(t).$$

Furthermore,

$$Y_i(t) = e^{\tilde{A}_i t} Q_i e^{\tilde{A}_i^* t} + \int_0^t e^{\tilde{A}_i\ (t-s)} \left(\sum_{\{j \ne i\}} \lambda_{ji} Y_j(s) \right) e^{\tilde{A}_i^*(t-s)} ds.$$

Now, it is a *fait accompli* that the above equation has a unique integrable solution $Y_i(t)$ that can be found by successive approximations as follows. Consider the sequence $\{Y_i^k(t) : k = 0, 1, \dots, t \in \mathbb{R}^+\}$ for $i \in \mathcal{S}$ obtained recursively as

$$\begin{aligned} Y_i^{(k+1)}(t) &= e^{\tilde{A}_i t} Q_i e^{\tilde{A}_i^* t} + \int_0^t e^{\tilde{A}_i\ (t-s)} \left(\sum_{\{j \ne i\}} \lambda_{ji} Y_j^k(s) \right) e^{\tilde{A}_i^*(t-s)} ds, \\ Y_i^0(t) &= 0, \quad i \in \mathcal{S}. \end{aligned}$$

Bearing in mind that $\lambda_{ij} \ge 0$ for $i \ne j$, $i \in \mathcal{S}$, it is a routine exercise to show that $Y_i^{(k+1)}(t) \ge Y_i^{(k)}(t) \ge 0$ for $k = 0, 1, \dots$ and any $t \in \mathbb{R}^+$. Inspired in [29], we prove that, for $i \in \mathcal{S}$, $\|Y_i^{(k)}(t)\| \le \ell_i(t)$ for every $k = 0, 1, \dots$ and any $t \in \mathbb{R}^+$, where

$$\begin{aligned} \dot{\ell}_i(t) &= 2\|\tilde{A}_i\|\ell_i(t) + \sum_{\{j \ne i\}} \lambda_{ji} \|\ell_j(t)\|, \\ \ell_i(0) &= \|Y_i(0)\|. \end{aligned}$$

This is carried out by induction as follows. First notice that the assertion above is obviously true for $k = 0$. Assuming now that it holds for some k, i.e., $\|Y_i^{(k)}(t)\| \le \ell_i(t)$ for $i \in \mathcal{S}$ and any $t \in \mathbb{R}^+$, we have

$$\begin{aligned} \ell_i(t) &= e^{2\|\tilde{A}_i\| t} \ell_i(0) + \int_0^t e^{2\|\tilde{A}_i\|(t-s)} \left(\sum_{\{j \ne i\}} \lambda_{ji} \|\ell_j(s)\| \right) ds \\ &\ge \left\| e^{\tilde{A}_i t} Q_i e^{\tilde{A}_i^* t} + \int_0^t e^{\tilde{A}_i\ (t-s)} \left(\sum_{\{j \ne i\}} \lambda_{ji} Y_j^k \right) e^{\tilde{A}_i^*(t-s)} ds \right\| \\ &= \|Y_i^{(k+1)}(t)\|, \end{aligned}$$

and the assertion follows. Finally, using the differential equation for the approximating sequence, we get that $\lim_{k\to\infty} Y_i^{(k)}(t) = Y_i(t) \geq 0$ for $i \in \mathcal{S}$ and any $t \in \mathbb{R}^+$, i.e., $Y_i(t) \in \mathbb{B}(\mathbb{C}^n)^+$ for $i \in \mathcal{S}$ and any $t \in \mathbb{R}^+$, and part (d) follows. □

Proof of Lemma 4.2. Bearing in mind that $d(1_{\{\theta_t=i\}}) := 1_{\{\theta_{t+dt}=i\}} - 1_{\{\theta_t=i\}}$, we have

$$\begin{aligned} E(f(t)d(1_{\{\theta_t=i\}})) &= E(f(t)1_{\{\theta_{t+dt}=i\}}) - E(f(t)1_{\{\theta_t=i\}}) \\ &= \sum_{j=1}^{\infty} E(E(f(t)1_{\{\theta_{t+dt}=i\}}1_{\{\theta_t=j\}})|\mathcal{F}_t) - E(f(t)1_{\{\theta_t=i\}}) \\ &= \sum_{j=1}^{\infty} \mathbb{P}(\theta_{t+dt}=i|\theta_t=j)f_j(t) - f_i(t) \\ &= \sum_{j=1}^{\infty} \lambda_{ji} f_j(t)dt \ + \ o(dt). \end{aligned}$$ □

Proof of Proposition 4.3. We show now that for some constants $c_1 > 0$ and $c_2 > 0$ we have, for any $H_i \in \mathbb{B}(\mathbb{C}^n)$, that $c_1\|\varphi(H_i)\| \leq \|H_i\| \leq c_2\|\varphi(H_i)\|$. Since all norms in finite-dimensional spaces are equivalent (see [44]), for any $H = (H_1, \dots) \in \mathbb{H}_1^n$, $c_1\|\hat{\varphi}(H)\|_1 \leq \|H\|_1 \leq c_2\|\hat{\varphi}(H)\|_1$. □

Proof of Proposition 4.5. The proof follows from the definition of $\hat{\varphi}$ in section 2, in conjunction with Proposition 4.3 and the results given by (2.7), bearing in mind the definition of the operators $\mathcal{F}$, $\mathcal{L}$, and $\mathcal{T}$ in (4.4)–(4.5). □

Proof of Lemma 4.6. (a) We begin by noticing that the solution of the above differential equation is given by $y(t) = e^{\mathcal{A}t}y(0)$. Then

$$y(t) = e^{\mathcal{A}t}\hat{\varphi}(Q) = \sum_{n=0}^{\infty} \frac{1}{n!}\mathcal{A}^n\hat{\varphi}(Q)t^n = \sum_{n=0}^{\infty} \frac{1}{n!}\hat{\varphi}(\mathcal{L}^n(Q))t^n = \hat{\varphi}(e^{\mathcal{L}t}(Q)),$$

where the third equality follows from Proposition 4.5(a). Part (b) follows from Lemma 4.1 and the definitions of $\hat{\varphi}$ and ${\hat{\varphi}_j}^{-1}$. □

Proof of Lemma 4.7. Consider (4.1) with initial time $t_0 = \tau$ and initial condition $x(t_0) = z$. It follows from (4.2) that over the set Υ,

$$x(t) = \Phi(t, \tau)z.$$

Recalling that $\mathbb{P}(\Upsilon) = 1$, $Q_i(t) = E(x(t)x(t)^* 1_{\{\theta_t=i\}})$, $Q_i(\tau) = E(zz^* 1_{\{\theta_\tau=i\}})$, it follows that

$$\begin{aligned} \|x(t)\|_2^2 = \|\Phi(t,\tau)z\|_2^2 &= \sum_{i=1}^{\infty} \mathrm{tr}(Q_i(t)) \\ &\leq n\sum_{i=1}^{\infty} \|Q_i(t)\| = n\|\hat{Q}(t)\|_1. \end{aligned} \tag{5.14}$$

From Proposition 4.4(b),

$$\hat{Q}(t) = e^{\mathcal{L}(t-\tau)}\hat{Q}(\tau). \tag{5.15}$$

Substituting (5.15) into (5.14) leads to

$$\begin{aligned} \|\Phi(t,\tau)z\|_2^2 &\leq n\|e^{\mathcal{L}(t-\tau)}\hat{Q}(\tau)\| \\ &\leq n\|e^{\mathcal{L}(t-\tau)}\|\|\hat{Q}(\tau)\|_1. \end{aligned} \tag{5.16}$$

The result follows from (5.16) after noticing that

$$\|\hat{Q}(\tau)\|_1 = \sum_{i=1}^{\infty} \|Q_i(\tau)\| \leq \sum_{i=1}^{\infty} \mathrm{tr}(Q_i(\tau)) = \|z\|_2^2. \qquad \square \tag{5.17}$$

Proof of Proposition 4.8. Since $\mathbb{R}_e\{\lambda(\mathcal{A})\} < 0$ it follows from Propositions 2.3 and 4.9 that $\|e^{\mathcal{L}t}\| \leq \mu e^{-bt}$ for some $\mu > 0, b > 0$. For the homogeneous systems

$$\dot{x}(t) = A(\theta_t)x(t), \qquad t \in \mathbb{R}^+,$$

we have from Proposition 4.4(b) and (3.12) that

$$\begin{aligned} E(\|x(t)\|^2) &= \sum_{j=1}^{\infty} \mathrm{tr}(E(x(t)x(t)^* 1_{\{\theta_t = j\}})) \\ &= \sum_{j=1}^{\infty} \mathrm{tr}(Q_j(t)) \leq n \sum_{j=1}^{\infty} \|Q_j(t)\| = n\|\hat{Q}(t)\|_1 \\ &= \|e^{\mathcal{L}t}(\hat{Q}(0))\|_1 \leq n\mu e^{-bt}\|\hat{Q}(0)\|_1 \leq n\mu e^{-bt} E(\|x(0)\|^2) \end{aligned}$$

and from Proposition 4.4(a)

$$\begin{aligned} \|e^{Ft}\hat{q}(0)\|_1 = \|\hat{q}(t)\|_1 = \sum_{j=1}^{\infty} \|q_j(t)\| &\leq \sum_{j=1}^{\infty} E(\|x(t)\| 1_{\{\theta_t = j\}}) = E(\|x(t)\|) \\ &\leq (E(\|x(t)\|^2))^{1/2} \leq (n\mu)^{1/2} e^{-\frac{b}{2}t} (E(\|x(0)\|^2))^{1/2}. \end{aligned} \tag{5.18}$$

From Proposition 2.2 and (5.18) we get $\mathbb{R}_e\{\lambda(F)\} < 0$. $\square$

Proof of Lemma 5.1. The proof follows the same arguments as in [48, pp. 119–120]. Given any $\epsilon > 0$ consider $t_\epsilon > 0$ such that $\int_{t_\epsilon}^{\infty} \|w(\tau)\|_2^2 d\tau \leq \epsilon^2$. Then for $t > t_\epsilon$,

$$\chi(t) = e^{-\lambda(t-t_\epsilon)} \int_0^{t_\epsilon} e^{-\lambda(t_\epsilon - \tau)} \|w(\tau)\|_2 d\tau + \int_{t_\epsilon}^{t} e^{-\lambda(t-\tau)} \|w(\tau)\|_2 d\tau. \tag{5.19}$$

We have from the Schwarz inequality that

$$\left(\int_0^{t_\epsilon} e^{-\lambda(t_\epsilon - \tau)} \|w(\tau)\|_2 d\tau \right)^2 \leq \int_0^{t_\epsilon} e^{-2\lambda(t_\epsilon - \tau)} d\tau \int_0^{t_\epsilon} \|w(\tau)\|_2^2 d\tau \leq \frac{1}{2\lambda} \|w\|_2^2 \tag{5.20}$$

and

$$\left(\int_{t_\epsilon}^{t} e^{-\lambda(t-\tau)} \|w(\tau)\|_2 d\tau \right)^2 \leq \int_{t_\epsilon}^{t} e^{-2\lambda(t-\tau)} d\tau \int_{t_\epsilon}^{\infty} \|w(\tau)\|_2^2 d\tau \leq \frac{1}{2\lambda} \epsilon^2. \tag{5.21}$$

Taking the limit as $t \to \infty$ in (5.19), and from (5.20) and (5.21) we obtain that $0 \leq \lim_{t\to\infty} \chi(t) \leq \frac{\epsilon}{(2\lambda)^{1/2}}$, showing the desired result. $\square$

Acknowledgments. The authors would like to express their gratitude to the referees for their suggestions and helpful comments. Authors are particularly grateful to two of the referees for their helpful comments on some technical issues and some important suggestions on the bibliography.

REFERENCES

[1] L. Arnold, *A formula connecting sample and moment stability of linear stochastic systems,* SIAM J. Appl. Math., 44 (1984), pp. 793–802.

[2] L. Arnold, E. Oeljeklaus, and E. Pardoux, *Almost sure and moment stability for linear Ito equations,* in Lyapunov Exponents, L. Arnold and V. Wihstutz, eds., Lecture Notes in Math. 1186, Springer-Verlag, New York, 1985, pp. 129–159.

[3] L. Arnold and V. Wihstutz, eds., *Lyapunov Exponents,* Lecture Notes in Math. 1186, Springer-Verlag, New York, 1986.

[4] G. Basak, A. Bisi, and M.K. Ghosh, *Stability of a random diffusion with linear drift,* J. Math. Anal. Appl., 202 (1996), pp. 604–622.

[5] H.A.P. Blom and Y. Bar-Shalom, *The interacting multiple model algorithm for systems with Markovian switching coefficients,* IEEE Trans. Automat. Control, 33 (1988), pp. 780–783.

[6] S. Bohacek and E.A. Jonckheere, *Relationships between linear dynamically varying systems and jump linear systems,* Math. Control Signals Syst., 16 (2003), pp. 207–224.

[7] N.K. Bose and Y.Q. Shi, *A simple general proof of the Kharitonov's generalized stability criterion,* IEEE Trans. Circuits Syst., CAS-34 (1987), pp. 1233–1237.

[8] E. Boukas and P. Shi, *Stochastic stability and guaranteed cost control of discrete-time uncertain systems with Markovian jumping parameters,* Internat. J. Robust Nonlinear Control, 8 (1998), pp. 1155–1167.

[9] J.W. Brewer, *Kronecker products and matrix calculus in system theory,* IEEE Trans. Circuits Syst., 25 (1978), pp. 772–781.

[10] F.M. Callier and C.A. Desoer, *Linear Systems Theory,* Springer-Verlag, New York, 1991.

[11] O.L.V. Costa and M.D. Fragoso, *Stability results for discrete-time linear systems with Markovian jumping parameters,* J. Math. Anal. Appl., 179 (1993), pp. 154–178.

[12] O.L.V. Costa and M.D. Fragoso, *On the discrete-time infinite coupled Riccati equations which arises in a certain optimal control problem,* IEEE Trans. Automat. Control, 40 (1995), pp. 2076–2088.

[13] O.L.V. Costa and M.D. Fragoso, *Comments on "Stochastic stability of jump linear systems,"* IEEE Trans. Automat. Control, 49 (2004), pp. 1414–1416.

[14] O.L.V. Costa, M.D. Fragoso, and R.P. Marques, *Discrete-Time Markov Jump Linear Systems,* Probability and Its Applications, Springer-Verlag, New York, 2004.

[15] O.L.V. Costa and J.B.R. Do Val, *Full information H^∞-control for discrete-time infinite Markov jump parameters systems,* J. Math. Anal. Appl., 202 (1996), pp. 578–603.

[16] R.F. Curtain, ed., *Stability of Stochastic Dynamical Systems,* Lecture Notes in Math. 294, Springer-Verlag, New York, 1972.

[17] C.E. de Souza and M.D. Fragoso, *H^∞ control for linear systems with Markovian jumping parameters,* Control Theory Adv. Tech., 9 (1993), pp. 457–466.

[18] J.B.R. Do Val, C. Nespoli, and Y.R.Z. Caceres, *Stochastic stability for Markovian jump linear systems associated with a finite number of jump times,* J. Math. Anal. Appl., 285 (2003), pp. 551–563.

[19] V. Dragan and T. Morozan, *Stability and robust stabilization to linear stochastic systems described by differential equations with Markovian jumping and multiplicative noise,* Stochastic Anal. Appl., 20 (2002), pp. 33–92.

[20] F. Dufour and R.J. Elliot, *Adaptive control of linear systems with Markov perturbations,* IEEE Trans. Automat. Control, 43 (1998), pp. 351–372.

[21] F. Dufour and P. Bertrand, *The filtering problem for continuous-time linear systems with Markovian switching coefficients,* Systems Control Lett., 23 (1994), pp. 453–461.

[22] Y. Fang, *A new general sufficient condition for almost sure stability of jump linear systems,* IEEE Trans. Automat. Control, 42 (1997), pp. 378–382.

[23] Y. Fang, K.A. Loparo, and X. Feng, *Almost sure and δ-moment stability of jump linear systems,* Internat. J. Control, 59 (1994), pp. 1281–1307.

[24] X. Feng, K.A. Loparo, Y. Ji, and Chizeck, *Stochastic stability properties of jump linear systems,* IEEE Trans. Automat. Control, 37 (1992), pp. 38–53.

[25] M.D. Fragoso and J. Baczynski, *Optimal control for continuous-time linear quadratic problems with infinite Markov jump parameters,* SIAM J. Control Optim., 40 (2001), pp. 270–297.

[26] M.D. Fragoso and J. Baczynski, *Stochastic versus mean square stability in continuous time linear infinite Markov jump parameter systems,* Stochastic Anal. Appl., 20 (2002), pp. 347–356.

[27] M.D. Fragoso and J. Baczynski, *Lyapunov coupled equations for continuous-time infinite Markov jump linear systems,* J. Math. Anal. Appl., 274 (2002), pp. 319–335.

[28] M.D. Fragoso and O.L.V. Costa, *A unified approach for mean square stability of continuous-time linear systems with Markovian jumping parameters and additive disturbances,* in Proceedings of the 39th Conference on Decision and Control, Sydney, Australia, 2000. See also LNCC Internal Report 11/99, 1999.

[29] M.D. Fragoso, O.L.V. Costa, and C.E. de Souza, *A new approach to linearly perturbed Riccati equations arising in stochastic control,* J. Appl. Math. Optim., 37 (1999), pp. 99–126.

[30] W.S. Gray and O. Gonzalez, *Modelling electromagnetic disturbances in closed-loop computer controlled flight systems,* in Proceedings of the 37th IEEE Conference on Decision and Control, Philadelphia, PA, 1998, pp. 359–364.

[31] R.J. Has'Minskij, *Stochastic Stability of Differential Equations,* Sijthoff & Noordhoff, Groningen, The Netherlands, 1980.

[32] D. Henrion, J. Ježek, and M. Šebek, *Discrete-time symmetric polynomial equations with complex coefficients,* Kybernetica, 38 (2002), pp. 113–139.

[33] M.W. Hirsch and S. Smale, *Differential Equations, Dynamical Systems, and Linear Algebra,* Academic Press, New York, 1974.

[34] Y. Ji and H.J. Chizeck, *Controllability, stabilizability and continuous-time Markovian jumping linear quadratic control,* IEEE Trans. Automat. Control, 35 (1990), pp. 777–788.

[35] Y. Ji and H.J. Chizeck, *Jump linear quadratic Gaussian control: Steady-state solution and testable conditions,* Control-Theory and Adv. Tech., 6 (1990), pp. 289–319.

[36] S. Karlin and H.M. Taylor, *A Second Course on Stochastic Processes,* Academic Press, New York, 1981.

[37] A. Leizarowits, *Estimates and exact expressions for Lyapunov exponents of stochastic linear differential equations,* Stochastics, 24 (1988), pp. 335–356.

[38] Z.G. Li, Y.C. Soh, and C.Y. Wen, *Sufficient conditions for almost sure stability of jump linear systems,* IEEE Trans. Automat. Control, 45 (2000), pp. 1325–1329.

[39] X. Mao, *Stability of stochastic differential equations with Markovian switching,* Stochastic Process. Appl., 79 (1999), pp. 45–67.

[40] M. Mariton, *Almost sure and moments stability of jump linear systems,* Systems Control Lett., 11 (1988), pp. 393–397.

[41] M. Mariton, *Jump Linear Systems in Automatic Control,* Marcel Dekker, New York, 1990.

[42] S.P. Meyn and R.L. Tweedie, *Stability of Markovian processes* II: *Continuous-time processes and sampled chains,* Adv. in Appl. Probab., 25 (1993), pp. 487–517.

[43] T. Morozan, *Optimal stationary control for dynamic systems with Markov perturbations,* Stochastic Anal. Appl., 1 (1983), pp. 219–225.

[44] A.W. Naylor and G.R. Sell, *Linear Operator Theory in Engineering and Science,* 2nd ed., Marcel Dekker, New York, 1990.

[45] W. Rudin, *Real and Complex Analysis,* 2nd ed., Tata McGraw-Hill, New Delhi, 1974.

[46] P. Shi and J.A. Filar, *Stability analysis and controller design for a class of uncertain systems with Markovian jumping parameters,* IMA J. Math. Control Inform., AC-17 (2000), pp. 179–190.

[47] J. Weidmann, *Linear Operators in Hilbert Spaces,* Springer-Verlag, New York, 1980.

[48] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations,* Springer-Verlag, New York, 1983.

[49] C. Yuan and X. Mao, *Asymptotic stability in distribution of stochastic differential equations with Markovian switching,* Stochastic Process. Appl., 103 (2003), pp. 277–291.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 4, pp. 1192–1214

CHARACTERIZATION OF GRADIENT CONTROL SYSTEMS*

JORGE CORTÉS†, ARJAN VAN DER SCHAFT‡, AND PETER E. CROUCH§

Abstract. Given a general nonlinear affine control system with outputs and a torsion-free affine connection defined on its state space, we investigate the gradient realization problem: we give necessary and sufficient conditions under which the control system can be written as a gradient control system corresponding to some pseudo-Riemannian metric whose Levi-Civita connection is equal to the given affine connection. The results rely on a suitable notion of compatibility of the system with respect to the given affine connection, and on the output behavior of the prolonged system and the gradient extension. The symmetric product associated with an affine connection plays a key role throughout the discussion.

Key words. gradient control systems, symmetric product, prolongation and gradient extension of a nonlinear system, externally equivalent systems

AMS subject classifications. 93C10, 93B29, 53B05, 93B15

DOI. 10.1137/S0363012903425568

1. Introduction.

A physically motivated class of nonlinear systems are *gradient control systems*; see [4, 5, 10, 22, 23, 25, 26, 27] and the references quoted therein. These systems are described in the following way: they are nonlinear affine control systems, which are endowed with a pseudo-Riemannian metric on the state-space manifold. The drift vector field of the system is the gradient vector field associated with an internal potential function with respect to the pseudo-Riemannian metric, and the input vector fields are the gradient vector fields associated with the output functions of the system. Examples of gradient control systems include nonlinear electrical RLC networks, and dissipative systems where the inertial effects are neglected. In the case of RL or RC networks, the pseudo-Riemannian metric is positive-definite, and thus is a usual Riemannian metric, while for general RLC networks the metric is indefinite. We refer to [4, 5, 10, 25, 26] for more background on the modeling of nonlinear networks as gradient systems.

Another relevant class of nonlinear systems is the family formed by the *Hamiltonian control systems*; see [13]. In this case, the state-space manifold is equipped with a symplectic form. The drift vector field and the input vector fields are the Hamiltonian vector fields associated, respectively, to an internal energy function and the output functions of the system with respect to the symplectic form. Hamiltonian equations are of central importance in the modeling of physical systems as they are the starting point to describe the dynamics of a very large class of phenomena, including mechanical, electrical, and electromagnetic systems.

*Received by the editors March 27, 2003; accepted for publication (in revised form) April 6, 2005; published electronically October 7, 2005. A short version [9] of this paper was presented at the IFAC Workshop on Lagrangian and Hamiltonian Methods for Nonlinear Control, Seville, Spain, 2003.
http://www.siam.org/journals/sicon/44-4/42556.html

†Department of Applied Mathematics and Statistics, University of California at Santa Cruz, 1156 High Street, Santa Cruz, CA 95064 (jcortes@ucsc.edu, http://www.ams.ucsc.edu/~jcortes). This author's work was partially supported by NSF grant CMS-0100162 and by the European Union Training and Mobility of Researchers Program, ERB FMRXCT-970137.

‡Department of Applied Mathematics, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands (a.j.vanderschaft@math.utwente.nl, http://www.math.utwente.nl/~twarjan).

§Department of Electrical and Computer Engineering, Arizona State University, Tempe, AZ 85287 (peter.crouch@asu.edu, http://www.eas.asu.edu/~sserc/people/crouch/crouch.html).

Apart from their physical and engineering importance, gradient and Hamiltonian systems also possess very peculiar mathematical properties. For instance, an observable and controllable linear input-state-output system is a Hamiltonian control system [6] (respectively, a gradient control system [27]) if and only if its impulse response matrix $W(t)$ satisfies $W(t) = -W^T(-t)$ (respectively, $W(t) = W^T(t)$). Although they are typically not amenable to linearization techniques, their rich geometric structure makes it possible to combine powerful tools from nonlinear control theory, differential geometry, and classical mechanics in the study of a variety of problems including stability and stabilization, input-output decoupling, structural synthesis, and interconnection.

Their theoretical and practical relevance, together with their meaningful geometric properties and the wide range of results available for them, make the classes of Hamiltonian and gradient systems *distinct* within the family of nonlinear affine control systems. This explains the interest in identifying those systems that can be written as either Hamiltonian or gradient. This *characterization problem* is motivated by the realization problem in systems theory and the inverse problem in mechanics. The realization problem addresses the question of when the input-output map of a system can be realized as the external behavior of a Hamiltonian (respectively, gradient) input-output system. The inverse problem, which has a longstanding history in mathematical physics, poses the question of when a second-order differential equation can be realized as the Euler–Lagrange equations corresponding to certain Lagrangian function. For further reference on these problems, the reader is referred to [11, 15, 20, 21].

In [12, 13], necessary and sufficient conditions were given under which a minimal nonlinear affine control system with an equal number of inputs and outputs is a Hamiltonian control system with respect to some symplectic structure, which turned out to be unique. A different but somehow related problem is considered in [24]: assuming the state space of the nonlinear affine control system is already endowed with a symplectic form, conditions are derived that guarantee that a feedback transformation exists making Hamiltonian the drift vector field of the transformed control system. As we discuss below, there are a number of key differences in the treatment of the characterization problem for the Hamiltonian and the gradient cases, which make the latter more involved. A fundamental observation is that, while every input-state-output system admits a natural extension to a Hamiltonian system living on the cotangent bundle of its state space, the construction of a gradient extension on the cotangent bundle relies on the selection of a torsion-free affine connection on the state space. This is why our starting point in the gradient setting is the selection of an *appropriate* torsion-free affine connection. This appropriateness is defined in terms of a novel *compatibility condition* of the given nonlinear system with the selected affine connection, guaranteeing an appropriate choice of the latter one. The compatibility condition is expressed as a relation of the symmetric products of the drift vector field and the input vector fields with the output functions of the system. As a further remark, the role played in the Hamiltonian setting by the Lie bracket and the Hamiltonian vector fields is taken here by the symmetric product associated with the given affine connection and the gradient vector fields.

The question solved by the main result of this paper (cf. Theorem 5.4) is the following: given a torsion-free affine connection which is compatible with the nonlinear control system, find necessary and sufficient conditions under which the system is gradient with respect to a pseudo-Riemannian metric whose Levi-Civita connection is the given affine connection. The question that still remains to be addressed in order

to solve the full characterization problem for gradient control systems is the following: given a nonlinear control system, when does an affine connection exist such that these necessary and sufficient conditions are satisfied?

The paper is organized as follows. In section 2 we present the class of nonlinear systems considered throughout the paper. We also introduce the notions of prolongation and gradient extension of a nonlinear system. The observability properties of these systems, studied in section 3, together with the concept of (weakly) externally equivalent systems, introduced in section 4, turn out to be key in establishing Theorem 5.4. In section 5, we introduce the important notion of compatibility between a nonlinear system and a given affine connection. At this point, we are ready to state and prove the main result of the paper, namely, the characterization of when a general nonlinear control system is gradient. This characterization can be roughly described as follows: under certain technical conditions, a nonlinear affine control system is gradient if and only if its prolongation and its gradient extension *behave* similarly (i.e., have the same input-output behavior). In section 6 we investigate the uniqueness (up to isometry) of gradient realizations with the same input-output behavior and we give an alternative proof of a result in [1, 2]. We present our conclusions in section 7. Finally, an appendix in section 8 contains a simplifying result concerning the checkability of the compatibility condition for a nonlinear affine control system.

2. Setting. Let M be an n-dimensional (real-)analytic manifold. We will denote by TM, T^*M the tangent and cotangent bundles of M, by $\mathfrak{X}(M)$ the set of analytic vector fields on M, by $\Omega^1(M)$ the set of analytic one-forms on M, and by $C^\omega(M)$ the set of analytic functions on M. Consider a nonlinear control system Σ with state space M, affine in the inputs, and with an equal number of inputs and outputs,

$$\Sigma : \begin{cases} \dot{x} = g_0(x) + \displaystyle\sum_{j=1}^{m} u_j g_j(x)\,, \\ y_j = V_j(x)\,, \quad j = 1, \dots, m\,, \end{cases} \tag{2.1}$$

where $x \in M$, $x(0) = x_0$, and $u = (u_1, \dots, u_m) \in U \subset \mathbb{R}^m$. The vector fields $g_0, g_1, \dots, g_m$ on M are assumed to be complete and $V_1, \dots, V_m$ are real-valued functions on M. The set U is the control space, which for simplicity is assumed to be an open subset of $\mathbb{R}^m$, containing 0. The function $t \mapsto u(t) = (u_1(t), \dots, u_m(t))$, which we will commonly denote as $u(\cdot)$, belongs to a certain class of functions of time, denoted by $\mathcal{U}$, called the *admissible controls*. For our purposes, we may restrict the admissible controls to be the piecewise constant right continuous functions.

An important subclass of the family of nonlinear systems (2.1) is formed by the Hamiltonian control systems; see [13]. Here, we will instead focus our attention on the family of gradient control systems. Let $\mathcal{G}$ be a pseudo-Riemannian metric on M, i.e., a nondegenerate symmetric (0,2)-tensor on M (not necessarily positive-definite); see [7]. Consider the "musical" isomorphisms associated with $\mathcal{G}$, $\flat_{\mathcal{G}} : \mathfrak{X}(M) \to \Omega^1(M)$, $\sharp_{\mathcal{G}} : \Omega^1(M) \to \mathfrak{X}(M)$ defined by

$$\flat_{\mathcal{G}}(X)(Y) = \mathcal{G}(X, Y)\,, \quad \sharp_{\mathcal{G}}(\mu) = \flat_{\mathcal{G}}^{-1}(\mu)\,,$$

where $X, Y \in \mathfrak{X}(M)$ and $\mu \in \Omega^1(M)$. The *gradient vector field* associated with a function $V \in C^\omega(M)$ is given by $\operatorname{grad}_{\mathcal{G}} V = \sharp_{\mathcal{G}}(dV)$. Reciprocally, a vector field $X \in \mathfrak{X}(M)$ is said to be *locally gradient* if the one-form $\flat_{\mathcal{G}}(X)$ is closed. By Poincaré's lemma, this is equivalent to saying that there exists a locally defined function $V \in C^\omega(M)$ such that $\flat_{\mathcal{G}}(X) = dV$. If this equality holds globally, X is called *gradient* and

will be denoted by $X = \operatorname{grad}_{\mathcal{G}} V$. Throughout the paper, we will drop the subindex when the pseudo-Riemannian metric used in the computation of the gradient vector field is clear from the context. If we fix coordinates $(x^1, \dots, x^n)$ on M, then the pseudo-Riemannian metric can be locally expressed as $\mathcal{G} = \mathcal{G}_{ab} dx^a \otimes dx^b$, where $(\mathcal{G}_{ab} = \mathcal{G}(\frac{\partial}{\partial x^a}, \frac{\partial}{\partial x^b}))$ is a symmetric matrix. The musical isomorphisms are then given by $\flat_{\mathcal{G}} = \mathcal{G}_{ab} dx^a \otimes dx^b$, $\sharp_{\mathcal{G}} = \mathcal{G}^{ab} \frac{\partial}{\partial x^a} \otimes \frac{\partial}{\partial x^b}$, where $(\mathcal{G}^{ab})$ is the inverse matrix of $(\mathcal{G}_{ab})$. Finally, the gradient vector field associated with V reads

$$\operatorname{grad}_{\mathcal{G}} V = \mathcal{G}^{ab} \frac{\partial V}{\partial x^b} \frac{\partial}{\partial x^a}.$$

Now, assume that the state space M in (2.1) is a pseudo-Riemannian manifold, $(M, \mathcal{G})$. Furthermore, assume that the drift vector field g_0 is locally gradient and the input vector fields g_j, $j = 1, \dots, m$, are gradient with respect to the functions $V_1, \dots, V_m$, i.e., $g_j = \operatorname{grad}_{\mathcal{G}} V_j$, $j = 1, \dots, m$. Then, the resulting system

$$\Sigma : \begin{cases} \dot{x} = g_0(x) + \sum_{j=1}^{m} u_j(t) \operatorname{grad}_{\mathcal{G}} V_j(x), \\ y_j = V_j(x), \quad j = 1, \dots, m, \end{cases} \tag{2.2}$$

is called a *locally gradient control system on* M. If the drift g_0 is a gradient vector field, then the system is called a *gradient control system on* M.

Given an affine connection on M, our objective is to characterize when a nonlinear system of the form (2.1) is a locally gradient control system (2.2), i.e., find necessary and sufficient conditions for the existence of a pseudo-Riemannian metric $\mathcal{G}$ on the state space M whose Levi-Civita connection is the given affine connection such that the system (2.1) equals system (2.2). These conditions will be given in terms of the output behavior of the so-called prolonged system and the gradient extension of Σ, which we describe next.

2.1. The prolongation of a nonlinear system. Given an initial state $x(0) = x_0$, take a coordinate neighborhood of M containing x_0. Let $t \in [0, T] \mapsto x(t)$ be the solution of (2.1) corresponding to the input function $t \in [0, T] \mapsto u(t) = (u_1(t), \dots, u_m(t))$ and the initial state $x(0) = x_0$, such that $x(t)$ remains within the selected coordinate neighborhood. Denote the resulting output by $t \in [0, T] \mapsto y(t) = (y_1(t), \dots, y_m(t))$, with $y_j(t) = V_j(x(t))$. Then the *variational system* along the input-state-output trajectory $t \in [0, T] \mapsto (x(t), u(t), y(t))$ is given by the following time-varying system:

$$\begin{aligned} \dot{v}(t) &= \frac{\partial g_0}{\partial x}(x(t))v(t) + \sum_{j=1}^{m} u_j(t) \frac{\partial g_j}{\partial x}(x(t))v(t) + \sum_{j=1}^{m} u_j^p g_j(x(t)), \\ y_j^p(t) &= \frac{\partial V_j}{\partial x}(x(t))v(t), \quad j = 1, \dots, m, \end{aligned} \tag{2.3}$$

where $v(0) = v_0 \in \mathbb{R}^n$, and $u^p = (u_1^p, \dots, u_m^p)$, $y^p = (y_1^p, \dots, y_m^p)$ denote the inputs and the outputs of the variational system. The reasoning behind the term "variational" comes from the following fact: let $(x(t, \epsilon), u(t, \epsilon), y(t, \epsilon))$, $t \in [0, T]$, be a family of input-state-output trajectories of (2.1) parameterized by $\epsilon \in (-\delta, \delta)$, with $x(t, 0) = x(t)$, $u(t, 0) = u(t)$, and $y(t, 0) = y(t)$, $t \in [0, T]$. Then, the infinitesimal variations

$$v(t) = \frac{\partial x}{\partial \epsilon}(t, 0), \quad u^p(t) = \frac{\partial u}{\partial \epsilon}(t, 0), \quad y^p(t) = \frac{\partial y}{\partial \epsilon}(t, 0)$$

satisfy (2.3). Additionally, if the initial state is the same for the whole family of trajectories, $x(0,\epsilon) = x_0$, then the variational state $v(0)$ at time 0 is necessarily 0.

The *prolongation or prolonged system* of (2.1) corresponds to considering together the original system (2.1) and the variational system

$$\begin{aligned} \dot{x} &= g_0(x) + \sum_{j=1}^{m} u_j g_j(x), \\ \dot{v}(t) &= \frac{\partial g_0}{\partial x}(x(t))v(t) + \sum_{j=1}^{m} u_j(t)\frac{\partial g_j}{\partial x}(x(t))v(t) + \sum_{j=1}^{m} u_j^p g_j(x(t))\,, \\ y_j &= V_j(x)\,, \quad y_j^p(t) = \frac{\partial V_j}{\partial x}(x(t))\,v(t)\,, \quad j = 1,\dots,m, \end{aligned} \tag{2.4}$$

with inputs u_j, u_j^p, outputs y_j, y_j^p, and state (x, v). To state a coordinate-free definition of the prolonged system (2.4) on the whole tangent space TM, we need to introduce the notions of vertical and complete lifts of functions and vector fields. We do this following [28]. Given a function V on M, the *complete lift* of V to TM, $V^{\mathrm{C}} : TM \to \mathbb{R}$, is defined by $V^{\mathrm{C}}(v) = \langle dV, v\rangle$. In the induced local coordinates on TM, $(x^1,\dots,x^n,v^1,\dots,v^n)$, this reads

$$V^{\mathrm{C}}(x,v) = \sum_{a=1}^{n} \frac{\partial V}{\partial x^a}(x)\, v_a\,.$$

The *vertical lift* of V to TM, $V^{\mathrm{V}} : TM \to \mathbb{R}$, is defined by $V^{\mathrm{V}} = V \circ \tau_M$, where τ_M denotes the tangent bundle projection. Given a vector field X on M, the *complete lift* of X to TM, $X^{\mathrm{C}} \in \mathfrak{X}(TM)$, is defined as the unique vector field verifying $X^{\mathrm{C}}(f^{\mathrm{C}}) = (Xf)^{\mathrm{C}}$ for any $f \in C^\omega(M)$. Alternatively, if $\Phi_t : M \to M$, $t \in [0,\epsilon)$, denotes the flow of X, then we can define X^{C} as the vector field whose flow is given by $(\Phi_t)_* : TM \to TM$. In local coordinates,

$$X^{\mathrm{C}}(x,v) = \sum_{a=1}^{n} X_a(x)\frac{\partial}{\partial x^a} + \sum_{a,b=1}^{n} \frac{\partial X_a}{\partial x^b}(x)v^b\frac{\partial}{\partial v^a}\,. \tag{2.5}$$

The *vertical lift* of X to TM, $X^{\mathrm{V}} \in \mathfrak{X}(TM)$, is the unique vector field such that $X^{\mathrm{V}}(f^{\mathrm{C}}) = (Xf)^{\mathrm{V}}$ for any $f \in C^\omega(M)$. In local coordinates,

$$X^{\mathrm{V}}(x,v) = \sum_{a=1}^{n} X_a(x)\frac{\partial}{\partial v^a}\,. \tag{2.6}$$

The following definition provides an intrinsic way of pasting together the system (2.1) with the variational systems associated with its input-state-output trajectories.

DEFINITION 2.1. *The prolonged system Σ^p of a nonlinear system Σ of the form* (2.1) *is defined by*

$$\Sigma^p : \begin{cases} \dot{x}_p = g_0^{\mathrm{C}}(x_p) + \displaystyle\sum_{j=1}^{m} u_j(t) g_j^{\mathrm{C}}(x_p) + \sum_{j=1}^{m} u_j^p(t) g_j^{\mathrm{V}}(x_p), \\ y_j = V_j^{\mathrm{V}}(x_p)\,, \quad y_j^p = V_j^{\mathrm{C}}(x_p)\,, \quad j = 1,\dots,m\,, \end{cases} \tag{2.7}$$

where $x_p = (x,v) \in TM$ and $x_p(0) = (x_0, v_0)$.

One can easily check that in the induced tangent bundle coordinates, the local expression of the system (2.7) is precisely (2.4).

Remark 2.2. In the same way as we have presented above, one can also introduce the notions of adjoint variational system and Hamiltonian extension of the nonlinear system (2.1). These notions play a key role in the characterization of when a general system admits a Hamiltonian description; see [13].

2.2. The gradient extension of a nonlinear system. When dealing with the Hamiltonian extension of a nonlinear system, one relies on the fact that the cotangent bundle is endowed with a canonical symplectic structure. However, this is not the case when treating gradient systems, since a canonical pseudo-Riemannian structure on the cotangent bundle does not exist. In order to define the gradient extension of a nonlinear system of the form (2.1), we will first select a torsion-free affine connection ∇ on M, and then consider its Riemannian extension to T^*M (cf. [19]).

Let us briefly present some basic notions on affine connections and Riemannian geometry. An *affine connection* [16] on a manifold M is defined as an assignment

$$\begin{aligned} \nabla : \mathfrak{X}(M) \times \mathfrak{X}(M) &\longrightarrow \mathfrak{X}(M), \\ (X, Y) &\longmapsto \nabla_X Y \end{aligned}$$

which is $\mathbb{R}$-bilinear and satisfies $\nabla_{fX} Y = f \nabla_X Y$ and $\nabla_X (fY) = f \nabla_X Y + X(f) Y$ for any X, $Y \in \mathfrak{X}(M)$, $f \in C^\omega(M)$. This implies that $\nabla_X Y(x)$ depends only on $X(x)$ and the value of Y along a curve which is tangent to X at x. Let $c : t \in [t_0, t_1] \mapsto c(t) = (x^1(t), \dots, x^n(t)) \in M$ be a curve on M and W a vector field along c, i.e., a map $W \colon [t_0, t_1] \to TM$ such that $\tau_M(W(t)) = c(t)$ for all $t \in [a, b]$. Let V be a vector field that satisfies $V(c(t)) = W(t)$. The *covariant derivative of* W *along* c is defined by

$$\frac{DW(t)}{dt} = \nabla_{\dot{c}(t)} W(t) = \nabla_{\dot{c}(t)} V(x)\big|_{x=c(t)}.$$

This definition makes sense because of the defining properties of the affine connection. Now, we may take $W(t) = \dot{c}(t)$ and set up $\nabla_{\dot{c}(t)} \dot{c}(t) = 0$. This equation is called the *geodesic equation*, and its solutions are termed the *geodesics* of ∇. In local coordinates, this condition can be expressed as $\ddot{x}^a + \Gamma^a_{bc}(x) \dot{x}^b \dot{x}^c = 0$, $1 \le a \le n$, where the $\Gamma^a_{bc}(x)$ are the *Christoffel symbols* of the affine connection, defined by

$$\nabla_{\frac{\partial}{\partial x^b}} \frac{\partial}{\partial x^c} = \Gamma^a_{bc}(x) \frac{\partial}{\partial x^a} \,.$$

The vector field S on TM describing the geodesic equation is called the *geodesic spray* associated with the affine connection ∇. In local coordinates,

$$S = v^a \frac{\partial}{\partial x^a} - \Gamma^a_{bc}(x) v^b v^c \frac{\partial}{\partial v^a} \,.$$

Therefore, the integral curves of the geodesic spray S are the solutions of the geodesic equation. The torsion tensor of an affine connection is defined by

$$\begin{aligned} T : \mathfrak{X}(M) \times \mathfrak{X}(M) &\longrightarrow \mathfrak{X}(M), \\ (X, Y) &\longmapsto \nabla_X Y - \nabla_Y X - [X, Y] \,. \end{aligned}$$

Locally, we have

$$T\left(\frac{\partial}{\partial x^a}, \frac{\partial}{\partial x^b} \right) = \left(\Gamma^c_{ab} - \Gamma^c_{ba} \right) \frac{\partial}{\partial x^c} \,.$$

An affine connection is *torsion-free* if T is identically zero. Given an affine connection, the *symmetric product* [18] of two vector fields $X, Y \in \mathfrak{X}(M)$ is defined by the operation

$$\langle X : Y \rangle = \nabla_X Y + \nabla_Y X \,.$$

The geometric meaning of the symmetric product is the following [17]: a distribution $\mathcal{D}$ on M is geodesically invariant (meaning that each geodesic of ∇ whose initial velocity is in $\mathcal{D}$ has all its velocities in $\mathcal{D}$) if and only if $\langle X : Y \rangle \in \mathcal{D}$ for all X, $Y \in \mathcal{D}$. The symmetric product plays a crucial role within the so-called affine connection formalism for mechanical control systems in the study of a variety of aspects such as controllability, series expansions, motion planning, and optimal control [7]. Note that if the affine connection ∇ is torsion-free, then $\nabla_X Y = \frac{1}{2}(\langle X : Y \rangle + [X, Y])$ for all $X, Y \in \mathfrak{X}(M)$, i.e., there is a one-to-one correspondence between the covariant derivative and the symmetric product.

Associated with the metric $\mathcal{G}$ there is a natural affine connection called the *Levi-Civita* connection. The Levi-Civita connection $\nabla^{\mathcal{G}}$ is determined by the formula

$$\begin{aligned} 2\,\mathcal{G}(\nabla^{\mathcal{G}}_X Y, Z) = {} & X(\mathcal{G}(Y,Z)) + Y(\mathcal{G}(Z,X)) - Z(\mathcal{G}(X,Y)) \\ & + \mathcal{G}(Y,[Z,X]) - \mathcal{G}(X,[Y,Z]) + \mathcal{G}(Z,[X,Y])\,, \quad X,Y,Z \in \mathfrak{X}(M)\,. \end{aligned}$$

One can compute the Christoffel symbols of $\nabla^{\mathcal{G}}$ to be

$$\Gamma^a_{bc} = \frac{1}{2}\mathcal{G}^{ad}\left(\frac{\partial \mathcal{G}_{db}}{\partial x^c} + \frac{\partial \mathcal{G}_{dc}}{\partial x^b} - \frac{\partial \mathcal{G}_{bc}}{\partial x^d}\right). \tag{2.8}$$

The Levi-Civita connection is torsion-free, that is, $T(X, Y) = 0$ for any X, $Y \in \mathfrak{X}(M)$.

Therefore, a pseudo-Riemannian metric on M defines a unique affine torsion-free connection on M. The converse is, however, not true. Also, note that given an affine torsion-free connection which is the Levi-Civita connection corresponding to some pseudo-Riemannian metric, then there exist many more metrics that give rise to the same affine connection. For instance, any constant metric on the Euclidean space gives rise to the affine connection with Christoffel symbols $\Gamma^a_{bc} = 0$, $1 \le a, b, c \le n$.

Given a pseudo-Riemannian metric $\mathcal{G}$ on M, we can define the so-called Beltrami bracket [10, 22] of functions on M,

$$\{f : g\}_{\mathcal{G}} = \mathcal{G}(\mathrm{grad}_{\mathcal{G}}\, f, \mathrm{grad}_{\mathcal{G}}\, g)\,, \quad f, g \in C^\omega(M)\,.$$

In local coordinates, one has the expression

$$\{f : g\}_{\mathcal{G}} = \frac{\partial f}{\partial x^a}\mathcal{G}^{ab}\frac{\partial g}{\partial x^b}\,.$$

It is interesting to note that the mapping

$$\mathrm{grad}_{\mathcal{G}} : (C^\omega(M), \{\cdot : \cdot\}_{\mathcal{G}}) \to (\mathfrak{X}(M), \langle \cdot : \cdot \rangle_{\nabla^{\mathcal{G}}})$$

is a homomorphism of symmetric algebras, i.e., $\mathrm{grad}_{\mathcal{G}}\{f : g\}_{\mathcal{G}} = \langle \mathrm{grad}_{\mathcal{G}}\, f : \mathrm{grad}_{\mathcal{G}}\, g \rangle_{\nabla^{\mathcal{G}}}$ for all $f, g \in C^\omega(M)$.

Remark 2.3. The latter observation is the gradient analogue of the following fact in the Hamiltonian setting: consider the mapping $(C^\omega(M), \{\cdot, \cdot\}) \to (\mathfrak{X}(M), [\cdot, \cdot])$ (where $\{\cdot, \cdot\}$ denotes the Poisson bracket and $[\cdot, \cdot]$ denotes the Lie bracket) associating to each function f its Hamiltonian vector field X_f. Then this mapping is a homomorphism of Lie algebras, i.e., $X_{\{f,g\}} = [X_f, X_g]$.

Let us now turn our discussion to the cotangent bundle of M. First, we introduce the construction that associates to each vector field X on M a function V^X on T^*M, defined by $V^X(p) = \langle p, X\rangle$. In the induced local coordinates $(x^1, \dots, x^n, p_1, \dots, p_n)$ on T^*M, this reads $V^X(x,p) = \sum_{a=1}^n p_a X_a(x)$. The *complete lift* of X to T^*M, $X^{\mathrm{C}} \in \mathfrak{X}(T^*M)$, is defined as the Hamiltonian vector field (with respect to the canonical symplectic form on T^*M) associated with the function V^X. In local coordinates,

$$X^{\mathrm{C}}(x,p) = \sum_{a=1}^n X_a \frac{\partial}{\partial x^a} - \sum_{a,b=1}^n \frac{\partial X_b}{\partial x^a}(x) p_b \frac{\partial}{\partial p_a}.$$

The notion of *vertical lift* of a function V on M to a function V^{V} on T^*M is given by $V^{\mathrm{V}} = V \circ \pi_M$, where π_M is the cotangent bundle projection. An object which will play a key role in the subsequent discussion is the *Riemannian extension* [19, 28] of a torsion-free affine connection. Let ∇ be a torsion-free affine connection on M. Then, ∇ defines a pseudo-Riemannian metric on T^*M, denoted $\mathcal{G}^\nabla$, as the unique (0,2)-tensor on T^*M which satisfies

$$\mathcal{G}^\nabla(X^{\mathrm{C}}, Y^{\mathrm{C}}) = -V^{\langle X:Y\rangle}\,.$$

The fact that this single equality completely determines the Riemannian extension $\mathcal{G}^\nabla$ is a consequence of the result in Proposition 4.2 in [28, Chapter VII], which asserts that any $(0,s)$-tensor field on T^*M is univocally defined by its action on the complete lifts of vector fields of M. The matrix representations of the musical isomorphisms defined by $\mathcal{G}^\nabla$ in local coordinates are given by

$$\flat_{\mathcal{G}^\nabla} \equiv \begin{pmatrix} -2p_c\Gamma^c_{ab} & I_n \\ I_n & 0 \end{pmatrix}, \quad \sharp_{\mathcal{G}^\nabla} \equiv \begin{pmatrix} 0 & I_n \\ I_n & 2p_c\Gamma^c_{ab} \end{pmatrix}. \tag{2.9}$$

As for the gradient vector fields associated with the functions V^X, $V^{\mathrm{V}} \in C^\omega(T^*M)$, $X \in \mathfrak{X}(M)$, $V \in C^\omega(M)$, one has the local expressions

$$\operatorname{grad}_{\mathcal{G}^\nabla} V^X = X^a \frac{\partial}{\partial x^a} + p_a \left(\frac{\partial X^a}{\partial x^b} + 2\Gamma^a_{bc} X^c \right) \frac{\partial}{\partial p_b}, \quad \operatorname{grad}_{\mathcal{G}^\nabla} V^{\mathrm{V}} = \frac{\partial V}{\partial x^a} \frac{\partial}{\partial p_a}. \tag{2.10}$$

Given a metric $\mathcal{G}$ on M, one can also verify that the pseudo-Riemannian metric on T^*M defined by $\mathcal{G}^{\nabla^{\mathcal{G}}}$ corresponds to the pullback by $\sharp_{\mathcal{G}}$ of the complete lift $\mathcal{G}^{\mathrm{C}}$ to TM of $\mathcal{G}$, i.e., $\mathcal{G}^{\nabla^{\mathcal{G}}} = \sharp^*_{\mathcal{G}}(\mathcal{G}^{\mathrm{C}})$ (see [28]).

Definition 2.4. *The gradient extension Σ^e of a nonlinear system Σ of the form* (2.1) *with respect to a torsion-free affine connection ∇ on M is given by*

$$\Sigma^e : \begin{cases} \dot{x}_e = \operatorname{grad}_{\mathcal{G}^\nabla} V^{g_0}(x_e) + \sum_{j=1}^m u_j(t) \operatorname{grad}_{\mathcal{G}^\nabla} V^{g_j}(x_e) + \sum_{j=1}^m u^e_j(t) \operatorname{grad}_{\mathcal{G}^\nabla} V^{\mathrm{V}}_j(x_e), \\ y_j = V^{\mathrm{V}}_j(x_e), \quad y^e_j = V^{g_j}(x_e), \quad j = 1, \dots, m, \end{cases} \tag{2.11}$$

where $x_e = (x,p) \in T^*M$, $x_e(0) = (x_0, p_0)$, $u = (u_1, \dots, u_m) \in U \subset \mathbb{R}^m$, and $u^e = (u^e_1, \dots, u^e_m) \in \mathbb{R}^m$.

Remark 2.5. Note that the gradient extension Σ^e is itself a gradient control system.

3. Observability of the prolongation and the gradient extension. In this section, we investigate the observability properties of the prolonged system and the gradient extension of a nonlinear system. Roughly speaking, the observability properties of a given system determine to what extent one can *observe* the actual state of the system from its input-output behavior, i.e., to what extent the knowledge of the input-output response allows us to infer things about the evolution of the state. This study will later be key in establishing the characterization of when a nonlinear control system can be written as a gradient control system.

We start by briefly reviewing some notions such as distinguishable points and local observability. Let $\mathcal{Y}$ denote the space of absolutely continuous functions defined on $\mathbb{R}_+ = [0, +\infty)$ with values in $\mathbb{R}^m$. For a nonlinear system of the form (2.1), the *input-output map* $\mathcal{R}_\Sigma : M \times \mathcal{U} \to \mathcal{Y}$, $\mathcal{R}_\Sigma(x_0, u(\cdot)) = y(\cdot)$ is defined by assigning to each initial condition $x_0 \in M$ and any admissible control $u \in \mathcal{U}$ the output of the system

$$y(\cdot) = (V_1(x(\cdot, x_0, u(\cdot))), \ldots, V_m(x(t, x_0, u(\cdot)))),$$

where $x(\cdot, x_0, u(\cdot))$ denotes the solution of $\dot{x}(t) = g_0(x(t)) + \sum_{j=1}^m u_j(t) g_j(x(t))$ starting at x_0. Now, two points x_1, $x_2 \in M$ are said to be *indistinguishable*, $x_1 \sim x_2$, if $\mathcal{R}_\Sigma(x_1, u(\cdot)) = \mathcal{R}_\Sigma(x_2, u(\cdot))$ for any $u(\cdot) \in \mathcal{U}$.

DEFINITION 3.1. *A system Σ is* observable *if for any x_1, $x_2 \in M$, one has that $x_1 \sim x_2 \Rightarrow x_1 = x_2$. Alternatively, for any $x_1 \neq x_2$, there exists an admissible control such that the output functions resulting from the initial conditions $x(0) = x_1$, respectively, $x(0) = x_2$, are different. The system is* locally observable *at x_0 if there exists a neighborhood $\mathcal{N}$ of x_0 such that this holds for points in $\mathcal{N}$.*

Denote by $\mathcal{H}$ the $\mathbb{R}$-linear space in $C^\omega(M)$ spanned by the functions of the form $\mathcal{L}_{X_1}\mathcal{L}_{X_2}\cdots\mathcal{L}_{X_s}V_j$, with $\{X_r\}_{r=1}^s \subset \{g_i \mid i = 0, 1, \ldots, m\}$, and $j \in \{1, \ldots, m\}$. Alternatively, we may take X_r to be arbitrary elements of the accessibility algebra corresponding to the vector fields $g_0, g_1, \ldots, g_m$. $\mathcal{H}$ is called the *observation space* of Σ. It follows from the analyticity assumption that the system is observable if and only if $\mathcal{H}$ distinguishes points in M, i.e., for every x_1, $x_2 \in M$ with $x_1 \neq x_2$, there exists $V \in \mathcal{H}$ such that $V(x_1) \neq V(x_2)$; cf. [14].

PROPOSITION 3.2 (see [13]). *Consider a nonlinear system Σ of the form (2.1), with observation space $\mathcal{H}$. Then, the observation space $\mathcal{H}^p$ of the prolongation Σ^p is given by $\mathcal{H}^p = \mathcal{H}^{\mathrm{C}} + \mathcal{H}^{\mathrm{V}}$, where $\mathcal{H}^{\mathrm{C}} = \{V^{\mathrm{C}} \mid V \in \mathcal{H}\}$ and $\mathcal{H}^{\mathrm{V}} = \{V^{\mathrm{V}} \mid V \in \mathcal{H}\}$.*

The following corollary is a modified statement of Corollary 3.3 in [13].

COROLLARY 3.3. *Assume the codistribution $d\mathcal{H}$ is of constant rank. Then the system Σ is (locally) observable if and only if its prolongation is (locally) observable.*

Proof. Following [14], Σ is locally observable if and only if $\mathrm{rk}(d\mathcal{H}) = \dim M$. In addition, the codistribution $d\mathcal{H}$ on M has constant rank if and only if the codistribution $d\mathcal{H}^p$ on TM has constant rank. Therefore, $\mathrm{rk}(d\mathcal{H}) = \dim M$ if and only if $\mathrm{rk}(d\mathcal{H}^p) = \dim TM$ if and only if Σ^p is locally observable. The statement regarding observability is proved as in Corollary 3.3 in [13]. ∎

Let us turn our attention to the observability properties of the gradient extension of a nonlinear system of the form (2.1). The following lemma will be most helpful.

LEMMA 3.4. *Let ∇ be a torsion-free affine connection on a manifold M, and let $\mathcal{G}^\nabla$ denote its Riemannian extension to T^*M. Then, for any vector fields X, $Y \in \mathfrak{X}(M)$, and any functions f, $g \in C^\omega(M)$, the following identities hold:*

(i) $(\mathrm{grad}_{\mathcal{G}^\nabla} V^X)(V^Y) = \{V^X : V^Y\}_{\mathcal{G}^\nabla} = V^{\langle X:Y\rangle} = -\mathcal{G}^\nabla(X^{\mathrm{C}}, Y^{\mathrm{C}})$;

(ii) $(\mathrm{grad}_{\mathcal{G}^\nabla} V^X)(f^{\mathrm{V}}) = (\mathrm{grad}_{\mathcal{G}^\nabla} f^{\mathrm{V}})(V^X) = \{V^X : f^{\mathrm{V}}\}_{\mathcal{G}^\nabla} = X(f)^{\mathrm{V}}$;

(iii) $(\mathrm{grad}_{\mathcal{G}^\nabla} f^{\mathrm{V}})(g^{\mathrm{V}}) = \{f^{\mathrm{V}} : g^{\mathrm{V}}\}_{\mathcal{G}^\nabla} = 0$.

Proof. The first equality in (i) is the definition of the Beltrami bracket associated with $\mathcal{G}^\nabla$. For the second one, we resort to the local expressions in (2.9) to compute

$$\begin{aligned}\{V^X : V^Y\}_{\mathcal{G}^\nabla} &= \left(p_a \frac{\partial X^a}{\partial x^b}, X^b\right) \begin{pmatrix} 0 & I \\ I & 2p_e\Gamma^e_{cd} \end{pmatrix} \left(p_a \frac{\partial Y^a}{\partial x^b}, Y^b\right)^T \\ &= p_a \left(\frac{\partial X^a}{\partial x^b} Y^b + \frac{\partial Y^a}{\partial x^b} X^b + 2\Gamma^a_{bc} X^b Y^c\right) = V^{\langle X:Y\rangle}\,.\end{aligned}$$

The third equality corresponds to the definition of $\mathcal{G}^\nabla$. The first and second equalities in (ii) follow again by definition. As for the third one, note that

$$\operatorname{grad}_{\mathcal{G}^\nabla} f^{\mathrm{V}}(V^X) = \frac{\partial f}{\partial x^a}\frac{\partial}{\partial p_a}\left(p_b X^b\right) = \frac{\partial f}{\partial x^a} X^a = X(f)^{\mathrm{V}}.$$

Finally, the equalities in (iii) are straightforward. ∎

Denote by S_0 the $\mathbb{R}$-linear space in $\mathfrak{X}(M)$ spanned by the vector fields of the form $\langle X_1 : \langle X_2 : \langle \cdots : \langle X_s : g_j\rangle\rangle \cdots \rangle\rangle$, with $\{X_r\}_{r=1}^s \subset \{g_i \mid i = 0, 1, \ldots, m\}$ and $j \in \{1, \ldots, m\}$. Alternatively, one can define S_0 as the smallest subspace of $\mathfrak{X}(M)$ such that (i) $g_1, \ldots, g_m \in S_0$ and (ii) if $X \in S_0$, then $\langle g_i : X\rangle \in S_0$ for all $i = 0, 1, \ldots, m$. We denote by $\mathcal{S}_0$ the distribution on M generated by the space S_0,

$$\mathcal{S}_0(x) = \operatorname{span}\{X(x) \mid X \in S_0\}\,, \quad x \in M\,. \tag{3.1}$$

PROPOSITION 3.5. *Consider a nonlinear system Σ of the form (2.1), with observation space $\mathcal{H}$. Let ∇ be a torsion-free affine connection on M. Then, the observation space $\mathcal{H}^e$ of the gradient extension Σ^e is given by $\mathcal{H}^e = V^{\mathcal{S}_0} + (\mathcal{H} + \mathfrak{h})^{\mathrm{V}}$, where $V^{\mathcal{S}_0} = \{V^X \mid X \in \mathcal{S}_0\}$ and $\mathfrak{h}$ is spanned by $\mathcal{L}_{X_1}\mathcal{L}_{X_2}\cdots\mathcal{L}_{X_s}\mathcal{L}_X V_j$, with X_r, $r = 1, \ldots, s$, equal to g_i, $i = 0, 1, \ldots, m$, $X \in \mathcal{S}_0$, and $j = 1, \ldots, m$.*

Proof. The observation space of the gradient extension of Σ is spanned by

$$\mathcal{L}_{X_1}\mathcal{L}_{X_2}\cdots\mathcal{L}_{X_s} V_j^{\mathrm{V}}\,, \quad \mathcal{L}_{X_1}\mathcal{L}_{X_2}\cdots\mathcal{L}_{X_s} V^{g_j}\,,$$

where X_r, $r = 1, \ldots, s$, is equal to $\operatorname{grad}_{\mathcal{G}^\nabla} V^{g_i}$, $\operatorname{grad}_{\mathcal{G}^\nabla} V_j^{\mathrm{V}}$, $i = 0, 1, \ldots, m$, $j = 1, \ldots, m$. Now, using Lemma 3.4, we have that

$$\begin{aligned} &\mathcal{L}_{\operatorname{grad}_{\mathcal{G}^\nabla} V^{g_i}} V_j^{\mathrm{V}} = (\mathcal{L}_{g_i} V_j)^{\mathrm{V}}\,, && \mathcal{L}_{\operatorname{grad}_{\mathcal{G}^\nabla} V^{g_i}} V^{g_j} = V^{\langle g_i : g_j\rangle}\,, \\ &\mathcal{L}_{\operatorname{grad}_{\mathcal{G}^\nabla} V_j^{\mathrm{V}}} V_k^{\mathrm{V}} = 0\,, && \mathcal{L}_{\operatorname{grad}_{\mathcal{G}^\nabla} V_j^{\mathrm{V}}} V^{g_k} = (\mathcal{L}_{g_k} V_j)^{\mathrm{V}}\,, \end{aligned}$$

with $i = 0, 1, \ldots, m$ and $j, k = 1, \ldots, m$. Considering the next step of Lie derivatives yields

$$\begin{aligned} &\mathcal{L}_{\operatorname{grad}_{\mathcal{G}^\nabla} V^{g_h}} V^{\langle g_i : g_j\rangle} = V^{\langle g_h : \langle g_i : g_j\rangle\rangle}\,, && \mathcal{L}_{\operatorname{grad}_{\mathcal{G}^\nabla} V^{g_h}} (\mathcal{L}_{g_i} V_j)^{\mathrm{V}} = (\mathcal{L}_{g_h}\mathcal{L}_{g_i} V_j)^{\mathrm{V}}\,, \\ &\mathcal{L}_{\operatorname{grad}_{\mathcal{G}^\nabla} V_k^{\mathrm{V}}} V^{\langle g_i : g_j\rangle} = \left(\mathcal{L}_{\langle g_i : g_j\rangle} V_k\right)^{\mathrm{V}}\,, && \mathcal{L}_{\operatorname{grad}_{\mathcal{G}^\nabla} V_k^{\mathrm{V}}} (\mathcal{L}_{g_i} V_j)^{\mathrm{V}} = 0\,, \end{aligned}$$

with $h = 0, 1, \ldots, m$. Further iterating this process, we get to the desired result. ∎

COROLLARY 3.6. *Consider a nonlinear system Σ of the form (2.1), with observation space $\mathcal{H}$. Assume the codistribution $d\mathcal{H}$ is of constant rank. Let ∇ be a torsion-free affine connection on M and further assume that the distribution $\mathcal{S}_0$ is full-rank. Then, that Σ is (locally) observable implies that Σ^e is (locally) observable.*

Proof. Since the codistribution $d\mathcal{H}$ has constant rank, Σ is locally observable if and only if $\dim d\mathcal{H}(x) = \dim M$. Since $\mathcal{S}_0$ is full-rank, it is clear that Σ locally

observable implies that $\mathcal{H}^e$ has constant maximal rank, and therefore Σ^e is locally observable. With respect to observability, let (x_1, p_1), $(x_2, p_2) \in T^*M$, and assume that $V^e(x_1, p_1) = V^e(x_2, p_2)$ for all $V^e \in \mathcal{H}^e$. Since $\mathcal{H}^{\mathrm{v}} \subset \mathcal{H}^e$, this yields $V(x_1) = V(x_2)$ for any $V \in \mathcal{H}$. So, under observability of Σ, we conclude that $x_1 = x_2 = x$. Then, we have that $V^X(x, p_1) = V^X(x, p_2)$ for all $X \in S_0$, which finally implies that $p_1 = p_2$. □

4. Externally equivalent systems. In this section we introduce the notion of (weakly) externally equivalent systems, which will be instrumental in the statement of the main result in section 5. Consider two nonlinear systems $\alpha = 1, 2$, of the form

$$\Sigma^\alpha : \begin{cases} \dot{x}^\alpha = g_0^\alpha(x^\alpha) + \displaystyle\sum_{j=1}^{m} u_j g_j^\alpha(x^\alpha), \quad x^\alpha \in M^\alpha, \\ y_j = V_j^\alpha(x^\alpha), \quad j = 1, \dots, m, \ u = (u_1, \dots, u_m) \in U \subset \mathbb{R}^m. \end{cases}$$

Denote by $\mathcal{H}^\alpha$, $\alpha = 1, 2$, the associated observation spaces. Take a function $H^1 \in \mathcal{H}^1$, $H^1 = \mathcal{L}_{X_1} \cdots \mathcal{L}_{X_s} V_j^1$, with $X_r = g_{i_r}^1$, $i_r \in \{0, 1, \dots, m\}$, $r = 1, \dots, s$, and $j \in \{1, \dots, m\}$. Consider the function in $\mathcal{H}^2$ defined by $H^2 = \mathcal{L}_{Y_1} \cdots \mathcal{L}_{Y_s} V_j^2$, with $Y_r = g_{i_r}^2$, $r = 1, \dots, s$. Then we say that H^1 and H^2 formally *correspond* to each other. This notion is useful in defining the concept of *weakly externally equivalent* systems.

DEFINITION 4.1. *The systems Σ^1 and Σ^2 are weakly externally equivalent if and only if for all $x^1 \in M^1$, there exists $x^2 \in M^2$ such that $H^1(x^1) = H^2(x^2)$ for all corresponding $H^1 \in \mathcal{H}^1$, $H^2 \in \mathcal{H}^2$, and reciprocally, for all $x^2 \in M^2$, there exists $x^1 \in M^1$ such that $H^1(x^1) = H^2(x^2)$ for all corresponding $H^1 \in \mathcal{H}^1$, $H^2 \in \mathcal{H}^2$.*

DEFINITION 4.2. *The systems Σ^1 and Σ^2 are externally equivalent if and only if for all $x^1 \in M^1$, there exists $x^2 \in M^2$ such that the input-output maps corresponding to x^1 and x^2 coincide, i.e., $\mathcal{R}_{\Sigma^1}(x^1, u(\cdot)) = \mathcal{R}_{\Sigma^2}(x^2, u(\cdot))$ for all $u(\cdot) \in \mathcal{U}$, and reciprocally, for all $x^2 \in M^2$, there exists $x^1 \in M^1$ such that $\mathcal{R}_{\Sigma^1}(x^1, u(\cdot)) = \mathcal{R}_{\Sigma^2}(x^2, u(\cdot))$ for all $u(\cdot) \in \mathcal{U}$.*

Equivalently, Σ^1 and Σ^2 are externally equivalent if and only if their behaviors are equal. Clearly, if two systems are externally equivalent, then they are weakly externally equivalent.

PROPOSITION 4.3. *Assume that Σ^1 and Σ^2 are weakly externally equivalent and observable and that the codistributions $d\mathcal{H}^\alpha$, $\alpha = 1, 2$, have constant rank. Then there exists a unique diffeomorphism $\varphi : M^1 \to M^2$ with $\varphi^*(\mathcal{H}^2) = \mathcal{H}^1$.*

Proof. Let $x^1 \in M^1$. By definition, there exists $x^2 \in M^2$ such that $H^1(x^1) = H^2(x^2)$ for all corresponding $H^1 \in \mathcal{H}^1$, $H^2 \in \mathcal{H}^2$. Since $\mathcal{H}^2$ distinguishes points in M^2, it follows that x^2 is unique. Define $\varphi : M^1 \to M^2$, $\varphi(x^1) = x^2$. Using $\dim d\mathcal{H}^2 = \dim M^2$ and the inverse function theorem, it follows that φ is smooth. Indeed, for each $x^2 \in M^2$, there exists a neighborhood V of M^2 at x_2 and $\dim M^2$ independent functions $H_1^2, \dots, H_{\dim M^2}^2$ on V such that φ is given by

$$x^2 = (H_1^2, \dots, H_{\dim M^2}^2)^{-1}(H_1^1, \dots, H_{\dim M^2}^1)(x^1).$$

Analogously, we can construct the inverse mapping $\varphi^{-1} : M^2 \to M^1$, making use of the fact that Σ_1 is observable, which concludes the proof. □

COROLLARY 4.4. *Let the systems Σ^1 and Σ^2 be observable and the codistributions $d\mathcal{H}^\alpha$, $\alpha = 1, 2$, have constant rank. Then Σ^1 and Σ^2 are weakly externally equivalent if and only if they are externally equivalent.*

Proof. We already know that if the systems are externally equivalent, then they are weakly externally equivalent. Conversely, assume that Σ^1 and Σ^2 are weakly externally equivalent. From Proposition 4.3, we have that there exists a diffeomorphism $\varphi : M^1 \to M^2$ with $\varphi^*(\mathcal{H}^2) = \mathcal{H}^1$. Using this latter fact, and since the vector fields g_0^i, g_j^i are determined by their action as derivations on $\mathcal{H}^\alpha$, $\alpha = 1, 2$, we conclude that $\varphi_* g_0^1 = g_0^2$, $\varphi_* g_j^1 = g_j^2$, $j = 1, \dots, m$. ∎

Remark 4.5. The map φ in the previous proof is called a *state-space diffeomorphism.*

5. Gradient realization of a nonlinear control system. This section contains the main result of the paper. Under certain technical conditions, Theorem 5.4 characterizes when a nonlinear control systems admits a gradient realization. Before stating this result, we need to introduce the novel notion of *compatibility* between a nonlinear system and an affine connection.

DEFINITION 5.1 (compatibility). *Let ∇ be an affine connection on M. A nonlinear control system Σ of the form* (2.1) *is compatible with ∇ if and only if the following two conditions hold:*

(a) *For all vector fields $X_1, \dots, X_{s_1}$, $Y_1, \dots, Y_{s_2} \in \{g_0, g_1, \dots, g_m\}$, and all indexes $j, k = 1, \dots, m$,*

$$\mathcal{L}_{\langle X_1 : \langle X_2 : \langle \cdots : \langle X_{s_1} : g_j \rangle \rangle \cdots \rangle \rangle} \left[\mathcal{L}_{Y_1} \mathcal{L}_{Y_2} \cdots \mathcal{L}_{Y_{s_2}} V_k \right] = \mathcal{L}_{\langle Y_1 : \langle Y_2 : \langle \cdots : \langle Y_{s_2} : g_k \rangle \rangle \cdots \rangle \rangle} \left[\mathcal{L}_{X_1} \mathcal{L}_{X_2} \cdots \mathcal{L}_{X_{s_1}} V_j \right].$$

(b) *For all vector fields $X_1, \dots, X_{s_1}$, $Y_1, \dots, Y_{s_2}$, $Z_1, \dots, Z_{s_3} \in \{g_0, g_1, \dots, g_m\}$, and all indexes $j, k, l = 1, \dots, m$,*

$$\mathcal{L}_{\langle \langle X_1 : \langle X_2 : \langle \cdots : \langle X_{s_1} : g_j \rangle \rangle \cdots \rangle \rangle : \langle Y_1 : \langle Y_2 : \langle \cdots : \langle Y_{s_2} : g_k \rangle \rangle \cdots \rangle \rangle \rangle} [\mathcal{L}_{Z_1} \mathcal{L}_{Z_2} \cdots \mathcal{L}_{Z_{s_3}} V_l] = \mathcal{L}_{\langle Z_1 : \langle Z_2 : \langle \cdots : \langle Z_{s_3} : g_l \rangle \rangle \cdots \rangle \rangle} [\mathcal{L}_{\langle X_1 : \langle X_2 : \langle \cdots : \langle X_{s_1} : g_j \rangle \rangle \cdots \rangle \rangle} [\mathcal{L}_{Y_1} \mathcal{L}_{Y_2} \cdots \mathcal{L}_{Y_{s_2}} V_k]].$$

Remark 5.2. In case the distribution $\mathcal{S}_0$ (cf. (3.1)) is full-rank, note that property (b) in the above definition implies property (a) up to a constant on each connected component of M. To see this, one can use the symmetry of the symmetric product to deduce from (b) that

$$\mathcal{L}_{\langle Z_1 : \langle Z_2 : \langle \cdots : \langle Z_{s_3} : g_l \rangle \rangle \cdots \rangle \rangle} [\mathcal{L}_{\langle X_1 : \langle X_2 : \langle \cdots : \langle X_{s_1} : g_j \rangle \rangle \cdots \rangle \rangle} [\mathcal{L}_{Y_1} \mathcal{L}_{Y_2} \cdots \mathcal{L}_{Y_{s_2}} V_k]] = \mathcal{L}_{\langle Z_1 : \langle Z_2 : \langle \cdots : \langle Z_{s_3} : g_l \rangle \rangle \cdots \rangle \rangle} [\mathcal{L}_{\langle Y_1 : \langle Y_2 : \langle \cdots : \langle Y_{s_2} : g_k \rangle \rangle \cdots \rangle \rangle} [\mathcal{L}_{X_1} \mathcal{L}_{X_2} \cdots \mathcal{L}_{X_{s_1}} V_j]].$$

Now, one concludes the result from the full-rankness of $\mathcal{S}_0$. Another interesting observation in this case is that the checkability of the compatibility condition can be performed taking a basis of vector fields in S_0, as we discuss later in Lemma 8.1.

Remark 5.3. Note that a locally gradient control system of the form (2.2) is compatible with the Levi-Civita connection associated with the pseudo-Riemannian metric $\mathcal{G}$. Indeed, let $\langle \cdot : \cdot \rangle$, $\{\cdot : \cdot\}$ denote, respectively, the symmetric product and the Beltrami bracket corresponding to $\nabla^{\mathcal{G}}$ and $\mathcal{G}$. Take $X_{r_1} = \operatorname{grad} V_{\alpha_{r_1}}$, $Y_{r_2} = \operatorname{grad} V_{\beta_{r_2}}$, $Z_{r_3} = \operatorname{grad} V_{\gamma_{r_3}}$, $r_i \in \{1, \dots, s_i\}$ (which can always be written at least locally); then

$$\begin{aligned}
&\mathcal{L}_{\langle X_1 : \langle X_2 : \langle \cdots : \langle X_{s_1} : g_j \rangle \rangle \cdots \rangle \rangle} \left[\mathcal{L}_{Y_1} \mathcal{L}_{Y_2} \cdots \mathcal{L}_{Y_{s_2}} V_k \right] \\
&\quad = \mathcal{L}_{\operatorname{grad}\{V_{\alpha_1} : \{V_{\alpha_2} : \{\cdots : \{V_{\alpha_{s_1}} : V_j\}\} \cdots\}\}} \left[\{V_{\beta_1} : \{V_{\beta_2} : \{\cdots : \{V_{\beta_{s_2}} : V_k\}\} \cdots\}\} \right] \\
&\quad = \mathcal{L}_{\operatorname{grad}\{V_{\beta_1} : \{V_{\beta_2} : \{\cdots : \{V_{\beta_{s_2}} : V_k\}\} \cdots\}\}} \left[\{V_{\alpha_1} : \{V_{\alpha_2} : \{\cdots : \{V_{\alpha_{s_1}} : V_j\}\} \cdots\}\} \right] \\
&\quad = \mathcal{L}_{\langle Y_1 : \langle Y_2 : \langle \cdots : \langle Y_{s_2} : g_k \rangle \rangle \cdots \rangle \rangle} \left[\mathcal{L}_{X_1} \mathcal{L}_{X_2} \cdots \mathcal{L}_{X_{s_1}} V_j \right]
\end{aligned}$$

and

$$\begin{aligned}
&\mathcal{L}_{\langle\langle X_1:\langle X_2:\langle\cdots:\langle X_{s_1}:g_j\rangle\rangle\cdots\rangle\rangle:\langle Y_1:\langle Y_2:\langle\cdots:\langle Y_{s_1}:g_k\rangle\rangle\cdots\rangle\rangle\rangle} \left[\mathcal{L}_{Z_1}\mathcal{L}_{Z_2}\cdots\mathcal{L}_{Z_{s_3}}V_l\right] \\
&\quad = \mathcal{L}_{\mathrm{grad}\{\{V_{\alpha_1}:\{V_{\alpha_2}:\{\cdots:\{V_{\alpha_{s_1}}:V_j\}\}\cdots\}\}:\{V_{\beta_1}:\{V_{\beta_2}:\{\cdots:\{V_{\beta_{s_2}}:V_k\}\}\cdots\}\}\}} \\
&\left[\{V_{\gamma_1}:\{V_{\gamma_2}:\{\cdots:\{V_{\gamma_{s_3}}:V_l\}\}\cdots\}\}\right] \\
&\quad = \mathcal{L}_{\mathrm{grad}\{V_{\gamma_1}:\{V_{\gamma_2}:\{\cdots:\{V_{\gamma_{s_3}}:V_l\}\}\cdots\}\}} \\
&\left[\{\{V_{\alpha_1}:\{V_{\alpha_2}:\{\cdots:\{V_{\alpha_{s_1}}:V_j\}\}\cdots\}\}:\{V_{\beta_1}:\{V_{\beta_2}:\{\cdots:\{V_{\beta_{s_2}}:V_k\}\}\cdots\}\}\}\right] \\
&\quad = \mathcal{L}_{\langle Z_1:\langle Z_2:\langle\cdots:\langle Z_{s_3}:g_l\rangle\rangle\cdots\rangle\rangle}[\mathcal{L}_{\langle X_1:\langle X_2:\langle\cdots:\langle X_{s_1}:g_j\rangle\rangle\cdots\rangle\rangle}[\mathcal{L}_{Y_1}\mathcal{L}_{Y_2}\cdots\mathcal{L}_{Y_{s_2}}V_k]]
\end{aligned}$$

as claimed.

Now, we come to the main result of the paper.

THEOREM 5.4. *Let Σ be a nonlinear control system of the form (2.1). Let ∇ be a torsion-free affine connection defined on the state manifold M. Assume that Σ is observable with $\dim d\mathcal{H}$ constant, compatible with ∇, and that the distribution $\mathcal{S}_0$ is full-rank. Then, Σ is a locally gradient control system with respect to a pseudo-metric whose Levi-Civita connection is ∇ if and only if its prolonged system Σ^p and its gradient extension Σ^e are weakly externally equivalent.*

Proof. Consider a locally gradient control system Σ on $(M,\mathcal{G})$ (cf. (2.2)), together with its prolongation Σ^p on TM and its gradient extension Σ^e on T^*M. Recall that in the induced bundle coordinates (x^a,v^a) on TM, (x^a,p_a) on T^*M, the musical isomorphisms associated with $\mathcal{G}$ read $\flat_{\mathcal{G}}(x^a,v^a)=(x^a,\mathcal{G}_{ab}v^b)$ and $\sharp_{\mathcal{G}}(x^a,p_a)=(x^a,\mathcal{G}^{ab}p_b)$. We are going to show that $\flat_{\mathcal{G}}$ is actually an isomorphism between the prolongation and the gradient extension, i.e., we will prove that $\flat_{\mathcal{G}}(x_p(\cdot))=x_e(\cdot)$ along the solutions of (2.7) and (2.11), respectively. This will be a consequence of the following equalities:

$$\begin{aligned}
&(\flat_{\mathcal{G}})_* g_i^{\mathrm{C}} = \mathrm{grad}_{\mathcal{G}^\nabla} V^{g_i}\circ\flat_{\mathcal{G}}\,, && V^{g_j}\circ\flat_{\mathcal{G}} = V_j^{\mathrm{C}}\,, \\
&(\flat_{\mathcal{G}})_* g_j^{\mathrm{V}} = \mathrm{grad}_{\mathcal{G}^\nabla} V_j^{\mathrm{V}}\circ\flat_{\mathcal{G}}\,, && V_j^{\mathrm{V}}\circ\flat_{\mathcal{G}} = V_j^{\mathrm{V}}
\end{aligned} \tag{5.1}$$

for all $i=0,1,\ldots,m$, $j=1,\ldots,m$. In order to show (5.1), we will make use of the following identities:

$$(\flat_{\mathcal{G}})_*\left(\frac{\partial}{\partial x^a}\right) = \frac{\partial}{\partial x^a} + \frac{\partial\mathcal{G}_{cb}}{\partial x^a}v^b\frac{\partial}{\partial p_c}\,, \quad (\flat_{\mathcal{G}})_*\left(\frac{\partial}{\partial v^a}\right) = \mathcal{G}_{ab}\frac{\partial}{\partial p_b}\,.$$

Let $g\in\mathfrak{X}(M)$. In local coordinates, $g=g^a\partial/\partial x^a$. Using (2.5), we get

$$(\flat_{\mathcal{G}})_*(g^{\mathrm{C}}) = g^a\frac{\partial}{\partial x^a} + \left\{g^c\frac{\partial\mathcal{G}_{ab}}{\partial x^c} + \mathcal{G}_{ac}\frac{\partial g^c}{\partial x^b}\right\}v^b\frac{\partial}{\partial p_a}\,.$$

On the other hand, we have that

$$\mathrm{grad}_{\mathcal{G}^\nabla} V^g\circ\flat_{\mathcal{G}} = g^a\frac{\partial}{\partial x^a} + \left\{\mathcal{G}_{bc}\frac{\partial g^c}{\partial x^a} + 2\mathcal{G}_{bc}\Gamma^c_{ad}g^d\right\}v^b\frac{\partial}{\partial p_a}\,.$$

Now, suppose that g is a locally gradient vector field. In local coordinates, this means that $\mathcal{G}_{ac}g^c=\partial V/\partial x^a$, for a certain function V, which in turn implies that $\partial\{\mathcal{G}_{ac}g^c\}/\partial x^b=\partial\{\mathcal{G}_{bc}g^c\}/\partial x^a$, that is,

$$\mathcal{G}_{ac}\frac{\partial g^c}{\partial x^b} = \frac{\partial\mathcal{G}_{bc}}{\partial x^a}g^c + \mathcal{G}_{bc}\frac{\partial g^c}{\partial x^a} - \frac{\partial\mathcal{G}_{ac}}{\partial x^b}g^c\,.$$

Substituting into the above expression for $(\flat_{\mathcal{G}})_*(g^{\mathrm{C}})$,

$$\begin{aligned}(\flat_{\mathcal{G}})_*(g^{\mathrm{C}}) &= g^a\frac{\partial}{\partial x^a}+\left\{g^c\frac{\partial \mathcal{G}_{ab}}{\partial x^c}+\frac{\partial \mathcal{G}_{bc}}{\partial x^a}g^c+\mathcal{G}_{bc}\frac{\partial g^c}{\partial x^a}-\frac{\partial \mathcal{G}_{ac}}{\partial x^b}g^c\right\}v^b\frac{\partial}{\partial p_a}\\ &= g^a\frac{\partial}{\partial x^a}+\left\{g^c\left(\frac{\partial \mathcal{G}_{ab}}{\partial x^c}+\frac{\partial \mathcal{G}_{bc}}{\partial x^a}-\frac{\partial \mathcal{G}_{ac}}{\partial x^b}\right)+\mathcal{G}_{bc}\frac{\partial g^c}{\partial x^a}\right\}v^b\frac{\partial}{\partial p_a}\\ &= g^a\frac{\partial}{\partial x^a}+\left\{2g^c\mathcal{G}_{bd}\Gamma^d_{ac}+\mathcal{G}_{bc}\frac{\partial g^c}{\partial x^a}\right\}v^b\frac{\partial}{\partial p_a}=\mathrm{grad}_{\mathcal{G}^\nabla}V^g\circ\flat_{\mathcal{G}}\,.\end{aligned}$$

Therefore, the first equality in (5.1) holds for every $i=0,1,\dots,m$. The equality $(\flat_{\mathcal{G}})_* g_j^{\mathrm{V}}=\mathrm{grad}_{\mathcal{G}^\nabla}V_j^{\mathrm{V}}\circ\flat_{\mathcal{G}}$, $j=1,\dots,m$, follows by considering (2.6) and the fact that the vector fields g_j are gradient by hypothesis,

$$(\flat_{\mathcal{G}})_*(g^{\mathrm{V}})=\mathcal{G}_{ab}g^b\frac{\partial}{\partial p_a}=\frac{\partial V}{\partial x^a}\frac{\partial}{\partial p_a}=\mathrm{grad}_{\mathcal{G}^\nabla}V^{\mathrm{V}}\circ\flat_{\mathcal{G}}\,.$$

As for $V^{g_j}\circ\flat_{\mathcal{G}}=V_j^{\mathrm{C}}$, for each $v\in T_xM$, we compute $V^{g_j}\circ\flat_{\mathcal{G}}(v)=\mathcal{G}_{ab}v^bg_j^a=\partial V_j/\partial x^b\cdot v^b=\langle dV_j,v\rangle=V_j^{\mathrm{C}}(v)$, where, with a slight abuse of notation, we have used the bracket $\langle\cdot,\cdot\rangle$ to denote the contraction between a covector and a vector. In the remainder of the paper, the intended use of $\langle\,.\,,\,.\,\rangle$ should be clear from the context. The last equality follows trivially. Consequently, the prolongation and the gradient extension of a nonlinear system Σ which is itself gradient are externally equivalent, in particular weakly externally equivalent systems.

To prove the converse implication, we need some intermediate steps that we describe in what follows. □

LEMMA 5.5. *Let Σ be a nonlinear system of the form* (2.1)*. Under the hypothesis of Theorem* 5.4*, assume that the prolongation Σ^p and the gradient extension Σ^e are weakly externally equivalent. Then there exists a unique diffeomorphism $\varphi:TM\to T^*M$ such that*

$$\begin{aligned}&(\varphi)_*g_i^{\mathrm{C}}=\mathrm{grad}_{\mathcal{G}^\nabla}V^{g_i}\circ\varphi\,, && V^{g_j}\circ\varphi=V_j^{\mathrm{C}}\,,\\ &(\varphi)_*g_j^{\mathrm{V}}=\mathrm{grad}_{\mathcal{G}^\nabla}V_j^{\mathrm{V}}\circ\varphi\,, && V_j^{\mathrm{V}}\circ\varphi=V_j^{\mathrm{V}}\end{aligned}\tag{5.2}$$

for all $i=0,1,\dots,m$, $j=1,\dots,m$. Moreover, φ is a bundle morphism over the identity $\mathrm{Id}_M:M\to M$, i.e., in natural coordinates $\varphi(x,v)=(x,\phi(x,v))$, for certain map $\phi:T_xM\to T_x^*M$, $x\in M$.

Proof. By Proposition 3.2 and Corollary 3.6, we have that both the prolongation and the gradient extension are observable systems. Since they are also weakly externally equivalent by assumption, Corollary 4.4 ensures that there exists a unique diffeomorphism $\varphi:TM\to T^*M$ verifying (5.2). Applying now Corollary 4.4 to $\Sigma^1=\Sigma=\Sigma^2$, we deduce that there exists a unique diffeomorphism from M to M mapping the original nonlinear system to itself, namely, the identity mapping. Using uniqueness and the fact that φ satisfies (5.2), it then follows that φ is of the form $\varphi(x,v)=(x,\phi(x,v))$, for certain map $\phi:T_xM\to T_x^*M$, $x\in M$. □

LEMMA 5.6. *Under the same assumptions as in Lemma* 5.5*, there exists a unique pseudo-Riemannian metric $\mathcal{G}$ on M such that $\flat_{\mathcal{G}}=\varphi$, i.e., $\flat_{\mathcal{G}}(v)=\phi(x,v)$ for all $v\in T_xM$.*

Proof. It follows from $V^{g_j}\circ\varphi=V_j^{\mathrm{C}}$ (cf. (5.2)) and the structure of the diffeomorphism φ that

$$\langle\phi(x,v),g_j(x)\rangle=\langle dV_j(x),v\rangle\quad\forall v\in T_xM\,,\ j=1,\dots,m\,.$$

Furthermore, from $(\varphi)_* g_i^{\mathrm{C}} = \operatorname{grad} V^{g_i} \circ \varphi$ (see (5.2)), it follows that

$$\mathcal{L}_{\operatorname{grad} V^{g_i}} V^{g_j} \circ \varphi = \mathcal{L}_{g_i^{\mathrm{c}}} V_j^{\mathrm{C}}, \quad i = 0, 1, \ldots, m, \; j = 1, \ldots, m.$$

Using now Lemma 3.4(i), we get $\langle \phi(x, v), \langle g_i : g_j \rangle(x) \rangle = \langle d\left(\mathcal{L}_{g_i} V_j\right)(x), v \rangle$. In general for all $v \in T_x M$,

$$\begin{aligned} &\langle \phi(x, v), \langle X_1 : \langle X_2 : \langle X_3, \cdots : \langle X_s : g_j \rangle \rangle \cdots \rangle \rangle(x) \rangle \\ &\quad = \langle d\left(\mathcal{L}_{X_1} \mathcal{L}_{X_2} \cdots \mathcal{L}_{X_s} V_j\right)(x), v \rangle, \end{aligned} \tag{5.3}$$

with the X_r, $r = 1, \ldots, s$, equal to some g_i, $i = 0, 1, \ldots, m$. Since the right-hand side of this equation is linear in v and the distribution generated by the space S_0 is full-rank by hypothesis, it follows that for each $x \in M$ there exists a unique matrix $\mathcal{G}(x)$ such that $\phi(x, v) = \mathcal{G}(x) v$. Since φ is a diffeomorphism, $\mathcal{G}(x)$ is nonsingular for every x and depends smoothly on the base point. Consider the adjoint mapping of φ, $\varphi^T : TM \to T^*M$, defined by $\langle \varphi(v), w \rangle = \langle v, \varphi^T(w) \rangle$, $v, w \in T_x M$, $x \in M$. Then, $\varphi^T(x, v) = (x, \mathcal{G}^T(x) v)$. It follows from (5.3) that $\mathcal{G}(x)$ satisfies

$$\varphi^T(\langle X_1 : \langle X_2 : \langle X_3, \cdots : \langle X_s : g_j \rangle \rangle \cdots \rangle \rangle(x)) = d\left(\mathcal{L}_{X_1} \mathcal{L}_{X_2} \cdots \mathcal{L}_{X_s} V_j\right)(x), \tag{5.4}$$

with the X_r as above. Let us see now that $\mathcal{G}(x) = \mathcal{G}^T(x)$. Note that in local coordinates $(\varphi)_* g_j^{\mathrm{V}} = \operatorname{grad}_{\mathcal{G}^\nabla} V_j^{\mathrm{V}} \circ \varphi$ yields

$$\begin{pmatrix} I & 0 \\ \frac{\partial}{\partial x}\left(\mathcal{G}(x) v\right) & \mathcal{G}(x) \end{pmatrix} \begin{pmatrix} 0 \\ g_j(x) \end{pmatrix} = \begin{pmatrix} 0 \\ \left(\frac{\partial V_j}{\partial x}\right)^T (x) \end{pmatrix}$$

or, equivalently, $\mathcal{G}(x) g_j(x) = \left(\partial V_j / \partial x\right)^T (x)$, $j = 1, \ldots, m$, which in intrinsic terms, can be written as $\varphi(g_j) = dV_j$. Now,

$$\begin{aligned} &\langle \varphi\left(\langle X_1 : \langle X_2 : \langle X_3, \cdots : \langle X_{s_1} : g_j \rangle \rangle \cdots \rangle \rangle\right), \langle Y_1 : \langle Y_2 : \langle Y_3, \cdots : \langle Y_{s_2} : g_k \rangle \rangle \cdots \rangle \rangle \rangle \\ &= \langle \langle X_1 : \langle X_2 : \langle X_3, \cdots : \langle X_{s_1} : g_j \rangle \rangle \cdots \rangle \rangle, \varphi^T\left(\langle Y_1 : \langle Y_2 : \langle Y_3, \cdots : \langle Y_{s_2} : g_k \rangle \rangle \cdots \rangle \rangle\right) \rangle. \end{aligned}$$

Using (5.4), the latter is equal to

$$\begin{aligned} &\langle \langle X_1 : \langle X_2 : \langle X_3, \cdots : \langle X_{s_1} : g_j \rangle \rangle \cdots \rangle \rangle, d\mathcal{L}_{Y_1} \mathcal{L}_{Y_2} \cdots \mathcal{L}_{Y_{s_2}} V_k \rangle \\ &\quad = \langle d\mathcal{L}_{X_1} \mathcal{L}_{X_2} \cdots \mathcal{L}_{X_{s_1}} V_j, \langle Y_1 : \langle Y_2 : \langle \cdots : \langle Y_{s_2} : g_k \rangle \rangle \cdots \rangle \rangle \rangle, \end{aligned}$$

where in the last equality we have used the property (a) of the compatibility definition between the nonlinear system Σ and the affine connection ∇. Finally,

$$\begin{aligned} &\langle \varphi\left(\langle X_1 : \langle X_2 : \langle X_3, \cdots : \langle X_{s_1} : g_j \rangle \rangle \cdots \rangle \rangle\right), \langle Y_1 : \langle Y_2 : \langle Y_3, \cdots : \langle Y_{s_2} : g_k \rangle \rangle \cdots \rangle \rangle \rangle \\ &\quad = \langle \varphi^T\left(\langle X_1 : \langle X_2 : \langle X_3, \cdots : \langle X_{s_1} : g_j \rangle \rangle \cdots \rangle \rangle\right), \langle Y_1 : \langle Y_2 : \langle \cdots : \langle Y_{s_2} : g_k \rangle \rangle \cdots \rangle \rangle \rangle. \end{aligned}$$

By the assumption on the full-rankness of the distribution $\mathcal{S}_0$, we conclude that

$$\begin{aligned} &\varphi\left(\langle X_1 : \langle X_2 : \langle X_3, \cdots : \langle X_{s_1} : g_j \rangle \rangle \cdots \rangle \rangle\right) \\ &\quad = \varphi^T\left(\langle X_1 : \langle X_2 : \langle X_3, \cdots : \langle X_{s_1} : g_j \rangle \rangle \cdots \rangle \rangle\right), \end{aligned}$$

which in turn implies that $\varphi(x) = \varphi^T(x)$, i.e., the matrix $\mathcal{G}(x)$ is symmetric. □

LEMMA 5.7. *Under the same assumptions as in Lemma* 5.5*, the torsion-free affine connection* ∇ *is the Levi-Civita connection corresponding to the pseudo-Riemannian metric* $\mathcal{G}$*.*

Proof. First of all, note that

$$
\begin{aligned}
(5.5)\quad & \langle V^{\langle\langle X_1:\langle X_2:\langle\cdots:\langle X_s:g_j\rangle\rangle\cdots\rangle\rangle:\langle Y_1:\langle Y_2:\langle\cdots:\langle Y_{s_2}:g_k\rangle\rangle\cdots\rangle\rangle\rangle}, d(\mathcal{L}_{Z_1}\mathcal{L}_{Z_2}\cdots\mathcal{L}_{Z_{s_3}}V_l)\rangle \\
&= \mathcal{L}_{\langle\langle X_1:\langle X_2:\langle\cdots:\langle X_s:g_j\rangle\rangle\cdots\rangle\rangle:\langle Y_1:\langle Y_2:\langle\cdots:\langle Y_{s_2}:g_k\rangle\rangle\cdots\rangle\rangle\rangle}\left[\mathcal{L}_{Z_1}\mathcal{L}_{Z_2}\cdots\mathcal{L}_{Z_{s_3}}V_l\right] \\
&= \mathcal{L}_{\langle Z_1:\langle Z_2:\langle\cdots:\langle Z_{s_3}:g_l\rangle\rangle\cdots\rangle\rangle}\left[\mathcal{L}_{\langle X_1:\langle X_2:\langle\cdots:\langle X_s:g_j\rangle\rangle\cdots\rangle\rangle}\left[\mathcal{L}_{Y_1}\mathcal{L}_{Y_2}\cdots\mathcal{L}_{Y_{s_2}}V_k\right]\right] \\
&= \langle\left(\mathcal{L}_{\langle X_1:\langle X_2:\langle\cdots:\langle X_s:g_j\rangle\rangle\cdots\rangle\rangle}\left[\mathcal{L}_{Y_1}\mathcal{L}_{Y_2}\cdots\mathcal{L}_{Y_{s_2}}V_k\right]\right)^{\mathrm{C}}\circ\varphi^{-1}, \\
&\quad d(\mathcal{L}_{Z_1}\mathcal{L}_{Z_2}\cdots\mathcal{L}_{Z_{s_3}}V_l)\rangle,
\end{aligned}
$$

where in the second equality we have used the property (b) of the compatibility definition between the nonlinear system Σ and the affine connection ∇. Since the observation space of the nonlinear system Σ is generated by the functions of the form $\mathcal{L}_{Z_1}\mathcal{L}_{Z_2}\cdots\mathcal{L}_{Z_{s_3}}V_l$, and Σ is observable by hypothesis, we conclude that

$$
\begin{aligned}
&V^{\langle\langle X_1:\langle X_2:\langle\cdots:\langle X_s:g_j\rangle\rangle\cdots\rangle\rangle:\langle Y_1:\langle Y_2:\langle\cdots:\langle Y_{s_2}:g_k\rangle\rangle\cdots\rangle\rangle\rangle}\circ\varphi \\
&\quad = \left(\mathcal{L}_{\langle X_1:\langle X_2:\langle\cdots:\langle X_s:g_j\rangle\rangle\cdots\rangle\rangle}\left[\mathcal{L}_{Y_1}\mathcal{L}_{Y_2}\cdots\mathcal{L}_{Y_{s_2}}V_k\right]\right)^{\mathrm{C}}.
\end{aligned}
$$

Given the structure of the mapping φ (cf. Lemmas 5.5 and 5.6) and (5.4), this equality can be rewritten as

$$
\begin{aligned}
&\mathcal{G}(\langle\langle X_1:\langle X_2:\langle\cdots:\langle X_s:g_j\rangle\rangle\cdots\rangle\rangle:\langle Y_1:\langle Y_2:\langle\cdots:\langle Y_{s_2}:g_k\rangle\rangle\cdots\rangle\rangle\rangle,\cdot) \\
&\quad = d\langle\varphi(\langle Y_1:\langle Y_2:\langle\cdots:\langle Y_{s_2}:g_k\rangle\rangle\cdots\rangle\rangle\rangle), \langle X_1:\langle X_2:\langle\cdots:\langle X_s:g_j\rangle\rangle\cdots\rangle\rangle\rangle \\
&\quad = d\left(\mathcal{G}(\mathrm{grad}_{\mathcal{G}}\left(\mathcal{L}_{Y_1}\mathcal{L}_{Y_2}\cdots\mathcal{L}_{Y_{s_2}}V_k\right), \mathrm{grad}_{\mathcal{G}}\left(\mathcal{L}_{X_1}\mathcal{L}_{X_2}\cdots\mathcal{L}_{X_s}V_j\right))\right) \\
&\quad = d\{\mathcal{L}_{Y_1}\mathcal{L}_{Y_2}\cdots\mathcal{L}_{Y_{s_2}}V_k : \mathcal{L}_{X_1}\mathcal{L}_{X_2}\cdots\mathcal{L}_{X_s}V_j\}_{\mathcal{G}}\,.
\end{aligned}
$$

Since $\mathrm{grad}_{\mathcal{G}}\{f:g\}_{\mathcal{G}} = \langle\mathrm{grad}_{\mathcal{G}}\, f : \mathrm{grad}_{\mathcal{G}}\, g\rangle_{\nabla^{\mathcal{G}}}$, we conclude

$$
\begin{aligned}
&\langle\langle X_1:\langle X_2:\langle\cdots:\langle X_s:g_j\rangle\rangle\cdots\rangle\rangle:\langle Y_1:\langle Y_2:\langle\cdots:\langle Y_{s_2}:g_k\rangle\rangle\cdots\rangle\rangle\rangle \\
&\quad = \langle\langle X_1:\langle X_2:\langle\cdots:\langle X_s:g_j\rangle\rangle\cdots\rangle\rangle:\langle Y_1:\langle Y_2:\langle\cdots:\langle Y_{s_2}:g_k\rangle\rangle\cdots\rangle\rangle\rangle_{\mathcal{G}}\,.
\end{aligned}
$$

Using the fact that $\mathcal{S}_0$ is full-rank, we deduce that $\langle X:Y\rangle = \langle X:Y\rangle_{\mathcal{G}}$ for all X, $Y\in\mathfrak{X}(M)$. Finally, using the fact that ∇ is torsion-free, we compute

$$
\nabla_X Y = \frac{1}{2}\left(\langle X:Y\rangle + [X,Y]\right) = \frac{1}{2}\left(\langle X:Y\rangle_{\mathcal{G}} + [X,Y]\right) = \nabla^{\mathcal{G}}_X Y \quad \forall X,Y\in\mathfrak{X}(M)\,,
$$

which concludes the result. □

We are now ready to conclude the proof of Theorem 5.4.

Proof of Theorem 5.4. Assume the prolongation Σ^p and the gradient extension Σ^e are weakly externally equivalent. From Lemmas 5.5, 5.6, and 5.7, we deduce the existence of a pseudo-Riemannian metric $\mathcal{G}$ on M such that $\nabla = \nabla^{\mathcal{G}}$ and the unique diffeomorphism between TM and T^*M relating Σ^p and Σ^e and verifying (5.2) is $\flat_{\mathcal{G}}$. From $(\flat_{\mathcal{G}})_* g_j^{\mathrm{V}} = \mathrm{grad}_{\mathcal{G}^\nabla} V_j^{\mathrm{V}}\circ\flat_{\mathcal{G}}$, we deduce $\flat_{\mathcal{G}}(g_j) = dV_j$, and hence $\mathrm{grad}_{\mathcal{G}}\, V_j = g_j$, $j=1,\ldots,m$. Finally, we show that g_0 is a locally gradient vector field. From $(\flat_{\mathcal{G}})_* g_0^{\mathrm{C}} = \mathrm{grad}_{\mathcal{G}^\nabla} V^{g_0}\circ\flat_{\mathcal{G}}$ and the local expression (2.8) of the Christoffel symbols of the Levi-Civita connection $\nabla^{\mathcal{G}}$, we deduce that

$$
\frac{\partial}{\partial x^b}\left(\mathcal{G}_{ac}g_0^c\right) = \frac{\partial}{\partial x^a}\left(\mathcal{G}_{bc}g_0^c\right) \quad \forall a,b=1,\ldots,n\,,
$$

which implies that the one-form $\flat_{\mathcal{G}}(g_0)$ is closed. □

Example 5.8. Consider a linear input-state-output system Σ on $M = \mathbb{R}^n$, i.e., $\dot{x} = Ax + Bu$, $y = Cx$, $x \in \mathbb{R}^n$, with A an $(n \times n)$-matrix, B an $(n \times m)$-matrix, and C an $(m \times n)$-matrix. Assume Σ is observable and controllable. Consider the trivial connection ∇ on $\mathbb{R}^n$ whose Christoffel symbols are given by $\Gamma^a_{bc} = 0$, $1 \leq a, b, c \leq n$. One can easily verify that Σ is compatible with ∇, and, using the hypothesis of controllability, that the distribution $\mathcal{S}_0$ has full-rank. The prolonged system consists of the system itself together with the variational equations $\dot{v} = Av + Bu^p$, $y^p = Cv$, and the gradient extension consists of the system itself together with the equations $\dot{p} = A^T p + C^T u^e$, $y^e = B^T p$. Hence the prolonged system and the gradient extension are weakly externally equivalent if and only if the impulse responses of $\dot{v} = Av + Bu^p$, $y^p = Cv$ and $\dot{p} = A^T p + C^T u^e$, $y^e = B^T p$ are equal, that is, $W(t) = W^T(t)$, with $W(t) := Ce^{At}B$. Thus from Theorem 5.4 we recover the classical result (see, e.g., [27]) that an observable and controllable linear system is a gradient system (with respect to the trivial connection) if and only if $W(t) = W^T(t)$.

Remark 5.9. Note that, given the torsion-free affine connection ∇, the pseudo-Riemannian metric $\mathcal{G}$ obtained in the proof of Theorem 5.4 is unique such that Σ is locally gradient with respect to it. In section 6 below, we investigate the uniqueness (up to isometry) of gradient realizations with the same input-output behavior.

Remark 5.10. In general, we cannot ensure that the drift vector field g_0 is globally gradient, unless we impose some additional conditions on the topology of the state space M (for instance, that the first Betti number of M is zero). This is analogous to the situation in the Hamiltonian setting [13].

Remark 5.11. As noted in section 2.2, one can verify that the pseudo-Riemannian metric on T^*M defined by $\mathcal{G}^\nabla$ corresponds to the pullback by $\sharp_{\mathcal{G}}$ of the complete lift $\mathcal{G}^{\mathrm{C}}$ to TM of the original metric $\mathcal{G}$ on M.

Remark 5.12. A different way to prove the same result which indeed keeps a closer parallelism with the proof for the Hamiltonian case [13] would be the following. Once one has proved Lemmas 5.5 and 5.6, instead of proving Lemma 5.7, one can show that

$$(\varphi)_* \langle X_1 : \langle X_2 : \langle \cdots : \langle X_s : g_j \rangle \rangle \cdots \rangle \rangle^{\mathrm{C}} = \operatorname{grad} V^{\langle X_1 : \langle X_2 : \langle \cdots : \langle X_s : g_j \rangle \rangle \cdots \rangle \rangle} \circ \varphi \tag{5.6}$$

for any $j \in \{1, \ldots, m\}$ and $X_r \in \{g_0, g_1, \ldots, g_m\}$, $r = 1, \ldots, s$. This can be done by considering the following vector fields on T^*M,

$$\begin{aligned} \mathcal{Z}_1 &= (\varphi)_* \langle X_1 : \langle X_2 : \langle \cdots : \langle X_s : g_j \rangle \rangle \cdots \rangle \rangle^{\mathrm{C}} \circ \varphi^{-1}, \\ \mathcal{Z}_2 &= \operatorname{grad} V^{\langle X_1 : \langle X_2 : \langle \cdots : \langle X_s : g_j \rangle \rangle \cdots \rangle \rangle} \end{aligned}$$

and showing that their action on the observation space $\mathcal{H}^e$ of Σ^e is the same. To see this, recall from Proposition 3.5 that $\mathcal{H}^e = V^{S_0} + (\mathcal{H} + \mathfrak{h})^{\mathrm{V}}$. Consider a function of the form $\mathcal{L}_{X_1} \mathcal{L}_{X_2} \cdots \mathcal{L}_{X_s} V_j$, with X_r, $r = 1, \ldots, s$, equal to g_i, $i = 0, 1, \ldots, m$, and $j = 1, \ldots, m$. Then,

$$\begin{aligned} &\mathcal{L}_{\mathcal{Z}_1} \left[(\mathcal{L}_{X_1} \mathcal{L}_{X_2} \cdots \mathcal{L}_{X_s} V_j)^{\mathrm{V}} \right] \\ &\quad = \left(\mathcal{L}_{\langle X_1 : \langle X_2 : \langle \cdots : \langle X_s : g_j \rangle \rangle \cdots \rangle \rangle^{\mathrm{C}}} \left[(\mathcal{L}_{X_1} \mathcal{L}_{X_2} \cdots \mathcal{L}_{X_s} V_j)^{\mathrm{V}} \circ \varphi \right] \right) \circ \varphi^{-1} \\ &\quad = \left(\mathcal{L}_{\langle X_1 : \langle X_2 : \langle \cdots : \langle X_s : g_j \rangle \rangle \cdots \rangle \rangle} \left[\mathcal{L}_{X_1} \mathcal{L}_{X_2} \cdots \mathcal{L}_{X_s} V_j \right] \right)^{\mathrm{V}}, \end{aligned}$$

where we have used twice the fact that φ is the identity mapping on the base manifold M. On the other hand,

$$\mathcal{L}_{\mathcal{Z}_2} \left[(\mathcal{L}_{X_1} \mathcal{L}_{X_2} \cdots \mathcal{L}_{X_s} V_j)^{\mathrm{V}} \right] = \left(\mathcal{L}_{\langle X_1 : \langle X_2 : \langle \cdots : \langle X_s : g_j \rangle \rangle \cdots \rangle \rangle} \left[\mathcal{L}_{X_1} \mathcal{L}_{X_2} \cdots \mathcal{L}_{X_s} V_j \right] \right)^{\mathrm{V}},$$

using property (ii) in Lemma 3.4. The same argument also guarantees that the action of $\mathcal{Z}_1$ and $\mathcal{Z}_2$ is the same over the vertical lifts of the functions spanning $\mathfrak{h}$. Finally, let $\langle Y_1 : \langle Y_2 : \langle \cdots : \langle Y_{s_2} : g_k \rangle \rangle \cdots \rangle \rangle \in S_0$ and consider the corresponding function on T^*M, $V^{\langle Y_1:\langle Y_2:\langle\cdots:\langle Y_{s_2}:g_k\rangle\rangle\cdots\rangle\rangle}$. Then,

$$
\begin{aligned}
(5.7)\qquad & \mathcal{L}_{\mathcal{Z}_1}\left[V^{\langle Y_1:\langle Y_2:\langle\cdots:\langle Y_{s_2}:g_k\rangle\rangle\cdots\rangle\rangle}\right] \\
&= \left(\mathcal{L}_{\langle X_1:\langle X_2:\langle\cdots:\langle X_s:g_j\rangle\rangle\cdots\rangle\rangle^{\mathrm{C}}}\left[V^{\langle Y_1:\langle Y_2:\langle\cdots:\langle Y_{s_2}:g_k\rangle\rangle\cdots\rangle\rangle}\circ\varphi\right]\right)\circ\varphi^{-1} \\
&= \left(\mathcal{L}_{\langle X_1:\langle X_2:\langle\cdots:\langle X_s:g_j\rangle\rangle\cdots\rangle\rangle^{\mathrm{C}}}\left[(\mathcal{L}_{Y_1}\mathcal{L}_{Y_2}\cdots\mathcal{L}_{Y_{s_2}}V_k)^{\mathrm{C}}\right]\right)\circ\varphi^{-1} \\
&= \left(\mathcal{L}_{\langle X_1:\langle X_2:\langle\cdots:\langle X_s:g_j\rangle\rangle\cdots\rangle\rangle}\left[\mathcal{L}_{Y_1}\mathcal{L}_{Y_2}\cdots\mathcal{L}_{Y_{s_2}}V_k\right]\right)^{\mathrm{C}}\circ\varphi^{-1},
\end{aligned}
$$

where we have used (5.4). In addition,

$$
(5.8)\quad \mathcal{L}_{\mathcal{Z}_2}\left[V^{\langle Y_1:\langle Y_2:\langle\cdots:\langle Y_{s_2}:g_k\rangle\rangle\cdots\rangle\rangle}\right] = V^{\langle\langle X_1:\langle X_2:\langle\cdots:\langle X_s:g_j\rangle\rangle\cdots\rangle\rangle:\langle Y_1:\langle Y_2:\langle\cdots:\langle Y_{s_2}:g_k\rangle\rangle\cdots\rangle\rangle\rangle},
$$

where we have used property (i) in Lemma 3.4. Now, (5.5) implies that (5.7) and (5.8) coincide. Therefore, $\mathcal{Z}_1$ and $\mathcal{Z}_2$ coincide over $\mathcal{H}^e$, and this concludes the proof of (5.6).

Now, one can proceed by taking local coordinates $(x^1, \ldots, x^n)$ in M such that every coordinate function x^i is of the form $\mathcal{L}_{X_1}\cdots\mathcal{L}_{X_s}V_j$ for a certain $j \in \{1, \ldots, m\}$ and certain vector fields $X_r \in \{g_0, g_1, \ldots, g_m\}$, $r = 1, \ldots, s$. It follows from (5.4) that there exists n independent vector fields $k^1, \ldots, k^n$ of the form $\langle X_1 : \langle X_2 : \langle \cdots : \langle X_s : g_j \rangle\rangle \cdots \rangle\rangle$ such that $\flat_{\mathcal{G}}(k^i) = dx^i$. Finally, spelling out (5.6) for the vector fields k^i and making use of the symmetry of $\mathcal{G}$, one obtains that the Christoffel symbols of the affine connection ∇ are precisely given by (2.8), which concludes the result.

6. Uniqueness of the gradient realization. In this section, we investigate the gradient analogue of the following well-known result for Hamiltonian systems: if two minimal Hamiltonian systems have the same input-output map, then they are symplectomorphic [3, 22]. We will see how the setting of Theorem 5.4 also provides sufficient conditions under which a similar result holds for gradient realizations.

In [25], Varaiya conjectured that if there exists a state-space diffeomorphism between two locally controllable gradient systems, then the diffeomorphism is actually an isometry between the underlying pseudo-Riemannian manifolds (see also [26]). Subsequently, in [1, 2], Basto Gonçalves produced an example of two locally controllable and observable gradient systems living on the same state space with state-space diffeomorphism given by the identity mapping, where, however, the Riemannian metrics are different; thus providing a counterexample to the conjecture by Varaiya. For the sake of completeness, we review it in the following.

Example 6.1 (see [1, 2]). Consider two gradient systems Σ^1 and Σ^2 on $M^1 = M^2 = \mathbb{R}^4$ with Riemannian metrics $\mathcal{G}^1$ and $\mathcal{G}^2$ given, respectively, by

$$
\begin{aligned}
\mathcal{G}^1(x_1, x_2, x_3, x_4) &= dx_1 \otimes dx_1 + e^{-x_4}dx_2 \otimes dx_2 + e^{-x_1}dx_3 \otimes dx_3 + e^{-x_3}dx_4 \otimes dx_4\,, \\
\mathcal{G}^2(x_1, x_2, x_3, x_4) &= dx_1 \otimes dx_1 + e^{-x_4}dx_2 \otimes dx_2 + \left(e^{-x_1} + e^{x_3}\right)dx_3 \otimes dx_3 \\
&\quad + e^{-x_3}\left(1 + e^{2x_1}\right)dx_4 \otimes dx_4 - e^{x_1}\left(dx_3 \otimes dx_4 + dx_4 \otimes dx_3\right).
\end{aligned}
$$

Furthermore, let Σ^1 and Σ^2 have both zero drift vector fields and the same output functions given by

$$
y_1 = V_1(x) := x_1\,, \quad y_2 = V_2(x) := x_2 + x_3 + x_4\,.
$$

From the definition of $\mathcal{G}^1$ and $\mathcal{G}^2$, it easily follows that the input vector fields of both systems are the same, i.e.,

$$\operatorname{grad}_{\mathcal{G}^1} V_1 = \operatorname{grad}_{\mathcal{G}^2} V_1 = \frac{\partial}{\partial x_1}\,,$$
$$\operatorname{grad}_{\mathcal{G}^1} V_2 = \operatorname{grad}_{\mathcal{G}^2} V_2 = e^{x_4}\frac{\partial}{\partial x_2} + e^{x_1}\frac{\partial}{\partial x_3} + e^{x_3}\frac{\partial}{\partial x_4}\,.$$

Therefore, Σ^1 and Σ^2 are externally equivalent with state-space diffeomorphism given by the identity mapping $\mathrm{Id} : \mathbb{R}^4 \to \mathbb{R}^4$. However, the metrics $\mathcal{G}^1$ and $\mathcal{G}^2$ are different, and hence the identity mapping is *not* an isometry. It should also be noted that Σ^1 and Σ^2 are both controllable and observable.

The following result shows that, under the hypotheses of Theorem 5.4, a state-space diffeomorphism linking two gradient systems is an isometry, *provided* the state-space diffeomorphism is already known to respect the affine connections determined by their respective pseudo-Riemannian metrics. A similar statement is already contained in [1, 2]. Here we make use of an argument given in [13, p. 58] for the case of Hamiltonian systems.

PROPOSITION 6.2. *Let Σ^1 and Σ^2 be two gradient systems with state spaces $\big(M^1, \mathcal{G}^1\big)$ and $\big(M^2, \mathcal{G}^2\big)$, respectively. For $i = 1, 2$, assume that Σ^i is observable with $\dim d\mathcal{H}^i$ constant, and that the distribution $\mathcal{S}_0^i$ is full-rank. Furthermore, let Σ^1 and Σ^2 be externally equivalent with the corresponding state-space diffeomorphism $\psi : M^1 \to M^2$ satisfying*

$$\psi_*(\nabla^{\mathcal{G}^1}_X Y) \circ \psi^{-1} = \nabla^{\mathcal{G}^2}_{\psi_* X \circ \psi^{-1}}(\psi_* Y \circ \psi^{-1}) \quad \forall X, Y \in \mathfrak{X}(M^1)\,. \tag{6.1}$$

Then $\psi^* \mathcal{G}^2 = \mathcal{G}^1$, that is, ψ is an isometry.

Proof. By Lemmas 5.5 and 5.6, the map $\varphi^i = \flat_{\mathcal{G}^i}$ is the unique diffeomorphism satisfying (5.2) for system Σ^i, $i = 1, 2$. It is easily checked that since Σ^1 and Σ^2 are externally equivalent with state-space diffeomorphism ψ, then their prolongations Σ^{1p} and Σ^{2p} are externally equivalent with uniquely determined state-space diffeomorphism given by $\psi_* : TM^1 \to TM^2$. Furthermore, it can be readily checked that the gradient extensions Σ^{1e} and Σ^{2e} are externally equivalent with state-space diffeomorphism $\psi^* : T^*M^2 \to T^*M^1$, *provided* ψ satisfies (6.1). This is because (6.1) implies that ψ^* respects the Riemannian extensions $\mathcal{G}^{\nabla^{\mathcal{G}^1}}$ and $\mathcal{G}^{\nabla^{\mathcal{G}^2}}$ determined, respectively, by the affine connections $\nabla^{\mathcal{G}^1}$ and $\nabla^{\mathcal{G}^2}$. Therefore, by the uniqueness of all these state-space diffeomorphisms, we obtain the following commutative diagram:

$$\begin{array}{ccc} TM^1 & \xrightarrow{\psi_*} & TM^2 \\ {\scriptstyle\varphi^1}\downarrow & & \downarrow{\scriptstyle\varphi^2} \\ T^*M^1 & \xleftarrow[\psi^*]{} & T^*M^2 \end{array}$$

that is,

$$\psi^* \circ \varphi^2 \circ \psi_* = \varphi^1\,. \tag{6.2}$$

Recalling that $\varphi^i = \flat_{\mathcal{G}^i}$, $i = 1, 2$, it is readily seen that (6.2) is equivalent to

$$\psi^* \mathcal{G}^2 = \mathcal{G}^1, \tag{6.3}$$

that is, $\psi : \big(M^1, \mathcal{G}^1\big) \to \big(M^2, \mathcal{G}^2\big)$ is an isometry. ∎

Remark 6.3. Note that in Example 6.1 the torsion-free connections determined by $\mathcal{G}^1$ and $\mathcal{G}^2$ are *different*, and hence the identity map does not respect them.

Remark 6.4. Since (6.2) is equivalent to (6.3), one may also conclude that under the conditions of Theorem 5.4, the state-space diffeomorphism $\psi : M^1 \to M^2$ is an isometry if and only if $\psi^* : T^*M^2 \to T^*M^1$ is a state-space diffeomorphism between Σ^{1e} and Σ^{2e}.

7. Conclusions. We have discussed necessary and sufficient conditions for a nonlinear control system to be realizable as a gradient control system with respect to a pseudo-Riemannian metric whose Levi-Civita connection coincides with a given affine connection. The results rely on a suitable notion of compatibility of the system with respect to the given affine connection, and on the input-output behavior of the prolonged system and the gradient extension. The symmetric product associated with an affine connection plays a key role in the discussion. We believe that the developments in this paper not only give insight in the system-theoretic properties of the physically motivated class of gradient control systems, but also shed light on the differential-geometric properties of gradient and Lagrangian control systems. Future work will include the investigation of necessary and sufficient conditions that guarantee the existence of an affine connection such that the hypothesis of Theorem 5.4 are satisfied, the development of equivalent characterizations in terms of the input-output behavior of the original nonlinear system, and the study of the application of the results to specific classes of nonlinear systems, such as bilinear, homogeneous, and polynomial systems.

8. Appendix. In this appendix we present a simplifying result concerning the compatibility hypothesis in the statement of Theorem 5.4. In general, checking conditions (a) and (b) in the definition of compatibility between the affine connection ∇ and the nonlinear system Σ cannot be performed for every possible choice of vector fields in $\{g_0, g_1, \dots, g_m\}$ and $\{V_1, \dots, V_m\}$. The following result shows that it is enough to check the compatibility condition on a basis of vector fields and the corresponding associated functions once we know that the prolongation and the gradient extension of Σ are weakly externally equivalent.

LEMMA 8.1. *Let ∇ be a torsion-free affine connection. Assume Σ is observable with $\dim d\mathcal{H}$ constant, and that the distribution $\mathcal{S}_0$ is full-rank. Assume the prolongation Σ^p and the gradient extension Σ^e of Σ are weakly externally equivalent. Then Σ is compatible with ∇ if and only if properties* (a) *and* (b) *are verified by a basis of vector fields in S_0.*

Proof. Let $R_1, \dots, R_n$ be linearly independent vector fields of the form $R_i = \langle X_1^i : \langle X_2^i : \langle \cdots : \langle X_{s_i}^i : g_{j_i} \rangle \rangle \cdots \rangle \rangle$, $i = 1, \dots, n$. Let V_{R_i} denote the function on M given by $\mathcal{L}_{X_1^i} \cdots \mathcal{L}_{X_{s_i}^i} V_{j_i}$. From (5.4), we know that $\varphi^T(R_i) = dV_{R_i}$. Assume properties (a) and (b) in the definition of the compatibility condition (cf. Definition 5.1) are verified by any combination of the vector fields $R_1, \dots, R_n$ and the functions $V_{R_1}, \dots, V_{R_n}$. Let $X = \langle X_1 : \langle X_2 : \langle \cdots : \langle X_s : g_k \rangle \rangle \cdots \rangle \rangle$ be any element of S_0, and $V_X = \mathcal{L}_{X_1} \mathcal{L}_{X_2} \cdots \mathcal{L}_{X_s} V_k$ the associated function on M. Since $\mathcal{S}_0$ is full-rank, we have that $X = \sum_{i=1}^n f_X^i R_i$. Then,

$$dV_X = d\mathcal{L}_{X_1} \mathcal{L}_{X_2} \cdots \mathcal{L}_{X_s} V_k = \varphi^T(X) = \sum_{i=1}^n f_X^i \varphi^T(R_i) = \sum_{i=1}^n f_X^i dV_{R_i}\,.$$

Now, let us see that properties (a) and (b) are naturally verified by all possible choices of vector fields in S_0 and generating functions in $\mathcal{H}$. First,

$$\begin{aligned}
&\mathcal{L}_{\langle X_1:\langle X_2:\langle\cdots:\langle X_{s_1}:g_j\rangle\rangle\cdots\rangle\rangle}\left[\mathcal{L}_{Y_1}\mathcal{L}_{Y_2}\cdots\mathcal{L}_{Y_{s_2}}V_k\right]\\
&\quad=\sum_{i=1}^n f_Y^i dV_{R_i}\left(\sum_{j=1}^n f_X^j R_j\right)=\sum_{i,j=1}^n f_Y^i f_X^j dV_{R_j}(R_i)=\sum_{j=1}^n f_X^j dV_{R_j}\left(\sum_{i=1}^n f_X^i R_i\right)\\
&\qquad\qquad=\mathcal{L}_{\langle Y_1:\langle Y_2:\langle\cdots:\langle Y_{s_2}:g_k\rangle\rangle\cdots\rangle\rangle}\left[\mathcal{L}_{X_1}\mathcal{L}_{X_2}\cdots\mathcal{L}_{X_{s_1}}V_j\right],
\end{aligned}$$

where we have used the fact that condition (a) is verified by the vector fields $R_1,\ldots,R_n$ and the functions $V_{R_1},\ldots,V_{R_n}$. Second,

$$\begin{aligned}
(8.1)\qquad &\mathcal{L}_{\langle\langle X_1:\langle X_2:\langle\cdots:\langle X_{s_1}:g_j\rangle\rangle\cdots\rangle\rangle:\langle Y_1:\langle Y_2:\langle\cdots:\langle Y_{s_2}:g_k\rangle\rangle\cdots\rangle\rangle\rangle}\left[\mathcal{L}_{Z_1}\mathcal{L}_{Z_2}\cdots\mathcal{L}_{Z_{s_3}}V_l\right]\\
&\quad=\sum_{i=1}^n f_Z^i dV_{R_i}\left(\sum_{j=1}^n f_{\langle X:Y\rangle}^j R_j\right)=\sum_{i,j=1}^n f_Z^i f_{\langle X:Y\rangle}^j dV_{R_j}(R_i)\\
&\quad=\sum_{j=1}^n f_{\langle X:Y\rangle}^j dV_{R_j}\left(\sum_{i=1}^n f_Z^i R_i\right)=\left\langle\sum_{j=1}^n f_{\langle X:Y\rangle}^j dV_{R_j},Z\right\rangle.
\end{aligned}$$

Let us compute the coefficients $f_{\langle X:Y\rangle}^j$. We have

$$\begin{aligned}
&\langle\langle X_1:\langle X_2:\langle\cdots:\langle X_{s_1}:g_j\rangle\rangle\cdots\rangle\rangle:\langle Y_1:\langle Y_2:\langle\cdots:\langle Y_{s_2}:g_k\rangle\rangle\cdots\rangle\rangle\rangle\\
&\quad=\sum_{i,j=1}^n\langle f_X^i R_i:f_Y^j R_j\rangle=\sum_{i,j=1}^n\left(f_X^i f_Y^j\langle R_i:R_j\rangle+f_X^i R_i[f_Y^j]R_j+f_Y^j R_j[f_X^i]R_i\right)\\
&\quad=\sum_{k=1}^n\left(\sum_{i,j=1}^n f_X^i f_Y^j f_{\langle R_i:R_j\rangle}^k+\sum_{i=1}^n f_X^i R_i[f_Y^k]+\sum_{j=1}^n f_Y^j R_j[f_X^k]\right)R_k\,.
\end{aligned}$$

Now, note that $\sum_{k=1}^n\langle f_{\langle R_i:R_j\rangle}^k dV_{R_k},R_l\rangle=\sum_{k=1}^n\langle f_{\langle R_i:R_j\rangle}^k dV_{R_l},R_k\rangle$ using condition (a) for the vector fields $R_1,\ldots,R_n$ and the functions $V_{R_1},\ldots,V_{R_n}$. Moreover, using condition (b), $\sum_{k=1}^n\langle f_{\langle R_i:R_j\rangle}^k dV_{R_l},R_k\rangle=\langle dV_{R_l},\langle R_i:R_j\rangle\rangle=\langle d(dV_{R_i}[R_j]),R_l\rangle$. Hence, $\sum_{k=1}^n f_{\langle R_i:R_j\rangle}^k dV_{R_k}=d(dV_{R_i}[R_j])$. On the other hand,

$$\begin{aligned}
f_X^i R_i[f_Y^k]dV_{R_k}&=f_X^i\langle df_Y^k,R_i\rangle dV_{R_k}\\
&=f_X^i\langle dV_{R_k},R_i\rangle df_Y^k+f_X^i\left(df_Y^k\wedge dV_{R_k}\right)(R_i,\cdot).
\end{aligned}$$

Since $f_X^i\left(df_Y^k\wedge dV_{R_k}\right)(R_i,\cdot)=f_X^i\left(d\left(f_Y^k dV_{R_k}\right)\right)(R_i,\cdot)=f_X^i\left(d\left(dV_Y\right)\right)(R_i,\cdot)=0$, we have

$$f_X^i R_i[f_Y^k]dV_{R_k}=f_X^i\langle dV_{R_k},R_i\rangle df_Y^k\,.$$

Analogously, one can see that $f_Y^j R_j[f_X^k]dV_{R_k} = f_Y^j\langle dV_{R_k}, R_j\rangle df_X^k$. Finally,

$$\begin{aligned}\sum_{k=1}^{n} f^k_{\langle X:Y\rangle} dV_{R_k} &= \sum_{k=1}^{n}\left(\sum_{i,j=1}^{n} f_X^i f_Y^j f^k_{\langle R_i:R_j\rangle} + \sum_{i=1}^{n} f_X^i R_i[f_Y^k] + \sum_{j=1}^{n} f_Y^j R_j[f_X^k]\right) dV_{R_k}\\ &= \sum_{i,j=1}^{n} f_X^i f_Y^j d(dV_{R_i}[R_j]) + \sum_{i,j=1}^{n} f_X^i\langle dV_{R_j}, R_i\rangle df_Y^j + \sum_{i,j=1}^{n} f_Y^j\langle dV_{R_i}, R_j\rangle df_X^i\\ &= d\left(\sum_{i,j=1}^{n} f_X^i f_Y^j dV_{R_i}[R_j]\right) = d(\mathcal{L}_Y[V_X])\,.\end{aligned}$$

Plugging this equality into (8.1), we get the desired result. □

Acknowledgment. The authors would like to thank the reviewers for their helpful remarks on how to improve the presentation of the paper.

REFERENCES

[1] J. Basto Gonçalves, *Equivalence of Gradient Systems,* Control Theory Centre Report 84, University of Warwick, UK, 1979.
[2] J. Basto Gonçalves, *Equivalencia de sistemas de gradiente*, Port. Math., 40 (1981), pp. 263–277.
[3] J. Basto Gonçalves, *Realization theory for Hamiltonian systems,* SIAM J. Control Optim., 25 (1987), pp. 63–73.
[4] R.K. Brayton and J.K. Moser, *A theory of nonlinear networks*, I, Quart. Appl. Math., 22 (1964), pp. 1–33.
[5] R.K. Brayton and J.K. Moser, *A theory of nonlinear networks*, II, Quart. Appl. Math., 22 (1964), pp. 81–104.
[6] R.W. Brockett and A. Rahimi, *Lie algebras and linear differential equations,* in Ordinary Differential Equations, L. Weiss, ed., Academic Press, New York, 1972, pp. 379–386.
[7] F. Bullo and A.D. Lewis, *Geometric Control of Mechanical Systems*, Text Appl. Math. 49, Springer-Verlag, New York, 2004.
[8] M.P. do Carmo, *Riemannian Geometry,* Birkhäuser, Basel, 1992.
[9] J. Cortés, A.J. van der Schaft, and P.E. Crouch, *Gradient realization of nonlinear control systems,* in Proceedings of the IFAC Workshop on Lagrangian and Hamiltonian Methods for Nonlinear Control, Seville, Spain, 2003, pp. 73–78.
[10] P.E. Crouch, *Geometric structures in systems theory,* Proc. IEE-D, 128 (1981), pp. 242–252.
[11] P.E. Crouch and M. Irving, *On finite Volterra series which admit Hamiltonian realizations,* Math. Systems Theory, 17 (1984), pp. 293–318.
[12] P.E. Crouch, F. Lamnabhi-Lagarrigue, and A.J. van der Schaft, *Adjoint and Hamiltonian input-output differential equations,* IEEE Trans. Automat. Control, AC-40 (1995), pp. 603–615.
[13] P.E. Crouch and A.J. van der Schaft, *Variational and Hamiltonian Control Systems,* Lectures Notes in Control and Inform. Sci. 101, Springer-Verlag, New York, 1987.
[14] R. Hermann and A.J. Krener, *Nonlinear controllability and observability*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 728–740.
[15] B. Jakubczyk, *Hamiltonian realizations of nonlinear systems,* in Theory and Applications of Nonlinear Control Systems, C.I. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam, 1986, pp. 261–271.
[16] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry,* Interscience Tracts in Pure and Applied Mathematics, John Wiley, New York, 1963.
[17] A.D. Lewis, *Affine connections and distributions with applications to nonholonomic mechanics,* Rep. Math. Phys., 42 (1998), pp. 135–164.
[18] A.D. Lewis and R.M. Murray, *Configuration controllability of simple mechanical control systems,* SIAM J. Control Optim., 35 (1997), pp. 766–790.
[19] E.M. Patterson and A.G. Walker, *Riemann extensions,* Quart. J. Math., 3 (1952), pp. 19–28.
[20] R.M. Santilli, *Foundations of Theoretical Mechanics* I, Springer-Verlag, New York, 1978.

[21] W. Sarlet, G. Thompson, and G.E. Prince, *The inverse problem of the calculus of variations: The use of geometrical calculus in Douglas's analysis,* Trans. Amer. Math. Soc., 354 (2002), pp. 2897–2919.
[22] A.J. van der Schaft, *System Theoretic Descriptions of Physical Systems,* in CWI Tract 3, CWI, Amsterdam, 1984.
[23] A.J. van der Schaft, *Linearization of Hamiltonian and gradient systems,* IMA J. Math. Control Inform., 1 (1984), pp. 185–198.
[24] P. Tabuada and G.J. Pappas, *From nonlinear to Hamiltonian via feedback,* IEEE Trans. Automat. Control, 45 (2003), pp. 1439–1442.
[25] P. Varaiya, *Equivalent non-linear networks,* in Mathematical Aspects of Electrical Network Analysis (Proc. Sympos. Appl. Math., New York, 1969), SIAM-AMS Proceedings, Vol. III, AMS, Providence, RI, 1971, pp. 141–147.
[26] T.P. Verma, *Equivalence of Nonlinear Networks,* Ph.D. dissertation, University of California, Berkeley, 1969.
[27] J.C. Willems, *Dissipative dynamical systems, Part* II*,* Arch. Ration. Mech. Anal., 45 (1972), pp. 352–292.
[28] K. Yano and S. Ishihara, *Tangent and Cotangent Bundles,* Marcel Dekker, New York, 1973.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 4, pp. 1215–1238

OPTIMAL CONTROL PROBLEMS WITH FINAL OBSERVATION GOVERNED BY EXPLOSIVE PARABOLIC EQUATIONS*

H. AMANN† AND P. QUITTNER‡

Abstract. We study optimal control problems with final observation. The governing parabolic equations or systems involve superlinear nonlinearities, and their solutions may blow up in finite time. Our proof of the existence, regularity, and optimality conditions for an optimal pair is based on uniform a priori estimates for the approximating solutions. Our conditions on the growth of the nonlinearity are essentially optimal. In particular, we also solve a long-standing open problem of Lions concerning singular systems.

Key words. optimal control problem, nonlinear parabolic equation, blow-up, final observation, optimality conditions, strong nonlinearities

AMS subject classifications. 49J20, 49K20, 49N60, 35K55

DOI. 10.1137/S0363012903433450

1. Introduction.

In his book [21], Lions studied several optimal control problems governed by nonlinear parabolic equations of the form

$$\partial_t y - \Delta y = y^\lambda + u, \qquad x \in \Omega,\ t \in [0,T], \tag{1.1}$$

where Ω is a bounded domain in $\mathbb{R}^n$, $\lambda \in \{2,3\}$, $u = u(x,t)$ is the control, and $y = y(x,t)$ is the state variable. Equation (1.1) is complemented by suitable boundary and initial conditions, for example,

$$y = 0 \ \text{ on } \partial\Omega \times (0,T), \qquad y(\cdot,0) = y_0, \tag{1.2}$$

where $y_0 \in L_\infty(\Omega)$. If u is regular enough, then the state problem (1.1)–(1.2) possesses a unique strong solution $y = y(u)$ defined on the maximal existence interval J_u (see section 2 for the definition of a strong solution). However, even for smooth controls u, the solution $y(u)$ need not be global—the interval J_u need not coincide with $[0,T]$. In this case, $y(u)$ *blows up* at the time $t(u) := \sup J_u$; i.e., it develops a singularity and leaves its natural regularity class. After blow-up, the solution either can be continued in a weak sense (the blow-up is *incomplete* [16]) or such continuation is not possible (the solution blows up *completely* [9]).

Let $\mathbb{U}_{\rm ad}$ denote the set of admissible controls,

$$\mathbb{U}_{\rm ad}^G := \{u \in \mathbb{U}_{\rm ad} : \text{the solution } y(u) \text{ is global}\},$$

and let $\mathbb{J} = \mathbb{J}(y,u)$ be the cost functional. A standard way to solve the optimal control problem

$$\text{minimize } \mathbb{J}(y(u),u) \text{ over } u \in \mathbb{U}_{\rm ad}^G \tag{1.3}$$

*Received by the editors August 15, 2003; accepted for publication (in revised form) January 22, 2005; published electronically October 7, 2005.
http://www.siam.org/journals/sicon/44-4/43345.html
†Institut für Mathematik, Universität Zürich, Winterthurerstr. 190, CH–8057 Zürich, Switzerland (herbert.amann@math.unizh.ch).
‡Department of Applied Mathematics and Statistics, Comenius University, SK–84248 Bratislava, Slovakia (quittner@fmph.uniba.sk). This author's research was supported in part by VEGA grant 1/0259/03 and by the Swiss National Science Foundation (Schweizerischer Nationalfonds).

is to consider controls u_k, $k = 1, 2, \dots$, such that $(\mathbb{J}(y(u_k), u_k))$ is a minimizing sequence for (1.3) and to show that a suitable subsequence of $((y(u_k), u_k))$ converges to an optimal pair $(y(u), u)$. Assume, for example, that $\mathbb{U}_{\rm ad}$ is a (weakly closed) subset of a reflexive Banach space. If $\mathbb{J}$ is coercive with respect to u (or $\mathbb{U}_{\rm ad}$ is bounded), then the sequence (u_k) is bounded and we may assume that $u_k \to u$ in the weak topology. Similarly, if $\mathbb{J}$ is coercive with respect to y (in a suitable space of functions defined in $Q := \Omega \times [0, T]$), then the sequence $(y(u_k))$ is bounded and standard compactness results for the state problem enable us to pass to the limit in order to find a minimizer for (1.3).

If we consider problems with final observation (where $\mathbb{J}$ depends just on u and the final value $y(\cdot, T)$), then the coerciveness of $\mathbb{J}$ provides a priori estimates for u_k and final values of $y(u_k)$. However, such estimates are, in general, not sufficient for the uniform boundedness of solutions $y(u_k)$ on the whole interval $[0, T]$. Consequently, we have to find sufficient conditions on λ and/or other parameters of the problem which guarantee a priori bounds for global solutions y of (1.1)–(1.2) depending only on suitable norms of u and $y(\cdot, T)$.

Let us discuss the question of a priori bounds for problems with final observation in the particular setting of [21, section I.10]. Fix $N > 0$, $q \ge 1$, $y_d \in L_q(\Omega)$, and set

$$\mathbb{J}(y, u) := \int_\Omega |y(x, T) - y_d(x)|^q \, dx + N \int_Q u^2(x, t) \, dx \, dt.$$

Assume also that $\mathbb{U}_{\rm ad} \subset L_2(Q)$ is closed and convex and that $\mathbb{U}_{\rm ad}^G \ne \emptyset$. If $\lambda = 2$, $q = 3$, and $n \le 3$, then [21, Theorem I.10.1] and its proof guarantee the required bounds for the solutions $y(u_k)$, hence the existence of an optimal pair (y, u). If, in addition, $n \le 2$, then optimality conditions for the optimal pair (y, u) were derived in [21, Theorem I.10.3]. On the other hand, the existence of an optimal pair in the case $\lambda = 3$, $q = 4$ (or $\lambda = 2$, $q < 3$) and the optimality conditions for $n = 3$ were left as open problems; see [21, Remarks I.10.1, I.10.2, and I.10.4]. Our results give, in particular, positive answers to all those open problems. In fact, we consider an arbitrary dimension n, exponents $q \ge 2$, $\lambda > 1$ (where either $y^\lambda := |y|^{\lambda-1} y$ or $y^\lambda := |y|^\lambda$), and controls $u \in L_r([0, T], L_2(\Omega))$, $r \ge 2$, and find sufficient conditions on q, λ, and r that guarantee the existence of optimal controls and the optimality conditions (see section 2 for precise statements of our results).

We also show that many of our conditions are essentially optimal. In particular, if $\mathbb{U}_{\rm ad} \subset L_\infty([0, T], L_2(\Omega))$, then our sufficient conditions on q and λ guaranteeing the existence of optimal controls have the form

$$\lambda < \frac{n+2}{(n-2)_+} \qquad \text{and} \qquad q \in \left((\lambda - 1)\frac{n}{2}, \frac{2n}{(n-4)_+} \right),$$

where $a_+ := \max(a, 0)$ and $a/b_+ := \infty$ if $a > 0$ and $b \le 0$. The upper bound for q is required by the (low) regularity of controls u: it guarantees $y(u) \in C([0, T], L_q(\Omega))$ so that $\mathbb{J}(y(u), u)$ is defined. If $q < (\lambda - 1)n/2$ or $\lambda > (n+2)/(n-2)_+$ and $n \le 10$, then we show that problem (1.3) need not be solvable even if the set $\mathbb{U}_{\rm ad}$ is a compact subset of $C^\infty(\bar\Omega \times [0, T])$ and $\mathbb{U}_{\rm ad}^G \ne \emptyset$; see Remark 3.4. This nonexistence result is due to the fact that the set $\mathbb{U}_{\rm ad}^G$ need not be closed in $\mathbb{U}_{\rm ad}$: if $u_k \in \mathbb{U}_{\rm ad}^G$, $u_k \to u \in \mathbb{U}_{\rm ad}$, then the limiting solution $y(u)$ may blow up at $t(u) < T$. The conditions on q show the importance of a good choice of the cost functional in order to control the equation. On the other hand, if $\lambda > (n+2)/(n-2)_+$, then (a strong) solvability of our control problem cannot be guaranteed for any q.

The solvability of (1.3) with $\mathbb{U}_{\rm ad}$, $\mathbb{J}$ as above was proved by Imanuvilov [18, Theorem 2.1] and Fursikov [15, Theorem 4.3] for $r = q = 2$ and any $\lambda > 1$, but their function $y(u)$ corresponding to the optimal control u need not be a strong solution in our sense. In fact, the results of [18], [15] also apply to the example of Remark 3.4(i), where $y(u)$ blows up at $t(u) < T$ (but can be continued in a weak sense). This lack of regularity causes serious problems in establishing the optimality conditions. In order to obtain these conditions, Imanuvilov and Fursikov have to assume $q = 2 \geq (\lambda - 1)n/2$; see [15, Theorem 5.1]. Note also that the proofs in [18], [15] substantially use the choice $q = 2$ and hence require $\lambda \leq 1 + 4/n$. In particular, if $n = 3$, then their method cannot be used in the case $\lambda = 3$, $q = 4$ mentioned above.

Our proof of a priori estimates is based on energy and perturbation arguments in [25], [27]. The same approach can be used for more general problems. For example, the case of general second-order elliptic operators and/or general nonlinearities with polynomial growth can be solved by adopting the proofs in [26]. Similarly, if one considers linear or nonlinear parabolic equations complemented by nonlinear Neumann boundary conditions of the form $\partial_\nu y = y^\lambda$ or $\partial_\nu y = y^\lambda + u$, then one can use estimates in [28] and [11].

In this paper we consider two modifications of the model problem (1.1)–(1.2): a problem with multiplicative control and a problem governed by a parabolic system.

In the case of multiplicative control we replace the state equation (1.1) by

$$\partial_t y - \Delta y = y^\lambda + uy, \qquad x \in \Omega,\ t \in [0,T], \tag{1.4}$$

and we prove the required a priori bounds by using the energy and perturbation arguments mentioned above. This study is motivated by the fact that multiplicative controls often appear in the literature.

In section 6 we investigate the existence of optimal controls for problems governed by the system

$$\left.\begin{aligned} \partial_t y_1 - \Delta y_1 &= \kappa y_1 y_2 - b y_1 + u, && x \in \Omega,\ t \in [0,T], \\ \partial_t y_2 - d\Delta y_2 &= a y_1, && x \in \Omega,\ t \in [0,T], \end{aligned}\right\} \tag{1.5}$$

which is complemented by suitable boundary and (nonnegative) initial conditions. Here $d \geq 0$, $\kappa, a > 0$, $b \in \mathbb{R}$, and u is a nonnegative control. System (1.5) (with $d = 0$ and $u = 0$) was derived in [19] as a model for the dynamics of a nuclear reactor close to a stationary state. The state variables y_1 and y_2 correspond to the neutron flux and the temperature, respectively, and the constant κ represents the temperature feedback (cf. also [29]). Since this system (with $d \geq 0$, $\kappa > 0$, and $u = 0$) possesses an interesting dynamics with possible blow-up in finite time, it became the object of study of many mathematical papers (see [10], [17], [23], [24], [31], [32], and the references therein). We consider the case $d = \kappa = 1$ and study the corresponding optimal control problem with final observation. Since the energy arguments used in the case of (1.1) or (1.4) cannot be applied, we use a different approach to the proof of a priori bounds.

This paper is organized as follows. In section 2 we formulate our main results (Theorems 2.3, 2.6, 2.8, and 2.10). Sections 3 and 4 are devoted to the proof of existence of optimal controls and optimality conditions, respectively, for the problem governed by the model equation (1.1). Problems governed by (1.4) and (1.5) are studied in sections 5 and 6, respectively. In the appendix we recall, for the reader's convenience, from [6] the basic existence, uniqueness, and stability results for semilinear parabolic equations which are the fundament for our investigations.

2. Main results. First we introduce some notation which will be used throughout this paper. If $a, b \in \mathbb{R}$, then we denote $a \vee b := \max(a, b)$ and $a \wedge b := \min(a, b)$. If $p \in (1, \infty)$, then p' is the dual exponent defined by $1/p + 1/p' = 1$. For $X \subset \mathbb{R}^n$ we write $\mathcal{D}(X)$ for the space of smooth functions with compact support in X. The symbols w and w^* are used to denote the weak and weak-star topology, respectively. By Ω we mean an open bounded subset of $\mathbb{R}^n$ having a smooth boundary Γ. We also set $Q := \Omega \times J$ and $\Sigma := \Gamma \times J$, where $J := [0, T]$ with a fixed $T > 0$. By $\mathcal{B}$ we denote one of the boundary operators γ, ∂_ν, where γ is the trace operator and ∂_ν is the derivative with respect to ν, the outer unit normal on Γ.

Let $s \in [-2, 2]$ and $1 < q < \infty$. We write $W_q^s := W_q^s(\Omega)$ for the usual Sobolev–Slobodeckiĭ spaces; hence $W_q^0 = L_q(\Omega)$. If $\mathcal{B} = \gamma$, then we set

$$W_{q,\mathcal{B}}^s := \begin{cases} \{\, u \in W_q^s;\ \mathcal{B}u = 0 \,\}, & 1/q < s \le 2, \\ W_q^s, & 0 \le s < 1/q, \\ (W_{q',\mathcal{B}}^{-s})', & -2 \le s < 0, \quad s \ne -1 + 1/q, \end{cases}$$

where X' denotes the dual space to X. If $\mathcal{B} = \partial_\nu$, then

$$W_{q,\mathcal{B}}^s := \begin{cases} \{\, u \in W_q^s;\ \mathcal{B}u = 0 \,\}, & 1 + 1/q < s \le 2, \\ W_q^s, & 0 \le s < 1 + 1/q, \\ (W_{q',\mathcal{B}}^{-s})', & -2 \le s < 0, \quad s \ne -2 + 1/q. \end{cases}$$

In either case the dual spaces are determined by means of the standard L_q-duality pairing. We also set $S_q := \{-2 + 1/q, -1 + 1/q, 1/q, 1 + 1/q\}$.

Weak and strong solutions. Consider the problem

$$\left.\begin{aligned} \partial_t y - \Delta y &= f && \text{in } Q, \\ \mathcal{B}y &= 0 && \text{on } \Sigma, \\ y(\cdot, 0) &= y^0 && \text{in } \Omega, \end{aligned}\right\} \tag{2.1}$$

where $y^0 \in L_1(\Omega)$ and $f \in L_1(Q)$.

DEFINITION 2.1. *Assume that $s \in [0, 2] \setminus S_q$ and $1 < p, q < \infty$. A* weak $L_p(W_q^s)$-solution *of* (2.1) on $[0, t]$, $0 < t \le T$, *is a function $y \in L_{p,\mathrm{loc}}([0, t), W_{q,\mathcal{B}}^s)$ such that*

$$\int_0^t \int_\Omega (-\partial_t \varphi - \Delta \varphi) y \, dx \, d\tau = \int_0^t \int_\Omega \varphi f \, dx \, d\tau + \int_\Omega \varphi(0) y^0 \, dx$$

for any $\varphi \in \mathcal{D}(\overline{\Omega} \times [0, t))$ satisfying $\mathcal{B}\varphi = 0$ on $\Gamma \times [0, t]$. It is global *if $t = T$ and $y \in L_p((0, T), W_{q,\mathcal{B}}^s)$.*

The differential operator $\mathcal{C} := 1 - \Delta$ defines an isomorphism between $W_{q,\mathcal{B}}^2$ and $L_q(\Omega)$, and this isomorphism admits a unique extension to an isomorphism $C = C_s$ between $W_{q,\mathcal{B}}^s$ and $W_{q,\mathcal{B}}^{s-2}$ for any $s \in [0, 2] \setminus S_q$ (see [1]). Moreover, $-A := 1 - C$ generates a strongly continuous analytic semigroup $\{e^{-tA};\ t \ge 0\}$ on $W_{q,\mathcal{B}}^r$ for $r \in [-2, s] \setminus S_q$, and

$$(t \mapsto e^{-tA} x) \in C\big([0, T], W_{q,\mathcal{B}}^r\big) \cap C\big((0, T], W_{q,\mathcal{B}}^s\big) \tag{2.2}$$

with

$$\|e^{-tA} x\|_{W_{q,\mathcal{B}}^s} \le c t^{(r-s)/2} \, \|x\|_{W_{q,\mathcal{B}}^r}, \qquad 0 < t \le T, \tag{2.3}$$

for $x \in W^r_{q,\mathcal{B}}$ (cf. [1, Theorem 5.2] and [2, Theorem V.2.1.3]). Then, provided $1 < q < n/(n-2)$ and $0 \le s < 2 - n/q'$, (the weak form of) problem (2.1) is equivalent to the abstract evolution equation

$$\dot{y} + Ay = f \text{ in } [0,T], \qquad y(0) = y^0, \tag{2.4}$$

(see [6] for details).

DEFINITION 2.2. *A weak $L_p(W^s_q)$-solution y of (2.1) on $[0,t]$ is a* strong $L_p(W^s_q)$-solution *if*

$$y \in W^1_{r,\mathrm{loc}}([0,t), W^{s-2}_{q,\mathcal{B}}) \cap L_{r,\mathrm{loc}}([0,t), W^s_{q,\mathcal{B}})$$

for some $r > 1$ and (2.4) is satisfied a.e. in $[0,t]$. If, in addition, $y \in C^\rho([0,t), W^s_{q,\mathcal{B}})$ for some $\rho \in [0,1)$ then y is called a strong $C^\rho(W^s_q)$-solution.

A model problem. Now we are ready to formulate the main results of this paper. First consider the optimal control problem (1.3) for the model state equation

$$\left.\begin{aligned} \partial_t y - \Delta y &= |y|^{\lambda-1}y + u && \text{in } Q,\\ \mathcal{B}y &= 0 && \text{on } \Sigma,\\ y(\cdot,0) &= y^0 && \text{in } \Omega, \end{aligned}\right\} \tag{2.5}$$

where $\mathcal{B} \in \{\gamma, \partial_\nu\}$. As already announced in the introduction, instead of the operator $-\Delta$ and the model nonlinearity $|y|^{\lambda-1}y$, we could handle a general second-order elliptic operator $\mathcal{A}$ and a general superlinear function $f(x,y)$ satisfying suitable growth conditions (see [26] for details).

In the following theorem we consider cost functionals $\mathbb{J}$ which depend on the final value of y and which satisfy the coercivity condition

$$\mathbb{J}(y,u) \ge c_1 \|y(\cdot,T)\|_{L_q(\Omega)} - c_2, \tag{2.6}$$

with positive constants c_1 and c_2.

THEOREM 2.3. *Let*

$$1 < \lambda < \frac{n+2}{(n-2)_+}, \tag{2.7}$$

$$q \in \left((\lambda-1)\frac{n}{2}, \frac{2n}{(n-4)_+}\right) \qquad \text{and} \qquad q \ge 2. \tag{2.8}$$

Suppose that $r \ge 2$ satisfies

$$\frac{1}{r} < 1 - \frac{n}{4} + \frac{n}{2q} \tag{2.9}$$

and

$$r > \frac{\lambda+1}{\lambda}\,\frac{\lambda n - (n+4)}{n+2-\lambda(n-2)} - \frac{2}{\lambda}. \tag{2.10}$$

Assume that $y^0 \in W^2_{q,\mathcal{B}}$ and $\mathbb{U}_{\mathrm{ad}}$ is a weakly compact subset of $L_r(J, L_2(\Omega))$. If $u \in \mathbb{U}_{\mathrm{ad}}$, then (2.5) has a unique strong $L_{r\lambda}(L_{2\lambda})$-solution defined on the maximal existence interval J_u.

Assume $\mathbb{U}^G_{\mathrm{ad}} \ne \emptyset$. Let (2.6) be true and assume that $\mathbb{J}$ can be written in the form $\mathbb{J}(y,u) = \mathbb{J}_T(y(\cdot,T),u)$, where $\mathbb{J}_T : L_q(\Omega) \times (L_r(J, L_2(\Omega)), w) \to \mathbb{R}$ is lower semicontinuous. Then the optimal control problem (1.3) governed by (2.5) has a solution.

Remark 2.4. (i) Theorem 2.3 remains true if we replace the nonlinearity $|y|^{\lambda-1}y$ with $|y|^{\lambda}$; see Remark 3.3.

(ii) Let λ, q, r satisfy (2.7)–(2.10), $y^0 \in W^2_{q,\mathcal{B}}$,

$$\mathbb{J}(y,u) := \int_\Omega |y(x,T) - y_d(x)|^q \, dx + N \int_0^T \left(\int_\Omega u^2(x,t)\, dx \right)^{r/2} dt, \tag{2.11}$$

where $y_d \in L_q(\Omega)$, $N \geq 0$, and let $\mathbb{U}_{\rm ad} \subset L_r(J, L_2(\Omega))$ be closed, convex, and bounded. Then all assumptions of Theorem 2.3 are satisfied provided $\mathbb{U}^G_{\rm ad} \neq \emptyset$. In addition, if $N > 0$, then $\mathbb{U}_{\rm ad}$ need not be bounded (we can replace the set $\mathbb{U}_{\rm ad}$ in problem (1.3) with $\tilde{\mathbb{U}}_{\rm ad} := \mathbb{U}_{\rm ad} \cap \mathbb{B}_R$, where $\mathbb{B}_R$ is a large closed ball in $L_r(J, L_2(\Omega))$).

(iii) If $r = 2$, then conditions (2.7)–(2.10) in Theorem 2.3 read

$$1 < \lambda < \frac{3n+8}{(3n-4)_+}, \qquad q \in \left((\lambda-1)\frac{n}{2}, \frac{2n}{(n-2)_+} \right), \qquad \text{and} \qquad q \geq 2.$$

In particular, if $n \leq 3$, then we may choose $\lambda = 3$ and $q = 4$ (cf. the open problems of Lions mentioned above). □

Example 2.5. Let $\lambda, q, r, y^0, \mathbb{J}, \mathbb{U}_{\rm ad}$ be as in Remark 2.4(ii). Assume $|y^0| \leq C_0$ for some $C_0 \geq 0$ and $\{u \in L_\infty(Q)\,;\, |u| \leq C_0^\lambda\} \subset \mathbb{U}_{\rm ad}$. Then $\mathbb{U}^G_{\rm ad} \neq \emptyset$; hence the optimal control problem (1.3) has a solution. In fact, the solution $\tilde{y}$ of the linear problem

$$\left.\begin{aligned} \partial_t \tilde{y} - \Delta \tilde{y} &= 0 && \text{in } Q, \\ \mathcal{B}\tilde{y} &= 0 && \text{on } \Sigma, \\ \tilde{y}(\cdot,0) &= y^0 && \text{in } \Omega, \end{aligned}\right\}$$

satisfies $|\tilde{y}| \leq C_0$ by the maximum principle; thus $u := -|\tilde{y}|^{\lambda-1}\tilde{y} \in \mathbb{U}^G_{\rm ad}$ (the function $y := \tilde{y}$ is a global solution of (2.5)).

Optimality conditions. In order to obtain the optimality conditions, we restrict ourselves to the case $r = 2$ and we also fix

$$\mathbb{J}(y,u) := \int_\Omega |y(x,T) - y_d(x)|^q \, dx + N \int_Q u^2(x,t)\, dx\, dt, \tag{2.12}$$

where $q > 1$, $y_d \in L_q(\Omega)$, and $N \geq 0$ are given. This particular choice of r and $\mathbb{J}$ corresponds to the setting of Lions in [21].

THEOREM 2.6. *Let the assumptions of Theorem* 2.3 *be fulfilled and let, moreover,* $r = 2$, $\mathbb{U}_{\rm ad}$ *be convex, and* $\mathbb{J}$ *be as in* (2.12). *If* (y,u) *is an optimal pair for problem* (1.3) *governed by* (2.5) *and* p *is the solution of*

$$\left.\begin{aligned} -\partial_t p - \Delta p &= \lambda |y|^{\lambda-1} p && \text{in } Q, \\ \mathcal{B}p &= 0 && \text{on } \Sigma, \\ p(\cdot,T) &= q|y(\cdot,T) - y_d|^{q-2}(y(\cdot,T) - y_d) && \text{in } \Omega, \end{aligned}\right\} \tag{2.13}$$

then

$$\int_Q (p + 2Nu)(v-u)\, dx\, dt \geq 0 \qquad \text{for all } v \in \mathbb{U}_{\rm ad}.$$

Remark 2.7. (a) The existence of an optimal pair (y,u) in Theorem 2.6 is guaranteed by Theorem 2.3. The solvability of (2.13) follows from Lemma 4.1 and Remark 4.2 below.

(b) As in Remark 2.4(ii), in Theorem 2.6 we can allow $\mathbb{U}_{\rm ad}$ to be any closed convex subset of $L_2(Q)$ if $N > 0$.

Multiplicative controls. Next we consider the optimal control problem (1.3) governed by the equation

$$\left.\begin{aligned} \partial_t y - \Delta y &= |y|^{\lambda-1}y + uy && \text{in } Q,\\ \mathcal{B}y &= 0 && \text{on } \Sigma,\\ y(\cdot,0) &= y^0 && \text{in } \Omega, \end{aligned}\right\} \tag{2.14}$$

where $\mathcal{B} \in \{\gamma, \partial_\nu\}$.

THEOREM 2.8. *Let* (2.6), (2.7), *and* (2.8) *be satisfied, let* $y^0 \in W^2_{q,\mathcal{B}}$, *let* $\mathbb{U}_{\rm ad} \subset L_\infty(Q)$ *be* w^**-sequentially compact, and let* $\mathbb{U}^G_{\rm ad} \neq \emptyset$. *Assume that* $\mathbb{J}$ *can be written in the form* $\mathbb{J}(y,u) = \mathbb{J}_T(y(\cdot,T),u)$, *where* $\mathbb{J}_T : L_q(\Omega) \times (L_\infty(Q), w^*) \to \mathbb{R}$ *is sequentially lower semicontinuous. Then problem* (1.3) *governed by* (2.14) *has a solution.*

Remark 2.9. Similarly as in Remark 2.4(ii) and Example 2.5, all assumptions of Theorem 2.8 concerning $\mathbb{U}_{\rm ad}$ and $\mathbb{J}$ are satisfied if, for example, $|y^0| \le C_0$, $D_1 \ge C_0^{\lambda-1}$, $D_2 \ge 0$, $N \ge 0$,

$$\mathbb{U}_{\rm ad} = \{u \in L_\infty(Q) : -D_1 \le u \le D_2\},$$

and

$$\mathbb{J}(y,u) = \int_\Omega |y(x,T) - y_d(x)|^q \, dx + N\|u\|_{L_\infty(Q)}.$$

Again, we may take $D_1 = \infty$ and/or $D_2 = \infty$ if $N > 0$.

Control of systems. Finally, let us formulate our result concerning the parabolic system

$$\left.\begin{aligned} \partial_t y_1 - \Delta y_1 &= y_1 y_2 - b y_1 + u && \text{in } Q,\\ \partial_t y_2 - \Delta y_2 &= a y_1 && \text{in } Q,\\ \mathcal{B}y_1 = \mathcal{B}y_2 &= 0 && \text{on } \Sigma,\\ y_1(\cdot,0) &= y_1^0 && \text{in } \Omega,\\ y_2(\cdot,0) &= y_2^0 && \text{in } \Omega, \end{aligned}\right\} \tag{2.15}$$

where $a > 0$, $b \in \mathbb{R}$, $\mathcal{B} \in \{\gamma, \partial_\nu\}$,

$$y_1^0, y_2^0 \ge 0, \qquad y_1^0, y_2^0 \in C^2(\bar\Omega), \qquad \mathcal{B}y_1^0 = \mathcal{B}y_2^0 = 0, \tag{2.16}$$

$$u \in L_r(J, L_z^+(\Omega)), \qquad r, z > 1, \quad \frac{1}{r} + \frac{n}{2z} < 1. \tag{2.17}$$

As usual, $L_r(J, L_z^+(\Omega))$ is the set of positive functions in $L_r(J, L_z(\Omega))$. The regularity assumption (2.17) guarantees that (2.15) possesses a unique strong solution (defined on the maximal existence interval J_u) and that this solution is Hölder continuous in both x and t.

THEOREM 2.10. *Consider problem* (2.15) *with* $a > 0$, $b \in \mathbb{R}$. *Let* (2.16) *and* (2.17) *be satisfied, where either* $\mathcal{B} = \partial_\nu$ *and* $n \le 3$ *or* $\mathcal{B} = \gamma$ *and* $n \le 2$. *Assume that* $\mathbb{U}_{\rm ad}$ *is a compact set in* $L_r(J, L_z^+(\Omega))$, $\mathbb{U}^G_{\rm ad} \neq \emptyset$, *and* $\mathbb{J}$ *can be written in the form* $\mathbb{J}(y,u) = \mathbb{J}_T(y_1(T),u)$, *where*

$$\mathbb{J}_T : L_q(\Omega) \times L_r(J, L_z^+(\Omega)) \to \mathbb{R} \quad \textit{is lower semicontinuous,}$$

$q \in [1,\infty]$, *and* $\mathbb{J}(y,u) \ge c_1\|y_1(T)\|_{L_q(\Omega)} - c_2$. *Then the optimal control problem* (1.3) *governed by* (2.15) *has a solution.*

Remark 2.11. As above, we can easily find examples of $\mathbb{U}_{\rm ad}$ and $\mathbb{J}$ satisfying the compactness and lower semicontinuity assumptions in Theorem 2.10. The assumption $\mathbb{U}^G_{\rm ad} \ne \emptyset$ is satisfied if, for example, $\mathcal{B} = \gamma$, $b \ge 0$, $0 \in \mathbb{U}_{\rm ad}$, and y_1^0, y_2^0 are small enough (e.g., in $L_\infty(\Omega)$). This is due to the fact that in this case, zero is an asymptotically stable equilibrium of (2.15) with $u = 0$. If $\mathcal{B} = \partial_\nu$, $y_1^0 = y_2^0 = 0$, and $0 \in \mathbb{U}_{\rm ad}$, then obviously $0 \in \mathbb{U}^G_{\rm ad}$.

3. Solvability of the model problem.

Proof of Theorem 2.3. Set $s := 0$, $q := 2\lambda$, and $p := r\lambda$. Since $r \ge 2$ and $1 < \lambda < \frac{n+2}{(n-2)_+}$, there exists $\sigma \notin S_q$ satisfying

$$\frac{2}{r\lambda'} < \sigma < \frac{2}{r} \wedge \Big(2 - \frac{n}{2\lambda'}\Big).$$

Now Theorem A.1 guarantees the existence of a unique $L_{r\lambda}(L_{2\lambda})$-solution y of (2.5) defined on the maximal existence interval J_u. Fixing $u \in \mathbb{U}^G_{\rm ad}$, this solution is global and $|y|^\lambda \in L_r(J, L_2(\Omega))$. The Sobolev maximal regularity for (2.5), [2, Theorem III.4.10.2] and interpolation theorems in [4] (also see [3, Theorem 3]) imply

$$y \in W^1_r(J, L_2(\Omega)) \cap L_r(J, W^2_{2,\mathcal{B}}) \hookrightarrow C(J, W^1_{2,\mathcal{B}}) \cap C(J, W^z_{q,\mathcal{B}}) \cap L_{r\lambda}(J, L_{2\lambda}(\Omega)) \tag{3.1}$$

for any

$$z < 2 - \frac{n}{2} + \frac{n}{q} - \frac{2}{r},$$

where the embedding into $C(J, W^z_{q,\mathcal{B}}) \cap L_{r\lambda}(J, L_{2\lambda}(\Omega))$ is compact.

Let (y_k, u_k) be a minimizing sequence for problem (1.3). We may assume $u_k \to u$ weakly in $L_r(J, L_2(\Omega))$ and $\|u_k\|_{L_r(J,L_2(\Omega))} \le C_r$. Part (a) of the proof of [6, Theorem 1.1] shows that there exists $t_0 > 0$ independent of k such that

$$y_k \text{ are uniformly bounded in } L_{r\lambda}([0,t_0], L_{2\lambda}(\Omega)). \tag{3.2}$$

Set $u_k(x,t) := 0$ for $t \in (T, 2T]$ and consider problem (2.5) with J replaced by $[0,2T]$. This problem possesses a unique $L_{r\lambda}(L_{2\lambda})$-solution $\tilde{y}_k$ defined on the maximal existence interval $J_{\tilde{y}_k} \subset [0,2T]$. The function $w_k(t) := \tilde{y}_k(T+t)$ is the $L_{r\lambda}(L_{2\lambda})$-solution of (2.5) with $u \equiv 0$, initial condition $w_k(0) = y_k(T)$, and the maximal existence interval $J_{w_k} \subset [0,T]$. The boundedness of $\mathbb{J}(y_k, u_k)$ implies a bound for $y_k(T)$ in $L_q(\Omega)$, and the well posedness of (2.5) in $L_q(\Omega)$, guaranteed by Lemma 3.1 below, shows the existence of $t_1 > 0$ such that $[0,t_1] \subset J_{w_k}$ for any k. Consequently, all solutions y_k can be continued on the interval $[T, T+t_1]$. Now Lemma 3.2 below implies $\|y_k(\tau)\|_{L_q(\Omega)} \le C_q$ for any $\tau \in [0,T]$.

Let $\tau^* = \tau^*(C_r, C_q)$ be from Lemma 3.1. Fixing $\delta \in (0, t_0 \wedge \tau^*)$ and using the last statement of Lemma 3.1 for $w_k(t) := y_k(\tau + t)$, $t \in [0, \tau^*]$, $\tau \in [t_0 - \delta, T - \tau^*]$, we get a uniform bound for y_k in $L_{r\lambda}([t_0, T], L_{2\lambda}(\Omega))$. This bound and (3.2) show the boundedness of $|y_k|^{\lambda-1}y_k$ in $L_r(J, L_2(\Omega))$. As in (3.1), we get that the sequence $(|y_k|^{\lambda-1}y_k)$ is compact in $L_r(J, L_2(\Omega))$ and $(y_k(T))$ is compact in $L_q(\Omega)$. Now it is easy to pass to the limit to get a solution of (1.3). ∎

Let λ, q be as in Theorem 2.3 and let $r \ge 2$ satisfy (2.9). These assumptions guarantee that there exists $s \notin S_q$ such that

$$0 \vee \Big(\frac{n}{q} - \frac{n}{\lambda}\Big) \vee \Big[\frac{n}{q} - \frac{1}{\lambda}\Big(2 + \frac{n}{q}\Big)\Big] < s < \frac{2}{\lambda} \wedge \Big(2 + \frac{n}{q} - \frac{n}{2} - \frac{2}{r}\Big) \wedge \Big[\frac{1}{\lambda}\Big(2 + \frac{n}{q}\Big) - \frac{2}{r} - \frac{n}{2} + 2\Big]. \tag{3.3}$$

LEMMA 3.1. *Let λ, q be as in Theorem 2.3 and let $r \geq 2$ satisfy (2.9). Assume $u \in L_r(J, L_2(\Omega))$. Then problem (2.5) is well posed in $L_q(\Omega)$. More precisely, if the norm of u in $L_r(J, L_2(\Omega))$ is bounded by a constant C_r, $y^0 \in L_q(\Omega)$, $\|y^0\|_{L_q(\Omega)} \leq C_q$, and s satisfies (3.3), then there exist $\tau^* = \tau^*(C_r, C_q) > 0$ and a unique solution*

$$y \in C([0, \tau^*], L_q(\Omega)) \cap C((0, \tau^*], W^s_{q,\mathcal{B}}). \tag{3.4}$$

In addition, this solution satisfies

$$\|y(t)\|_{L_q(\Omega)} + t^{s/2}\|y(t)\|_{W^s_{q,\mathcal{B}}} \leq C, \qquad t \in (0, \tau^*], \tag{3.5}$$

where C depends only on s, C_r, C_q (and q, r, λ, Ω). If $\hat{q} \geq q$ satisfies

$$\hat{q} < \frac{2n}{(n-4)_+} \qquad \text{and} \qquad \frac{1}{r} < 1 - \frac{n}{4} + \frac{n}{2\hat{q}}, \tag{3.6}$$

then

$$y \in C((0, \tau^*], L_{\hat{q}}(\Omega)) \tag{3.7}$$

and

$$\|y(t)\|_{L_{\hat{q}}(\Omega)} \leq C(\delta, \hat{q}, C_r, C_q), \qquad t \in [\delta, \tau^*], \quad \delta \in (0, \tau^*). \tag{3.8}$$

Finally,

$$y \in C([\delta, \tau^*], W^1_{2,\mathcal{B}}(\Omega)) \cap L_{r\lambda}([\delta, \tau^*], L_{2\lambda}(\Omega)) \tag{3.9}$$

for any $\delta > 0$, and the norm of y in this space can be bounded by $C(\delta, C_r, C_q)$.

Proof. The proof of the first part is an easy modification of [8, Theorem 4.1]. In fact, let X be the Banach space of all functions

$$y \in C([0, \tau^*], L_q(\Omega)) \cap C((0, \tau^*], W^s_{q,\mathcal{B}})$$

for which

$$\|y\|_X := \sup_{t \in (0, \tau^*]} (\|y(t)\|_{L_q(\Omega)} + t^{s/2}\|y(t)\|_{W^s_{q,\mathcal{B}}}) < \infty.$$

Then it is sufficient to use the Banach fixed point theorem for the mapping

$$Ky(t) = e^{-At}y^0 + \int_0^t e^{-A(t-\tau)}(|y(\tau)|^{\lambda-1}y(\tau) + u(\tau))\, d\tau$$

in a large closed ball $\mathbb{B}$ of X with radius R, where A is as in (2.4). For example, assume that $y \in \mathbb{B}$ and denote by $\|\cdot\|_s$ the norm in $W^s_{q,\mathcal{B}}$. Fixing s satisfying (3.3), there exists

$$z \in (1, q] \quad \text{such that} \qquad \lambda\left(\frac{n}{q} - s\right) \vee \lambda\left(\frac{2}{r} + \frac{n}{2} - 2\right) < \frac{n}{z} < 2 + \frac{n}{q} - s\lambda. \tag{3.10}$$

Choose $\sigma_1 \in (s\lambda, 2+n/q-n/z)$ and $\sigma_2 \in (s+2/r, 2+n/q-n/2)$, $\sigma_1, \sigma_2 \notin S_q$. Then we have $L_z(\Omega) \hookrightarrow W^{\sigma_1-2}_{q,\mathcal{B}}$ and $L_2(\Omega) \hookrightarrow W^{\sigma_2-2}_{q,\mathcal{B}}$; hence it follows from (2.3) that

$$\begin{aligned}
t^{s/2}\|Ky(t)\|_s &\le C(C_q) + Ct^{s/2}\int_0^t (t-\tau)^{(\sigma_1-s)/2-1}\||y(\tau)|^{\lambda-1}y(\tau)\|_{\sigma_1-2}\,d\tau \\
&\quad + Ct^{s/2}\int_0^t (t-\tau)^{(\sigma_2-s)/2-1}\|u(\tau)\|_{\sigma_2-2}\,d\tau \\
&\le C(C_q) + Ct^{s/2}\int_0^t (t-\tau)^{(\sigma_1-s)/2-1}\|y(\tau)\|_s^{\lambda}\,d\tau \\
&\quad + Ct^{s/2}\int_0^t (t-\tau)^{(\sigma_2-s)/2-1}\|u(\tau)\|_{L_2(\Omega)}\,d\tau \\
&\le C(C_q) + CR^{\lambda}t^{s/2}\int_0^t (t-\tau)^{(\sigma_1-s)/2-1}\tau^{-s\lambda/2}\,d\tau \\
&\quad + CC_rt^{s/2}\left(\int_0^t (t-\tau)^{r'[(\sigma_2-s)/2-1]}\,d\tau\right)^{1/r'},
\end{aligned}$$

which shows $t^{s/2}\|Ky(t)\|_s < R/2$ if $R = R(C_q)$ is large enough and $t = t(R, C_r)$ is small enough. Similar arguments show the same bound for $\|Ky(t)\|_{L_q(\Omega)}$ and the fact that K is a contraction. Obviously, the fixed point of K is a solution of our problem. Uniqueness of this solution in the class (3.4) can be proved in the same way as in [7, pp. 295–296].

We have $W^s_{q,\mathcal{B}} \hookrightarrow L_{q_1}(\Omega)$ whenever $n/q_1 > n/q - s$. Due to the upper bound for s in (3.3), q_1 is restricted by the conditions

$$\frac{n}{q_1} > -2 + \frac{2}{r} + \frac{n}{2} \quad \text{and} \quad \frac{n}{q_1} > \frac{n}{q} - \varepsilon(q), \tag{3.11}$$

where

$$\varepsilon(q) := \frac{2}{\lambda} + \frac{n}{\lambda q} + 2 - \frac{2}{r} - \frac{n}{2} > 0.$$

Let $\hat q \ge q$ satisfy (3.6). If $n/\hat q > n/q - \varepsilon(q)$, then $W^s_{q,\mathcal{B}} \hookrightarrow L_{\hat q}(\Omega)$ since the second inequality in (3.6) guarantees that the first condition in (3.11) is satisfied with $q_1 = \hat q$. Consequently, (3.7) and (3.8) follow from (3.4) and (3.5). If $n/\hat q \le n/q - \varepsilon(q)$, then we fix $q_1 > q$ satisfying (3.11) (this is possible due to (2.9)). Now the first part of the lemma with q replaced by q_1 (and $t = 0$ replaced by $t = \delta_1$, where $\delta_1 > 0$ is small) implies $y \in C((\delta_1, \tau^*], W^{s_1}_{q_1,\mathcal{B}})$. Similarly as above, $W^{s_1}_{q_1,\mathcal{B}} \hookrightarrow L_{q_2}(\Omega)$, where

$$\frac{n}{q_2} > -2 + \frac{2}{r} + \frac{n}{2} \quad \text{and} \quad \frac{n}{q_2} > \frac{n}{q_1} - \varepsilon(q_1).$$

Repeating this bootstrapping argument finitely many times, we obtain (3.7) and (3.8).

It remains to prove (3.9) and the corresponding bound. Fix $\delta \in (0, \tau^*)$ and set $t_0 := \delta/2$, $J_0 := [t_0, \tau^*]$, and $J^* := [\delta, \tau^*]$. Taking $R > 1$ large and $\hat q$ close to its upper bound, we have $\hat q > \lambda$ and $|y|^{\lambda} \in L_R(J_0, L_{\hat q/\lambda}(\Omega))$. Set $f_1 := |y|^{\lambda-1}y$ and $f_2 := u$. Writing $y = y_1 + y_2 + y_3$, where $\mathcal{B}y_i = 0$, $i = 1, 2, 3$, and

$$\left.\begin{aligned}
&\partial_t y_1 - \Delta y_1 = f_1 \quad \text{in } \Omega \times J_0, \qquad y_1(t_0) = 0, \\
&\partial_t y_2 - \Delta y_2 = f_2 \quad \text{in } \Omega \times J_0, \qquad y_1(t_0) = 0, \\
&\partial_t y_3 - \Delta y_3 = 0 \quad \text{in } \Omega \times J_0, \qquad y_1(t_0) = y(t_0),
\end{aligned}\right\} \tag{3.12}$$

the maximal Sobolev regularity implies

$$y_1 \in W^1_R(J_0, L_{\hat{q}/\lambda}(\Omega)) \cap L_R(J_0, W^2_{\hat{q}/\lambda,\mathcal{B}}) \hookrightarrow C(J_0, W^1_{2,\mathcal{B}})$$

since we can take $\hat{q} > 2\lambda n/(n+2)$ and R arbitrarily large. Similar arguments guarantee $y_2 \in C(J_0, W^1_{2,\mathcal{B}})$ and $y_3 \in C(J^*, W^1_{2,\mathcal{B}})$; hence $y \in C(J^*, W^1_{2,\mathcal{B}})$ (and the corresponding estimate in this space is valid).

Choose $k > 1$ such that $\hat{q} > (\lambda - 1/k)n/2$ and fix $m \in \mathbb{N}$ such that $k^m\hat{q} > 2\lambda$. Choose also $R > r\lambda^{m+1}$. Set $t_i := \delta/2 + i\delta/(2m+2)$, $J_i := [t_i, \tau^*]$, and $\hat{q}_i := k^i\hat{q}$, $i = 1, 2, \dots, m$. Notice that $y_2, y_3 \in L_{r\lambda}(J_1, L_{2\lambda}(\Omega))$ and $W^2_{\hat{q}/\lambda,\mathcal{B}} \hookrightarrow L_{\hat{q}_1}(\Omega)$; hence $y_1 \in L_R(J_1, L_{\hat{q}_1}(\Omega))$. Consequently, $|y|^\lambda$ can be written in the form

$$|y|^\lambda = \tilde{f}_1 + \tilde{f}_2, \qquad \tilde{f}_1 \in L_{R/\lambda}(J_1, L_{\hat{q}_1/\lambda}(\Omega)), \quad \tilde{f}_2 \in L_r(J_1, L_2(\Omega)).$$

Writing $y = \tilde{y}_1 + \tilde{y}_2 + \tilde{y}_3$, where $\mathcal{B}\tilde{y}_i = 0$, $i = 1, 2, 3$, and $\tilde{y}_1, \tilde{y}_2, \tilde{y}_3$ satisfy (3.12) with f_1, f_2, J_0, t_0 replaced by $\tilde{f}_1, \tilde{f}_2, J_1, t_1$, respectively, we obtain as above $\tilde{y}_2, \tilde{y}_3 \in L_{r\lambda}(J_2, L_{2\lambda}(\Omega))$ and $\tilde{y}_1 \in L_{R/\lambda}(J_2, L_{\hat{q}_2}(\Omega))$. Repeating this argument m times we get

$$y \in L_{r\lambda}(J^*, L_{2\lambda}(\Omega)) + L_{R/\lambda^m}(J^*, L_{\hat{q}_m}(\Omega)) = L_{r\lambda}(J^*, L_{2\lambda}(\Omega))$$

(and the corresponding estimates), which concludes the proof. □

LEMMA 3.2. *Let λ, q, r be as in Theorem 2.3. Let $t_1 > 0$, $u \in L_r([0, T + t_1], L_2(\Omega))$, and let its norm in this space be bounded by a positive constant C_r. Assume that y is a global solution of (2.5) (with J replaced by $[0, T + t_1]$) and $y^0 \in W^2_q(\Omega)$, $\|y^0\|_{W^2_q(\Omega)} \le C_q$. Then there exists a constant $C = C(C_r, C_q, t_1)$ such that $\|y(t)\|_{L_q(\Omega)} \le C$ for any $t \in [0, T]$.*

Proof. The proof is a modification of the proof of the main result in [25] (cf. also [26] and [27, proof of Theorem 5.1]).

All our constants (and bounds) in this proof may change from line to line and may depend on C_r, C_q, t_1. First we deduce from Lemma 3.1 and the beginning of the proof of Theorem 2.3 that $y \in C([0, T + t_1], W^1_{2,\mathcal{B}})$ and there exists $\tau > 0$ such that

$$y \text{ is bounded in } C([0, \tau], L_q(\Omega)) \text{ by a constant } C = C(C_q, C_r). \tag{3.13}$$

Denote

$$V(t) = \frac{1}{2}\int_\Omega |\nabla y(x,t)|^2\,dx - \frac{1}{\lambda+1}\int_\Omega |y(x,t)|^{\lambda+1}\,dx.$$

If u is smooth, then

$$V'(t) = -\int_\Omega (\partial_t y)^2\,dx + \int_\Omega u\partial_t y\,dx \le \frac{1}{2}\int_\Omega u^2\,dx - \frac{1}{2}\int_\Omega (\partial_t y)^2\,dx;$$

hence

$$V(\tau_2) - V(\tau_1) \le C - \frac{1}{2}\int_{\tau_1}^{\tau_2}\int_\Omega (\partial_t y)^2\,dx\,dt. \tag{3.14}$$

Now let u be general. Approximating u by smooth functions u_k we see that (3.14) remains true for any $u \in L_r([0, T + t_1], L_2(\Omega))$.

We will show that $V(t)$ is bounded for $t \in [0, T]$. The upper estimate for $V(t)$ follows immediately from (3.14). To prove the lower estimate we assume on the

contrary that $V(t_0) \le -(C+K)$ for some $t_0 \in [0,T]$, where C is from (3.14) and $K \gg 1$. Then (3.14) guarantees $V(t) \le -K$ for all $t \ge t_0$. Multiplying the equation in (2.5) by y and integrating over Ω we obtain

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\int_\Omega y^2\,dx &= -2V(t) + c_1\int_\Omega |y|^{\lambda+1}\,dx + \int_\Omega uy\,dx \\ &\ge K + c_2\left(\int_\Omega y^2\,dx\right)^{(\lambda+1)/2} - C_2\int_\Omega u^2\,dx, \end{aligned} \tag{3.15}$$

where the inequality is true for all $t \ge t_0$. Denote $Y(t) = \int_{t_0}^t \int_\Omega y^2\,dx\,dt$. Then integrating estimate (3.15) we get

$$Y' \ge c_3 Y^{(\lambda+1)/2} + 2K(t-t_0) - C_3.$$

Let $K \ge 10C_3/t_1$. Integrating the inequality $Y' \ge 2K(t-t_0) - C_3$ on $[t_0, t_0+t_1/2]$ we obtain

$$Y(t_0+t_1/2) \ge K\frac{t_1^2}{4} - C_3\frac{t_1}{2} \ge K\frac{t_1^2}{5}.$$

We also have

$$Y' \ge c_3 Y^{(\lambda+1)/2} \quad \text{for} \quad t \ge t_0 + \frac{t_1}{2}.$$

Since the solution of the equation $Z'(t) = c_3 Z^{(\lambda+1)/2}(t)$ for $t \ge 0$, $Z(0) = Kt_1^2/5$, blows up at $t < t_1/2$ if K is large enough, the function $Y(t) \ge Z(t-t_0-t_1/2)$ blows up at some $t < T + t_1$, which yields a contradiction. Hence K has to be bounded by a constant depending only on c_3, C_3, t_1, and λ. Consequently, V is bounded on $[0,T]$ and (3.14) provides a bound for y in the space $W_2^1([0,T], L_2(\Omega))$.

If $\lambda < 1 + 4/n$, then Lemma 3.1 with q replaced by $\tilde{q} := 2$ and $\hat{q}$ replaced by q guarantees a bound for y in $L_\infty([\tau,T], L_q(\Omega))$ which (together with (3.13)) implies the assertion.

Let $\lambda \ge 1 + 4/n$. Since y is bounded in $W_2^1([0,T], L_2(\Omega)) \hookrightarrow L_\infty([0,T], L_2(\Omega))$, we have

$$\int_0^T \|uy\|_{L_1(\Omega)}^z\,dt \le C\int_0^T \|u\|_{L_2(\Omega)}^z\,dt \le C, \qquad z \le r.$$

Using this bound and the boundedness of V on $[0,T]$, we obtain from the equality in (3.15)

$$\int_0^T \|y(t)\|_{L_{\lambda+1}(\Omega)}^{z(\lambda+1)}\,dt \le C\left(1 + \int_0^T \|\partial_t y(t) y(t)\|_{L_1(\Omega)}^z\,dt\right).$$

In particular, if $z = 2$, then this estimate, the bound for y in $W_2^1([0,T], L_2(\Omega))$ and

$$\|\partial_t y(t) y(t)\|_{L_1(\Omega)} \le \|\partial_t y(t)\|_{L_2(\Omega)} \|y(t)\|_{L_2(\Omega)} \le C\|\partial_t y\|_{L_2(\Omega)}$$

guarantee a uniform bound for y in

$$X_z := L_{z(\lambda+1)}([0,T], L_{\lambda+1}(\Omega)).$$

Interpolating between the bound of y in X_z and in $W_2^1([0,T],L_2(\Omega))$ yields a bound in $L_\infty([0,T],L_m(\Omega))$ provided

$$m < \lambda + 1 - \frac{\lambda - 1}{z+1} \tag{3.16}$$

(cf. [25, (12)]). If $r > 2$, then we will use the bootstrapping procedure in [25] in order to get these estimates for some $z > 2$. Replacing u by y, p by λ, q by z, $\tilde{q}$ by $\tilde{z}$, and λ by m in [25], denoting

$$\lambda_1 := (\lambda+1)/\lambda, \qquad \theta := \frac{\lambda+1}{\lambda-1}\,\frac{m-2}{m}, \qquad \beta := \frac{2}{(1-\theta)\tilde{z}},$$

and assuming the estimate in X_z for some $z \ge 2$, we get for $\tilde{z} > z$

$$\begin{aligned}
\int_0^T \|y(t)\|_{L_{\lambda+1}(\Omega)}^{\tilde{z}(\lambda+1)}\,dt &\le C\left(1 + \int_0^T \|\partial_t y(t) y(t)\|_{L_1(\Omega)}^{\tilde{z}}\,dt\right)\\
&\le C\left(1 + \int_0^T \|\partial_t y(t)\|_{L_{m'}(\Omega)}^{\tilde{z}}\,dt\right)\\
&\le C\left(1 + \int_0^T \|\partial_t y(t)\|_{L_{\lambda_1}(\Omega)}^{\theta\tilde{z}} \|\partial_t y(t)\|_{L_2(\Omega)}^{(1-\theta)\tilde{z}}\,dt\right)\\
&\le C\left(1 + \left(\int_0^T \|\partial_t y(t)\|_{L_{\lambda_1}(\Omega)}^{\theta\beta'\tilde{z}}\,dt\right)^{1/\beta'}\right)\\
&\le C\left(1 + \left(\int_0^T \|y(t)\|_{L_{\lambda+1}(\Omega)}^{\theta\beta'\tilde{z}\lambda}\,dt\right)^{1/\beta'}\right),
\end{aligned}$$

provided $\tilde{z} \le r$ and

$$u \in L_{\theta\beta'\tilde{z}}(J, L_{\lambda_1}(\Omega)). \tag{3.17}$$

Recall from [25] that the bootstrap condition $\theta\beta' \le \lambda_1$ is satisfied if m is chosen close to its upper bound and $\tilde{z}$ is close to z. For such m and $\tilde{z}$, one can even check that $\theta\beta' < (\lambda+1)r/(\lambda r+2)$ provided $\tilde{z} < r$. Consequently, $\theta\beta'\tilde{z} \vee \tilde{z} < r$ (and (3.17) is true) whenever $\tilde{z} < (\lambda r+2)/(\lambda+1)$. Hence, we obtain a bound for y in X_z for any

$$z < (\lambda r+2)/(\lambda+1). \tag{3.18}$$

Recall that this guarantees a bound in $L_\infty([0,T],L_m(\Omega))$ for any m satisfying (3.16). Using (2.10) we can find z satisfying (3.18) and $m \in ((\lambda-1)n/2, q]$ such that (3.16) is true. Now we can use Lemma 3.1 with q replaced by m and $\hat{q}$ replaced by q to get a bound for y in $L_\infty([\tau,T],L_q(\Omega))$ which (together with (3.13)) concludes the proof. □

Remark 3.3. We announced in Remark 2.4(i) that Theorem 2.3 remains true if we replace the nonlinearity $|y|^{\lambda-1}y$ with $|y|^\lambda$. Let us sketch the proof of this statement.

Since y satisfies

$$\partial_t y - \Delta y = |y|^\lambda + u \ge u \qquad \text{in } Q,$$

the parabolic maximum principle implies $y \geq y_L$, where y_L is the solution of the linear problem

$$\left.\begin{aligned} \partial_t y_L - \Delta y_L &= u && \text{in } Q,\\ \mathcal{B} y_L &= 0 && \text{on } \Sigma,\\ y_L(\cdot,0) &= y^0 && \text{in } \Omega. \end{aligned}\right\}$$

Using the same arguments as in (3.1) we see that $y_L \in L_{r\lambda}(J, L_{2\lambda}(\Omega))$ and that the norm of y_L in this space can be bounded by the norm of u in $L_r(J, L_2(\Omega))$ and a suitable norm of y^0. Notice that

$$|y|^\lambda = |y|^{\lambda-1} y + 2|y^-|^{\lambda-1} y^-,$$

where $y^- := -\min(0, y)$ is bounded above by $|y_L|$; hence $2|y^-|^{\lambda-1}y^-$ is bounded in $L_r(J, L_2(\Omega))$. Consequently, replacing u with $\tilde{u} := u + 2|y^-|^{\lambda-1}y^-$, we can repeat word by word the proof of Theorem 2.3.

Optimality of the growth bounds.

Remark 3.4. (i) Consider problem (2.5) with Ω being the unit ball in R^n, $n \geq 3$, $\mathcal{B} = \gamma$, and $\lambda > (n+2)/(n-2)_+$. If $n > 10$, then assume also

$$\lambda < 1 + 4\frac{n - 4 + 2\sqrt{n-1}}{(n-2)(n-10)}. \tag{3.19}$$

Choose a smooth radial, radially decreasing function $\psi : \bar{\Omega} \to \mathbb{R}^+$ satisfying $\psi(0) > 0$ and $\psi(x) = 0$ for $x \in \Gamma$ and denote by w_α the (classical) solution of (2.5) with $u = 0$ and $y^0 = \alpha\psi$, $\alpha \geq 0$. We deduce from [22] and an obvious modification of [20] that there exists $\alpha^* > 0$ with the following property: if $\alpha < \alpha^*$, then $w_\alpha(t)$ exists for all $t \in \mathbb{R}^+$ and $w_\alpha(t) \to 0$ as $t \to \infty$; if $\alpha > \alpha^*$, then this solution blows up in finite time completely.

From now on fix $y^0 = \alpha^*\psi$. Let y_k be the solution of (2.5) with $u = 0$ and the nonlinearity y^λ replaced by $\min(y^\lambda, k)$, $k = 1, 2, \ldots$. Then y_k are globally defined classical solutions, $y_{k+1} \geq y_k$. Set $y^*(t) = \lim_{k\to\infty} y_k(t)$. The results in [22] and [16] guarantee that $y^* \in L_{p,\mathrm{loc}}([0,\infty), L_p(\Omega))$ is a weak solution of (2.5) with $u = 0$ and that there exists $T^* \in (0,\infty)$ such that y^* is a classical solution on $(0, T^*)$ but it blows up at $t = T^*$ in the $L_\infty(\Omega)$-norm. In particular, $w_{\alpha^*} = y^*|_{[0,T^*)}$. Next [12] shows that y^* is a classical solution for all t except for finitely many points $T_0 = T^* < T_1 < \cdots < T_k$. Choose $T > T^*$ such that $T \neq T_j$ for any j and let $y_d(x) := y^*(x, T)$. Choose also $0 < t_1 < t_2 < T^*$ and a smooth function $U : \bar{\Omega} \times [0, T] \to [0,\infty)$ with support $K_U \subset \Omega \times [t_1, t_2]$, $K_U \neq \emptyset$, and denote by y^*_β the solution of (2.5) with $u = \beta U$ and $y^0 = \alpha^*\psi$.

Since $y^* > 0$ in K_U and $y^*_{-\beta} \to y^*$ uniformly in K_U as $\beta \to 0+$, fixing $b > 0$ small we have $|y^*_{-b}|^{\lambda-1} y^*_{-b} - bU \geq 0$ in K_U. Consequently, the maximum principle implies $y^*_{-b} \geq 0$. Choose $\beta \in (0, b]$. Since $y^*(t_2) - y^*_{-\beta}(t_2)$ belongs to the interior of the positive cone in $C^1(\bar{\Omega})$ and $w_\alpha(t_2) \to w_{\alpha^*}(t_2) = y^*(t_2)$ in $C^1(\bar{\Omega})$ as $\alpha \to \alpha^*-$, there exists $\alpha < \alpha^*$ such that $y^*_{-\beta}(t_2) \leq w_\alpha(t_2)$. Now the maximum principle implies $y^*_{-\beta}(t) \leq w_\alpha(t)$ for any $t \geq t_2$ and $y^*_\beta \geq y^*_{-b} \geq 0$ for any $t \geq 0$; hence $y^*_{-\beta}$ is a global nonnegative classical solution. On the other hand, if $\beta \geq 0$, then $y^*_\beta \geq y^*$; hence y^*_β blows up at finite time $T_\beta \leq T^*$ in the $L_\infty(\Omega)$-norm and, consequently, in the $L_q(\Omega)$-norm for any $q > n(\lambda-1)/2$ (cf. [14], [30]).

Let $\mathbb{U}_{\rm ad} = \{\beta U\,;\,\beta \in [-b,b]\}$. Fix $q > n(\lambda-1)/2$ and set $\mathbb{J}(y,u) = \int_\Omega |y(T) - y_d|^q\,dx$. The above arguments show that y^*_β is a global $L_\infty(L_q)$-solution of (2.5) if and only if $\beta < 0$. Moreover, $\beta \mapsto \mathbb{J}(y^*_\beta, \beta U)$ is decreasing on $[-b,0)$. Hence the optimal control problem (1.3) does not have a solution with $y \in L_\infty(J, L_q(\Omega))$.

(ii) Consider problem (2.5) with Ω being the unit ball in R^n and $\mathcal{B} = \gamma$ and let $1 \le q < (\lambda-1)n/2$. Then there exists a smooth radial positive function y^0 such that the solution y of (2.5) with $u = 0$ blows up at $t = T$ in the L_∞-norm and satisfies $\partial_t y \ge 0$, $y_d := y(\cdot,T) \in L_q(\Omega)$ (see [14]). Let U be a smooth nonnegative function with support $K \subset \{(x,t)\,;\,|x| < 1/2\}$, $K \ne \emptyset$, and $u_c := cU$. Then there exists $\varepsilon > 0$ such that the solution y of (2.5) with u replaced by $u_{-\varepsilon}$ remains positive. Let $\mathbb{U}_{\rm ad} = \{u_c\,;\,c \in [-\varepsilon, 0]\}$ and

$$\mathbb{J}(y,u) = \left| \int_\Omega |y(x,T)|^q\,dx - \int_\Omega y_d^q\,dx \right|.$$

Then $(y(u_{-1/k}), u_{-1/k})$, $k \ge k_0$, is obviously a minimizing sequence for the control problem (1.3), but $y(u_0)$ is not a (classical) global solution of (2.5).

4. Proof of the optimality conditions. We start with the following technical lemma concerning linear problems.

LEMMA 4.1. *Suppose that $\beta > 2 \vee (n+2)/2$ and $2 \le q < 2n/(n-2)_+$. Given $a \in L_\beta(Q)$, $u \in L_2(Q)$, and $y^0 \in L_{q'}(\Omega)$, the problem*

$$\left.\begin{aligned} \partial_t y - \Delta y &= ay + u && \text{in } Q,\\ \mathcal{B}y &= 0 && \text{on } \Sigma,\\ y(\cdot,0) &= y^0 && \text{in } \Omega \end{aligned}\right\} \tag{4.1}$$

has a unique solution

$$y \in C([0,T], L_{q'}(\Omega)) \cap C((0,T], L_q(\Omega)) \cap L_2(Q).$$

The map

$$L_\beta(Q) \times L_2(Q) \times L_{q'}(\Omega) \to L_2(Q) \times L_q(\Omega), \qquad (a,u,y^0) \mapsto (y, y(T)),$$

is analytic and bounded on bounded sets.

Proof. (i) Writing (4.1) in the abstract form

$$\dot y + Ay = ay + u \quad \text{in } (0,T], \qquad y(0) = y^0,$$

and denoting $U(t) := e^{-tA}$, we see that we have to prove the unique solvability of

$$y = U * (ay) + U * u + Uy^0 \tag{4.2}$$

in appropriate spaces.

(ii) Fix $s \in [0,1) \setminus \{1/q'\}$ such that $q \le 2n/(n-2s)_+$. Then $W^s_{q',\mathcal{B}} \hookrightarrow L_2(\Omega)$. Hence we infer from (2.3) (with q replaced by q' and $r := 0$) that

$$\|U(t)y^0\|_{L_2(\Omega)} \le c\|U(t)y^0\|_{W^s_{q',\mathcal{B}}} \le ct^{-s/2}\|y^0\|_{L_{q'}(\Omega)}, \qquad 0 < t \le T.$$

Since $s < 1$ it follows that

$$(y^0 \mapsto Uy^0) \in \mathcal{L}\big(L_{q'}(\Omega), L_2((0,T), L_2(\Omega))\big) = \mathcal{L}\big(L_{q'}(\Omega), L_2(Q)\big),$$

where $\mathcal{L}(X,Y)$ denotes the space of continuous linear operators from X to Y.

(iii) It is easy to see that

$$(u \mapsto U * u) \in \mathcal{L}(L_2(Q)).$$

(iv) Put $1/r := 1/\beta + 1/2 < 1$ and note that

$$L_r(\Omega) \hookrightarrow W^{-2+\gamma}_{2,\mathcal{B}} \quad \text{if} \quad 1/2 \geq 1/r + (\gamma - 2)/n,$$

that is, if $0 \leq \gamma \leq 2 - n/\beta$.

(v) For $m \in \mathbb{R}$ we write $L_{2,m}(Q)$ for $L_2(Q)$ endowed with the equivalent norm

$$y \mapsto \left(\int_0^T e^{-2mt} \|y(t)\|^2_{L_2(\Omega)}\, dt \right)^{1/2}.$$

From (iv), Hölder's inequality, and (2.3) (with $q = 2$ and $r := \gamma - 2$) we infer that

$$\begin{aligned} \|U * (ay)(t)\|_{L_2(\Omega)} &\leq c \int_0^t (t-\tau)^{\gamma/2-1} \|a(\tau)\|_{L_\beta(\Omega)} \|y(\tau)\|_{L_2(\Omega)}\, d\tau \\ &= ce^{mt} \int_0^t (t-\tau)^{\gamma/2-1} e^{-m(t-\tau)} \|a(\tau)\|_{L_\beta(\Omega)} e^{-m\tau} \|y(\tau)\|_{L_2(\Omega)}\, d\tau. \end{aligned}$$

Thus, by Young's inequality for convolutions (cf. the proof of [5, Lemma 3]), followed by Hölder's inequality,

$$\begin{aligned} \|U * (ay)\|_{L_{2,m}(Q)} &\leq cI(m) \left(\int_0^T \big(\|a(\tau)\|_{L_\beta(\Omega)} e^{-m\tau} \|y(\tau)\|_{L_2(\Omega)} \big)^r\, d\tau \right)^{1/r} \\ &\leq cI(m) \|a\|_{L_\beta(Q)} \|y\|_{L_{2,m}(Q)}, \end{aligned}$$

where

$$I(m) := \left(\int_0^T t^{(\gamma/2-1)\beta'} e^{-\beta' mt}\, dt \right)^{1/\beta'},$$

provided $\gamma > 2/\beta$. Such a choice is possible by (iv), thanks to $2/\beta < 2 - n/\beta$.

(vi) For $a \in L_\beta(Q)$ set $T_a(y) := U * (ay)$. Then (v) implies

$$(a \mapsto T_a) \in \mathcal{L}\big(L_\beta(Q), \mathcal{L}(L_{2,m}(Q))\big)$$

and

$$\|T_a\|_{\mathcal{L}(L_{2,m}(Q))} \leq cI(m) \|a\|_{L_\beta(Q)}.$$

Note that, by Lebesgue's theorem, $I(m) \to 0$ as $m \to \infty$. Thus, given $R > 0$, there exists $m := m_R > 0$ such that $\|T_a\|_{\mathcal{L}(L_{2,m}(Q))} \leq 1/2$ for all $a \in L_\beta(Q)$ satisfying $\|a\|_{L_\beta(Q)} \leq R$. Consequently, $1 - T_a$ has a bounded inverse on $L_{2,m}(Q)$, and the map $a \mapsto (1 - T_a)^{-1}$ is analytic for $\|a\|_{L_\beta(Q)} \leq R$. Hence, by (4.2),

$$y = (1 - T_a)^{-1}(U * u + Uy^0) \in L_2(Q)$$

for $\|a\|_{L_\beta(Q)} \le R$, thanks to (ii) and (iii), and the map

$$L_\beta(Q) \times L_2(Q) \times L_{q'}(\Omega) \to L_2(Q), \quad (a, u, y^0) \mapsto y$$

is analytic and bounded on bounded sets.

(vii) Let

$$q' \le q_1 \le 2 \le q_2 \le q \quad \text{with} \quad \frac{1}{n} > \frac{1}{q_1} - \frac{1}{q_2}.$$

Choose s such that

$$1 + n\left(\frac{1}{2} - \frac{1}{q_1}\right) > s > n\left(\frac{1}{2} - \frac{1}{q_2}\right). \tag{4.3}$$

Then there exists $\xi \in (1/2, 1)$ such that $2 - 2\xi + n(1/2 - 1/q_1) > s$. This choice of s, ξ guarantees

$$W^s_{2,\mathcal{B}} \hookrightarrow L_{q_2}(\Omega) \tag{4.4}$$

and

$$L_{q_1}(\Omega) \hookrightarrow W^{s-2+2\xi}_{2,\mathcal{B}}. \tag{4.5}$$

(viii) Let q_1, q_2, s, ξ be as in (vii). For $m \in \mathbb{R}$ we denote by $C_{1-\xi,m}((0,T], L_{q_2}(\Omega))$ the Banach space of all $v \in C((0,T], L_{q_2}(\Omega))$ such that $\sup_{0<t\le T} t^{1-\xi}\|v(t)\|_{L_{q_2}(\Omega)} < \infty$, endowed with the norm

$$\|v\|_{C_{1-\xi,m}} := \sup_{0<t\le T} t^{1-\xi} e^{-mt} \|v(t)\|_{L_{q_2}(\Omega)}.$$

It is an easy consequence of (2.3), (4.4), and (4.5) that

$$(y^0 \mapsto Uy^0) \in \mathcal{L}(L_{q_1}(\Omega), C_{1-\xi,m}((0,T], L_{q_2}(\Omega))).$$

(ix) Let q_1, q_2, s, ξ be as in (vii). Using (2.3) we get

$$\begin{aligned} \|U * u(t)\|_{L_{q_2}(\Omega)} &\le c\|U * u(t)\|_{W^s_{2,\mathcal{B}}} \le c \int_0^t (t-\tau)^{-s/2} \|u(\tau)\|_{L_2(\Omega)}\, d\tau \\ &\le ct^{(1-s)/2}\|u\|_{L_2(Q)} \le c\|u\|_{L_2(Q)} \end{aligned}$$

for $0 < t \le T$. In particular,

$$(u \mapsto U * u) \in \mathcal{L}(L_2(Q), C_{1-\xi,m}((0,T], L_{q_2}(\Omega))).$$

(x) Let q_1, q_2, s, ξ be as in (vii) such that s also satisfies

$$2 - \frac{n+2}{\beta} > s - n\left(\frac{1}{2} - \frac{1}{q_2}\right).$$

Then there exists $\eta > 1/\beta$ such that

$$2 - \frac{n}{\beta} - 2\eta > s - n\left(\frac{1}{2} - \frac{1}{q_2}\right).$$

Hence

$$L_r(\Omega) \hookrightarrow W^{s-2+2\eta}_{2,\mathcal{B}}, \tag{4.6}$$

where $1/r := 1/\beta + 1/q_2$. With this choice it follows that

$$\begin{aligned} e^{-mt}\|U*(ay)(t)\|_{L_{q_2}(\Omega)} &\le ce^{-mt}\|U*(ay)(t)\|_{W^s_{2,\mathcal{B}}} \\ &\le ce^{-mt}\int_0^t (t-\tau)^{\eta-1}\|a(\tau)\|_{L_\beta(\Omega)}\|y(\tau)\|_{L_{q_2}(\Omega)}\,d\tau \\ &\le c\int_0^t (t-\tau)^{\eta-1}\tau^{\xi-1}e^{-m(t-\tau)}\|a(\tau)\|_{L_\beta(\Omega)}\,d\tau\ \|y\|_{C_{1-\xi,m}} \end{aligned}$$

for $0 < t \le T$. Thus, by Hölder's inequality,

$$t^{1-\xi}e^{-mt}\|U*(ay)(t)\|_{L_{q_2}(\Omega)} \le cK(t,m)\|a\|_{L_\beta(Q)}\|y\|_{C_{1-\xi,m}},$$

where

$$\begin{aligned} K(m,t) &:= t^{1-\xi}\left(\int_0^t (t-\tau)^{(\eta-1)\beta'}\tau^{(\xi-1)\beta'}e^{-\beta' m(t-\tau)}\,d\tau\right)^{1/\beta'} \\ &= t^{\eta-1/\beta}\left(\int_0^1 (1-\sigma)^{(\eta-1)\beta'}\sigma^{(\xi-1)\beta'}e^{-\beta' mt(1-\sigma)}\,d\sigma\right)^{1/\beta'}. \end{aligned}$$

Fix any $\delta \in (0,T)$. Then $K(t,m) \to 0$ as $m \to \infty$ by Lebesgue's theorem, uniformly with respect to $t \in [\delta, T]$. If $0 < t \le \delta$, then

$$K(t,m) \le c\delta^{\eta-1/\beta}.$$

Thus, given $R > 0$, it follows that we can fix $m > 0$ such that

$$\|T_a\|_{\mathcal{L}(C_{1-\xi,m}((0,T],L_{q_2}(\Omega)))} \le 1/2$$

for all $a \in L_\beta(Q)$ satisfying $\|a\|_{L_\beta(Q)} < R$. Now we infer from (viii) and (ix) that

$$y = (1-T_a)^{-1}(U*u + Uy^0) \in C_{1-\xi,m}((0,T],L_{q_2}(\Omega))$$

for $y^0 \in L_{q_1}(\Omega)$ and $\|a\|_{L_\beta(Q)} < R$ and that the map

$$L_\beta(Q)\times L_2(Q)\times L_{q_1}(\Omega) \to C_{1-\xi,m}((0,T],L_{q_2}(\Omega)), \qquad (a,u,y^0)\mapsto y$$

is analytic and bounded on bounded sets. Using this property for the couple $(q_1,q_2) := (q',2)$ and, subsequently, for $(q_1,q_2) := (2,q)$, we see that the map

$$L_\beta(Q)\times L_2(Q)\times L_{q'}(\Omega) \to L_q(\Omega), \qquad (a,u,y^0)\mapsto y(T)$$

is analytic and bounded on bounded sets. This concludes the proof. □

Remark 4.2. Lemma 4.1 guarantees the solvability of (2.13): notice that $r = 2$ and (2.9) imply $q < 2n/(n-2)_+$, that $a := \lambda|y|^{\lambda-1} \in L_\beta(Q)$ for some $\beta > 2\vee(n+2)/2$ due to $y \in L_{2\lambda}(Q)$ and $\lambda < (n+2)/(n-2)_+$, and that $p(\cdot,T) \in L_{q'(\Omega)}$ due to $y(\cdot,T) \in L_q(\Omega)$.

Proof of Theorem 2.6. Choose $v \in \mathbb{U}_{\rm ad}$, $\mu \in [0,1]$ and let y_μ be the solution of (2.5) with u replaced by $u + \mu(v-u)$. If μ is small enough, say $\mu \le \mu_0$, then due to

the stability estimates in Theorem A.1 and the regularity results in Theorem 2.3, the solution y_μ is global and satisfies

$$\|y_\mu - y\|_{L_{2\lambda}(Q)} + \|y_\mu(\cdot,T) - y(\cdot,T)\|_{L_q(\Omega)} \le C\mu\|v-u\|_{L_2(Q)}. \tag{4.7}$$

Assume $\mu \le \mu_0$ and set $z_\mu := (y_\mu - y)/\mu$. Then z_μ solves the problem

$$\left.\begin{aligned} \partial_t z_\mu - \Delta z_\mu &= a_\mu z_\mu + (v-u), && x\in\Omega,\ t\in J,\\ \mathcal{B}z_\mu &= 0, && x\in\Gamma,\ t\in J,\\ z_\mu(\cdot,0) &= 0, \end{aligned}\right\} \tag{4.8}$$

where $a_\mu := \lambda\int_0^1 |y+\theta(y_\mu - y)|^{\lambda-1}\,d\theta$. Let z be the solution of

$$\left.\begin{aligned} \partial_t z - \Delta z &= az + (v-u), && x\in\Omega,\ t\in J,\\ \mathcal{B}z &= 0, && x\in\Gamma,\ t\in J,\\ z(\cdot,0) &= 0, \end{aligned}\right\}$$

where $a := \lambda|y|^{\lambda-1}$. Set $\beta := 2\lambda/(\lambda-1)$. Since $a_\mu \to a$ in $L_\beta(Q)$ as $\mu\to 0$, Lemma 4.1 implies

$$z_\mu(\cdot,T) \to z(\cdot,T) \quad \text{in } L_q(\Omega). \tag{4.9}$$

Set

$$\mathcal{I}_1(\mu) := \int_\Omega |y_\mu(\cdot,T) - y^*|^q\,dx, \qquad \mathcal{I}_2(\mu) := N\int_Q (u+\mu(v-u))^2\,dx\,dt.$$

The mapping $L_q(\Omega)\to\mathbb{R} : \varphi \mapsto \int_\Omega |\varphi - y^*|^q\,dx$ is convex. Hence

$$\begin{aligned} q\int_\Omega |y(\cdot,T)-y^*|^{q-2}(y(\cdot,T)-y^*)z_\mu(\cdot,T)\,dx &\le \frac{\mathcal{I}_1(\mu)-\mathcal{I}_1(0)}{\mu}\\ &\le q\int_\Omega |y_\mu(\cdot,T)-y^*|^{q-2}(y_\mu(\cdot,T)-y^*)z_\mu(\cdot,T)\,dx. \end{aligned}$$

Since (4.7) implies

$$|y_\mu(\cdot,T)-y^*|^{q-2}(y_\mu(\cdot,T)-y^*) \to |y(\cdot,T)-y^*|^{q-2}(y(\cdot,T)-y^*)$$

in $L_{q'}(\Omega)$ and (4.9) is true, we see that $\mathcal{I}_1$ is right differentiable at 0 and $\mathcal{I}_1'(0+) = \int_\Omega p(\cdot,T)z(\cdot,T)\,dx$. We also have $\mathcal{I}_2'(0) = 2N\int_Q u(v-u)\,dx\,dt$ and

$$(\mathcal{I}_1+\mathcal{I}_2)(\mu) = \mathbb{J}(y_\mu, u+\mu(v-u)) \ge J(y,u) = (\mathcal{I}_1+\mathcal{I}_2)(0);$$

hence

$$\int_\Omega p(\cdot,T)z(\cdot,T)\,dx + 2N\int_Q u(v-u)\,dx\,dt \ge 0.$$

Consequently, it is sufficient to show that

$$\int_\Omega p(\cdot,T)z(\cdot,T)\,dx = \int_Q p(v-u)\,dx\,dt.$$

Let $\varphi_k \in \mathcal{D}(\Omega)$ be such that $\varphi_k \to p(\cdot,T)$ in $L_{q'}(\Omega)$ and $a_k \in \mathcal{D}(Q)$ be such that $a_k \to a$ in $L_\beta(Q)$. Let p_k be the solution of (2.13) with $a = \lambda|y|^{\lambda-1}$ replaced by a_k and the final condition replaced by $p_k(\cdot,T) = \varphi_k$. Then p_k is smooth and $p_k \to p$ in $L_2(Q)$ due to Lemma 4.1. Notice that $z \in L_{2\lambda}(Q)$ due to Theorem A.1 (cf. the beginning of the proof of Theorem 2.3); hence $az \in L_2(Q)$, and the maximal Sobolev regularity implies $\partial_t z, \Delta z \in L_2(Q)$. We have

$$\begin{aligned}\int_Q p_k(v-u)\,dx\,dt &= \int_Q p_k(\partial_t z - \Delta z - az)\,dx\,dt\\ &= \int_Q (-\partial_t p_k - \Delta p_k - a p_k) z\,dx\,dt + \int_\Omega \varphi_k z(\cdot,T)\,dx\\ &= \int_Q (a_k - a) p_k z\,dx\,dt + \int_\Omega \varphi_k z(\cdot,T)\,dx \to \int_\Omega p(\cdot,T) z(\cdot,T)\,dx,\end{aligned}$$

since the p_k stay bounded in $L_2(Q)$ due to Lemma 4.1. Now

$$\int_Q p_k(v-u)\,dx\,dt \to \int_Q p(v-u)\,dx\,dt$$

concludes the proof. □

5. The case of a multiplicative control.

Proof of Theorem 2.8. The proof is almost the same as in Theorem 2.3 (but the solutions y are more regular now). The only nontrivial modification is required in the estimate of the function V and the $L_2(Q)$-norm of $\partial_t y$ in the proof of Lemma 3.2.

Hence, assume $y \in C([0,T+t_1], W^1_{2,\mathcal{B}})$ is a solution of (2.14), where $t_1 > 0$ is fixed. Since $\mathbb{U}_{\rm ad}$ is bounded in $L_\infty(Q)$, there exists a constant M such that $\|u\|_{L_\infty(Q)} \le M$ for all $u \in \mathbb{U}_{\rm ad}$. Let V be defined as in the proof of Lemma 3.2. Then

$$\begin{aligned}V'(t) &= -\int_\Omega (\partial_t y)^2(t)\,dx + \int_\Omega u y \partial_t y(t)\,dx\\ &\le \frac{M^2}{2}\int_\Omega y^2(t)\,dx - \frac{1}{2}\int_\Omega (\partial_t y)^2(t)\,dx.\end{aligned} \tag{5.1}$$

Let $\tau < 1$, $\tau \le T + t_1$, and $t \in [0,\tau]$. Denoting $C_0 := \int_\Omega y^2(x,0)\,dx$, we have

$$\int_\Omega y^2(t)\,dx = C_0 + 2\int_0^t\int_\Omega y\partial_t y\,dx\,dt \le C_0 + \int_0^\tau\int_\Omega y^2\,dx\,dt + \int_0^\tau\int_\Omega (\partial_t y)^2\,dx\,dt.$$

Integrating this estimate over $t \in [0,\tau]$, we get

$$\int_0^\tau\int_\Omega y^2\,dx\,dt \le C_0\tau + \tau\int_0^\tau\int_\Omega y^2\,dx\,dt + \tau\int_0^\tau\int_\Omega (\partial_t y)^2\,dx\,dt;$$

hence

$$\int_0^\tau\int_\Omega y^2\,dx\,dt \le \frac{C_0\tau}{1-\tau} + \frac{\tau}{1-\tau}\int_0^\tau\int_\Omega (\partial_t y)^2\,dx\,dt. \tag{5.2}$$

Let $\tau_1 \in (0,1)$ be defined by $\frac{\tau_1}{1-\tau_1}M^2 = \frac{1}{2}$ and $\tau \in [0,\tau_1]$ (enlarging M we may assume $\tau_1 \le T + t_1$). Then integrating (5.1) and using (5.2) we arrive at

$$V(\tau) - V(0) \le \frac{C_0}{4} - \frac{1}{4}\int_0^\tau\int_\Omega (\partial_t y)^2\,dx\,dt, \qquad \tau \in [0,\tau_1]. \tag{5.3}$$

This estimate guarantees $V(t) \le V(0) + C_0/4$ on $[0,\tau_1]$.

Fix $\delta \in (0, t_1 \wedge \tau_1)$ and assume $V(t_0) \ll -1$ for some $t_0 \in [0, \tau_1 - \delta]$. Then (5.3) implies $V(t) \le -K \ll -1$ for all $t \in [\tau_1 - \delta, \tau_1]$. As in (3.15) we have

$$\frac{1}{2}\frac{d}{dt}\int_\Omega y^2\,dx = -2V(t) + c_1\int_\Omega |y|^{\lambda+1}\,dx + \int_\Omega uy^2\,dx$$
$$\ge K + c_2\left(\int_\Omega y^2\,dx\right)^{(\lambda+1)/2}$$

for any $t \in [\tau_1 - \delta, \tau_1]$. In the same way as in the proof of Lemma 3.2, this inequality yields a contradiction if $K = K(\lambda, c_2, \delta)$ is large enough. Consequently,

$$V(t) \ge -C \qquad \text{for all } t \in [0, \tau_1 - \delta]. \tag{5.4}$$

Now (5.3) implies $\int_0^{\tau_1-\delta}\int_\Omega (\partial_t y)^2\,dx\,dt \le C$; hence $\int_\Omega y^2(t)\,dx \le C$ for t belonging to $[0, \tau_1 - \delta]$. In particular, $\int_\Omega y^2(\tau_1 - \delta)\,dx \le C_1$, where C_1 does not depend on u.

Repeating the estimates above on the interval $[\tau_1 - \delta, 2\tau_1 - \delta]$ instead of $[0, \tau_1]$ and then on $[2\tau_1 - 2\delta, 3\tau_1 - 2\delta]$, etc., we obtain the desired bounds for $V(t)$, $\|y(t)\|_{L_2(\Omega)}$, $t \in J$, and $\|\partial_t y\|_{L_2(Q)}$. □

6. Parabolic systems.

Proof of Theorem 2.10. Let $\varphi_1 > 0$ be an eigenfunction corresponding to the first eigenvalue μ_1 of the problem $-\Delta\varphi = \mu\varphi$ in Ω, $\mathcal{B}_1\varphi = 0$ on Γ. Notice that φ_1 is a positive constant if $\mathcal{B}_1 = \partial_\nu$; hence the weighted Lebesgue space $L_1(\Omega, \varphi_1(x)\,dx)$ equals $L_1(\Omega)$ in this case.

We shall prove that

(i) any bound of $y_1(t)$ in $L_p(\Omega, \varphi_1(x)\,dx)$ or $L_p(\Omega)$, $p \ge 1$, implies a bound of $y_2(t)$ in the same space;

(ii) the space $X := L_1(\Omega, \varphi_1(x)\,dx) \times L_1(\Omega, \varphi_1(x)\,dx)$ is a *continuation space* for problem (2.15); that is, if the solution y is defined on $[0, T^*]$, $T^* > 0$, and $\|y(T^*)\|_X \le M$, then this solution can be continued for $t \in [T^*, T^* + \tau]$, where $\tau = \tau(M) > 0$. In addition, $\|u(t)\|_{L_\infty(\Omega)\times L_\infty(\Omega)} \le C(\delta, M)$ for any $t \in [T^* + \delta, T^* + \tau]$ and $\delta > 0$;

(iii) all global solutions of problem (2.15) with u bounded in $L_r(J, L_z^+(\Omega))$ and $y_1(T)$ bounded in $L_q(\Omega)$ are uniformly bounded in $L_\infty(Q)$.

Then the conclusion follows similarly as in the proof of Theorem 2.3.

(i) Let $u \in \mathbb{U}_{\rm ad}$. Set $w := y_2^2/2 - by_2 - ay_1$. One can easily verify

$$\partial_t w - \Delta w \le -au \le 0;$$

hence the comparison principle guarantees $w \le C$ in Q, where C does not depend on u. This estimate implies

$$y_2^2 \le C(1 + y_1), \tag{6.1}$$

and the conclusion follows.

(ii) Set $z := ay_1 + by_2$. Then

$$\partial_t z - \Delta z = ay_1y_2 + au \le C(1 + z^{3/2}) + au. \tag{6.2}$$

Since $n < 4$ if $\mathcal{B} = \partial_\nu$ and $n < 3$ if $\mathcal{B} = \gamma$, the problem $\partial_t\tilde z - \Delta\tilde z = C(1 + |\tilde z|^{3/2}) + au$, $\mathcal{B}\tilde z = 0$, is well posed in $X_1 := L_1(\Omega, \varphi_1(x)\,dx)$ due to [30] and [13], respectively. More precisely, if $\|\tilde z(0)\|_{X_1} \le M$, then there exists $\tau = \tau(M) > 0$ such that the solution $\tilde z$ exists on $[0, \tau]$ and satisfies $\|\tilde z(t)\|_{L_\infty(\Omega)} \le C(\delta, M)$ for any $t \in [\delta, \tau]$ and $\delta > 0$.

A comparison argument shows that the same estimate is true for the function z. In particular, the space X is a continuation space for (2.15) in the sense described above.

(iii) Now assume that u belongs to a bounded set in $\mathbb{U}_{\rm ad}^G \subset L_r(J, L_z^+(\Omega))$ and $y_1(T)$ is bounded in $L_1(\Omega)$. The above arguments show that the solution y can be continued for $t \in [0, T+\tau]$, where $\tau > 0$ does not depend on u and $u(x,t) := 0$ if $t > T$. Multiplying the second equation in (2.15) with φ_1 and using (6.1) we obtain

$$\partial_t \int_\Omega y_2\varphi_1\,dx + \mu_1 \int_\Omega y_2\varphi_1\,dx = a\int_\Omega y_1\varphi_1\,dx \ge c\int_\Omega y_2^2\varphi_1\,dx - C$$
$$\ge c\left(\int_\Omega y_2\varphi_1\,dx\right)^2 - C$$

for any $t \in [0, T+\tau]$. Using standard blow-up arguments (cf. the arguments following (3.15) in the proof of Lemma 3.2), this estimate guarantees a uniform bound for $y_2(t)$, $t \in [0, T+\tau/2]$, in the weighted space $L_1(\Omega, \varphi_1(x)\,dx)$. Integrating the second equation in (2.15) we now obtain

$$\int_0^{T+\tau/2}\int_\Omega y_1\varphi_1\,dx\,dt \le C. \tag{6.3}$$

The first equation in (2.15) implies

$$\int_\Omega y_1\varphi_1\,dx\Big|_{t_1}^{t_2} + (\mu_1 + b)\int_{t_1}^{t_2}\int_\Omega y_1\varphi_1\,dx\,dt \ge 0;$$

hence using (6.3) we deduce

$$\int_\Omega y_1(t_2)\varphi_1\,dx \ge \int_\Omega y_1(t_1)\varphi_1\,dx - C \tag{6.4}$$

for any $t_1, t_2 \in [0, T+\tau/2]$, $t_2 > t_1$.

Obviously, (6.3) and (6.4) imply a uniform estimate for $y_1(t)$, $t \in J$, in the space $L_1(\Omega, \varphi_1(x)\,dx)$. Now (i) and (ii) imply uniform bounds for y_1, y_2 in $L_\infty([\delta, T]\times\Omega)$ for any $\delta > 0$. Since the bounds for y_1, y_2 in $L_\infty([0,\delta]\times\Omega)$ for $\delta > 0$ small enough are guaranteed by the well posedness of (2.15) in $L_\infty(\Omega)\times L_\infty(\Omega)$ and the boundedness of u in $L_r(J, L_z(\Omega))$, the conclusion follows. □

Appendix: The basic existence, uniqueness, and stability theorem for semilinear problems. For the reader's convenience we collect here the main existence, uniqueness, and stability results for strong solutions of the semilinear problem

$$\dot y + Ay = F(y) \text{ in } [0,T], \qquad y(0) = y^0, \tag{A.1}$$

where $A = A_s$ is the isomorphism between $W^s_{q,\mathcal{B}}$ and $W^{s-2}_{q,\mathcal{B}}$ mentioned in section 2. They follow from [6, Theorems 3.3 and 3.4] and [5, Theorems 5 and 7(ii)]. Analogous results are true in the case of systems.

We write $C_b^{1-}(Y, X)$ for the space of all maps from Y into X which are uniformly Lipschitz continuous on bounded sets. If X and Y are paces of functions defined on $[0,T]$, then $F : X \to Y$ is said to possess the Volterra property if, given any $u \in X$ and $t \in (0,T)$, the restriction of $F(u)$ to $[0,t]$ depends on the values of $u\,|\,[0,t]$ only.

THEOREM A.1. *Assume*

$$s, \sigma \notin S_q, \quad 0 \le s < \sigma < 2, \tag{A.2}$$

and suppose that $r > 1$, $r \neq 2/(\sigma - s)$, $\sigma - 2/r \notin S_q$, $y^0 \in Y^0 := W_{q,\mathcal{B}}^{\sigma-2/r}$. *Denote* $X_t := L_r([0,t], W_{q,\mathcal{B}}^{\sigma-2})$.

If $r < 2/(\sigma - s)$, *fix* $p \in [1, 2/(s - \sigma + 2/r))$ *and set* $Y_t := L_p([0,t], W_{q,\mathcal{B}}^s)$,
if $r > 2/(\sigma - s)$, *fix* $\rho \in [0, (\sigma - s - 2/r)/2)$ *and set* $Y_t := C^\rho([0,t], W_{q,\mathcal{B}}^s)$.

Let $F \in C_b^{1-}(Y_T, X_T)$ *have the Volterra property. If* $r < 2/(\sigma - s)$ *or* $r > 2/(\sigma - s)$, *then* (A.1) *has a unique strong* $L_p(W_q^s)$- *or* $C^\rho(W_q^s)$-*solution* $y(y^0, F)$, *respectively, defined on the maximal existence interval* $[0, t(y^0, F))$. *If* $y(y^0, F) \in Y_{t(y^0,F)}$ *or* $F(y(y^0, F)) \in X_{t(y^0,F)}$, *then* $y(y^0, F)$ *is global.*

The map $(y^0, F) \mapsto y(y^0, F)$ *is Lipschitz continuous in the following sense: Fix* $t < t(y^0, F)$ *(we can take* $t = t(y^0, F) = T$ *if* $y(y^0, F)$ *is global). Let* $\omega_1 > 0$, *and let* $\omega_2 : \mathbb{R}^+ \to \mathbb{R}^+$ *be an increasing function*

$$\left.\begin{aligned} &\|y^0\|_{Y^0} + \|F(0)\|_{X_T} \le \omega_1, \\ &\|F(y_1) - F(y_2)\|_{X_T} \le \omega_2(R)\|y_1 - y_2\|_{Y_T} \end{aligned}\right\} \tag{A.3}$$

for any $R > 0$ *and* $y_1, y_2 \in Y_T$ *whose norms are bounded by* R. *Fix* $R > \|y(y^0, F)\|_{Y_t}$. *Then there exist positive constants* ε, c *(depending only on* R, t, ω_1, ω_2*) with the following property: If* $\tilde{y}^0 \in Y^0$, $\tilde{F} \in C_b^{1-}(Y_T, X_T)$ *has the Volterra property,* $\tilde{y}^0$ *and* $\tilde{F}$ *satisfy* (A.3), *and*

$$\|y^0 - \tilde{y}^0\|_{Y^0} + \sup_{\|y\|_{Y_T} \le R} \|(F - \tilde{F})(y)\|_{X_T} \le \varepsilon,$$

then $t \le t(\tilde{y}^0, \tilde{F})$, $y(\tilde{y}^0, \tilde{F}) \in Y_t$, *and*

$$\|y(y^0, F) - y(\tilde{y}^0, \tilde{F})\|_{Y_t} \le c\left(\|y^0 - \tilde{y}^0\|_{Y^0} + \sup_{\|y\|_{Y_T} \le R} \|(F - \tilde{F})(y)\|_{X_T}\right).$$

If $y = y(y^0, F)$ *is global, then*

$$y \in L_r(J, W_{q,\mathcal{B}}^{\tilde{\sigma}}) \cap W_r^1(J, W_{q,\mathcal{B}}^{\tilde{\sigma}-2}) \tag{A.4}$$

for any $\tilde{\sigma} < \sigma$, *and the norm of* y *in this space can be estimated by a constant* $C = C(\|F(y)\|_{X_T}, \|y^0\|_{Y^0})$.

REFERENCES

[1] H. AMANN, *Nonhomogeneous linear and quasilinear elliptic and parabolic boundary value problems*, in Function Spaces, Differential Operators and Nonlinear Analysis, H.J. Schmeisser and H. Triebel, eds., Teubner, Stuttgart, Leipzig, 1993, pp. 9–126.

[2] H. AMANN, *Linear and Quasilinear Parabolic Problems, Volume* I: *Abstract linear theory*, Monogr. Math. 89, Birkhäuser Boston, Boston, 1995.

[3] H. AMANN, *Linear and Quasilinear Parabolic Problems,* Vol. II, in preparation.

[4] H. AMANN, *Compact embeddings of vector-valued Sobolev and Besov spaces*, Glas. Mat. Ser. III, 35 (2000), pp. 161–177.

[5] H. Amann, *Linear parabolic problems involving measures*, RACSAM Rev. R. Acad. Cienc. Exactas F. Nat. Ser. A. Mat., 95 (2001), pp. 85–119.
[6] H. Amann and P. Quittner, *Semilinear parabolic equations involving measures and low regularity data*, Trans. Amer. Math. Soc., 356 (2004), pp. 1045–1119.
[7] J. M. Arrieta and A. N. Carvalho, *Abstract parabolic problems with critical nonlinearities and applications to Navier-Stokes and heat equations*, Trans. Amer. Math. Soc., 352 (1999), pp. 285–310.
[8] J. M. Arrieta, P. Quittner, and A. Rodriguez-Bernal, *Parabolic problems with nonlinear dynamical boundary conditions and singular initial data*, Differential Integral Equations, 14 (2001), pp. 1487–1510.
[9] P. Baras and L. Cohen, *Complete blow-up after T_{max} for the solution of a semilinear heat equation*, J. Funct. Anal., 71 (1987), pp. 142–174.
[10] H. Chen, *Positive steady-state solutions of a non-linear reaction-diffusion system*, Math. Methods Appl. Sci., 20 (1997), pp. 625–634.
[11] M. Chipot and P. Quittner, *Equilibria, connecting orbits and a priori bounds for semilinear parabolic equations with nonlinear boundary conditions*, J. Dynam. Differential Equations, 16 (2004), pp. 91–138.
[12] M. Fila, H. Matano, and P. Poláčik, *Immediate regularization after blow-up*, SIAM J. Math. Anal., to appear.
[13] M. Fila, Ph. Souplet, and F. B. Weissler, *Linear and nonlinear heat equations in L^q_δ spaces and universal bounds for global solutions*, Math. Ann., 320 (2001), pp. 87–113.
[14] A. Friedman and B. McLeod, *Blow-up of positive solutions of semilinear heat equations*, Indiana Univ. Math. J., 34 (1985), pp. 425–447.
[15] A. V. Fursikov, *Lagrange principle for problems of optimal control of ill-posed or singular distributed systems*, J. Math. Pures Appl. (9), 71 (1992), pp. 139–195.
[16] V. Galaktionov and J. L. Vázquez, *Continuation of blow-up solutions of nonlinear heat equations in several space dimensions*, Comm. Pure Appl. Math., 50 (1997), pp. 1–67.
[17] Y. G. Gu and M. X. Wang, *Existence of positive stationary solutions and threshold results for a reaction-diffusion system*, J. Differential Equations, 130 (1996), pp. 277–291.
[18] O. Yu. Imanuvilov, *Existence theorems for a solution of problems of controlling singular distributed systems*, Mat. Sb., 181 (1990), pp. 321–333 (in Russian); Math. USSR-Sb., 69 (1991), pp. 341–355 (in English).
[19] W. E. Kastenberg and P. L. Chambré, *On the stability of nonlinear space-dependent reactor kinetics*, Nucl. Sci. Eng., 31 (1968), pp. 67–79.
[20] A. A. Lacey and D. Tzanetis, *Complete blow-up for a semilinear diffusion equation with a sufficiently large initial condition*, IMA J. Appl. Math., 41 (1988), pp. 207–215.
[21] J.-L. Lions, *Contrôle des systèmes distribués singuliers*, Gauthier-Villars, Paris, 1983.
[22] W.-M. Ni, P. E. Sacks, and J. Tavantzis, *On the asymptotic behavior of solutions of certain quasilinear parabolic equations*, J. Differential Equations, 54 (1984), pp. 97–120.
[23] C. V. Pao, *Bifurcation analysis on a nonlinear diffusion system in reactor dynamics*, Appl. Anal., 9 (1979), pp. 107–119.
[24] P. Quittner, *Transition from decay to blow-up in a parabolic system*, Arch. Math. (Brno), 34 (1998), pp. 199–206.
[25] P. Quittner, *A priori bounds for global solutions of a semilinear parabolic problem*, Acta Math. Univ. Comenian., 68 (1999), pp. 195–203.
[26] P. Quittner, *Continuity of the blow-up time and a priori bounds for solutions in superlinear parabolic problems*, Houston J. Math., 29 (2003), pp. 757–799.
[27] P. Quittner, *Multiple equilibria, periodic solutions and a priori bounds for solutions in superlinear parabolic problems*, NoDEA Nonlinear Differential Equations Appl., 11 (2004), pp. 237–258.
[28] P. Quittner and Ph. Souplet, *Bounds of global solutions of parabolic problems with nonlinear boundary conditions*, Indiana Univ. Math. J., 52 (2003), pp. 875–900.
[29] W. M. Stacey, *Nuclear Reactor Physics*, John Wiley & Sons, New York, 2001.
[30] F. B. Weissler, *Semilinear evolution equations in Banach spaces*, J. Funct. Anal., 32 (1979), pp. 277–296.
[31] Z. Yan, *The global existence and blowing-up property of solutions for a nuclear model*, J. Math. Anal. Appl., 167 (1992), pp. 74–83.
[32] Z. Yan, *Mathematical analysis of a model for nuclear reactor dynamics*, J. Math. Anal. Appl., 186 (1994), pp. 623–633.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 4, pp. 1239–1258

OPTIMAL SOLUTION OF INVESTMENT PROBLEMS VIA LINEAR PARABOLIC EQUATIONS GENERATED BY KALMAN FILTER*

NIKOLAI DOKUCHAEV†

Abstract. We consider optimal investment problems for a diffusion market model with nonobservable random drifts that evolve as an Itô's process. Admissible strategies do not use direct observations of the market parameters, but rather use historical stock prices. For a nonlinear problem with a general performance criterion, the optimal portfolio strategy is expressed via the solution of a scalar minimization problem and a linear parabolic equation with coefficients generated by the Kalman filter.

Key words. optimal portfolio, nonobservable parameters, Kalman filter

AMS subject classifications. 49K45, 60G15, 93E20

DOI. 10.1137/S036301290342557X

1. Introduction. The paper investigates an optimal investment problem for a market which consists of a locally risk free asset, bond or bank account with interest rate $r(t)$, and a finite number n of risky stocks. We assume that the vector of stock prices $S(t)$ evolves according to an Itô stochastic differential equation $dS_i(t) = S_i(t)[a_i(t)\ dt + \sum_j \sigma_{ij}(t)\ dw_j(t)]$, $i = 1, \dots, n$, with a vector of appreciation rates $a(t)$ and a volatility matrix $\sigma(t)$. The problem goes back to Merton [26], who found strategies which solve the optimization problem in which $\mathbf{E}U(X(T))$ is to be maximized, where $X(T)$ represents the wealth at the final time T, and where $U(\cdot)$ is a utility function. If the market parameters are observed, then the optimal strategies (i.e., current vector of stock holdings) are functions of the current vector $(r(t), a(t), \sigma(t), S(t), X(t))$ (see, e.g., the survey in Hakansson [15] and Karatzas and Shreve [18]). But in practice, $a(t)$ and $\sigma(t)$ have to be estimated from historical stock prices or some other observation process. There are many papers devoted to estimation of $(a(\cdot), \sigma(\cdot))$, mainly based on modifications of Kalman–Bucy filtering or the maximum likelihood principle (see, e.g., Lo [25], Chen and Scott [3], Pearson and Sun [27]). Unfortunately, the process $a(\cdot)$ is usually hard to estimate in real-time markets, because the drift term, $a(\cdot)$, is usually overshadowed by the diffusion term, $\sigma(\cdot)$. On the other hand, $\sigma(t)$ can, in principle, be found from stock prices. Thus, there remains the problem of optimal investment with unobservable $a(\cdot)$.

In fact, the problem is one of linear filtering. If $R_i(t)$ is the return on the ith stock, then $dR(t) = a(t)dt + \sigma(t)dw(t)$, so the estimation of $a(t)$ given $\{R(\tau), \tau < t\}$ (or $\{S(\tau), \tau < t\}$) is a linear filtering problem. If $a(\cdot)$ is conditionally Gaussian, then the Kalman filter provides the estimate which minimizes the error in the mean square sense.

A popular tool in optimal control and filtering theory is the separation theorem. This theorem has an analog in portfolio theory: it is the so-called "certainty equivalence principle": agents who know the solution of the optimal investment problem

*Received by the editors March 22, 2004; accepted for publication (in revised form) January 21, 2005; published electronically October 7, 2005. Supported by NSERC of Canada under NCE Grant 30354 and Research Grant 88051.

http://www.siam.org/journals/sicon/44-4/42557.html

†Department of Mathematics and Statistics, University of Limerick, Ireland (Nikolai.Dokuchaev@ul.ie).

for the case of directly observable $a(t)$ can solve the problem with unobservable $a(t)$ by substituting $\mathbf{E}\{a(t)|S(\tau), \tau < t\}$ (see, e.g., Gennotte [14]). Unfortunately, this principle does not hold in the general case of nonlog utilities (see Kuwana [21]). Note that this principle is unrelated to the much more recent notion of "certainty equivalent value" to be found in the work of Frittelli [13].

Williams [30], Detemple [4], Dothan and Feldman [11], Gennotte [14], and Brennan [2] solved the investment problem using the Kalman–Bucy filter and dynamic programming. By this method, the optimal strategy can be calculated via solution of the Bellman parabolic equation; this equation is nonlinear.

Karatzas [16], Karatzas and Zhao [20], Dokuchaev and Zhou [10], and Dokuchaev and Teo [9] have obtained optimal portfolio strategies in general non-Gaussian setting, but only for case of time independent coefficients.

An approach based on Malliavin calculus gives a possibility of considering a more general setting. Lakner [22], [23] assumes that $S(\cdot)$ and $w(\cdot)$ have equal dimension (as we do), and that $r(\cdot)$ and $\sigma(\cdot)$ are deterministic. This again guarantees that the filtration of $S(\cdot)$ is Brownian. Results from filtering theory give a representation of the optimal portfolio, which is explicit in terms of a conditional expectation of a Malliavin derivative when the $a_i(\cdot)$ are Ornstein–Uhlenbeck processes independent of $w(\cdot)$. Karatzas and Xue [19] assume that there are more Brownian motions than stocks. They assume that $r(\cdot)$ and $\sigma(\cdot)$ are adapted to the observable $S(\cdot)$. After projecting onto an n-dimensional Brownian motion which generates the same filtration as $S(\cdot)$, they obtain a reduced, completely observable model; existence of an optimal portfolio follows, but the optimal strategy is, as usual, defined only implicitly.

We also consider the optimal investment problem with random and unobservable $a(\cdot)$. Following Lakner [23] and Rishel [28], we assume that $a(t)$ is a Gaussian process modelled by a system of linear Itô's equations. However, we consider a more general case when $(a(\cdot), r(\cdot))$ may depend on the realized returns (i.e., $b(\cdot) \neq 0$ in (2.4) below, and $r(\cdot)$ is correlated with $S(\cdot)$). We express the optimal strategy via solution of a Cauchy problem (4.3), (4.8) for a *linear* parabolic equation in $(n+1)$-dimensional vector space. Thus, we propose a simpler method than dynamic programming: the *nonlinear* parabolic Bellman equation is replaced for a linear parabolic equation. Note that the solution in Lakner [23] expresses the optimal strategy via a conditional expectation of a random claim that depends on $w(\cdot)$; the solution presented below is also based on the martingale method but is more constructive, provided we can solve the Cauchy problem (4.3), (4.8). Using the technique of backward stochastic partial differential equations, we prove existence and uniqueness of the solution for this Cauchy problem. Furthermore, the most restrictive condition in Lakner [23] was that the initial covariance of $a(0)$ is small enough (condition (3.5)). We replace it by another condition (4.9) that depends on U: it is less restrictive than (3.5) for some U's and more restrictive for others U's. For some problems, our condition (4.9) is automatically satisfied. In addition, we allow correlated $a(\cdot)$ and $w(\cdot)$.

2. The model and definitions. Consider a diffusion model of a market consisting of a locally risk free bank account or bond with price $B(t)$, $t \geq 0$, and n risky stocks with prices $S_i(t)$, $t \geq 0$, $i = 1, 2, \ldots, n$, where $n < +\infty$ is given. The prices of the stocks evolve according to the following equations:

$$dS_i(t) = S_i(t)\left(a_i(t)dt + \sum_{j=1}^{n} \sigma_{ij}(t)dw_j(t)\right), \quad t > 0, \tag{2.1}$$

where $w_i(t)$ are standard independent Wiener processes, $a_i(t)$ are appreciation rates, and $\sigma_{ij}(t)$ are volatility coefficients. The initial price $S_i(0) > 0$ is a given nonrandom constant. The price of the bond evolves according to the following equation:

$$B(t) = B(0) \exp\left(\int_0^t r(t) dt \right), \tag{2.2}$$

where $B(0)$ is a given constant which we take to be 1 without loss of generality, and $r(t)$ is the random process of the risk free interest rate.

We are given a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, where Ω is a set of elementary events, $\mathcal{F}$ is a complete σ-algebra of events, and $\mathbf{P}$ is a probability measure.

We introduce the vector processes ($^\top$ denoted transpose)

$$w(t) = (w_1(t), \dots, w_n(t))^\top, \quad S(t) = (S_1(t), \dots, S_n(t))^\top, \quad a(t) = (a_1(t), \dots, a_n(t))^\top,$$

and the matrix process $\sigma(t) = \{\sigma_{ij}(t)\}_{i,j=1}^n$.

Let $\mathbf{1} \triangleq (1, \dots, 1)^\top \in \mathbf{R}^n$, and $\widetilde{a}(t) \triangleq a(t) - r(t)\mathbf{1}$.

We define the return to time t by $dR_i(t) = dS_i(t)/S_i(t)$, $R_i(0) = 0$, and introduce the vector of returns $R(t) = (R_1(t), \dots, R_n(t))^\top$ and of excess returns $\widetilde{R}_i(t) = R_i(t) - \int_0^t r(s)\, ds$.

Let $\{\mathcal{F}_t^{S,r}\}_{0 \le t \le T}$ be the filtration generated by the process $(r(t), S(t))$ completed with the null sets of $\mathcal{F}$.

Set $\widetilde{S}(t) \triangleq \exp\left(- \int_0^t r(s) ds \right) S(t)$.

We denote by $|x|$ the Euclidean norm of a vector $x \in \mathbf{R}^k$. For an Euclidean space E, we denote by $\mathrm{B}([0,T]; E)$ the set of bounded measurable functions $f(t) : [0,T] \to E$. We denote by I_n the identity matrix in $\mathbf{R}^{n \times n}$. As usual, we say that $A < B$ for symmetric matrices if the matrix $B - A$ is definitely positive. We denote $\phi^- \triangleq \max(0, -\phi)$, and we denote by $\mathbb{I}_{\{\cdot\}}$ the indicator function.

The model for r, σ, and a. To describe the distribution of $\widetilde{a}(t)$, we shall use the model introduced in Lakner [23, p. 84], generalized for our case of random r, nonconstant coefficients for the equation for $\widetilde{a}$, and correlated r, $\widetilde{a}$, and w. We assume that we are given measurable deterministic processes $\alpha(t)$, $\beta(t)$, $b(t)$, and $\delta(t)$ such that

$$d\widetilde{a}(t) = \alpha(t)[\delta(t) - \widetilde{a}(t)]dt + b(t) d\widetilde{R}(t) + \beta(t) dW(t), \tag{2.3}$$

where $\alpha(t) \in \mathbf{R}^{n \times n}$, $\beta(t) \in \mathbf{R}^{n \times n}$, $b(t) \in \mathbf{R}^{n \times n}$, $\delta(t) \in \mathbf{R}^n$, and where W is an n-dimensional Wiener process in $(\Omega, \mathcal{F}, P)$. We assume that $\alpha(t)$, $\beta(t)$, $b(t)$, and $\delta(t)$ are continuous in t and such that the matrix $\beta(t)$ is invertible and $|\beta(t)^{-1}| \le c$, where $c > 0$ is a constant. Further, we assume that $\widetilde{a}(0)$ follows an n-dimensional normal distribution with mean vector m_0 and covariance matrix γ_0. The vector m_0 and the matrix γ_0 are assumed to be known. We note that this setting covers the case when $\widetilde{a}$ is an n-dimensional Ornstein–Uhlenbeck process with mean-reverting drift.

Clearly, (2.3) can be rewritten as

$$d\widetilde{a}(t) = \Big(\alpha(t)\delta(t) + [b(t) - \alpha(t)]\widetilde{a}(t) \Big) dt + b(t)\sigma(t) dw(t) + \beta(t) dW(t). \tag{2.4}$$

In addition, it can be seen that $\widetilde{R}_i(t)$ evolves as

$$d\widetilde{R}_i(t) = \widetilde{a}_i(t) dt + \sum_{j=1}^n \sigma_{ij}(t) dw_j(t), \quad t > 0. \tag{2.5}$$

We assume that the process $\sigma(t)$ is continuous in t, nonrandom, and such that $\sigma(t)\sigma(t)^\top \geq c_\sigma I_n$, where $c_\sigma > 0$ is a constant.

Further, we assume that $r(\cdot) = \phi_r(\widetilde{R}(\cdot), \Theta)$, where Θ is a random element in a metric space $\mathcal{X}_r$, and where $\phi_r : C([0,T];\mathbf{R}^n) \times \mathcal{X}_r \to \mathrm{B}([0,T];\mathbf{R})$ is a measurable function, and Θ does not depend on $(w(\cdot), W(\cdot), \widetilde{a}(0))$. In addition, we assume that the process $r(t)$ is adapted to the filtration generated by $(\widetilde{R}(t), \Theta)$. Note that the closed system (2.4)–(2.5) for the pair $(\widetilde{a}(t), \widetilde{R}(t))$ does not include $r(\cdot)$, and $(\widetilde{a}(\cdot), \widetilde{R}(\cdot))$ does not depend on Θ. Therefore, the market model is well defined. The assumptions for measurability of r don't look very natural. However, they cover generic models when r is independent on $\widetilde{R}$ or nonrandom, and we can still consider some models with correlated r and $\widetilde{R}$.

Under these assumptions, the solution of (2.1) is well defined, but the market is incomplete.

Let $\widetilde{\phi}_m(t,s)$, $m = 0, 1$, be the solution of the matrix equation

$$\begin{cases} \frac{d\widetilde{\phi}_m}{dt}(t,s) = [m \cdot b(t) - \alpha(t)]\widetilde{\phi}_m(t,s), \\ \widetilde{\phi}_m(s,s) = I_n. \end{cases}$$

Let

$$\widetilde{K}_m(t) \triangleq \int_0^t \widetilde{\phi}_m(t,s) b(s)\sigma(s)\sigma(s)^\top b(s)^\top \widetilde{\phi}_m(t,s)^\top ds, \quad m = 0, 1. \tag{2.6}$$

We have that

$$\widetilde{a}(t) = \widetilde{\phi}_1(t,0)\widetilde{a}(0) + \int_0^t \widetilde{\phi}_1(t,s)[\alpha(s)\delta(s)ds + b(s)\sigma(s)dw(s) + \beta(s)dW(s)].$$

It follows that $\widetilde{K}_1(t)$ is the covariance matrix for $\widetilde{a}(t)$ calculated with $\beta(t) \equiv 0$ and $\widetilde{a}(0) = 0$. By the linearity of (2.4), it follows that $\widetilde{K}_1(t)$ is the conditional covariance for $\widetilde{a}(t)$ given $(W(\cdot)|_{[0,t]}, \widetilde{a}(0))$ or $(W(\cdot)|_{[0,T]}, \widetilde{a}(0))$.

Note that $\widetilde{K}_m(t)$ can be found as solutions of linear equations that one can easily derive from (2.4) and (4.1) (see, e.g., Arnold [1, Chapter 8]).

We assume that b is "small." More precisely, we assume that there exists $\varepsilon > 0$ such that

$$T\widetilde{K}_m(t) + \varepsilon I_n < \sigma(t)\sigma(t)^\top \quad \forall t \in [0,T], \quad m = 0, 1. \tag{2.7}$$

The risk neutral probability measure. Set $Q(t) \triangleq (\sigma(t)\sigma(t)^\top)^{-1}$, and set

$$\mathcal{Z} \triangleq \exp\left(\int_0^T [\sigma(t)^{-1}\widetilde{a}(t)]^\top dw(t) + \frac{1}{2}\int_0^T \widetilde{a}(t)^\top Q(t)\widetilde{a}(t)dt \right). \tag{2.8}$$

Proposition 2.1.

$$\mathbf{E}\Big\{ \exp \frac{1}{2}\int_0^T \widetilde{a}(t)^\top Q(t)\widetilde{a}(t)dt \,\Big|\, W(\cdot), \widetilde{a}(0) \Big\} < +\infty \qquad \textit{a.s.} \tag{2.9}$$

By this proposition, the Novikov's condition is satisfied conditionally, and $\mathbf{E}\{\mathcal{Z}^{-1} \,|\, W(\cdot), \widetilde{a}(0)\} = 1$, then $\mathbf{E}\mathcal{Z}^{-1} = 1$.

Define the (equivalent) probability measure $\mathbf{P}_*$ by $d\mathbf{P}_*/d\mathbf{P} = \mathcal{Z}^{-1}$. Let $\mathbf{E}_*$ be the corresponding expectation.

The wealth and strategies. Let $X_0 > 0$ be the initial wealth at time $t = 0$, and let $X(t)$ be the wealth at time $t > 0$, $X(0) = X_0$. We assume that

$$X(t) = \pi_0(t) + \sum_{i=1}^{n} \pi_i(t), \tag{2.10}$$

where the pair $(\pi_0(t), \pi(t))$ describes the portfolio at time t. The process $\pi_0(t)$ is the investment in the bond, $\pi_i(t)$ is the investment in the ith stock $\pi(t) = (\pi_1(t), \dots, \pi_n(t))^\top$, $t \geq 0$.

DEFINITION 2.2. *The process* $\widetilde{X}(t) \triangleq \exp\left(-\int_0^t r(s)ds\right) X(t)$ *is called the normalized (or discounted) wealth.*

Let $\mathbf{S}(t) \triangleq \operatorname{diag}(S_1(t), \dots, S_n(t))$ and $\widetilde{\mathbf{S}}(t) \triangleq \operatorname{diag}(\widetilde{S}_1(t), \dots, \widetilde{S}_n(t))$ be diagonal matrices with the corresponding diagonal elements. The portfolio is said to be self-financing if

$$dX(t) = \pi(t)^\top \mathbf{S}(t)^{-1} dS(t) + \pi_0(t) r(t) dt = \pi(t)^\top dR(t) + \pi_0(t) r(t) dt. \tag{2.11}$$

It follows from (2.10) that for such portfolios

$$\begin{aligned} dX(t) &= r(t)X(t)\,dt + \pi(t)^\top \left(\widetilde{a}(t)\,dt + \sigma(t)\,dw(t)\right), \\ d\widetilde{X}(t) &= B(t)^{-1}\pi(t)^\top d\widetilde{R}(t), \end{aligned} \tag{2.12}$$

so π alone suffices to specify the portfolio; the process π_0 is uniquely defined by π via (2.10), (2.12); π it is called a self-financing strategy.

DEFINITION 2.3. *Let* $\bar{\Sigma}$ *be the class of all* $\mathcal{F}_t^{S,r}$*-predictable processes* $\pi(\cdot)$ *such that*

- $\int_0^T \left(|\pi(t)^\top \widetilde{a}(t)|^2 + |\pi(t)^\top \sigma(t)|^2\right) dt < \infty$ *a.s.*
- *there exists a constant* q_π *such that* $\mathbf{P}\left(\widetilde{X}(t) - X_0 \geq q_\pi \ \forall t \in [0,T]\right) = 1$.

A process $\pi(\cdot) \in \bar{\Sigma}$ is said to be an *admissible* strategy with corresponding wealth $X(\cdot)$.

For an admissible strategy $\pi(\cdot)$, $X(t, \pi(\cdot))$ denotes the corresponding total wealth, and $\widetilde{X}(t, \pi(\cdot))$ the corresponding normalized total wealth. It follows that $\widetilde{X}(t, \pi(\cdot))$ is a $\mathbf{P}_*$-supermartingale with $\mathbf{E}_* \widetilde{X}(t, \pi(\cdot)) \leq X_0$ and $\mathbf{E}_* |\widetilde{X}(t, \pi(\cdot))| \leq |X_0| + 2|q_\pi|$.

Note that by definition, admissible strategies from $\bar{\Sigma}$ use observations of $r(t)$ and $S(t)$ only. For these strategies, the processes $X(t)$ and $\widetilde{X}(t)$ are $\mathcal{F}_t^{S,r}$-adapted.

The following definition is standard.

DEFINITION 2.4. *Let* ξ *be a given random variable. An admissible strategy* $\pi(\cdot)$ *is said to replicate the claim* ξ *if* $X(T, \pi(\cdot)) = \xi$ *a.s.*

3. Problem statement and preliminary results. Let $T > 0$, let $\widehat{D} \subset \mathbf{R}$ be convex and bounded below, and let $X_0 \in \widehat{D}$ be given. Let $U(\cdot) : \widehat{D} \to \mathbf{R} \cup \{-\infty\}$ be such that $U(X_0) > -\infty$.

We may state our general problem as follows: Find an admissible self-financing strategy $\pi(\cdot)$ which solves the following optimization problem:

$$\text{Maximize} \quad \mathbf{E}U(\widetilde{X}(T, \pi(\cdot))) \quad \text{over} \quad \pi(\cdot) \in \bar{\Sigma} \tag{3.1}$$

$$\text{subject to} \begin{cases} \widetilde{X}(0, \pi(\cdot)) = X_0, \\ \widetilde{X}(T, \pi(\cdot)) \in \widehat{D} \quad \text{a.s.} \end{cases} \tag{3.2}$$

The condition $\widetilde{X}(T,\pi(\cdot))\in\widehat{D}$ may represent a requirement for a minimal normalized terminal wealth if $\widehat{D}=[k,+\infty)$, $k>0$. This condition may also represent a requirement for the normalized terminal wealth in goal achieving problems if $\widehat{D}=[k_0,k_1]$, $k_0<k_1$.

We assume that U, X_0, and $\widehat{D}$ satisfy the following two conditions.

CONDITION 3.1. *There exists a measurable set $\Lambda\subseteq[0,\infty)$, and a measurable function $F(\cdot,\cdot):(0,\infty)\times\Lambda\to\widehat{D}$ such that for each $z>0$, $\widehat{x}=F(z,\lambda)$ is a solution of the optimization problem*

$$\text{Maximize}\quad zU(x)-\lambda x\quad\text{over } x\in\widehat{D}. \tag{3.3}$$

Note that the usual concavity hypotheses imply this condition, but more general utility functions are also covered. For example, this condition is satisfied for the goal achieving problem when $U(x)$ is a step function (see, e.g., Karatzas [16], Dokuchaev and Zhou [10]).

Let $\bar{\mathcal{Z}}\triangleq\mathbf{E}\{\mathcal{Z}|\mathcal{F}_T^{S,r}\}$. Since $(\widetilde{R}(\cdot),\widetilde{a}(\cdot))$ does not depend on Θ, we have that $\mathcal{Z}$ does not depend on Θ, and $\bar{\mathcal{Z}}=\mathbf{E}\{\mathcal{Z}|\widetilde{R}(\cdot)\}$. Let $F(\cdot)$ be as in Condition 3.1.

CONDITION 3.2. *There exists $\widehat{\lambda}\in\Lambda$ such that $\mathbf{E}_*|F(\bar{\mathcal{Z}},\widehat{\lambda})|<+\infty$ and $\mathbf{E}_*F(\bar{\mathcal{Z}},\widehat{\lambda})=X_0$.*

We solve our problem in two steps using the martingale approach. First we show that $\mathbf{E}U(F(\bar{\mathcal{Z}},\widehat{\lambda}))$ is an upper bound for the expected utility of normalized terminal wealth for $\pi(\cdot)\in\bar{\Sigma}$. Then we find a portfolio $\widehat{\pi}(\cdot)$ which replicates the claim $B(T)F(\bar{\mathcal{Z}},\widehat{\lambda})$. This establishes the optimality of $\widehat{\pi}(\cdot)$.

The optimal claim. The following theorem is a reformulation of Theorem 2.5 from Lakner [23] under slightly more general conditions that allow discontinuous functions F and U such as step functions.

THEOREM 3.1. *(Dokuchaev and Haussmann [8]). With $\widehat{\lambda}$ as in Condition 3.2, let $\widehat{\xi}\triangleq F(\bar{\mathcal{Z}},\widehat{\lambda})$. Then*

(i) $\mathbf{E}\,U^-(\widehat{\xi})<\infty$, $\widehat{\xi}\in\widehat{D}$ *a.s.;*

(ii) $\mathbf{E}U(\widehat{\xi})\geq\mathbf{E}U(\widetilde{X}(T,\pi(\cdot)))\ \forall\pi(\cdot)\in\bar{\Sigma}$*;*

(iii) *The claim $B(T)\widehat{\xi}$ is attainable in $\bar{\Sigma}$, and there exists a replicating strategy in $\bar{\Sigma}$. This strategy is optimal for problem* (3.1)–(3.2).

This theorem uses the duality approach for constrained optimization that goes back to Lagrange, and $\widehat{\lambda}$ as the corresponding Lagrange multiplier.

Remark 3.1. Theorem 2.5 from Lakner [23] was stated under some additional assumptions that can be formulated in our notations as

(i) $b(t)\equiv 0$, r is nonrandom, r, σ, α, β, δ are constant, and $\widehat{D}=(0,+\infty)$;

(ii) U is strictly concave and continuously differentiable on $(0,+\infty)$, and $\lim_{x\to+\infty}U'(x)=0$;

(iii) there exists a function $J(\cdot):\widehat{D}\to\mathbf{R}$ such that $J(\lambda/x)\equiv F(x,\lambda)$;

(iv) $\mathbf{E}_*J(\lambda/\bar{\mathcal{Z}})<+\infty$ for any $\lambda>0$.

Solution via conditional expectation. Let

$$\widehat{a}(t)\triangleq\mathbf{E}\big\{\widetilde{a}(t)\,\big|\,\mathcal{F}_t^{S,r}\big\}.$$

Set $\widetilde{\alpha}(t)\triangleq\alpha(t)-b(t)$ and $m_0\triangleq\mathbf{E}\widetilde{a}(0)$.

Let $\gamma(t) \in \mathbf{R}^{n\times n}$ be the unique solution (in the class of symmetric nonnegative definite matrices) of the deterministic Riccati's equation

$$\begin{cases} \frac{d\gamma}{dt}(t) = -[b(t)\sigma(t)^\top + \gamma(t)]Q(t)[b(t)\sigma(t)^\top + \gamma(t)]^\top \\ \qquad\qquad -\widetilde{\alpha}(t)\gamma(t) - \gamma(t)\widetilde{\alpha}(t)^\top + \beta(t)\beta(t)^\top, \\ \gamma(0) = \gamma_0. \end{cases} \tag{3.4}$$

Here $\gamma_0 \triangleq \mathbf{E}[\widetilde{a}(0)-m_0][\widetilde{a}(0)-m_0]^\top$. In fact, $\gamma(t) = \mathbf{E}\left\{[\widetilde{a}(t) - \widehat{a}(t)][\widetilde{a}(t) - \widehat{a}(t)]^\top \big| \mathcal{F}_t^{S,r}\right\}$.

Let $A(t) \triangleq -\widetilde{\alpha}(t) - \gamma(t)Q(t)$, and let $\phi(t)$ be the solution of the matrix equation

$$\begin{cases} \frac{d\phi}{dt}(t) = A(t)\phi(t), \\ \phi(0) = I_n, \end{cases}$$

where I_n is the unit matrix in $\mathbf{R}^{n\times n}$.

The following theorem is a reformulation of Theorem 4.3 from Lakner [23]. It gives the solution of the investment problem via conditional expectation of future values of some processes with known evolution.

THEOREM 3.2. *(Lakner [23]). Let conditions* (i)–(iv) *in Remark* 3.1 *hold, let* $U(x)$ *be twice differentiable on* $(0, +\infty)$*, and let*

$$\operatorname{tr}\gamma_0 + T\|\beta\|^2 < K_1, \quad K_1 = \frac{1}{360T\|\sigma^{-1}\|^2 K_0}, \quad K_0 = \max_{t\in[0,T]} \|e^{-\alpha t}\|^2, \tag{3.5}$$

where $\|\cdot\|$ *denotes the Frobenius matrix norm, i.e.,* $\|\sigma^{-1}\|^2 = \operatorname{tr}[\sigma^{-1}\sigma^{-1^\top}]$*. Further, let*

$$J(x) < K(1 + x^{-5}), \quad -J'(x) < K(1 - x^{-2}) \tag{3.6}$$

for some $K > 0$*. Then the optimal strategy is*

$$\pi(t)^\top = H(t)\bar{\mathcal{Z}}(t)\mathbf{E}\left\{J'(\widehat{\lambda}\bar{\mathcal{Z}})\bar{\mathcal{Z}}^{-2}\left[-\gamma(t)[\phi(t)^\top]^{-1}\int_t^T \phi(s)^\top[\sigma^\top]^{-1}d\widehat{w}(s) - \widehat{a}(t)\right]\Big| \mathcal{F}_t^{S,r}\right\},$$

where $H(t) \triangleq \widehat{\lambda}e^{r(t-T)}Q$ *and* $\widehat{w}(t) \triangleq w(t) - \int_0^t \sigma^{-1}\widehat{a}(s)ds$*.*

We propose below another solution such that the optimal strategy is presented via solution of a linear deterministic parabolic equation. We replace conditions (3.5) by condition (4.9) which can be less restrictive and is always satisfied if $\widehat{D}$ is bounded. In addition, we dropped condition (3.6) and the condition that (r, a) and w are independent: we allow $b(\cdot) \neq 0$ and $r = \phi_r(\widetilde{R}(\cdot), \Theta)$.

4. Main results: Solution via linear parabolic equation.

Let $y(t) = (y_1(t), \dots, y_{n+1}(t)) = (\widehat{a}(t), y_{n+1}(t))$ be a process in $\mathbf{R}^{n+1}$, where

$$\widehat{a}(t) = \mathbf{E}\{\widetilde{a}(t)|\mathcal{F}_t^{S,r}\},$$
$$y_{n+1}(t) = -\frac{1}{2}\int_0^t \widehat{a}(s)^\top Q(s)\widehat{a}(s)ds + \int_0^t \widehat{a}(s)^\top Q(s)\, d\widetilde{R}(s).$$

Let functions $f(\cdot) : \mathbf{R}^{n+1} \times [0,T] \to \mathbf{R}^{n+1}$ and $g(\cdot) : \mathbf{R}^{n+1} \times [0,T] \to \mathbf{R}^{(n+1)\times n}$ be such that

$$f(x,t) \triangleq \begin{pmatrix} [A(t) - b(t)\sigma(t)^\top Q(t)]\widehat{x} + \alpha(t)\delta(t) \\ -\frac{1}{2}\widehat{x}^\top Q(t)\widehat{x} \end{pmatrix},$$

$$g(x,t) \triangleq \begin{pmatrix} [b(t)\sigma(t)^\top + \gamma(t)]Q(t) \\ \widehat{x}^\top Q(t) \end{pmatrix}.$$

Here $A(t)$ and $\gamma(t)$ are matrices defined above, $\gamma(t)$ is the solution of (3.4), and

$$x=(x_1,\dots,x_{n+1})^\top=\begin{pmatrix}\widehat{x}\\ x_{n+1}\end{pmatrix},\qquad \widehat{x}=(x_1,\dots,x_n)^\top.$$

By Theorem 10.3 from Liptser and Shiryaev [24, p. 396], the equation for $\widehat{a}(t)$ is

$$\begin{cases} d\widehat{a}(t)=[A(t)\widehat{a}(t)-b(t)\sigma(t)^\top Q(t)\widehat{a}(t)+\alpha(t)\delta(t)]dt+[b(t)\sigma(t)^\top+\gamma(t)]Q(t)d\widetilde{R}(t),\\ \widehat{a}(0)=m_0.\end{cases}$$

(4.1)

By (4.1)–(4.6), it follows that $y(\cdot)$ is the solution of the Itô's equation

$$\begin{cases} dy(t)=f(y(t),t)dt+g(y(t),t)\,d\widetilde{R}(t),\\ y(0)=y_0,\end{cases}\tag{4.2}$$

with

$$y_0=\begin{pmatrix}m_0\\0\end{pmatrix}\in\mathbf{R}^{n+1},\qquad m_0=\mathbf{E}\widetilde{a}(0).$$

The function $f(y,t)$ here does not satisfy Lipschitz condition with respect to $y\in\mathbf{R}^{n+1}$. However, the solution of this equation is uniquely defined. (It is shown in the proof of Lemma 4.1 below that the solution of (4.2) can be presented as a part of the unique solution of some Itô's equation with coefficients that are affine with respect to the state variable.)

LEMMA 4.1. *Let a function $\Phi(\cdot):\mathbf{R}^{n+1}\to\mathbf{R}$ be such that*

(i) $\mathbf{E}_*\Phi(y(T))=X_0$;

(ii) $\Phi(x)$ *is continuously twice differentiable;*

(iii) $\mathbf{E}_*\Phi(y(T))^2<+\infty$.

Then there exists a unique classical solution $V:\mathbf{R}^{n+1}\times[0,T]\to\mathbf{R}$ of the boundary value problem

$$\frac{\partial V}{\partial t}(x,t)+\frac{\partial V}{\partial x}(x,t)f(x,t)+\frac{1}{2}\mathrm{tr}\left\{\frac{\partial^2 V}{\partial x^2}(x,t)\,g(x,t)\sigma(t)\sigma(t)^\top g(x,t)^\top\right\}=0,\tag{4.3}$$

$$V(x,T)=\Phi(x).\tag{4.4}$$

Further, the processes $\widetilde{X}(t,\pi(\cdot))\triangleq V(y(t),t)$ and $\pi(t)^\top\triangleq B(t)\frac{\partial V}{\partial x}(y(t),t)g(y(t),t)$, are uniquely defined as elements of the spaces $C([0,T],L_2(\Omega,\mathcal{F},P_*))$ and $L_2([0,T],L_2(\Omega,\mathcal{F},P_*))$, respectively, and there exists a constant $C>0$ such that

$$\sup_{t\in[0,T]}\mathbf{E}_*|\widetilde{X}(t,\pi(\cdot))|^2+\mathbf{E}_*\int_0^T B(t)^{-2}|\pi(t)|^2dt\le C\mathbf{E}_*|\Phi(y(T))|^2\tag{4.5}$$

for all these Φ. Further, the strategy $\pi(t)=(\pi_1(t),\dots,\pi_n(t))$ belongs to $\bar{\Sigma}$ and replicates the claim $B(T)\Phi(y(T))$ given the initial wealth X_0 with the normalized wealth $\widetilde{X}(t)=V(y(t),t)$.

Note that estimate (4.5) restates the Krylov–Ficera estimate (see Theorem 5.3.3 from Rozovskii [29]) or its modification from Dokuchaev [5]).

Further, we have that

$$d\bar{\mathcal{Z}}(t)=\widehat{a}(t)\bar{\mathcal{Z}}(t)d\widetilde{R}(t).\tag{4.6}$$

Formula (4.6) was derived in Theorem 3.1 from Lakner [23] for the case when σ is constant and $b = 0$. The proof for a nonconstant $\sigma(t)$ and $b \neq 0$ can be found in Dokuchaev and Haussmann [8] and in Chapter 9 from Dokuchaev [6]. It follows that

$$y_{n+1}(t) = \ln \bar{\mathcal{Z}}(t). \tag{4.7}$$

Introduce the function $e(\cdot) : \mathbf{R}^{n+1} \to \mathbf{R}$ such that $e(y) = \exp[y_{n+1}]$ for $y = (y_1, \dots, y_{n+1})^\top$. Note that $\bar{\mathcal{Z}} = e(y(T))$.

Let $V = V(x, t, \lambda) : \mathbf{R}^{n+1} \times [0, T] \times \Lambda \to \mathbf{R}$ be the solution of partial differential equation (4.3) with the condition

$$V(x, T, \lambda) = F(e(x), \lambda). \tag{4.8}$$

The following result is now immediate.

THEOREM 4.2. *Let $\widehat{\lambda}$ be such as in Condition* 3.2. *Assume that the function $F(\cdot, \widehat{\lambda}) : \mathbf{R} \to \mathbf{R}$ is such that conditions* (i)–(ii) *of Lemma* 4.1 *are satisfied with $\Phi(x) \triangleq F(e(x), \widehat{\lambda})$, and*

$$\mathbf{E}_* F(\bar{\mathcal{Z}}, \widehat{\lambda})^2 < +\infty. \tag{4.9}$$

Then there exists an unique classical solution V of problem (4.3)–(4.8) *for $\lambda = \widehat{\lambda}$, and there exists an admissible self-financing strategy $\pi(\cdot) \in \bar{\Sigma}$ which replicates the claim $B(T)F(\bar{\mathcal{Z}}, \widehat{\lambda})$. This strategy is an optimal solution of problem* (3.1)–(3.2) *and*

$$\widehat{\pi}(t)^\top = B(t) \frac{\partial V}{\partial x}(y(t), t, \widehat{\lambda}) b(y(t), t), \quad \widetilde{X}(t, \pi(\cdot)) = V(y(t), t, \widehat{\lambda}). \tag{4.10}$$

Note that it is possible that condition (4.9) is not satisfied but the optimal claim $F(\bar{\mathcal{Z}}, \widehat{\lambda})$ is still replicable in the class of strategies $\bar{\Sigma}$. For example, let $U(x) \equiv \log x$, $X_0 = 1$, and $(0, +\infty) \subseteq \widehat{D}$, then $\Lambda = (0, \infty)$, $F(z, \lambda) = z/\lambda$, $\widehat{\lambda} = 1$, and the strategy is $\pi(t)^\top = B(t)\widehat{a}(t)^\top \bar{\mathcal{Z}}(t) Q(t)$ is replicating (and optimal) even in the case when (4.9) is not satisfied.

5. Special cases. Note that conditions (3.5) were imposed in Lakner [23] with the only purpose to ensure that

$$\mathbf{E}_* \bar{\mathcal{Z}}^5 < +\infty, \qquad \mathbf{E}_* \bar{\mathcal{Z}}^{-4} < +\infty. \tag{5.1}$$

Our condition (4.9) for examples (i)–(iii) listed below is satisfied if $\mathbf{E}_* \bar{\mathcal{Z}}^\mu < +\infty$ for some $\mu \in \mathbf{R}$. For example (i), condition (4.9) is less restrictive than (5.1) if $l < 5/2$ and more restrictive if $l > 5/2$. For example (ii), condition (4.9) is less restrictive than (5.1) if $l < 2$ and more restrictive if $l > 2$. For example (iii), condition (4.9) is always less restrictive than (5.1). These examples are from Dokuchaev and Haussmann [7]:

(i) *Power utility.* Assume $\widehat{D} = [0, +\infty)$, $X_0 > 0$, $U(x) = d^{-1}x^d$, where either $d \in (0, 1)$ or $d < 0$. Then $\Lambda = (0, \infty)$, $F(z, \lambda) = (z/\lambda)^l$, and $\widehat{\lambda} = X_0^{-1/l}(\mathbf{E}_* \bar{\mathcal{Z}}^l)^{1/l}$, where $l = 1/(1 - d)$.

(ii) Assume $\widehat{D} = [0, +\infty)$, $U(x) = -x^d + x$, where $d = 1 + 1/l$, and $l > 0$ is an integer, $X_0 > d^{-l}$. Then $\Lambda = [0, \infty)$, $F(z, \lambda) = (1 + \lambda/z)^l d^{-l}$, $\widehat{\lambda}$ is a root of a polynomial of degree l.

(iii) *Mean-variance utility.* Assume $\widehat{D} = \mathbf{R}$, $U(x) = -kx^2 + cx$, where $k \in \mathbf{R}$ and $c \geq 0$, $X_0 > 0$, then $F(z, \lambda) = (c - \lambda/z)/(2k)$.

We present below some sufficient conditions that ensure $\mathbf{E}_*\bar{\mathcal{Z}}^\mu < +\infty$ and, therefore, can be useful for verifying (6.2).

Let $\widetilde{K}(t)$ be the covariance for $\widetilde{a}(t)$ under the probability measure $\mathbf{P}_*$, and let $\widehat{K}(t)$ be the covariance for $\widehat{a}(t)$ under $\mathbf{P}_*$.

LEMMA 5.1. *If $\mu \in [0,1]$, then $\mathbf{E}_*\bar{\mathcal{Z}}^\mu < +\infty$. Let $\mu < 0$ or $\mu > 1$. Then $\mathbf{E}_*\bar{\mathcal{Z}}^\mu < +\infty$ if there exist $\varepsilon > 0$ and $p > 1$ such that at least one of the following conditions holds:*

(i) *$\kappa(p)\widehat{K}(t) < \sigma(t)\sigma(t)^\top - \varepsilon I_n$ for $t \in [0,T]$, where $\kappa(p) \triangleq qT(\mu^2 p - \mu) > 0$ with $q \triangleq p(p-1)^{-1}$.*

(ii) *$\kappa(p)\widetilde{K}(t) < \sigma(t)\sigma(t)^\top - \varepsilon I_n$ for $t \in [0,T]$.*

It follows from Proposition 7.2 that $\widetilde{K}(t)$ and $\widehat{K}(t)$ are the covariances of the processes defined by (2.4) and (4.1), respectively, with $\widetilde{R}(\cdot)$ replaced by $\widetilde{R}_*(\cdot)$. Thus, these covariances can be found as solutions of linear deterministic equations that one can easily derive from (2.4) and (4.1) (see, e.g., Arnold [1, Chapter 8]).

6. Case of discontinuous F. To proceed further, we shall need a special weighted L_2-space with a weight defined via some parabolic equation. First, we introduce the operator

$$\mathcal{M}(t)p \triangleq -\sum_{i=1}^{n+1}\frac{\partial}{\partial x_i}\left(p(x)f_i(x,t)\right) + \frac{1}{2}\sum_{i,j=1}^{n+1}\frac{\partial^2}{\partial x_i \partial x_j}\left(p(x)\widehat{g}_{ij}(x,t)\right),$$

where $\widehat{g} \triangleq g\sigma\sigma^\top g^\top$.

Let $\rho_i \in L_2(\mathbf{R}^{n+1}) \cap C^2(\mathbf{R}^{n+1})$, $i = 0, 1$, be given such that $\rho_i(x) > 0$ for all $x \in \mathbf{R}^{n+1}$ and $\int_{\mathbf{R}^{n+1}} \rho_i(x)dx = 1$.

We consider the following parabolic equation:

$$\begin{cases} \frac{\partial p}{\partial t}(x,t) = \mathcal{M}(t)p(x,t) + \rho_1(x), \quad t \in [0,T], \\ p(x,0) = \rho_0(x). \end{cases} \tag{6.1}$$

This boundary value problem has the unique classical solution $p(x,t)$ that is continuous in $\mathbf{R}^{n+1} \times [0,T]$. Let

$$\rho(x) \triangleq \min_{t\in[0,T]} p(x,t).$$

We have that

$$p(\cdot,t) = \mathrm{G}(t,0)\rho_0 + \int_0^t \mathrm{G}(t,s)\rho_1 ds,$$

where $\mathrm{G}(t,s)$ is the semigroup operator generated by (6.1) (with $\rho_1 \equiv 0$) and such that $\mathrm{G}(s,s)\rho_i \equiv \rho_i$. We have that $(\mathrm{G}(t,s)\rho_i)(x) > 0$ for $t \in [s, s+\varepsilon)$ for some $\varepsilon = \varepsilon(x,s) > 0$. Hence $p(x,t) > 0$ for all x, t, and $\rho(x) > 0$ for all $x \in \mathbf{R}^{n+1}$. We shall use this ρ as a weight function.

We have that $\rho \in L_2(\mathbf{R}^{n+1}) \cap L_1(\mathbf{R}^{n+1})$, since $|\rho(x)| \le |\rho_0(x)|$. We introduce the weighted space $L_{2,\rho}(\mathbf{R}^{n+1})$ with the norm

$$\|u\|_{L_{2,\rho}(\mathbf{R}^{n+1})} \triangleq \left(\int_{\mathbf{R}^{n+1}} \rho(x)|u(x)|^2 dx\right)^{1/2}.$$

We introduce the space $\mathcal{Y}_k$ of functions $u = \{u_i(x,t)\}_{i=1}^k : \mathbf{R}^{n+1} \times [0,T] \to \mathbf{R}^k$ with the norm

$$\|u\|_{\mathcal{Y}_k} \triangleq \left(\sum_{i=1}^k \int_0^T \|u_i(\cdot,t)\|^2_{L_{2,\rho}(\mathbf{R}^{n+1})}\, dt \right)^{1/2}.$$

Further, we introduce the space $\mathcal{W}^1$ of functions $u = u(x,t) : \mathbf{R}^{n+1} \times [0,T] \to \mathbf{R}$ with the norm

$$\|u\|_{\mathcal{W}^1} \triangleq \|u\|_{\mathcal{Y}_1} + \left\| \frac{\partial u}{\partial x}\, g \right\|_{\mathcal{Y}_n}.$$

Finally, we introduce the space $\mathcal{W}^1_C$ consisting of all functions $u(\cdot) \in \mathcal{W}^1$ such that $u(\cdot) \in C([0,T]; L_{2,\rho}(\mathbf{R}^{n+1}))$ with the norm

$$\|u\|_{\mathcal{W}^1_C} \triangleq \sup_{t\in[0,T]} \|u(\cdot,t)\|_{L_{2,\rho}(\mathbf{R}^{n+1})} + \|u\|_{\mathcal{W}^1}.$$

The above space is a Banach space, since the weighted space $L_{2,\rho}(\mathbf{R}^{n+1})$ is a Hilbert space.

In fact, the spaces $\mathcal{Y}_k$, $\mathcal{W}^1$, and $\mathcal{W}^1_C$, are the completions in the corresponding norms of the set of smooth functions $u : \mathbf{R}^{n+1} \times [0,T] \to \mathbf{R}^k$ or $u : \mathbf{R}^{n+1} \times [0,T] \to \mathbf{R}$, respectively, that have finite support.

THEOREM 6.1. *Let p be the solution of* (6.1)*, and let $\mathcal{W}^1_C$ be the corresponding space defined via the weight $\rho(x) = \min_{t\in[0,T]} p(x,t)$. Let $\Phi(\cdot) : \mathbf{R}^{n+1} \to \mathbf{R}$ be a measurable function such that*

$$\int_{\mathbf{R}^{n+1}} p(x,T)\Phi(x)^2 dx < +\infty. \tag{6.2}$$

Then boundary value problem (4.3)–(4.4) *admits a unique solution $V \in \mathcal{W}^1_C$. Moreover, there exists a constant $C > 0$ independent on $\Phi(\cdot)$ and such that*

$$\|V\|^2_{\mathcal{W}^1_C} \le C \int_{\mathbf{R}^{n+1}} p(x,T)\Phi(x)^2 dx. \tag{6.3}$$

Note that condition (6.2) allows discontinuous Φ.

Remark 6.1. The definition of $\mathcal{W}^1_C$ ensures that problem (4.3)–(4.4) can be stated in $\mathcal{W}^1_C$. The functions V and $(\partial V/\partial x)g$ are measurable and $L_{2,\rho}$-integrable. The equality in (4.4) is the equality for elements of the space $L_{2,\rho}(\mathbf{R}^{n+1})$, it is meaningful since $V(\cdot,t)$ is continuous in t in $L_{2,\rho}(\mathbf{R}^{n+1})$. The equality in (4.3) is the equality for elements of the dual space $\mathcal{W}^{1*}$, since all components of $\frac{\partial^2 V}{\partial x^2}(x,t)\, g(x,t)\sigma(t)\sigma(t)^\top g(x,t)^\top$ belong to $\mathcal{W}^{1*}$.

It follows from the proof of Theorem 6.1 below that $\|p(\cdot,T)\|_{L_1(\mathbf{R}^{n+1})} = 2$. Hence (6.2) is satisfied for any bounded Φ. In addition, it can be shown that $\|p(\cdot,T)\|_{L_2(\mathbf{R}^{n+1})} \le C \sum_{i=1,2} \|\rho_i(\cdot)\|_{L_2(\mathbf{R}^{n+1})}$, where $C > 0$ is a constant that does not depend on ρ_i. Therefore, (6.2) is satisfied for any $\Phi \in L_4(\mathbf{R}^{n+1})$.

Theorem 6.1 gives the possibility to present the optimal investment strategy via the solution of (4.3)–(4.4) for the case of discontinuous F. An example is the goal-achieving problem, when $\widehat{D} = [0,\infty)$, $X_0 \in (0,\alpha)$, and $U(x) = 0$ if $0 \le x < \alpha$, $U(x) = 1$ if $x \ge \alpha$. Then $\Lambda = (0,\infty)$, $F(z,\lambda) = \alpha$ if $0 < \lambda \le z/\alpha$, $F(z,\lambda) = 0$ if $\lambda > z/\alpha$, and (6.2) holds for $\Phi(x) = F(e(x),\lambda)$ $(\forall\lambda)$.

7. Appendix: Proofs.

Proof of Proposition 2.1. By Jensen's inequality, it follows that

$$\begin{aligned}\mathbf{E}\Big\{\exp\tfrac{1}{2}\textstyle\int_0^T \widetilde{a}(t)^\top Q(t)\widetilde{a}(t)dt \,\Big|\, W(\cdot),\widetilde{a}(0)\Big\}&\\ = \mathbf{E}\Big\{\exp\tfrac{1}{T}\textstyle\int_0^T \tfrac{T}{2}\widetilde{a}(t)^\top Q(t)\widetilde{a}(t)dt \,\Big|\, W(\cdot),\widetilde{a}(0)\Big\}&\\ \le \tfrac{1}{T}\textstyle\int_0^T \mathbf{E}\Big\{\exp\tfrac{T}{2}\widetilde{a}(t)^\top Q(t)\widetilde{a}(t)dt \,\Big|\, W(\cdot),\widetilde{a}(0)\Big\}.&\end{aligned}$$

For definitely positive matrices we have that if $A > B > 0$, then $B^{-1} > A^{-1}$. By condition (2.7) with $m = 1$, it follows that

$$\begin{aligned}\widetilde{K}_0(t)^{-1} &> T[\sigma(t)\sigma(t)^\top - \varepsilon I_n]^{-1} = TQ(t)[I_n - \varepsilon Q(t)]^{-1}\\ &= TQ(t)\left[I_n + \textstyle\sum_{k=1}^{+\infty}\{\varepsilon Q(t)\}^k\right] > TQ(t) + T\varepsilon Q(t)^2 > TQ(t) + M,\end{aligned}\tag{7.1}$$

where $M = M(\varepsilon) > 0$ is a definitely positive constant matrix. Clearly, we can take $\varepsilon > 0$ small enough to ensure convergency of the series in (7.1).

To complete the proof, we shall use the following fact. Let ξ be a Gaussian n-dimensional vector, $K_\xi \triangleq \mathbf{E}(\xi - \mathbf{E}\xi)(\xi - \mathbf{E}\xi)^\top > 0$. It is known that the probability density function for ξ is $C\exp[-\frac{1}{2}(x-\mathbf{E}\xi)^\top K_\xi^{-1}(x-\mathbf{E}\xi)]$, where $C > 0$ is a constant. It follows that $\mathbf{E}\exp(\frac{1}{2}\xi^\top P\xi) < +\infty$ for any matrix $P \in \mathbf{R}^{n\times n}$ such that $0 < P < K_\xi^{-1}$. Then the proof follows from (7.1).

We introduce the process

$$\widetilde{R}_*(t) \triangleq \int_0^t \sigma(s)\, dw(s).$$

Let n-dimensional vector random process $\widetilde{a}_*(t)$ be defined as the solution of

$$d\widetilde{a}_*(t) = \Big(\alpha(t)\delta(t) - \alpha(t)\widetilde{a}_*(t)\Big)dt + b(t)d\widetilde{R}_*(t) + \beta(t)dW(t), \quad \widetilde{a}_*(0) = \widetilde{a}(0).$$

Set

$$\mathcal{Z}_* \triangleq \exp\left(\int_0^T [\sigma(t)^{-1}\widetilde{a}_*(t)]^\top dw(t) - \frac{1}{2}\int_0^T \widetilde{a}_*(t)^\top Q(t)\widetilde{a}_*(t)dt\right).\tag{7.2}$$

Proposition 7.1. *There exists a measurable function* $\psi : C([0,T];\mathbf{R}^n)\times \mathrm{B}([0,T];\mathbf{R}^n) \to \mathbf{R}$ *such that* $\mathcal{Z}_* = \psi(\widetilde{R}_*(\cdot),\widetilde{a}_*(\cdot))$ *and* $\mathcal{Z} = \psi(\widetilde{R}(\cdot),\widetilde{a}(\cdot))$.

Proof. Clearly, ψ is defined by

$$\log \mathcal{Z}_* = \int_0^T \widetilde{a}_*(t)^\top Q(t)\left(d\widetilde{R}_*(t) - \frac{1}{2}\widetilde{a}_*(t)dt\right).\tag{7.3}$$

Let $r_*(\cdot) \triangleq \phi_r(\widetilde{R}_*(\cdot),\Theta)$ and $B_*(t) \triangleq B(0)\exp\left(\int_0^t r_*(s)ds\right)$ (ϕ_r is defined in section 2). Let

$$\bar{\mathcal{Z}}_* \triangleq \mathbf{E}\{\mathcal{Z}_*|\widetilde{R}_*(\cdot), r_*(\cdot)\}.\tag{7.4}$$

Let $\mathcal{T} \triangleq C([0,T];\mathbf{R}^n)\times\mathbf{R}^n$. Clearly, there exists a measurable mapping $\mathcal{A} : [0,T]\times C([0,T];\mathbf{R}^n)\times\mathcal{T} \to C([0,T];\mathbf{R}^n)$ such that $\widetilde{a}_*(t) = \mathcal{A}(t,\widetilde{R}_*(\cdot),W(\cdot),\widetilde{a}(0))$ and $\widetilde{a}(t) = \mathcal{A}(t,\widetilde{R}(\cdot),W(\cdot),\widetilde{a}(0))$.

We have that $\bar{\mathcal{Z}}_* = \mathbf{E}\{\mathcal{Z}_*|\widetilde{R}_*(\cdot)\} = \bar{\psi}(\widetilde{R}_*(\cdot))$ and

$$\bar{\mathcal{Z}}_* = \mathbf{E}\{\psi([\widetilde{R}_*(\cdot), \widetilde{a}_*(\cdot))]|\widetilde{R}_*(\cdot)\} = \mathbf{E}\{\psi[\widetilde{R}_*(\cdot), \mathcal{A}\big(\cdot, \widetilde{R}_*(\cdot), W(\cdot), \widetilde{a}(0)\big)]|\widetilde{R}_*(\cdot)\}.$$

By Proposition 7.1, it follows that

$$\bar{\mathcal{Z}} = \mathbf{E}\{\psi[\widetilde{R}(\cdot), \widetilde{a}(\cdot)]|\widetilde{R}(\cdot)\} = \mathbf{E}\{\psi[\widetilde{R}(\cdot), \mathcal{A}\big(\cdot, \widetilde{R}(\cdot), W(\cdot), \widetilde{a}(0)\big)]|\widetilde{R}(\cdot)\}.$$

Hence there exists a measurable mapping $\bar{\psi}(\cdot) : C([0,T];\mathbf{R}^n) \to \mathbf{R}$ such that

$$\bar{\mathcal{Z}} = \bar{\psi}(\widetilde{R}(\cdot)), \quad \bar{\mathcal{Z}}_* = \bar{\psi}(\widetilde{R}_*(\cdot)). \tag{7.5}$$

PROPOSITION 7.2. *Let a function* $\phi : C([0,T];\mathbf{R}^n)\times \mathrm{B}([0,T];\mathbf{R}^n)\times \mathrm{B}([0,T];\mathbf{R}) \to \mathbf{R}$ *be such that* $\mathbf{E}\phi^-(\widetilde{R}(\cdot), \widetilde{a}(\cdot), r(\cdot)) < +\infty$. *Further, let a function* $\widehat{\phi} : C([0,T];\mathbf{R}^n) \times \mathrm{B}([0,T];\mathbf{R}) \to \mathbf{R}$ *be such that* $\mathbf{E}\widehat{\phi}^-(\widetilde{R}(\cdot), r(\cdot)) < +\infty$. *Then*

$$\mathbf{E}\phi(\widetilde{R}(\cdot), \widetilde{a}(\cdot), r(\cdot)) = \mathbf{E}\mathcal{Z}_*\phi(\widetilde{R}_*(\cdot), \widetilde{a}_*(\cdot), r_*(\cdot)), \tag{7.6}$$

$$\mathbf{E}\widehat{\phi}(\widetilde{R}(\cdot), r(\cdot)) = \mathbf{E}\bar{\mathcal{Z}}_*\widehat{\phi}(\widetilde{R}_*(\cdot), r_*(\cdot)), \tag{7.7}$$

$$\mathbf{E}_*\widehat{\phi}(\widetilde{R}(\cdot), r(\cdot)) = \mathbf{E}\widehat{\phi}(\widetilde{R}_*(\cdot), r_*(\cdot)). \tag{7.8}$$

Proof. By assumption $(\Theta, W(\cdot), \widetilde{a}(0))$ is independent of $w(\cdot)$. To prove (7.6) it suffices to prove

$$\mathbf{E}\left\{\phi(\widetilde{R}(\cdot), \widetilde{a}(\cdot), r(\cdot))\Big|\Theta, W(\cdot), \widetilde{a}(0)\right\} = \mathbf{E}\left\{\mathcal{Z}_*\phi(\widetilde{R}_*(\cdot), \widetilde{a}_*(\cdot), r_*(\cdot))\Big|\Theta, W(\cdot), \widetilde{a}(0)\right\} \quad \text{a.s.} \tag{7.9}$$

Thus, for the next paragraph, without loss of generality, we shall suppose that $(\Theta, W(\cdot), \widetilde{a}(0))$ is deterministic, since for each value of $(\Theta, W(\cdot), \widetilde{a}(0))$ we can construct $\widetilde{R}, \widetilde{R}_*, \widetilde{a}, \widetilde{a}_*$.

By the linearity of (2.4), it follows that $\widetilde{K}_0(t)$ defined by (2.6) is the conditional covariance for $\widetilde{a}_*(t)$ given $(W(\cdot), \widetilde{a}(0))$. Similar to the proof of Proposition 2.1, it can be shown that (2.7) with $m = 0$ ensures that $\mathbf{E}\{\mathcal{Z}_*|\Theta, W(\cdot), \widetilde{a}(0)\} = 1$ and $\mathbf{E}\mathcal{Z}_* = 1$. We define the probability measure $\bar{\mathbf{P}}$ by $d\bar{\mathbf{P}}/d\mathbf{P} = \mathcal{Z}_*$. (Each value of $(\Theta, W(\cdot), \widetilde{a}(0))$ generates its own $\bar{\mathbf{P}}$.) By Girsanov's theorem, the process

$$\bar{w}(t) \triangleq w(t) - \int_0^t \sigma(s)^{-1}\widetilde{a}_*(s)ds$$

is a Wiener process under $\bar{\mathbf{P}}$. From this we obtain

$$\begin{aligned} d\widetilde{R}(t) &= \mathcal{A}(t, \widetilde{R}(\cdot), W(\cdot), \widetilde{a}(0))dt + \sigma(t)dw(t), \\ d\widetilde{R}_*(t) &= \mathcal{A}(t, \widetilde{R}_*(\cdot), W(\cdot), \widetilde{a}(0))dt + \sigma(t)d\bar{w}(t). \end{aligned}$$

Then for each value of $(\Theta, W(\cdot), \widetilde{a}(0))$ the processes $(\widetilde{R}(\cdot), \widetilde{a}(\cdot), r(\cdot))$ and $(\widetilde{R}_*(\cdot), \widetilde{a}_*(\cdot), r_*(\cdot))$ have the same distribution on the probability spaces defined by $\mathbf{P}$ and $\bar{\mathbf{P}}$, respectively, and (7.9), hence (7.6) follows.

Further, (7.7) follows by taking conditional expectation in (7.6). Finally, using Proposition 7.1 and (7.6),

$$\begin{aligned} \mathbf{E}_*\widehat{\phi}(\widetilde{R}(\cdot), r(\cdot)) &= \mathbf{E}\mathcal{Z}^{-1}\widehat{\phi}(\widetilde{R}(\cdot), r(\cdot)) = \mathbf{E}\psi(\widetilde{R}(\cdot), \widetilde{a}(\cdot))^{-1}\widehat{\phi}(\widetilde{R}(\cdot), r(\cdot)) \\ &= \mathbf{E}\mathcal{Z}_*\psi(\widetilde{R}_*(\cdot), \widetilde{a}_*(\cdot))^{-1}\widehat{\phi}(\widetilde{R}_*(\cdot), r_*(\cdot)) = \mathbf{E}\widehat{\phi}(\widetilde{R}_*(\cdot), r_*(\cdot)). \end{aligned}$$

We turn now to Theorem 3.1. Define $\widehat{\xi}_* \triangleq F(\bar{\mathcal{Z}}_*, \widehat{\lambda})$. It follows from (7.5) that if we define $\widetilde{\phi}$ by $\widehat{\xi} = \widetilde{\phi}(\widetilde{R}(\cdot))$, then $\widehat{\xi}_* = \widetilde{\phi}(\widetilde{R}_*(\cdot))$.

Proof of Theorem 3.1. Let us show that $\mathbf{E}U^-(\widehat{\xi}) < \infty$ so that $\mathbf{E}U(\widehat{\xi})$ is well defined. For $k = 1, 2, \dots$, we introduce the random events

$$\Omega_*^{(k)} \triangleq \big\{-k \le U(\widehat{\xi}_*) \le 0\big\}, \quad \Omega^{(k)} \triangleq \big\{-k \le U(\widehat{\xi}) \le 0\big\},$$

along with their indicator functions, $\mathbb{I}_*^{(k)}$ and $\mathbb{I}^{(k)}$, respectively. The number $\widehat{\xi}_*$ provides the unique maximum of the function $\bar{\mathcal{Z}}_* U(\xi_*) - \widehat{\lambda}\xi_*$ over $\widehat{D}$, and $X_0 \in \widehat{D}$. By Proposition 7.2, we have for all $k = 1, 2, \dots,$

$$\begin{aligned}\mathbf{E}\mathbb{I}^{(k)}U(\widehat{\xi}) - \mathbf{E}\mathbb{I}_*^{(k)}\widehat{\lambda}\widehat{\xi}_* &= \mathbf{E}\mathbb{I}_*^{(k)}\left(\bar{\mathcal{Z}}_* U(\widehat{\xi}_*) - \widehat{\lambda}\widehat{\xi}_*\right) \ge \mathbf{E}\mathbb{I}_*^{(k)}\left(\bar{\mathcal{Z}}_* U(X_0) - \widehat{\lambda}X_0\right)\\ &= \mathbf{E}\mathbb{I}^{(k)}U(X_0) - \widehat{\lambda}X_0\mathbf{P}(\Omega_*^{(k)}) \ge -|U(X_0)| - |\widehat{\lambda}X_0| > -\infty.\end{aligned}$$

Further, we have that $\mathbf{E}|\widehat{\xi}_*| = \mathbf{E}_*|\widehat{\xi}| < +\infty$. Hence $\mathbf{E}U^-(\widehat{\xi}) < \infty$.

Now observe that for any $\pi \in \bar{\Sigma}$ we can apply (7.7) and (7.8) to $U(\widetilde{X}^\pi(T))$ (and use (7.5)) to obtain

$$\begin{aligned}\mathbf{E}U(\widetilde{X}^\pi(T)) = \mathbf{E}_*\{\bar{\mathcal{Z}}U(\widetilde{X}^\pi(T))\} \quad &\le \mathbf{E}_*\{\bar{\mathcal{Z}}U(\widetilde{X}^\pi(T)) - \widehat{\lambda}\widetilde{X}^\pi(T)\} + \widehat{\lambda}X_0\\ &\le \mathbf{E}_*\{\bar{\mathcal{Z}}U(\widehat{\xi}) - \widehat{\lambda}\widehat{\xi}\} + \widehat{\lambda}X_0 = \mathbf{E}_*\bar{\mathcal{Z}}U(\widehat{\xi}) = \mathbf{E}U(\widehat{\xi}).\end{aligned}$$

Thus (ii) is satisfied.

Let us show (iii). Since σ is nonrandom, hence w-adapted, then $\widehat{\xi}_* = \widehat{\phi}(w(\cdot))$, where $\widehat{\phi}(\cdot) : \mathrm{B}([0,T];\mathbf{R}^n) \to \mathbf{R}$ is a measurable functions. By the Martingale representation theorem,

$$\widehat{\xi}_* = \mathbf{E}\widehat{\xi}_* + \int_0^T f(t, w(\cdot)|_{[0,t]})^\top dw(t),$$

where $f(t,\cdot) : \mathrm{B}([0,t];\mathbf{R}^n) \to \mathbf{R}^n$ is a measurable function such that $\int_0^T |f(t, w(\cdot)|_{[0,t]})|^2 dt < +\infty$ a.s. There exists a unique measurable function $f_0(t,\cdot) : \mathrm{B}([0,t];\mathbf{R}^n) \to \mathbf{R}^n$ such that $f(t, w(\cdot)|_{[0,t]}) \equiv f_0(t, \widetilde{R}_*(\cdot)|_{[0,t]})$. Thus,

$$\widehat{\xi}_* = \mathbf{E}\widehat{\xi}_* + \int_0^T f_0(t, \widetilde{R}_*(\cdot)|_{[0,t]})^\top dw(t) = \mathbf{E}\widehat{\xi}_* + \int_0^T f_0(t, \widetilde{R}_*(\cdot)|_{[0,t]})^\top \sigma(t)^{-1} d\widetilde{R}_*(t).$$

Proposition 7.2 implies that $\mathbf{E}\widehat{\xi}_* = \mathbf{E}_*\widehat{\xi} = X_0$, and

$$\widehat{\xi} = X_0 + \int_0^T f_0(t, \widetilde{R}(\cdot)|_{[0,t]})^\top \sigma(t)^{-1} d\widetilde{R}(t).$$

Hence the strategy $\widehat{\pi}(t)^\top = B(t) f_0(t, \widetilde{R}(\cdot)|_{[0,t]})^\top \sigma(t)^{-1}$ replicates $B(T)\widehat{\xi}$. It belongs to $\bar{\Sigma}$; in particular, since w and $\widetilde{R}$ generate the same sigma-algebra and $\widehat{D}$ is convex, then $\widetilde{X}(t, \pi(\cdot)) = \mathbf{E}\left\{\widehat{\xi}\,|\,\widetilde{R}(\cdot)|_{[0,t]}\right\} \in \widehat{D}$, hence bounded below. This completes the proof of Theorem 3.1.

Proof of Lemma 4.1. Let $\mathcal{V} \triangleq \mathbf{R}^n \times \mathbf{R} \times \mathbf{R}^{\frac{n(n+1)}{2}}$. Clearly, $\mathcal{V}$ is an $\widetilde{n}$-dimensional linear vector space, where $\widetilde{n} \triangleq n + 1 + n(n+1)/2$. Let $\widetilde{y}(t) = (\widetilde{y}_1(t), \widetilde{y}_2(t), \widetilde{y}_3(t))$ be a process in $\mathcal{V}$ such that

$$\widetilde{y}(t) = (y(t), \widetilde{y}_3(t)) = (\widetilde{y}_1(t), \widetilde{y}_2(t), \widetilde{y}_3(t)) = \big(\widehat{a}(t), \ln \bar{\mathcal{Z}}(t), \widehat{a}(t)\widehat{a}(t)^\top\big).$$

The last equality is satisfied by (4.7). It can be seen that the equation for $\widetilde{y}(t)$ is linear:

$$\begin{aligned}
d\widetilde{y}_1(t) &= [\widehat{A}(t)\widetilde{y}_1(t) + v(t)]dt + E(t)\,d\widetilde{R}(t),\\
d\widetilde{y}_2(t) &= -\tfrac{1}{2}\mathrm{Tr}\{Q(t)\widetilde{y}_3(t)\}\,dt + \widetilde{y}_1(t)^\top Q(t)\,d\widetilde{R}(t),\\
d\widetilde{y}_3(t) &= [\widehat{A}(t)\widetilde{y}_3(t) + \widetilde{y}_3(t)^\top \widehat{A}(t)^\top + v(t)\widetilde{y}_2(t)^\top\\
&\qquad + \widetilde{y}_2(t)v(t)^\top + \tfrac{1}{2}\{E(t)\sigma(t)\sigma(t)^\top E(t)^\top\}]dt\\
&\qquad + E(t)\,d\widetilde{R}(t)\widetilde{y}_2(t)^\top + y_2(t)\,d\widetilde{R}(t)^\top E(t)^\top.
\end{aligned}$$

Here $\widehat{A}(t)$, $v(t)$, $E(t)$ are known deterministic functions in $\mathbf{R}^{n\times n}$, $\mathbf{R}^n$ and $\mathbf{R}^{n\times n}$, respectively. In particular, $\widehat{A}(t) = A(t) - b(t)\sigma(t)^\top Q(t)$. Thus, the equation for $\widehat{y}(t)$ can be rewritten as

$$\left\{\begin{array}{l} d\widetilde{y}(t) = \widetilde{f}(\widetilde{y}(t),t)dt + \sum_{i=1}^n \widetilde{g}_i(\widetilde{y}(t),t)\,d\widetilde{R}_i(t),\\ \widetilde{y}(0) = \widetilde{y}_0, \end{array}\right. \tag{7.10}$$

with $\widetilde{y}_0 = \big(m_0, 0, m_0 m_0^\top\big)$, and with some functions $\widetilde{f}(\widetilde{x},t) : \mathcal{V}\times[0,T]\to\mathcal{V}$ and $\widetilde{g}_i(\widetilde{x},t) : \mathcal{V}\times[0,T]\to\mathcal{V}$, $i = 1,\dots,n$, that are affine in $\widetilde{x}\in\mathcal{V}$ with continuous in t coefficients. In particular, $\partial\widetilde{f}(\widetilde{x},t)/\partial\widetilde{x}$ and $\partial\widetilde{g}_i(\widetilde{x},t)/\partial\widetilde{x}$ depend only on t, and they are uniformly bounded. Hence (7.10) has a unique solution. Therefore, (4.2) has the unique solution $y(t)$.

Let $\widetilde{V}(\widetilde{x},t) \triangleq \mathbf{E}_*\Phi(\bar{y}^{\widetilde{x},s}(T))$, where the process $\bar{y}^{\widetilde{x},s}(\cdot)$ takes values in $\mathbf{R}^{n+1}$ and is such that $\widetilde{y}^{\widetilde{x},s}(\cdot) = (\bar{y}^{\widetilde{x},s}(\cdot), \widetilde{y}_3^{\widetilde{x},s}(\cdot))$ is the solution of (7.10) given the initial condition $\widetilde{y}(s) = \widetilde{x}\in\mathcal{V}$. Then $\widetilde{V}(\widetilde{x},t)$ is the classical solution of the boundary value problem for the corresponding backward Kolmogorov's equation

$$\left\{\begin{array}{l} \frac{\partial\widetilde{V}}{\partial t}(\widetilde{x},t) + \mathcal{L}(t)\widetilde{V}(\widetilde{x},t) = 0, \quad t\in[0,T],\\ \widetilde{V}(\widetilde{x},T) = \Phi(\widetilde{x}_1,\widetilde{x}_2), \end{array}\right. \tag{7.11}$$

where $\widetilde{x} = (\widetilde{x}_1,\widetilde{x}_2,\widetilde{x}_3)\in\mathbf{R}^n\times\mathbf{R}\times\mathbf{R}^{\frac{n(n+1)}{2}}$, and where $\mathcal{L}(t)$ is the second order differential operator on functions $v : \mathcal{V}\to\mathbf{R}$ generated by the Markov process $\widetilde{y}(t)$.

Let $y^{x,s}(\cdot) = (y_1^{x,s}(\cdot),\dots,y_{n+1}^{x,s}(\cdot))$ be the solution of (4.2), and let $V(x,t) \triangleq \mathbf{E}_*\Phi(y^{x,t}(T))$. Clearly,

$$\widetilde{y}^{\widetilde{x},s}(t) \equiv (y^{x,s}(t), \widehat{y}^{x,s}(t)\widehat{y}^{x,s}(t)^\top) = \big(y_1^{x,s}(t),\dots,y_n^{x,s}(t), y_{n+1}^{x,s}(t),\, \widehat{y}^{x,s}(t)\widehat{y}^{x,s}(t)^\top\big)\,,$$

if

$$\begin{aligned}
&\widetilde{x} = (x,\widetilde{x}_3) = (\widehat{x}, x_{n+1}, \widehat{x}\widehat{x}^\top), \quad \widehat{x} = (x_1,\dots,x_n),\\
&x = (x_1,\dots,x_n,x_{n+1}) = (\widehat{x},x_{n+1})\in\mathbf{R}^{n+1},\\
&\widetilde{x}_3 = \widehat{x}\widehat{x}^\top\in\mathbf{R}^{n(n+1)/2}, \quad \widehat{y}^{x,s}(\cdot) = (y_1^{x,s}(\cdot),\dots,y_n^{x,s}(\cdot)).
\end{aligned}$$

In that case, $V(x,t)\equiv\widetilde{V}(x_1,\widehat{x}_2,\widehat{x}_2\widehat{x}_2^\top)$, where $x = (\widehat{x},x_{n+1})$, $\widehat{x}\in\mathbf{R}^n$. Therefore, $V(x,t)$ is the classical solution of problem (4.3)–(4.4).

Let $y_*(\cdot)$ denote the solution of (4.2) with $\widetilde{R}(\cdot)$ replaced by $\widetilde{R}_*(\cdot) = \int_0^\cdot \sigma(t)\,dw(t)$.

Set $\widetilde{X}_*(t) \triangleq V(y_*(t),t)$. From (4.3) and Itô's lemma, it follows that

$$\widetilde{X}_*(T) = \widetilde{X}_*(t) + \int_t^T B_*(s)^{-1}\pi_*(s)^\top d\widetilde{R}_*(s),$$

where $\pi_*(t)^\top \triangleq B_*(t)\frac{\partial V}{\partial x}(y_*(t),t)g(y_*(t),t)$. It follows that $\widetilde{X}_*(0) = V(y_*(0),0) = \mathbf{E}V(y_*(T),T) = X_0$ and

$$d\widetilde{X}_*(t) = B_*(t)^{-1}\pi_*(t)^\top d\widetilde{R}_*(t), \quad \widetilde{X}_*(T) = \Phi(y_*(T)). \tag{7.12}$$

Then $\widetilde{X}_*(t) = \widetilde{\psi}(t,\widetilde{R}_*)$ for some measurable $\widetilde{\psi}$, and the result follows if we observe that $\widetilde{X}(t) = \widetilde{\psi}(t,\widetilde{R})$ replicates the claim as desired for $\pi(t)^\top \triangleq B(t)\frac{\partial V}{\partial x}(y(t),t)g(y(t),t)$.

To continue, we require some a priori estimates. Let $\zeta_*(t) \triangleq B_*(t)^{-1}\sigma(t)^\top\pi_*(t)$.

We consider the conditional probability space given in $(\Theta, W(\cdot), \widetilde{a}(0))$. With respect to the conditional probability space, it follows from (7.12) that

$$\begin{cases} d\widetilde{X}_*(t) = \zeta_*(t)^\top dw(t), \\ \widetilde{X}_*(T) = \Phi(y_*(T)). \end{cases} \tag{7.13}$$

By Proposition 2.2 from El Karoui, Peng, and Quenez [12], the (unique) solution $(\zeta_*(t), \widetilde{X}_*(t))$ of linear stochastic backward equation (7.13) is a process in $L_2([0,T], L_2(\Omega,\mathcal{F},P)) \times C([0,T], L_2(\Omega,\mathcal{F},P))$, and there exists a constant c_0, independent of $(\Phi(\cdot), \Theta, W(\cdot), \widetilde{a}(0))$, and such that

$$\begin{aligned} \sup_{t\in[0,T]} \mathbf{E}\left\{|\widetilde{X}_*(t)|^2 \middle| \Theta, W(\cdot), \widetilde{a}(0)\right\} \quad &+\mathbf{E}\left\{\int_0^T |\zeta_*(t)|^2 dt \middle| \Theta, W(\cdot), \widetilde{a}(0)\right\} \\ &\le c_0 \mathbf{E}\left\{\Phi(y_*(T))^2 \middle| \Theta, W(\cdot), \widetilde{a}(0)\right\} \quad \text{a.s.} \end{aligned}$$

Hence

$$\sup_{t\in[0,T]} \mathbf{E}|\widetilde{X}_*(t)|^2 + \mathbf{E}\int_0^T B_*(t)^{-2}|\pi_*(t)|^2 dt \le c_1 \mathbf{E}\Phi(y_*(T))^2, \tag{7.14}$$

where $c_1 > 0$ is a constant that does not depend on $\Phi(\cdot)$. Then (4.5) follows. This completes the proof. □

Proof of Theorem 4.2. Clearly, the equation for $y(t)$ is

$$\begin{cases} d\widehat{a}(t) = [A(t)\widehat{y}(t) - b(t)\sigma(t)^\top Q(t) + \alpha(t)\delta(t)]dt + \gamma(t)Q(t)\,d\widetilde{R}(t), \\ dy_{n+1}(t) = \frac{1}{2}\widehat{a}(t)^\top Q(t)\widehat{a}(t)dt - \widehat{a}(t)^\top Q(t)\,d\widetilde{R}(t). \end{cases}$$

As in the proof above, it can be shown that $\widetilde{X}(t) = V(y(t),t,\widehat{\lambda})$ is the solution of (7.12), i.e., it is the normalized wealth. Then the proof follows.

Let $\mathcal{N}_2$ be the set of all Gaussian processes $\bar{a}(t) : [0,T]\times\Omega \to \mathbf{R}^n$ which are progressively measurable with respect to the filtration generated by $[a(0), w(t), W(t)]$ and such that $\mathbf{E}\int_0^T |\bar{a}(t)|^2 dt < +\infty$. For $\bar{a}(\cdot) \in \mathcal{N}_2$, let

$$Z(t,\bar{a}(\cdot)) \triangleq \exp\left[\int_0^t \bar{a}(s)^\top Q(s) d\widetilde{R}(s) - \frac{1}{2}\int_0^t \bar{a}(s)^\top Q(s)\bar{a}(s)ds\right].$$

Proposition 7.3. *Let $\bar{a}(\cdot) \in \mathcal{N}_2$, let $p \in (1,+\infty)$, and let $\mu \in \mathbf{R}$, $\mu < 0$ or $\mu > 1$. Let $\bar{K}(t)$ be the covariance matrix of $\bar{a}(t)$ under $\mathbf{P}_*$, and let $\kappa(p) \triangleq qT(\mu^2 p - \mu)$, where $q \triangleq p(p-1)^{-1}$. Let $\kappa(p)\bar{K}(t) < \sigma(t)\sigma(t)^\top - \varepsilon I_n$, where $\varepsilon > 0$ is a constant. Then $\mathbf{E}_* Z(t,\bar{a}(\cdot))^\mu < +\infty$.*

Proof of Proposition 7.3. If $\mu \in [0,1]$, then $\mathbf{E}_* Z(t,\bar{a}(\cdot))^\mu < +\infty$ (see Lakner [23, p. 93]). Therefore, we can assume without loss of generality that $\mu < 0$ or $\mu > 1$. Clearly,

$$Z(t,\bar{a}(\cdot))^\mu = \exp\left[\mu \int_0^t \bar{a}(s)^\top Q(s) d\widetilde{R}(s) - \tfrac{\mu}{2}\int_0^t \bar{a}(s)^\top Q(s)\bar{a}(s)ds\right] = \zeta(t)\zeta_0(t),$$

where

$$\zeta(t) \triangleq \exp\left[\mu \int_0^t \bar{a}(s)^\top Q(s) d\widetilde{R}(s) - \frac{\mu^2 p}{2}\int_0^t \bar{a}(s)^\top Q(s)\bar{a}(s)ds\right],$$

and

$$\zeta_0(t) \triangleq \exp\left[\frac{\mu^2 p - \mu}{2}\int_0^t \bar{a}(s)^\top Q(s)\bar{a}(s)ds\right].$$

By Hölder inequality, $\mathbf{E}_* \mathcal{Z}^\mu \le [\mathbf{E}_* \zeta(T)^p]^{1/p} [\mathbf{E}_* \zeta_0(T)^q]^{1/q}$.

Similar to the proof of Lemma A.1 from Lakner [23], we have that $\mathbf{E}_* \zeta(T)^p < +\infty$ because $\zeta(t)^p$ is a positive local martingale with respect to $\mathbf{P}_*$, thus by Fatou's lemma it is a supermartingale.

By Jensen's inequality,

$$\begin{aligned}\mathbf{E}_* \zeta_0(T)^q &= \mathbf{E}_* \exp\left[q\tfrac{\mu^2 p-\mu}{2}\int_0^T \bar{a}(s)^\top Q(s)\bar{a}(s)ds\right] \\ &= \mathbf{E}_* \exp\left[\tfrac{1}{2T}\kappa(p)\int_0^T \bar{a}(s)^\top Q(s)\bar{a}(s)ds\right] \le \tfrac{1}{T}\int_0^T \mathbf{E}_* \exp\left[\tfrac{1}{2}\kappa(p)\bar{a}(s)^\top Q(s)\bar{a}(s)\right] ds.\end{aligned} \tag{7.15}$$

Remember that $Q \triangleq (\sigma\sigma^\top)^{-1}$, and $\kappa(p) > 0$. Similar to (7.1), we obtain

$$\begin{aligned}\bar{K}(t)^{-1} &> \kappa(p)[\sigma(t)\sigma(t)^\top - \varepsilon I_n]^{-1} = \kappa(p)Q(t)[I_n - \varepsilon Q(t)]^{-1} \\ &= \kappa(p)Q(t)\left[I_n + \textstyle\sum_{k=1}^{+\infty}\{\varepsilon Q(t)\}^k\right] > \kappa(p)Q(t) + \kappa(p)\varepsilon Q(t)^2 \\ &> \kappa(p)Q(t) + M_1,\end{aligned} \tag{7.16}$$

where $M_1 = M_1(\varepsilon) > 0$ is a definitely positive constant matrix. (We can take $\varepsilon > 0$ small enough to ensure convergency.) Similar to the proof of Proposition 2.1, it follows from (7.15), (7.16) that $\mathbf{E}_* \zeta_0(T)^q < +\infty$ and $\mathbf{E}_* Z(t,\bar{a}(\cdot))^\mu < +\infty$. □

Proof of Lemma 5.1. If $\mu \in [0,1]$, then $\mathbf{E}_* \bar{\mathcal{Z}}^\mu < +\infty$ (see Lakner [23, p. 93]). Therefore, we can assume without loss of generality that $\mu < 0$ or $\mu > 1$. Note that $\widehat{a}(\cdot) \in \mathcal{N}_2$. By Proposition 7.3, if (i) is satisfied, then $\mathbf{E}_* \bar{\mathcal{Z}}^\mu < +\infty$.

Further, let (ii) be satisfied. Clearly, $\widetilde{a}(\cdot) \in \mathcal{N}_2$. By Proposition 7.3 again, $\mathbf{E}_* Z(T,\widetilde{a}(\cdot))^\mu < +\infty$. By (7.4), $\bar{\mathcal{Z}}_* = \mathbf{E}\{Z(T,\widetilde{a}_*(\cdot))|\widetilde{R}_*(\cdot), r_*(\cdot)\}$. Hence, by Jensen's inequality $\mathbf{E}_* \bar{\mathcal{Z}}^\mu \le \mathbf{E}_* Z(T,\widetilde{a}(\cdot))^\mu < +\infty$.

Proof of Theorem 6.1. Let τ be a random variable that takes values in $[0,T]$ and such that $\mathbf{P}(\tau = 0) = 1/2$ and $\mathbf{P}(\tau \in (t_1,t_2]) = (t_2 - t_1)/(2T)$ for $0 < t_1 < t_2 \le T$. Let $\eta_i \in L_2(\Omega,\mathcal{F},\mathbf{P},\mathbf{R}^{n+2})$ be random vectors such that they have the probability density functions $\rho_i(x)$, $i = 0,1$. We assume that $\tau, \eta_0, \eta_1, w, \Theta, W(\cdot), \widetilde{a}(0)$ are mutually independent.

Let $\eta \triangleq \eta_0 \mathbb{I}_{\{\tau=0\}} + \eta_1 \mathbb{I}_{\{\tau>0\}}$, and let $\eta_*(\cdot)$ be the solution of the Itô's equation

$$\begin{cases} d\eta_*(t) = f(\eta_*(t),t)dt + g(\eta_*(t),t)d\widetilde{R}_*(t), & t > \tau, \\ \eta_*(\tau) = \eta. \end{cases} \tag{7.17}$$

Equation (6.1) is the forward Kolmogorov's equation for the case when time of birth is distributed as τ, and the vector $\eta_*(t)$ has the conditional probability density function $p(x,t)/2$ in the sense that $\mathbf{P}(\eta_*(t)\in\Gamma,\ t\geq\tau)=1/2\int_\Gamma p(x,t)dx$ for any domain $\Gamma\subset\mathbf{R}^{n+1}$, where p is the solution of (6.1).

Note that we need random τ with the selected probability density on $(0,T]$ to generate the free term in parabolic equation (6.1).

Assume that $\Phi(\cdot)\in C^2(\mathbf{R}^{n+1})$ and it has finite support. Let $V(x,t)\triangleq\mathbf{E}_*\Phi(y^{x,t}(T))$, where $y^{x,s}(\cdot)$ is the solution of (4.2). Then $V(x,t)$ is the classical solution of problem (4.3)–(4.4). Set $\widetilde{Y}_*(t)\triangleq V(\eta_*(t),t)$. From (4.3) and Itô's lemma, it follows that

$$\widetilde{Y}_*(T)=\widetilde{Y}_*(t)+\int_t^T B_*(s)^{-1}\varrho_*(s)^\top d\widetilde{R}_*(s),\quad \tau\leq t\leq T,$$

where $\varrho_*(t)^\top\triangleq B_*(t)\frac{\partial V}{\partial x}(\eta_*(t),t)g(\eta_*(t),t)$. Hence

$$d\widetilde{Y}_*(t)=B_*(t)^{-1}\varrho_*(t)^\top d\widetilde{R}_*(t),\quad \widetilde{Y}_*(T)=\Phi(\eta_*(T)). \tag{7.18}$$

To continue, we require some estimates. Let $\widehat{\zeta}_*(t)\triangleq B_*(t)^{-1}\sigma(t)^\top\varrho_*(t)$.

Consider the conditional probability space given in $(\tau,\eta,\Theta,W(\cdot),\widetilde{a}(0))$. With respect to the conditional probability space, it follows from (7.18) that

$$\begin{cases} d\widetilde{Y}_*(t)=\widehat{\zeta}_*(t)^\top dw(t),\\ \widetilde{Y}_*(T)=\Phi(\eta_*(T)).\end{cases} \tag{7.19}$$

By Proposition 2.2 from El Karoui, Peng, and Quenez [12] again, the (unique) solution $(\widehat{\zeta}_*(t),\widetilde{Y}_*(t))$ of stochastic backward equation (7.19) is a process in $L_2([\tau,T],L^2(\Omega,\mathcal{F},P))\times C([\tau,T],L^2(\Omega,\mathcal{F},P))$ given $(\tau,\eta,\Theta,W(\cdot),\widetilde{a}(0))$, and there exists a constant C_0 that is independent of $(\Phi(\cdot),\tau,\eta,\Theta,W(\cdot),\widetilde{a}(0))$, and such that

$$\begin{aligned}&\sup_{t\in[0,T]}\mathbf{E}\mathbb{I}_{\{t\geq\tau\}}\left\{|\widetilde{Y}_*(t)|^2\big|\,\tau,\eta,\Theta,W(\cdot),\widetilde{a}(0)\right\}\\ &+\mathbf{E}\big\{\textstyle\int_0^T\mathbb{I}_{\{t\geq\tau\}}|\widehat{\zeta}_*(t)|^2dt\,\big|\,\tau,\eta,\Theta,W(\cdot),\widetilde{a}(0)\big\}\\ &=\sup_{t\in[\tau,T]}\mathbf{E}\left\{|\widetilde{Y}_*(t)|^2\,\big|\,\tau,\eta,\Theta,W(\cdot),\widetilde{a}(0)\right\}+\mathbf{E}\big\{\textstyle\int_\tau^T|\widehat{\zeta}_*(t)|^2dt\Big|\,\tau,\eta,\Theta,W(\cdot),\widetilde{a}(0)\big\}\\ &\qquad\leq C_0\mathbf{E}\left\{\Phi(\eta_*(T))^2\,\big|\,\tau,\eta,\Theta,W(\cdot),\widetilde{a}(0)\right\}\quad\text{a.s.}\end{aligned}$$

Hence there exists a constant c_0, independent of $\Phi(\cdot)$ and such that

$$\sup_{t\in[0,T]}\mathbf{E}\mathbb{I}_{\{t\geq\tau\}}|\widetilde{Y}_*(t)|^2+\mathbf{E}\int_0^T\mathbb{I}_{\{t\geq\tau\}}B_*(t)^{-2}|\varrho_*(t)|^2dt\leq c_0\mathbf{E}\Phi(\eta_*(T))^2. \tag{7.20}$$

Let $\Phi(\cdot)$ be a general measurable function satisfying the conditions specified in the theorem. Then, there exists a sequence $\{\Phi^{(i)}(\cdot)\}$, where $\Phi^{(i)}(\cdot)\in C^2(\mathbf{R}^{n+1})$ are such that they all have finite support and

$$\mathbf{E}|\Phi^{(i)}(\eta_*(T))-\Phi(\eta_*(T))|^2=\textstyle\int_{\mathbf{R}^{n+2}}p(x,T)|\Phi^{(i)}(x)-\Phi(x)|^2dx\to0\quad\text{as}\quad i\to\infty. \tag{7.21}$$

Let $\widetilde{Y}_*^{(i)}(\cdot)$, $\varrho_*^{(i)}(\cdot)$, and $V^{(i)}(\cdot)$ be the corresponding processes and functions. Let

$$\Psi_{i,j}\triangleq\sup_{t\in[0,T]}\mathbf{E}\mathbb{I}_{\{t\geq\tau\}}|\widetilde{Y}_*^{(i)}(t)-\widetilde{Y}_*^{(j)}(t)|^2+\mathbf{E}\int_0^T\mathbb{I}_{\{t\geq\tau\}}B_*(t)^{-2}|\varrho_*^{(i)}(t)-\varrho_*^{(j)}(t)|^2dt.$$

By (7.20)–(7.21) and the linearity of (7.19), it follows that

$$\Psi_{i,j} \le c_0 \mathbf{E}|\Phi^{(i)}(\eta_*(T)) - \Phi^{(j)}(\eta_*(T))|^2 \to 0 \quad \text{as} \quad i,j \to \infty.$$

We have that $V^{(j)} \in \mathcal{W}_C^1$, since they are bounded together with their partial derivatives with respect to $x_1, \dots, x_{n+1}$. Remember that $0 < \rho(x) \le p(x,t)$ for all x, t. Further, we have that

$$\Psi_{i,j} = \sup_{t\in[0,T]} \int_{\mathbf{R}^{n+1}} p(x,t)|V^{(i)}(x,t) - V^{(j)}(x,t)|^2 dx + \int_s^T dt \int_{\mathbf{R}^{n+1}} p(x,t) \left| \left[\frac{\partial V^{(i)}}{\partial x}(x,t) - \frac{\partial V^{(j)}}{\partial x}(x,t) \right] g(x,t) \right|^2 dx.$$

Hence $\|V^{(i)} - V^{(j)}\|^2_{\mathcal{W}_C^1} \le \Psi_{i,j} \to 0$ as $i,j \to \infty$. Therefore, $V^{(i)}$ is a Cauchy sequence in $\mathcal{W}_C^1$, and it has the limit V in $\mathcal{W}_C^1$. This V is the desired solution, and (6.3) is satisfied. This completes the proof. □

Note that it follows from the proof above that the sequences $\{\widetilde{Y}_*^{(i)}(\cdot)\}_{i=1}^\infty$ and $\{\varrho_*^{(i)}(\cdot)\}_{i=1}^\infty$ are Cauchy sequences in the spaces $C([\tau,T]; L^2(\Omega, \mathcal{F}, \mathbf{P}\{\cdot \mid \tau\}))$ and $L_2([\tau,T]; L^2(\Omega, \mathcal{F}, \mathbf{P}\{\cdot \mid \tau\}))$, respectively. Hence the corresponding limits $\widetilde{Y}_*(\cdot)$, $\varrho_*(\cdot)$ exist and belong to these spaces given τ.

This paper presents the development of some results and ideas that grew from our collaboration with Ulrich Haussmann during the author's stay at Pacific Institute for the Mathematical Sciences, Vancouver (see, e.g., Dokuchaev and Haussmann [8]).

Acknowledgments. The author wishes to thank Prof. U. Haussmann for the support and useful discussion. The author also wishes to thank the anonymous referees for their insightful comments which greatly strengthened the paper.

REFERENCES

[1] L. Arnold, *Stochastic Differential Equations. Theory and Applications*, Wiley-Interscience. New York, 1973.

[2] M. J. Brennan, *The role of learning in dynamic portfolio decisions*, European Finance Rev., 1 (1998), pp. 295–306.

[3] R.-R. Chen and L. Scott, *Maximum likelihood estimation for a multifactor equilibrium model of the term structure of interest rates*, J. Fixed Income, 4 (1993), pp. 14–31.

[4] J. B. Detemple, *Asset pricing in an economy with incomplete information*, J. Finance, 41 (1986), pp. 369–382.

[5] N. G. Dokuchaev, *Probability distributions of Ito's processes: Estimations for density functions and for conditional expectations of integral functionals*, Theory of Probability and its Applications, 39 (1995), pp. 662–670.

[6] N. G. Dokuchaev, *Dynamic Portfolio Strategies: Quantitative Methods and Empirical Rules for Incomplete Information*, Kluwer Academic Publishers, Boston, 2002.

[7] N. G. Dokuchaev and U. Haussmann, *Optimal portfolio selection and compression in an incomplete market*, Quant. Finance, 1 (2001), pp. 336–345.

[8] N. G. Dokuchaev and U. Haussmann, *Adaptive portfolio selection based on Historical Prices*, in Quantitative Risk Management in Finance, Carnegie Mellon University, Pittsburgh, 2000.

[9] N. G. Dokuchaev and K. L. Teo, *Optimal hedging strategy for a portfolio investment problem with additional constraints*, Dyn. Contin. Discrete Impuls. Systems, 7 (2000), pp. 385–404.

[10] N. G. Dokuchaev and X. Y. Zhou, *Optimal investment strategies with bounded risks, general utilities, and goal achieving*, J. Math. Econom., 35 (2001), pp. 289–309.

[11] U. Dothan and D. Feldman, *Equilibrium interest rates and multiperiod bonds in a partially observable economy*, J. Finance, 41 (1986), pp. 369–382.

[12] N. El Karoui, S. Peng, and M. C. Quenez, *Backward stochastic differential equations in finance*, Math. Finance, 7 (1997), pp. 1–71.
[13] M. Frittelli, *Introduction to a theory of value coherent with the no-arbitrage principle*, Finance Stoch., 3 (2000), pp. 275–297.
[14] G. Gennotte, *Optimal portfolio choice under incomplete information*, J. Finance, 41 (1986), pp. 733–749.
[15] N. Hakansson, *Portfolio Analysis*, W. W. Norton, New York, 1997.
[16] I. Karatzas, *Adaptive control of a diffusion to a goal and a parabolic Monge-Ampère-type equation*, Asian J. Math., 1 (1997), pp. 295–313.
[17] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd ed., Springer-Verlag, New York, 1991.
[18] I. Karatzas and S. E. Shreve, *Methods of Mathematical Finance*, Springer-Verlag, New York, 1998.
[19] I. Karatzas and X.-X. Xue, *A note on utility maximization under partial observations*, Math. Finance, 1 (1991), pp. 57–70.
[20] I. Karatzas and X. Zhao, *Bayesian adaptive portfolio optimization*, Hand. Math. Finance, Cambridge University Press, Cambridge, 2001, pp. 632–669.
[21] Y. Kuwana, *Certainty equivalence and logarithmic utilities in consumption/investment problems*, Math. Finance, 5 (1995), pp. 297–309.
[22] P. Lakner, *Utility maximization with partial information*, Stochastic Processes Appl., 56 (1995), pp. 247–273.
[23] P. Lakner, *Optimal trading strategy for an investor: The case of partial information*, Stochastic Processes Appl., 76 (1998), pp. 77–97.
[24] R. S. Liptser and A. N. Shiryaev, *Statistics of Random Processes.* I. *General Theory*, 2nd ed., Springer-Verlag, Berlin, 2001.
[25] A. W. Lo, *Maximum likelihood estimation of generalized Itô processes with discretely sampled data*, Econometrics Theory, 4 (1988), pp. 231–247.
[26] R. Merton, *Lifetime portfolio selection under uncertainty: The continuous-time case*, Rev. Econom. Statist., 51 (1969), pp. 247–257.
[27] N. D. Pearson and T.-S. Sun, *Exploiting the conditional density in estimating the term structure: An application to the Cox, Ingersoll, and Ross model*, J. Finance, 49 (1994), pp. 1279–1304.
[28] R. W. Rishel, *Optimal portfolio management with partial observations and power utility function*, Stochastic Analysis, Control, Optimization and Applications, W. M. McEneaney, G. Yin, and Q. Zhang, eds., Birkhäuser, Boston, 1999, pp. 605–619.
[29] B. L. Rozovskii, *Stochastic Evolution Systems.* Linear Theory and Applications to Nonlinear Filtering. Kluwer Academic Publishers, Dordrecht, 1990.
[30] J. T. Williams, *Capital asset prices with heterogeneous beliefs*, J. Financial Economics, 5 (1977), pp. 219–240.
[31] J. Yong and X. Y. Zhou, *Stochastic controls: Hamiltonian systems and HJB equations*, Springer-Verlag, New York, 1999.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 4, pp. 1259–1288

HYBRID CONTROL SYSTEMS AND VISCOSITY SOLUTIONS*

SHEETAL DHARMATTI† AND MYTHILY RAMASWAMY‡

Abstract. We investigate a model of hybrid control system in which both discrete and continuous controls are involved. In this general model, discrete controls act on the system at a given set interface. The state of the system is changed discontinuously when the trajectory hits predefined sets, namely, an autonomous jump set A or a controlled jump set C where the controller can choose to jump or not. At each jump, the trajectory can move to a different Euclidean space. We prove the continuity of the associated value function V with respect to the initial point. Using the dynamic programming principle satisfied by V, we derive a quasi-variational inequality satisfied by V in the viscosity sense. We characterize the value function V as the unique viscosity solution of the quasi-variational inequality by the comparison principle method.

Key words. dynamic programming principle, viscosity solution, quasi-variational inequality, hybrid control

AMS subject classifications. 34H05, 34K35, 49L20, 49L25

DOI. 10.1137/040618072

1. Introduction.

Many complicated control systems, like flight control and transportation, perform computer coded checks and issue logical as well as continuous control commands. The interaction of these different types of dynamics and information leads to hybrid control problems. Thus hybrid control systems are those having continuous and discrete dynamics and continuous and discrete controls. Many control systems, which involve both logical decision making and continuous evolution, are of this type. Typical examples of such systems are constrained robotic systems [1] and automated highway systems [8]. See [5], [6], and the references therein for more examples of such systems.

In [5], Branicky, Borkar, and Mitter presented a model for the most general hybrid control system in which continuous controls are present and, in addition, discrete controls act at a given set interface, which corresponds to the logical decision making process as in the above examples. The state of the system is changed discontinuously when the trajectory hits these predefined sets, namely, an autonomous jump set A or a controlled jump set C where the controller can choose to jump or not. They prove right continuity of the value function corresponding to this hybrid control problem. Using the dynamic programming principle they arrive at the partial differential equation satisfied by the value function, which turns out to be the quasi-variational inequality, referred hereafter as QVI.

In [4], Bensoussan and Menaldi study a similar system and prove that the value function u is close to a certain u_ε which they mention to be continuous indicating the use of the basic ordinary differential equation estimate for continuous trajectories and the continuity of the first hitting time (see [4, Theorem 2.5 and Remark 3.5]). They

*Received by the editors November 1, 2004; accepted for publication (in revised form) March 24, 2005; published electronically October 7, 2005. This work was partially supported by DRDO 508 and ISRO 050 grants to Nonlinear Studies Group, IISc.

http://www.siam.org/journals/sicon/44-4/61807.html

†Department of Mathematics, Indian Institute of Science, Bangalore 560012, India (sheetal@math.iisc.ernet.in). This author is a UGC Research Fellow and the financial support from UGC is gratefully acknowledged.

‡IISc-TIFR Mathematics Program, TIFR Center, P.O. Box 1234, Bangalore 560012, India (mythily@math.tifrbng.res.in).

prove its uniqueness as a viscosity solution of the QVI in a certain special case where the autonomous jump set is empty and the controlled jump set is the whole space.

In our work, we study this problem in a more general case in which the autonomous jump set is nonempty and the controlled jump set can be arbitrary. Our model is based on that of [5]. Our main aim is to prove uniqueness in the most general case when the sets A and C are nonempty and also to obtain precise estimates to improve the earlier continuity results. Our motivation comes from the fact that in all the real-life models mentioned above, logical decision making is always involved as well as the continuous control. This will correspond to a nonempty autonomous jump set A.

Here we prove the local Hölder continuity of the value function under a transversality condition, the same as the one assumed in [5] and [4] (see (2.36) in [4]). For this we need to follow the trajectories starting from two neighboring points, through their continuous evolution, and through their discrete jumps since the autonomous jump set is nonempty. This involves careful estimation of the distance between the trajectories in various time intervals and summing up these terms to show that the distance remains small for initial points sufficiently close enough. Although the basic estimates used are similar to those available in the literature (e.g., [3], [4]), the crucial point in our proof is the convergence of the above summation. This also allows us to get the precise Hölder exponent for the continuity of the value function.

As in [5] and [4], using the dynamic programming principle, we arrive at the QVI satisfied by the value function. Then we show that the value function is the unique viscosity solution of the QVI. Our proof is very different from [4]. Their approach using a fixed point method does not seem to be suitable, as it is for the general case of a nonempty autonomous jump set. Our approach is based on the comparison principle in the class of bounded continuous functions. It is inspired by earlier work on impulse and switching control and game theoretic problems in the literature, namely, [2], [7], [9], particularly the idea of defining a sequence of new auxiliary functions. But the presence of the autonomous and controlled jump sets leads to different equations on these sets, and hence some new ideas are needed to arrive at the conclusion.

2. Notation and assumptions. In a hybrid control system, as in [5], the state vector during continuous evolution is given by the solution of the following problem:

$$\dot{X}(t) = f(X(t), u(t)), \tag{2.1}$$

$$X(0) = x, \tag{2.2}$$

where $X(t) \in \Omega := \bigcup_i \Omega_i \times \{i\}$, with each Ω_i a closed connected subset of $\mathbb{R}^{d_i}$, i, $d_i \in Z_+$; $x \in \Omega$; and $f : \Omega \times \mathcal{U} \to \Omega$. Actually, $f = f_i$ with the understanding that $\dot{X}(t) = f_i(X(t), u(t))$ whenever $x \in \Omega_i$. $\mathcal{U}$ is the continuous control set

$$\mathcal{U} = \{u : [\,0, \infty) \to U \mid u \ \text{ measurable, } U \text{ compact metric space}\}\,.$$

The trajectory also undergoes discrete jumps when it hits predefined sets A, the autonomous jump set, and C, the controlled jump set. A predefined set D is the destination set for both autonomous jumps as well as controlled jumps:

$$\begin{aligned} A &= \bigcup_i A_i \times \{i\}, \quad A_i \subseteq \Omega_i \subseteq \mathbb{R}^{d_i}, \\ C &= \bigcup_i C_i \times \{i\}, \quad C_i \subseteq \Omega_i \subseteq \mathbb{R}^{d_i}, \\ D &= \bigcup_i D_i \times \{i\}, \quad D_i \subseteq \Omega_i \subseteq \mathbb{R}^{d_i}\ . \end{aligned}$$

The trajectory starting from $x \in \Omega_i$, on hitting A, that is the respective $A_i \subseteq \Omega_i$, jumps to the destination set D according to the given transition map g. g uses discrete controls from the discrete control set V_1 and can move the trajectory from A_i to $D_j \subseteq \Omega_j \subseteq \mathbb{R}^{d_j}$. The trajectory then will continue its evolution under f_j till it again hits A or C, in particular A_j or C_j. On hitting C the controller can choose either to jump or not to jump. If the controller chooses to jump, then the trajectory is moved to a new point in D. In this case the controller can also move from Ω_i to any of the Ω_j.

This gives rise to a sequence of hitting times of A, which we denote by σ_i, and a sequence of hitting times of C, where the controller chooses to make a jump which is denoted by ξ_i. Thus σ_i and ξ_i are the times when continuous and discrete dynamics interact. Hence the trajectory of this problem is composed of continuous evolution given by (2.1) between two hitting times and discrete jumps at the hitting times. We denote $(X(\sigma_i^-), u(\cdot))$ by x_i and $g(X(\sigma_i^-), v)$ by x_i' and the destination of $X(\xi_i^+, u(\cdot))$ by $X(\xi_i)'$. In general we take the trajectory to be left continuous so that $X_x(\sigma_i)$ means $X_x(\sigma_i^-)$ and $X_x(\xi_i)$ means $X_x(\xi_i^-)$, whereas $X_x(\sigma_i^+)$ will be denoted by x_i' and $X_x(\xi_i^+)$ will be denoted by $X_x(\xi_i)'$.

We give the inductive limit topology on Ω, namely,

$$(x_n, i_n) \in \Omega \text{ converges to } (x, i) \in \Omega \quad \text{if for some } N \text{ large and } \forall n \geq N,$$

$$i_n = i, \quad x, x_n \in \Omega_i, \quad \Omega_i \subseteq \mathbb{R}^{d_i} \quad \text{for some } i, \quad \text{and } \|x_n - x\|_{\mathbb{R}^{d_i}} < \varepsilon.$$

With the understanding of the above topology we suppress the second variable i from Ω. We follow the same for A, C, and D. We make the following basic assumptions on the sets A, C, D, and on functions f and g.

(A1): Each Ω_i is the closure of a connected, open subset of $\mathbb{R}^{d_i}$.

(A2): A_i, C_i, D_i are closed, ∂A_i,∂C_i are C^2. For all i and for all $x \in D_i$, $|x| < R$, and $\partial A_i \supseteq \partial \Omega_i$ for all i.

(A3): $g : A \times V_1 \to D$ is a bounded, uniformly Lipschitz continuous map, with Lipschitz constant G with the understanding that $g = \{g_i\}$ and $g_i : A_i \times V \to D_j$.

(A4): Vector field f is Lipschitz continuous with Lipschitz constant L in the state variable x and uniformly continuous in control variable u. Also,

$$|f(x, u)| \leq F \quad \forall x \in \Omega \quad \text{and} \quad \forall u \in U. \tag{2.3}$$

(A5): We assume ∂A_i is compact for all i, and for some $\xi_0 > 0$, following transversality condition holds

$$f(x_0, u) \cdot \eta(x_0) \leq -2\xi_0 \ \ \forall x_0 \ \in \ \partial A_i \ \ \forall u \in U, \tag{2.4}$$

where $\eta(x_0)$ is the unit outward normal to ∂A_i at x_0. We assume a similar transversality condition on ∂C_i.

(A6):

$$\inf_i d(A_i, C_i) \geq \beta \quad \text{and} \quad \inf_i \ d(A_i, D_i) \geq \beta > 0, \tag{2.5}$$

where d is the appropriate Euclidean distance. Note that the above rules out infinitely many jumps in finite time.

(A7): We assume the control sets U and V_1 to be compact metric spaces.

Now $(u(\cdot), v, \xi_i, X(\xi_i)')$ is the control, and the total discounted cost is given by

$$J(x, u(\cdot), v, \xi_i, X(\xi_i)') = \int_0^\infty K(X_x(t), u(t))e^{-\lambda t}dt + \sum_{i=0}^{\infty} C_a(X(\sigma_i), v)e^{-\lambda\sigma_i} + \sum C_c(X(\xi_i), X(\xi_i)')e^{-\lambda\xi_i}, \tag{2.6}$$

where λ is the discount factor, $K : \Omega \times \mathcal{U} \to \mathbb{R}_+$ is the running cost, $C_a : A \times V_1 \to \mathbb{R}_+$ is the autonomous jump cost, and $C_c : C \times D \to \mathbb{R}_+$ is the controlled jump cost. The value function V is then defined as

$$V(x) = \inf_{\theta \in (\mathcal{U} \times V_1 \times [0,\infty) \times D)} J(x, u(\cdot), v, \xi_i, X(\xi_i)'). \tag{2.7}$$

We assume the following conditions on the cost functionals.

(C1): K is Lipschitz continuous in the x variable with Lipschitz constant K_1 and is uniformly continuous in the u variable. Moreover, K is bounded by K_0.

(C2): C_a and C_c are uniformly continuous in both variables and bounded below by $C' > 0$. Moreover, C_a is Lipschitz continuous in the x variable with Lipschitz constant C_1 and is bounded above by C_0. Also we assume

$$C_c(x, y) < C_c(x, z) + C_c(z, y) \quad \forall x \in C_i, z \in D \cap C_j, y \in D.$$

We now give two simple examples of hybrid control systems. For more examples, see [5].

Example 2.1 (collisions). Consider the ball of mass m which is moving in vertical and horizontal directions in a room under gravity with gravitational constant g. The dynamics can be given as

$$\begin{aligned} \dot{x} &= v_x, \quad \dot{v_x} = 0, \\ \dot{y} &= v_y, \quad \dot{v_y} = -mg. \end{aligned}$$

On hitting the boundaries of the room $A_1 = \{(x, y)|y = 0, \text{ or } y = R_1\}$ we instantly set v_y to $-\rho v_y$ for some $\rho \in [0, 1]$, the coefficient of restitution. Similarly we reset v_x to $-\rho v_x$ on hitting the boundary $A_2\{(x, y)|x = 0 \text{ or } x = R_2\}$. Thus in this case the sets A_1 and A_2 are autonomous jump sets. We can generalize the above system by allowing dynamics to occur in different $\mathbb{R}^d$ after hitting.

The next example illustrates the importance of the transversality condition, in the absence of which the optimal trajectory and hence the optimal control may fail to exist.

Example 2.2. Consider the dynamical system in $\mathbb{R}^2$ given by

$$\begin{aligned} \dot{x_1}(t) &= 1, \quad x_1(0) = 0, \\ \dot{x_2}(t) &= u, \quad x_2(0) = 0, \end{aligned}$$

where $u \in [0, 1]$, and when the trajectory hits the set A given by $A = \{(x_1, x_2)|(x_1 - 1)^2 + (x_2 + 1)^2 = 1\}$ it jumps to $(10^{10}, 10^{10})$. The cost is given by $\int_0^\infty e^{-t} \min\{|x_1(t) + x_2(t)|, 210^{10}\}$.

Here the vector field $(u, 1)$ is not transversal to the boundary at $(1, 0)$ for $u = 0$. Hence optimal trajectory does not exist and, moreover, the value function is discontinuous at $(1, 0)$.

In the following sections we are interested in exploring the value function of the hybrid control problem defined in (2.7). In section 2 we show that the value function is bounded and locally Hölder continuous with respect to the initial point. In section 3, we use viscosity solution techniques and the dynamic programming principle to derive a partial differential equation satisfied by V in the viscosity sense, which turns out to be the Hamilton–Jacobi–Bellman QVI. Section 4 deals with uniqueness of the solution of the QVI. We give a comparison principle proof characterizing the value function as unique viscosity solution of the QVI.

3. Continuity of the value function. Let the trajectory given by the solution of (2.1) and starting from the point x be denoted by $X_x(t, u(\cdot))$. Since $x \in \Omega$, it belongs in particular to some Ω_i. Then we have from the theory of ordinary differential equations

$$|X_x(t, u(\cdot)) - X_z(t, u(\cdot))| \le e^{Lt}|x - z|, \tag{3.1}$$

$$|X_x(t, u(\cdot)) - X_x(\bar{t}, u(\cdot))| \le F|t - \bar{t}|, \tag{3.2}$$

where F and L are as in (A4).

Define the first hitting time of the trajectory as

$$T(x) = \inf_u \{t > 0 \mid X_x(t, u) \in A\}.$$

Notice that this $T(x)$ is in particular with respect to A_i as $x \in \Omega_i$. By assuming a suitable transversality condition on ∂A_i and ∂C_i we prove the continuity of T in the topology of $\mathbb{R}^{d_i}$. This is equivalent to proving the continuity of T on Ω with respect to the inductive limit topology on Ω. Hereafter by convention we assume the topology to be of that Ω_i, in which the respective points belong.

THEOREM 3.1. *Assume* (A1)–(A7). *Let $X(t)$ be the trajectory given by the solution of* (2.1)*. Let the first hitting time $T(x)$ be finite. Then it is locally Lipschitz continuous, i.e., there exists a $\delta_1 > 0$ depending on f, ξ_0, and the distance function from ∂A_i such that for all y, $\bar{y}$ in $B(x, \delta_1)$, a δ_1 neighborhood of x in Ω*

$$|T(y) - T(\bar{y})| < C|y - \bar{y}|, \quad \textit{where } C \textit{ depends on } \xi_0.$$

Proof. Step 1. *Estimates for points near ∂A.* First we show that there exist $\delta > 0$ and $C > 0$ such that

$$T(x) < C\, d(x) \quad \forall x \in B(A_i, \delta) \setminus \overset{\circ}{A},$$

where $B(A_i, \delta)$ is a δ neighborhood of A_i and $d(x)$ is a signed distance of x from ∂A_i given by

$$d(x) = \begin{cases} -\operatorname{dist}(x, \partial A_i) & \text{if } x \in \overset{\circ}{A}_i, \\ 0 & \text{if } x \in \partial A_i, \\ \operatorname{dist}(x, \partial A_i) & \text{if } x \in \bar{A}_i^c. \end{cases}$$

For simplicity of notation we drop the suffix i from now on, remembering that the distances are in $\mathbb{R}^{d_i}$. It is possible to choose $R > 0$ such that in a small neighborhood of ∂A, say $B(\partial A, R)$, the above signed distance function d is C^1, thanks to our assumption (A2).

Now for $x_0 \in \partial A$ choose u_0 in $\mathcal{U}$ such that $u_0(t) = u_0$ for all t and $r_0 < R$ such that

$$f(x, u_0) \cdot Dd(x) < -\xi_0 \quad \forall x \in B(x_0, r_0). \tag{3.3}$$

Observe that we can choose r_0 independent of x_0 by using compactness of ∂A. Now consider the trajectory starting from x, given by

$$\begin{aligned} \dot{X}(t) &= f(X(t), u_0), \\ X(0) &= x, \end{aligned}$$

where $x \in B(x_0, r_0)$. Then

$$\begin{aligned} d(X(s)) - d(x) = &\int_0^s Dd(x) \cdot f(x, u_0)\ d\tau + \int_0^s \big(Dd(X(\tau)) - Dd(x)\big) \cdot f(X(\tau), u_0)\ d\tau \\ &+ \int_0^s Dd(x) \cdot (f(X(\tau), u_0) - f(x, u_0))\ d\tau. \end{aligned}$$

By using (3.3) and (2.3),

$$\begin{aligned} d(X(s)) - d(x) \le &\int_0^s -\xi_0\ d\tau + F \int_0^s (Dd(X(\tau)) - Dd(x))\ d\tau \\ &+ \int_0^s Dd(x) \cdot (f(X(\tau), u_0) - f(x, u_0))\ d\tau. \end{aligned}$$

Let c be the bound on Dd on $B(\partial A, r_0)$. Restricting s to be small so that $X(\tau)$ is in the r_0 neighborhood of ∂A, we are assured that Dd is continuous. So is f. Thus

$$\begin{aligned} d(X(s)) - d(x) &\le -\xi_0 s + o(Fs)\ + o(cLs) \\ &< -\frac{1}{2}\xi_0 s \quad \text{for } \ 0 < s < \bar{s} \end{aligned}$$

for some $\bar{s}$ dependent only on modulus of continuity of f and Dd and independent of x. Choose $\delta = \min\{r_0, \frac{\bar{s}\xi_0}{2}\}$. If x is in the δ ball around x_0, then $d(x) < \frac{\bar{s}\xi_0}{2}$ and, choosing $s_x = 2\frac{d(x)}{\xi_0}$, will imply

$$s_x < \bar{s} \quad \text{and hence} \quad d\,(X(s_x)) < 0.$$

Thus by our definition of d, $X(s_x) \in \overset{\circ}{A}$, which implies

$$T(x) < s_x = 2\frac{d(x)}{\xi_0}.$$

Then for $C = \frac{2}{\xi_0}$ we have

$$T(x) < Cd(x) \quad \forall x \in B(x_0, \delta) \setminus \overset{\circ}{A}\,.$$

Step 2. *Estimate for any two points in* Ω. In this step we estimate $|T(x) - T(\bar{x})|$ for any $x, \bar{x} \in \Omega$. Define

$$t(\bar{x}, \bar{u}) = \inf\{t > 0 \mid X(t) \in A, \dot{X}(t) = f(X(t), \bar{u}), X(0) = \bar{x}\}.$$

For given $0 < \epsilon < 1$, and $\bar{x} \in \Omega$ by the definition of $T(\bar{x})$, we can choose $\bar{u} \in \mathcal{U}$ such that

$$\bar{t} = t(\bar{x}, \bar{u}) < T(\bar{x}) + \epsilon. \tag{3.4}$$

Using estimate (3.1),

$$|X_{\bar{x}}(\bar{t}, \bar{u}) - X_x(\bar{t}, \bar{u})| \leq |\bar{x} - x| e^{L\bar{t}} \leq |\bar{x} - x| e^{L(T(\bar{x})+\varepsilon)}. \tag{3.5}$$

Define $\delta_1 = \delta e^{-L(T(\bar{x})+1)}$, where δ is as in Step 1. Let us choose x such that $|x - \bar{x}| < \delta_1$. Then

$$|X_{\bar{x}}(\bar{t}, \bar{u}) - X_x(\bar{t}, \bar{u})| \leq |\bar{x} - x| e^{L\bar{t}} < |\bar{x} - x| e^{L(T(\bar{x})+1)} < \delta.$$

Also we have $X_{\bar{x}}(\bar{t}, \bar{u}) \in \partial A$. Hence, $X_x(\bar{t}, \bar{u}) \in B(\partial A, \delta) \setminus \overset{\circ}{A}$. Therefore, by Step 1,

$$T(X_x(\bar{t}, \bar{u})) < C d(X_x(\bar{t}, \bar{u})). \tag{3.6}$$

We claim that

$$T(x) \leq \bar{t} \ + T(X_x(\bar{t}, \bar{u})). \tag{3.7}$$

For given $\varepsilon_1 > 0$, choose $u_1 \in \mathcal{U}$ such that

$$T(X_x(\bar{t}, \bar{u})) \geq t(X_x(\bar{t}, \bar{u}), u_1) - \varepsilon_1.$$

Define a new control u_2 by

$$u_2(s) = \begin{cases} \bar{u}(s) & \text{if } s \leq \bar{t}, \\ u_1(s - \bar{t}) & \text{if } s > \bar{t}. \end{cases}$$

Then

$$T(x) \leq t(x, u_2) \leq \bar{t} + t\left(X_x(\bar{t}, \bar{u}), u_1\right) \leq \bar{t} + T\left(X_x(\bar{t}, \bar{u})\right) + \varepsilon_1.$$

Since ε_1 is arbitrary, this proves (3.7). Using (3.4) and (3.7) for $x \in B(\bar{x}, \delta_1)$ we get

$$\begin{aligned} T(x) &\leq T(\bar{x}) + T(X_x(\bar{t}, \bar{u})) + \epsilon \\ &\leq T(\bar{x}) + C\ d(X_x(\bar{t}, \bar{u})) + \epsilon \quad \text{by (3.6)}. \end{aligned}$$

Notice that $d(X_x(\bar{t}, \bar{u})) \leq |X_x(\bar{t}, \bar{u}) - X_{\bar{x}}(\bar{t}, \bar{u})|$. So by (3.5)

$$T(x) \leq T(\bar{x}) + C\ |x - \bar{x}|\ e^{L(T(\bar{x})\,+\,\varepsilon)} + \epsilon.$$

Interchanging the roles of x and $\bar{x}$ we get

$$|T(x) - T(\bar{x})| \leq C\ |x - \bar{x}|\ e^{L(T(\bar{x}) \vee T(x))} \tag{3.8}$$

as ϵ tends to 0, where $T(\bar{x}) \vee T(x) = \max\{T(\bar{x}),\ T(x)\}$. Also observe that

$$\begin{aligned} T(x) &\leq T(\bar{x}) + C\ |x - \bar{x}|\ e^{L(T(\bar{x})+\varepsilon)} + \epsilon \\ &\leq T(\bar{x}) + C\delta + \varepsilon \leq T(\bar{x}) + C\delta + 1 \\ &\leq T(\bar{x}) + 2. \end{aligned}$$

Hence for all x belonging to $B(\bar{x}, \delta_1)$, T is bounded. Let this bound be T_0. Then we have

$$|T(x) - T(\bar{x})| < C|x - \bar{x}|e^{LT_0}.$$

Hence we conclude that the first hitting time of trajectory is locally Lipschitz continuous with respect to the initial point. □

Now we take up the issue of continuity of the value function. For this proof we need some estimates on hitting times of trajectories starting from two nearby points. We prove these estimates in the following lemmas. We fix the controls $\bar{u}$ and $\bar{v}$ and suppress them in the following calculations.

LEMMA 3.2. *Let σ_1 and Σ_1 be the first hitting times of trajectories evolving with fixed controls $\bar{u}$ and $\bar{v}$ according to* (2.1) *starting from x and z, respectively. Let x_1 and z_1 be points where these trajectories hit A for the first time:*

$$x_1 = X_x(\sigma_1), \quad z_1 = X_z(\Sigma_1), \quad x_1, z_1 \in \partial A.$$

If $|x - z| < \delta_1$, where δ_1 is as in Theorem 3.1*, then*

$$|x_1 - z_1| \leq (1 + FC)e^{L(\Sigma_1 \vee \sigma_1)}|x - z|. \tag{3.9}$$

Proof. Note here that by Theorem 3.1 we have the estimate on $|\sigma_1 - \Sigma_1|$ given by (3.8),

$$|\sigma_1 - \Sigma_1| < Ce^{L(\Sigma_1 \vee \sigma_1)}\ |x - z|. \tag{3.10}$$

Using this we estimate $|x_1 - z_1|$. Without loss of generality we assume that $\Sigma_1 > \sigma_1$,

$$\begin{aligned} |x_1 - z_1| &= |X_x(\sigma_1) - X_z(\Sigma_1)| \\ &\leq |X_x(\sigma_1) - X_z(\sigma_1)| + |X_z(\sigma_1) - X_z(\Sigma_1)|. \end{aligned}$$

Using (3.1) we get

$$|X_x(\sigma_1) - X_z(\sigma_1)| < e^{L\sigma_1}|x - z|,$$

while (3.2) and (3.10) lead to

$$|X_z(\sigma_1) - X_z(\Sigma_1)| \leq F|\sigma_1 - \Sigma_1| \leq FCe^{L\Sigma_1}|x - z|.$$

Combining these estimates, we get

$$|x_1 - z_1| \leq e^{L\Sigma_1}|x - z|(1 + FC) \quad \text{for} \quad z \in B(x, \delta_1). \qquad \square$$

Observe that the destination points of x_1 and z_1, which are denoted by ${x_1}' = g(x_1, \bar{v})$ and ${z_1}' = g(z_1, \bar{v})$, may belong to $\Omega_j \subseteq \mathbb{R}^{d_j}$. Without loss of generality we assume that ${x_1}', {z_1}' \in \Omega_2 \subseteq \mathbb{R}^{d_2}$, and the evolution of trajectories takes place in Ω_2 till the next hitting time. Let σ_2 and Σ_2 be the next hitting times of the trajectories when they hit A once again. The next lemma deals with the estimate of $|\sigma_2 - \Sigma_2|$.

LEMMA 3.3. *Let the first hitting time of trajectories starting from x and z, and evolving with fixed control $\bar{u}$, be σ_1 and Σ_1, and the second hitting times are σ_2 and Σ_2. Then there exists δ_2 such that for $|x - z| < \delta_2$,*

$$|\sigma_2 - \Sigma_2| \leq Ce^{(\Sigma_2 \vee \sigma_2)}(FC + G(FC + 1))|x - z| \tag{3.11}$$

and if we denote

$$x_2 = X_{x_1'}(\sigma_2 - \sigma_1), \quad x_2' = g(x_2),$$
$$z_2 = X_{z_1'}(\Sigma_2 - \Sigma_1), \quad z_2' = g(z_2),$$

then

$$|x_2 - z_2| \le (FC+1)e^{L(\Sigma_2 \vee \sigma_2)}(FC + G(FC+1))|x - z|. \tag{3.12}$$

Proof. Without loss of generality let $\sigma_1 < \Sigma_1$. Observe that σ_2 and Σ_2 are the first hitting times of trajectories starting from points $X_{x_1'}(\Sigma_1 - \sigma_1)$ and z_1' at time $t = \Sigma_1$. Then

$$T(z_1') = (\Sigma_2 - \Sigma_1) \quad \text{and} \quad T(X_{x_1'}(\Sigma_1 - \sigma_1)) = \sigma_2 - \Sigma_1.$$

Hence by (3.8)

$$|\sigma_2 - \Sigma_2| \le Ce^{L(\Sigma_2 - \Sigma_1)}|X_{x_1'}(\Sigma_1 - \sigma_1) - z_1'|$$

whenever $|X_{x_1'}(\Sigma_1 - \sigma_1) - z_1'| \le \delta_1$. Now

$$|X_{x_1'}(\Sigma_1 - \sigma_1) - z_1'| \le |X_{x_1{}'}(\Sigma_1 - \sigma_1) - x_1'| + |x_1' - z_1'|.$$

Hence by using estimate (3.2) and (3.10) for the first term we have

$$|X_{x_1'}(\Sigma_1 - \sigma_1) - x_1'| \le F|\Sigma_1 - \sigma_1| \le FCe^{L\Sigma_1}|x - z|,$$

whereas using Lipschitz continuity of g and (3.9) for the second term we get

$$|x_1' - z_1'| \le G|x_1 - z_1| \le G(FC+1)e^{L\Sigma_1}|x - z| \quad \text{for} \quad z \in B(x, \delta_1).$$

Combining the above two estimates we have

$$|X_{x_1'}(\Sigma_1 - \sigma_1) - z_1'| \le e^{L\Sigma_1}(FC + G(FC+1))|x - z| \tag{3.13}$$

and by our choice of $\delta_2 = \min\{\delta_1, \frac{\delta_1 e^{-L\Sigma_1}}{FC+G(FC+1)}\}$, $|X_{x_1'}(\Sigma_1 - \sigma_1) - z_1'| < \delta_1$. Using (3.13) in the estimate of $|\sigma_2 - \Sigma_2|$ for $z \in B(x, \delta_2)$ we have

$$|\sigma_2 - \Sigma_2| \le Ce^{L\Sigma_2}(FC + G(FC+1))|x - z|. \tag{3.14}$$

Now we estimate $|x_2 - z_2|$:

$$\begin{aligned} |x_2 - z_2| &= |X_{x_1'}(\sigma_2 - \sigma_1) - X_{z_1'}(\Sigma_2 - \Sigma_1)| \\ &\le |X_{x_1'}(\sigma_2 - \sigma_1) - X_{z_1'}(\sigma_2 - \Sigma_1)| + |X_{z_1'}(\sigma_2 - \Sigma_1) - X_{z_1'}(\Sigma_2 - \Sigma_1)|. \end{aligned}$$

Observe that by the semigroup property

$$X_{x_1'}(\sigma_2 - \sigma_1) = X_{X_{x_1'}(\Sigma_1 - \sigma_1)}(\sigma_2 - \Sigma_1).$$

Hence

$$|X_{x_1'}(\sigma_2 - \sigma_1) - X_{z_1'}(\sigma_2 - \Sigma_1)| = |X_{X_{x_1'}(\Sigma_1 - \sigma_1)}(\sigma_2 - \Sigma_1) - X_{z_1'}(\sigma_2 - \Sigma_1)|$$

and by (3.1)

$$|X_{x_1'}(\sigma_2-\sigma_1)-X_{z_1'}(\sigma_2-\Sigma_1)| \le e^{L(\sigma_2-\Sigma_1)}|X_{x_1'}(\Sigma_1-\sigma_1)-z_1'|. \tag{3.15}$$

From (3.2) and (3.14) we get

$$\begin{aligned}|X_{z_1'}(\sigma_2-\Sigma_1)-X_{z_1'}(\Sigma_2-\Sigma_1)| &\le F|\sigma_2-\Sigma_1-(\Sigma_2-\Sigma_1)|\\ &\le FCe^{L\Sigma_2}(FC+G(FC+1))|x-z|.\end{aligned} \tag{3.16}$$

Together these estimates yield, for $z \in B(x,\delta_2)$,

$$|x_2-z_2| \le e^{L\Sigma_2}(FC+1)(FC+G(FC+1))|x-z|. \qquad \square$$

Let σ_i and Σ_i be the ith hitting times of trajectories starting from x and z, respectively. With the above notation we assume that $x_i', z_i' \in \Omega_{i+1} \subseteq \mathbb{R}^{d_{i+1}}$. We apply Theorem 3.1 and the above lemmas recursively to find estimates on successive hitting times and points where trajectories hit A. We generalize the above estimates for the ith hitting times of trajectories when they hit A. For simplicity of calculations we denote $FC+G(FC+1)$ by P hereafter.

REMARK 3.4. *Let the control $\bar{u}$ be fixed. Let σ_i and Σ_i be the ith consecutive hitting time of the trajectory starting from x and z, respectively, when they hit A, and let x_i, z_i be the points on ∂A where trajectories hit A. Then proceeding along lines similar to those of Lemmas* 3.2 *and* 3.3 *we get the estimates for $|\sigma_i-\Sigma_i|$ and $|x_i-z_i|$ which are given by*

$$\begin{aligned}|\sigma_i-\Sigma_i| &\le Ce^{L\Sigma_i}P^{i-1}|x-z|,\\ |x_i-z_i| &\le e^{L\Sigma_i}(FC+1)P^{i-1}|x-z|\end{aligned}$$

whenever $|x-z|<\delta_i$, where $\delta_i := \min\{\delta_1,\delta_2,\ldots,\frac{\delta_1 e^{-L\Sigma_i}}{P^{i-1}}\}$.

THEOREM 3.5 (continuity of the value function). *Under the assumptions of Theorem* 3.1*, value function V of hybrid control problem defined by* (2.7) *is bounded and locally Hölder continuous with respect to the initial point.*

Proof. First we show that the value function is bounded. For any $u \in \mathcal{U}$ and $v \in V_1$,

$$V(x) \le \int_0^\infty K(X_x(t),u(t))e^{-\lambda t}dt + \sum_{i=0}^{\infty} C_a(X(\sigma_i),v)e^{-\lambda\sigma_i}.$$

By our assumptions (C1) and (C2),

$$V(x) \le K_0\int_0^{+\infty} e^{-\lambda t}\,dt \;+\sum_{i=1}^{+\infty} C_0e^{\lambda\sigma_i} \le \frac{K_0}{\lambda} \;+C_0\sum_{i=1}^{+\infty} e^{-\lambda\sigma_i}.$$

From (A5), recalling that $\beta = \inf\ d(A_i,D_i)$,

$$\sigma_{i+1} \;\ge\; \sigma_i + \frac{\beta}{\sup|f(x,u)|} \;\ge\; \sigma_i+\beta/F. \tag{3.17}$$

Hence we get

$$\sum_{i=1}^{\infty} e^{-\lambda\sigma_i} \le e^{-\lambda\sigma_1}\sum_{i=1}^{\infty}\left(e^{-\lambda\beta/F}\right)^i \le e^{-\lambda\sigma_1}\frac{1}{1-e^{-\lambda\beta/F}}, \tag{3.18}$$

leading to

$$V(x) \le \frac{K}{\lambda} + C_0 e^{-\lambda\sigma_1} \frac{1}{1 - e^{-\lambda\beta/F}}.$$

This proves $V(x)$ is bounded.

We now show that V defined in (2.7) is locally Hölder continuous with respect to the initial point. Let $x, z \in \Omega$. Regarding $V(x)$ as in (2.7), we assume that the controller chooses not to make any controlled jumps. Note that the controller has this choice because in the interior of C he can always choose not to jump. On the boundary of C that is ∂C by the transversality condition, vector field is nonzero and hence he can continue the evolution without jumping. Thus in any case he can choose not to jump. Then given $\varepsilon > 0$, we can choose the controls $\overline{u}, \overline{v}$ depending on ε such that

$$V(z) \ge \int_0^\infty K(X_z(t), \overline{u}(t)) e^{-\lambda t} dt + \sum_{i=1}^\infty C_a(X_z(\Sigma_i), \overline{v}) e^{-\lambda\Sigma_i} - \varepsilon.$$

Also

$$V(x) \le \int_0^\infty K(X_x(t), \bar{u}(t)) e^{-\lambda t} dt + \sum_{i=1}^\infty C_a(X_x(\sigma_i), \bar{v}) e^{-\lambda\sigma_i}.$$

Hence

$$\begin{aligned} V(x) - V(z) \le & \int_0^\infty |K(X_x(t), \overline{u}(t)) - K(X_z(t), \overline{u}(t))| e^{-\lambda t} dt \\ & + \sum_{i=1}^\infty |C_a(X_x(\sigma_i), \overline{v}) - C_a\big(X_z(\Sigma_i), \overline{v}\big)| e^{-\lambda(\sigma_i \vee \Sigma_i)} + \varepsilon, \end{aligned}$$

where $\sigma_i \vee \Sigma_i = \max\{\sigma_i, \Sigma_i\}$. Now for T large to be chosen precisely later on we split the integral and summation as follows:

$$\begin{aligned} V(x) - V(z) \le & \int_0^T |K(X_x(t), \overline{u}(t)) - K(X_z(t), \overline{u}(t))| e^{-\lambda t} dt \\ & + \sum_{i=1}^N |C_a(X_x(\sigma_i), \overline{v}) - C_a(X_z(\Sigma_i), \overline{v})| e^{-\lambda(\sigma_i \vee \Sigma_i)} \\ & + \int_T^\infty |K(X_x(t), \overline{u}(t)) - K(X_z(t), \overline{u}(t))| e^{-\lambda t} dt \\ & + \sum_{i=N+1}^\infty |C_a(X_x(\sigma_i), \overline{v}) - C_a(X_z(\Sigma_i), \overline{v})| e^{-\lambda(\sigma_i \vee \Sigma_i)} + \varepsilon, \end{aligned} \tag{3.19}$$

where T will be chosen so that the tail end of the integral and summation become small and T is in between the Nth and $(N+1)$th hitting times of the trajectories. By using the bound K_0 on K given by (C1) we get

$$\int_T^\infty |K(X_x(t), \overline{u}(t)) - K(X_z(t), \overline{u}(t))| e^{-\lambda t} dt \le \frac{2K_0}{\lambda} e^{-\lambda T} \tag{3.20}$$

and by using bound C_0 on C_a given by (C2) and doing calculations along lines similar to those of (3.18) we get the estimate

(3.21)
$$\sum_{i=N+1}^{\infty} |C_a(X_x(\sigma_i),\overline{v}) - C_a(X_z(\Sigma_i),\overline{v})|e^{-\lambda(\sigma_i\vee\Sigma_i)} \le 2C_0\big(e^{-\lambda\beta/F}\big)^N \frac{1}{1-e^{-\lambda\beta/F}}.$$

Now we calculate $\int_0^T |K(X_x(t),\overline{u}(t)) - K(X_z(t),\overline{u}(t))|e^{-\lambda t}dt$. We will show that there exists $\bar{\delta} > 0$ such that if $|x-z| < \bar{\delta}$, then the sequence of σ_i and Σ_i can be, for example,

$$0 \le \sigma_1 \le \Sigma_1 \le \sigma_2 \le \Sigma_2 \le \cdots \le \sigma_n \le \Sigma_n \le T \tag{3.22}$$

$$\text{or} \quad 0 \le \Sigma_1 \le \sigma_1 \le \cdots \le \Sigma_n \le \sigma_n \le T.$$

That is, every A hitting time of trajectory starting from x is followed by A hitting time of trajectory starting from z.

Without loss of generality let us assume $\sigma_1 < \Sigma_1$. If $\Sigma_1 < \sigma_1$, the following calculations go through with appropriate changes and hence we split this integral, assuming (3.22) as follows:

$$\begin{aligned}\int_0^T Ie^{-\lambda t}dt \le{}& \int_0^{\sigma_1} Ie^{-\lambda t}dt + \int_{\sigma_1}^{\Sigma_1} Ie^{-\lambda t}dt + \int_{\Sigma_1}^{\sigma_2} Ie^{-\lambda t}dt + \cdots \\ &+ \int_{\sigma_n}^{\Sigma_n} Ie^{-\lambda t}dt + \int_{\Sigma_n}^{\sigma_{n+1}} Ie^{-\lambda t}dt,\end{aligned} \tag{3.23}$$

where $I = |K(X_x(t),\overline{u}(t)) - K(X_z(t),\overline{u}(t))|$. In this there are two types of integrals:

1. $\int_{\sigma_i}^{\Sigma_i} Ie^{-\lambda t}dt$;
2. $\int_{\Sigma_i}^{\sigma_{i+1}} Ie^{-\lambda t}dt$.

If $|x-z| < \delta_N$, where $\delta_N = \min\{\delta_1,\delta_2,\ldots,\frac{\delta_1 e^{-L\Sigma_N}}{P^{N-1}}\}$, we can estimate the above integrals using Lemmas 3.2 and 3.3 and Remark 3.4. We use the bound on K to evaluate the first integral.

$$\int_{\sigma_i}^{\Sigma_i} Ie^{-\lambda t}dt \le \frac{2K_0}{\lambda}\big(e^{-\lambda\sigma_i} - e^{-\lambda\Sigma_i}\big) \le \frac{2K_0}{\lambda}\,\lambda|\sigma_i - \Sigma_i|.$$

Using Remark 3.4,

$$\int_{\sigma_i}^{\Sigma_i} Ie^{-\lambda t}dt \le 2K_0CP^{i-1}e^{L\Sigma_i}. \tag{3.24}$$

To evaluate the second integral we use the Lipschitz continuity of K.

$$\begin{aligned}\int_{\Sigma_i}^{\sigma_{i+1}} Ie^{-\lambda t}dt &= \int_{\Sigma_i}^{\sigma_{i+1}} |K\big(X_{x_i'}(t-\sigma_i)\big) - K\big(X_{z_i'}(t-\Sigma_i)\big)|e^{-\lambda t}dt \\ &\le K_1\int_{\Sigma_i}^{\sigma_{i+1}} |X_{x_i'}(t-\sigma_i) - X_{z_i'}(t-\Sigma_i)|e^{-\lambda t}dt.\end{aligned} \tag{3.25}$$

By the semigroup property,

$$|X_{x_i'}(t-\sigma_i) - X_{z_i'}(t-\Sigma_i)| = |X_{X_{x_i'}(\Sigma_i-\sigma_i)}(t-\Sigma_i) - X_{z_i'}(t-\Sigma_i)|$$
$$\leq e^{L(t-\Sigma_i)}|X_{x_i'}(\Sigma_i-\sigma_i) - z_i'| \quad \text{by (3.1).}$$

Now by generalizing the estimate in (3.13) we get

$$|X_{x_i'}(\Sigma_i-\sigma_i) - z_i'| \leq P^i e^{L\Sigma_i}|x-z|. \tag{3.26}$$

Hence substituting the above estimates in (3.25), we get

$$\int_{\Sigma_i}^{\sigma_{i+1}} Ie^{-\lambda t}dt \leq K_1 e^{-L\Sigma_i}P^i e^{L\Sigma_i}|x-z| \int_{\Sigma_i}^{\sigma_{i+1}} e^{(L-\lambda)t}dt.$$

For $L \neq \lambda$,

$$\int_{\Sigma_i}^{\sigma_{i+1}} Ie^{-\lambda t}dt \leq K_1 P^i\ |x-z|\ \frac{e^{(L-\lambda)(\sigma_{i+1})} - e^{(L-\lambda)\Sigma_i}}{L-\lambda} \tag{3.27}$$
$$\leq K_1 P^i |x-z|\ \frac{e^{(L-\lambda)T}-1}{L-\lambda}$$

and for $L = \lambda$,

$$\int_{\Sigma_i}^{\sigma_{i+1}} Ie^{-\lambda t}dt \leq K_1 e^{-L\Sigma_i}P^i e^{L\Sigma_i}|x-z| \int_{\Sigma_i}^{\sigma_{i+1}} dt \tag{3.28}$$
$$\leq K_1 P^i |x-z|\ |\sigma_{i+1}-\Sigma_i|$$
$$\leq K_1 P^i\ |x-z|\ \ 2T.$$

For $L \neq \lambda$, by using (3.24), (3.27), $\int_0^T Ie^{-\lambda t}dt$ becomes

$$\int_0^T Ie^{-\lambda t}dt \leq \sum_{i=1}^N 2K_0 C P^{i-1}e^{LT}|x-z| + \sum_{i=1}^N \frac{K_1}{L-\lambda}P^i\big(e^{(L-\lambda)T}-1\big)|x-z|.$$

Hence

$$\left.\begin{aligned}\int_0^T Ie^{-\lambda t}dt \leq\ & 2K_0 C\left[\tfrac{P^N-1}{P-1}\right]|x-z| \\ & + K_1\left[\tfrac{P^N-1}{P-1}\right]\tfrac{e^{(L-\lambda)T}-1}{L-\lambda}|x-z|\end{aligned}\right\} \quad \text{for } L \neq \lambda \tag{3.29}$$

and for $L = \lambda$, using (3.24) and (3.28),

$$\int_0^T Ie^{-\lambda t}dt \leq \sum_{i=1}^N 2K_0|\sigma_i - \Sigma_i| + \sum_{i=1}^N K_1 T P^i |x-z|$$
$$\leq \sum_{i=1}^N 2K_0 C P^{i-1}|x-z| + \sum_{i=1}^N K_1 T P^i |x-z|.$$

Thus

$$\left.\begin{aligned}\int_0^T Ie^{-\lambda t}dt \leq\ & 2K_0 C\left(\tfrac{P^N-1}{P-1}\right)|x-z| \\ & + 2K_1 T\left(\tfrac{P^N-1}{P-1}\right)|x-z|\end{aligned}\right\} \quad \text{for } L = \lambda. \tag{3.30}$$

Furthermore, by using (C2) and Remark 3.4 we get

$$\sum_{i=1}^{N} |C_a(x_i,\overline{v}) - C_a(z_i,\overline{v})| e^{-\lambda(\sigma_i \vee \Sigma_i)} \leq \sum_{i=1}^{N} 2C_1 |x_i - z_i| e^{-\lambda(\sigma_i \vee \Sigma_i)} \leq 2C_1 \sum_{i=1}^{N} (FC+1) e^{LT} P^{i-1} |x-z|,$$

(3.31)
$$\sum_{i=1}^{N} |C_a(x_i,\overline{v}) - C_a(z_i,\overline{v})| e^{-\lambda(\sigma_i \vee \Sigma_i)} \leq 2C_1(FC+1) e^{LT} |x-z| \left(\frac{P^{N-1}-1}{P-1} \right).$$

Since P is a constant, without loss of generality we can assume

$$\frac{P^N}{P-1} < 2P^N. \tag{3.32}$$

Also observe that $\sigma_i - \sigma_{i+1} \geq \beta/F$ implies that $T \geq \sigma_{N+1} - \sigma_1 \geq N\beta/F$ and hence

$$N < TF/\beta. \tag{3.33}$$

Using (3.20), (3.21), (3.29), (3.31), (3.32), (3.33) in (3.19) for $L \neq \lambda$ we have

$$\begin{aligned} V(x) - V(z) \leq\ & 4K_0 C e^{LT} P^{TF/\beta}\ |x-z| + 2K_1 P^{TF/\beta}\ \frac{e^{(L-\lambda)T}-1}{L-\lambda} |x-z| \\ & + \frac{2K_0}{\lambda} e^{-\lambda T} + 2C_1 e^{LT} P^{TF/\beta}\ |x-z| \\ & + 2C_0 \left(e^{-\lambda\beta/F}\right)^{TF/\beta}\ \frac{1}{1-e^{-\lambda\beta/F}}. \end{aligned}$$

Now we further restrict $|x-z| < (\delta_1)^{\frac{1}{1-\theta}}$ for some θ such that $0 < \theta < 1$. Then choose T such that

$$P^{TF/\beta} e^{LT} = |x-z|^{-\theta}.$$

This gives

$$T = \frac{-\theta \log |x-z|}{\lambda + F \log P/\beta}. \tag{3.34}$$

This together with the choice of $|x-z|$ implies

$$\delta_N = \frac{\delta_1}{e^{L\Sigma_N} P^{N-1}} > \frac{\delta_1}{e^{LT} P^{TF/\beta}}\ =\ \delta_1 |x-z|^{\theta}\ >\ |x-z|. \tag{3.35}$$

Thus $|x-z| < \delta_N$ and hence the above estimate holds true for our choice of T. Then substituting the value of T in the above estimate, for $L \neq \lambda$, we get

$$\begin{aligned} V(x) - V(z) \leq\ & 4K_0 C |x-z|^{1-\theta} + \frac{K_1}{L-\lambda} |x-z|^{1-\theta} + C_1 |x-z|^{1-\theta} \\ & + \frac{2K_0}{\lambda} |x-z|^{\frac{\lambda\theta}{(F\log P/\beta)+L}}\ +\ 2C_0 |x-z|^{\frac{\lambda\theta}{(F\log P/\beta)+L}}. \end{aligned}$$

Here we have used the fact that $e^{(L-\lambda)T}-1<e^{LT}$. Thus we have proved that in the $\delta_1^{\frac{1}{1-\theta}}$ ball around x,

$$V(x)-V(z)<C_1|x-z|^{\theta_1} \quad \text{for some constant } C_1,$$

where

$$\theta_1=\min\left\{1-\theta,\ \frac{\lambda\ \theta}{(F\log P/\beta)+L}\right\} \quad \text{for } 0<\theta<1.$$

For $L=\lambda$, using (3.20), (3.21), (3.30), (3.31), (3.32), and (3.34) in (3.19), we have

$$\begin{aligned}V(x)-V(z)\le{}& 4K_0C|x-z|^{1-\theta}+2\frac{K_1}{(F\log P/\beta)+L}\log(|x-z|)|x-z|^{1-\theta}\\&+2C_1(FC+1)|x-z|^{1-\theta}+\frac{2K_0}{L}|x-z|^{\frac{L\theta}{(F\log P/\beta)+L}}\\&+2C_0|x-z|^{\frac{L\theta}{(F\log P/\beta)+L}}.\end{aligned}$$

Since $|x-z|^{1-\theta}$ goes to 0 faster than $\log(|x-z|)$ goes to $-\infty$ as $|x-z|\to 0$, all terms on the right-hand side (RHS) go to 0. The modulus of continuity of V is the same as that of $\log(r)r^{1-\theta}$. This suggests that in the $\delta_1^{\frac{1}{1-\theta}}$ ball around x,

$$V(x)-V(z)<C_1|x-z|^{\theta_1} \quad \text{for some constant } C_1$$

and for all θ_1 such that

$$\theta_1<\min\left\{1-\theta,\ \frac{L\theta}{(F\log P/\beta)+L}\right\} \quad \text{for } \ 0<\theta<1.$$

Thus in any case we have shown that (for θ_1 chosen depending on $L\neq\lambda$ or $L=\lambda$)

$$V(x)-V(z)\le C_1|x-z|^{\theta_1} \quad \text{for some constant } C_1.$$

Interchanging the roles of x and z we will get

$$V(z)-V(x)\le C_2|x-z|^{\theta_1} \quad \text{for some constant } C_2.$$

Together these will give

$$|V(x)-V(z)|\le C|x-z|^{\theta_1} \quad \text{for some constant } C.$$

This proves the Hölder continuity of V.

Now we want to justify our claim in (3.22), i.e., if $\sigma_1<\Sigma_1$, we can choose $|x-z|$ small enough such that (3.22) holds. If we restrict $|x-z|$ such that $|x-z|\le\min(\frac{\beta}{4FC}\ ,(\frac{\beta}{4CF})^{\frac{1}{1-\theta}})$, then by Remark 3.4,

$$|\Sigma_i-\sigma_i|\le Ce^{LT}(FC+G(FC+1))^{TF/\beta}|x-z|.$$

By our choice of T,

$$|\Sigma_i-\sigma_i|\le C|x-z|^{1-\theta}\le\frac{1}{4}\frac{\beta}{F}<\frac{1}{2}|\sigma_i-\sigma_{i+1}|$$

and this together with the assumption $\sigma_1<\Sigma_1$ implies $\sigma_i<\Sigma_i<\sigma_{i+1}$ for all i. So our claim is justified. □

4. Dynamic programming principle and the QVI. Under our assumptions (A1)–(A7), an optimal trajectory exists for any initial condition as shown in [5, Theorem 6.4]. The following dynamic programming principle and derivation of the QVI is also found in the literature [5], [4]. For the sake of completeness we prove it in detail here.

THEOREM 4.1 (dynamic programming principle). *Let V be the value function of the hybrid control problem as given in* (2.7). *If t_1 is the first hitting time of A, then*

$$V(x) = \inf_u \left\{ \int_0^{t_1} K(X(t), u(t)) e^{-\lambda t} dt + e^{-\lambda t} MV(X_x(t_1)) \right\}, \tag{DPPA}$$

where

$$M\phi(x) = \inf_{v \in \mathcal{V}} \{\phi(g(x,v)) + C_a(x,v)\}$$

and if t_1 is the first hitting time of C, then

$$V(x) = \inf_u \left\{ \int_0^{t_1} K(X(t), u(t)) e^{-\lambda t} dt + e^{-\lambda t} NV(X_x(t_1)) \right\}, \tag{DPPC}$$

where

$$N\phi(x) = \inf_{x' \in D} \{(\phi(x')) + C_c(x,x')\}.$$

For any $T > 0$,

$$\begin{aligned} V(x) = \inf_{u,v,\xi_i, X(\xi_i)'} \Bigg\{ & \int_0^T K(X_x(t), u(t)) e^{-\lambda t} dt + \sum_{\sigma_i < T} e^{-\lambda \sigma_i} C_a(X(\sigma_i), v) \\ & + \sum_{\xi_i < T} e^{-\lambda \xi_i} C_c\big(X(\xi_i), X(\xi_i)'\big) + e^{-\lambda T} V(X_x(T)) \Bigg\}. \end{aligned} \tag{DPP}$$

Proof. Let t_1 be the first hitting time of trajectory when it hits $A \cup C$. If t_1 is a first hitting time of A, we denote it by σ_1,

$$\begin{aligned} V(x) \leq & \int_0^{\sigma_1} K(X(t), u(t)) e^{-\lambda t} dt + C_a(X(\sigma_1), v) e^{-\lambda \sigma_1} \\ & + \Bigg[\int_{\sigma_1}^{\infty} K(X(t), u(t)) e^{-\lambda t} dt + \sum_{i=2}^{\infty} C_a(X(\sigma_i), v) e^{-\lambda \sigma_i} \\ & + \sum_{i=1}^{\infty} C_c\big(X(\xi_i), X(\xi_i)'\big) e^{-\lambda \xi_i} \Bigg]. \end{aligned}$$

We change the variable $t' = t - \sigma_1$ in the square bracket. Then taking the infimum in the square brackets over the control variables we get a value function of the trajectory starting from the point $g(X_x(\sigma_1), v)$. Hence,

$$\begin{aligned} V(x) \leq & \int_0^{\sigma_1} K(X(t), u(t)) e^{-\lambda t} dt + e^{-\lambda \sigma_1} C_a(X(\sigma_1), v) \\ & + e^{-\lambda \sigma_1} V(g(X_x(\sigma_1), v). \end{aligned}$$

Now taking the infimum over discrete controls v belonging to $\mathcal{V}$ in the last two terms we get

$$V(x) \le \int_0^{\sigma_1} K(X(t), u(t))e^{-\lambda t}dt + MV(X_x(\sigma_1)).$$

Further taking the infimum over continuous controls u in $\mathcal{U}$ we have the one-way inequality in (DPPA). For the reverse inequality, let $\varepsilon > 0$ be given. Choose the control $\theta_\varepsilon = (u_\varepsilon, v_\varepsilon, \xi_{i_\varepsilon}, X(\xi_i)'_\varepsilon)$ such that

$$\begin{aligned} V(x) + \varepsilon \ge & \int_0^{\sigma_1} K(X(t), u_\varepsilon(t))e^{-\lambda t}dt + C_a(X(\sigma_1), v_\varepsilon)e^{-\lambda\sigma_1} \\ & + e^{-\lambda\sigma_1}\left[\int_{\sigma_1}^{\infty} K(X(t), u_\varepsilon(t))e^{-\lambda t}dt + \sum_{i=2}^{\infty} C_a(X(\sigma_i), v_\varepsilon)e^{-\lambda\sigma_i}\right. \\ & \left. + \sum_{i=1}^{\infty} C_c\big(X(\xi_{i_\varepsilon}), X(\xi_i)'_\varepsilon\big)e^{-\lambda\xi_{i_\varepsilon}}\right] \end{aligned}$$

with calculations similar to those earlier, we can conclude that

$$\begin{aligned} V(x) + \varepsilon &\ge \int_0^{\sigma_1} K(X(t), u(t))e^{-\lambda t}dt + MV(X_x(\sigma_1)) \\ &\ge \inf_u \int_0^{\sigma_1} K(X(t), u(t))e^{-\lambda t}dt + MV(X_x(\sigma_1)). \end{aligned}$$

Hence as $\varepsilon \to 0$ we have other way inequality. Thus (DPPA) is proved. Now we proceed to prove (DPPC). Let t_1 be the first hitting time of C where the controller chooses to jump. In this case we write $t_1 = \xi_1$. Then

$$\begin{aligned} V(x) \le & \int_0^{\xi_1} K(X(t), u(t))e^{-\lambda t}dt + C_c\big(X(\xi_1), X(\xi_1)'\big)e^{-\lambda\xi_1} \\ & + \left[\int_{\xi_1}^{\infty} K(X(t), u(t))e^{-\lambda t}dt + \sum_{i=1}^{\infty} C_a(X(\sigma_i), v)e^{-\lambda\sigma_i}\right. \\ & \left. + \sum_{i=2}^{\infty} C_c\big(X(\xi_i), X(\xi_i)'\big)e^{-\lambda\xi_i}\right]. \end{aligned}$$

Doing the change of variables $t' = t - \xi_1$ in the square brackets and taking the infimum over the control variables, it is the value function of trajectory starting from $(X_x(\xi_1))'$. Hence,

$$V(x) \le \int_0^{\xi_1} K(X(t), u(t))e^{-\lambda t}dt + e^{-\lambda\xi_1}C_c(X(\xi_1), X(\xi_1)') + e^{-\lambda\xi_1}V(X_x(\xi_1)').$$

Now taking the infimum over $(X_x(\xi_1))' \in D$ in the last two terms we get

$$V(x) \le \int_0^{\xi_1} K(X(t), u(t))e^{-\lambda t}dt + NV(X_x(\xi_1)),$$

and taking the infimum over u in $\mathcal{U}$ on the RHS we will get the one-way inequality of (DPPC).

For the reverse inequality, given $\varepsilon > 0$ choose $\theta_\varepsilon = (u_\varepsilon, v_\varepsilon, \xi_{i_\varepsilon}, X(\xi_i)'_\varepsilon)$ such that

$$\begin{aligned} V(x) + \varepsilon &\geq \int_0^{\xi_{1_\varepsilon}} K(X(t), u_\varepsilon(t)) e^{-\lambda t} dt + NV(X_x(\xi_{1_\varepsilon})) \\ &\geq \inf_u \int_0^{\xi_{1_\varepsilon}} K(X(t), u(t)) e^{-\lambda t} dt + NV(X_x(\xi_{1_\varepsilon})). \end{aligned}$$

As $\varepsilon \to 0$ we will get

$$V(x) = \inf_u \left\{ \int_0^{\xi_1} K(X(t), u(t)) e^{-\lambda t} dt + NV(X_x(\xi_1)) \right\},$$

which proves (DPPC). The proof of (DPP) for any $T > 0$ follows similarly, which we skip here. □

THEOREM 4.2 (quasi-variational inequality). *Under the assumptions* (A1)–(A7) *and* (C1), (C2)*, the value function* V *described in* (2.7) *satisfies the following the QVI in the viscosity sense:*

$$\text{(QVI)} \qquad V(x) = \begin{cases} MV(x) & \forall x \in A, \\ \min\{NV(x), -H(x, DV(x))\} & \forall x \in C, \\ -H(x, DV(x)) & \forall x \in \Omega \setminus A \cup C, \end{cases}$$

where H *is the Hamiltonian given by*

$$H(x,p) = \sup_{u \in U} \left\{ \frac{-K(x,u) - f(x,u) \cdot p}{\lambda} \right\}.$$

Proof. Let $x \in A$. In this case we have to show that $V(x) = MV(x)$. Since $x \in A$, the first hitting time of trajectory is $\sigma_1 = 0$. Hence, by (DPPA) we get $V(x) = MV(x)$.

Now we consider the case $x \in \Omega \setminus A \cup C$. In this case we want to show that V satisfies the Hamilton–Jacobi–Bellman (HJB) equation in the viscosity sense. For we need to show the following: for all $\phi \in C^1(\Omega)$ and x local maximum of $V - \phi$

$$V(x) + H(x, D\phi(x)) \leq 0$$

and for all $\phi \in C^1(\Omega)$ and x local minimum of $V - \phi$

$$V(x) + H(x, D\phi(x)) \geq 0.$$

Let $r = \min\{\mathbf{d}(x, \partial A), \mathbf{d}(x, \partial C)\}$. Choose $R < r$. Then in the ball $B(x, R)$ no impulses are applied. Now V is continuous at x, and assume that $V - \phi$ has local maximum at x. Choose τ small enough such that $X_x(\tau) \in B(x, R)$. By our choice of R and τ, τ is less than the first hitting time. Then, since x is the local maximum of $V - \phi$,

$$\begin{aligned} \phi(x) - \phi(X_x(\tau)) &\leq V(x) - V(X_x(\tau)) \\ &\leq \int_0^\tau K(X_x(t), u(t)) e^{-\lambda t} dt + (e^{-\lambda\tau} - 1) V(X_x(\tau)), \end{aligned}$$

where the second inequality follows by (DPP), since $\tau < \sigma_1$ and $\tau < \xi_1$. Dividing by τ and taking the limit as $\tau \to 0$ we get

$$-D\phi(x) \cdot f(x) \leq K(x, u(0)) - \lambda V(x),$$

which implies

$$V(x) + \frac{-K(x, u(0) - D\phi(x) \cdot f(x)}{\lambda} \leq 0.$$

Taking the supremum over all $u \in \mathcal{U}$ we will get

$$V(x) + H(x, D\phi(x)) \leq 0.$$

Hence V is a viscosity subsolution of HJB equation.

To show that V is a viscosity supersolution, let $V - \phi$ have local minimum at x. Then for τ such that $X_x(\tau) \in B(x, R)$,

$$\begin{aligned}\phi(X_x(\tau)) - \phi(x) &\leq V(X_x(\tau)) - V(x) \\ &\leq (1 - e^{-\lambda\tau})V(X_x(\tau)) - \int_0^\tau K(X_x(t), u(t))e^{-\lambda t}dt \quad \text{by (DPP).}\end{aligned}$$

Dividing by τ and taking the limit as $\tau \to 0$ we get

$$\lambda V(x) - K(x, u(0)) - D\phi(x) \cdot f(x) \geq 0,$$

$$V(x) + \frac{-K(x, u(0)) - D\phi(x) \cdot f(x)}{\lambda} \geq 0.$$

Taking the supremum over all u we will get

$$V(x) + H(x, D\phi(x)) \geq 0.$$

Hence V is a viscosity supersolution of the HJB equation. Thus we have shown that in the case $x \in \Omega \setminus A \cup C$, V satisfies the HJB equation in the viscosity sense.

Now consider the case $x \in C$. We observe that if $x \in C$, and the controller chooses to jump, then by (DPPC), V should satisfy $NV(x)$. Whereas if the controller decides not to jump, then the system undergoes some continuous evolution and we can analyze as before to conclude that V satisfies the HJB equation in the viscosity sense. In this case we have to show that V satisfies the following equation in the viscosity sense:

$$\min\{V(x) - NV(x), V(x) + H(x, DV(x))\} = 0.$$

For this we need to show that, for all $\phi \in C^1(\Omega), x$ local minimum of $V - \phi$

$$\min\{V(x) - NV(x), V(x) + H(x, DV(x))\} \geq 0,$$

and for all $\phi \in C^1(\Omega), x$ local maximum of $V - \phi$,

$$\min\{V(x) - NV(x), V(x) + H(x, DV(x))\} \leq 0.$$

Now if $V(x) = NV(x)$, there is nothing to prove.

Suppose $V(x) < NV(x)$; then we need to show that V satisfies the HJB equation in the viscosity sense. We show that whenever $V(x) < NV(x)$ there exists $r > 0$ and a ball $B(x, r)$ around x such that it is not optimal to apply any impulses on $B(x, r)$. Then we can do the analysis in this ball to conclude as in the case of $x \in \Omega \setminus A \cup C$. For we claim that there exists $\varepsilon > 0$ such that

$$V(x) = \inf_{u, v, \xi_i, X(\xi_i)'} \left\{ \int_0^{t_1} K(X_x(t), u(t))e^{-\lambda t}dt + NV(X_x(t_1)) \mid t_1 > \varepsilon \right\}.$$

Suppose not; then $\varepsilon = 0$, which implies $\xi_1 = 0$, which by (DPPC) implies $V(x) = NV(x)$; this is a contradiction of our hypothesis $V(x) < NV(x)$. Hence $\varepsilon > 0$. Choose $r < \min\{d(x, X_x(\varepsilon)), d(A, C)\}$. Then in the ball $B(x, r)$, no impulses are applied. So we can do the analysis in this ball around x and conclude as in the earlier case. This proves the QVI for the case $x \in C$. ∎

5. Uniqueness. We take up the issue of uniqueness of the viscosity solutions of (QVI) in this section. Inspired by the earlier work on impulse control problem (see [2], [9]), we prove the comparison between any two solutions of the QVI.

THEOREM 5.1. *Assume* (A1)–(A7) *and* (C1), (C2). *Let* $u_1, u_2 \in BC(\Omega)$, *bounded continuous functions on* Ω, *be two viscosity solutions of the QVI given by* (QVI). *Then* $u_1 = u_2$.

Proof. The idea of the proof is to show that $u_1(x) \le u_2(x)$ for all $x \in \Omega$. We define the following auxiliary function Φ on $\bigcup_{i=1}^{\infty} (\Omega_i \times \Omega_i)$ that is Φ^i on each $\Omega_i \times \Omega_i$ by

$$\Phi^i(x, y) = u_1(x) - u_2(y) - \frac{1}{\epsilon}|x - y|^2 - \kappa\big(|x|^2 + |y|^2\big), \tag{5.1}$$

where ε and κ are small positive parameters to be chosen suitably later on. Observe that for each i, Φ^i attains its supremum over $\Omega_i \times \Omega_i$, thanks to the last two terms, which become large negative as $|x|, |y|$ goes to 0. We prove the theorem in two steps. In the first step of the proof we show that $\sup_i \sup_{\Omega_i \times \Omega_i} \Phi^i(x, y) \le 0$. In the next step we prove the uniqueness using Step 1.

Step 1. Let

$$\sup_i \sup_{\Omega_i \times \Omega_i} \Phi^i(x, y) = C > 0.$$

Fix $\kappa > 0$ such that $\kappa < \min\{\frac{C}{2}, \frac{C'}{2}\}$. If the above supremum is achieved at some (x_0, y_0), the following proof gets simplified. If not, corresponding to this κ we can choose (x_κ, y_κ) in some $\Omega_i \times \Omega_i$, say, $\Omega_1 \times \Omega_1$, such that

$$\Phi^1(x_\kappa, y_\kappa) > C - \kappa > \frac{C}{2}. \tag{5.2}$$

Let Φ^1 attain its supremum at some finite point, say, at (x_0, y_0) in $\Omega_1 \times \Omega_1$. Then

$$\sup_{\Omega_1 \times \Omega_1} \Phi^1(x, y) = \Phi^1(x_0, y_0) > C - \kappa > \frac{C}{2}. \tag{5.3}$$

Since x_0 and y_0 can lie in different sets in Ω_1, $u_1(x_0)$ and $u_2(y_0)$ will satisfy different equations from the QVI. We list below the different cases which arise:

1. $(x_0, y_0) \in A \times C$ or $C \times A$.
2. $(x_0, y_0) \in \Omega \setminus (A \cup C) \times \Omega \setminus (A \cup C)$.
3. $x_0, y_0 \notin A$ and one of x_0 or $y_0 \in C$. This takes care of $(x_0, y_0) \in C \times \Omega \setminus (A \cup C), (x_0, y_0) \in \Omega \setminus (A \cup C) \times C$, $(x_0, y_0) \in C \times C$.
4. $x_0, y_0 \notin C$ and one of the x_0 or $y_0 \in A$, i.e., $(x_0, y_0) \in A \times A$ or $(x_0, y_0) \in A \times \Omega \setminus (A \cup C)$, $(x_0, y_0) \in \Omega \setminus (A \cup C) \times A$.

Our idea is to show that in any of these cases, $u_1(x) - u_2(x)$ is arbitrarily small for ε and κ small. For this we will estimate $u_1(x_0) - u_2(y_0)$ at the maximum point (x_0, y_0) of Φ^1 or $u_1(x_n) - u_2(y_n)$ at the maximum point (x_n, y_n) of ψ_n, a suitably defined auxiliary function. The crucial point in our proof is that after at most finitely many

steps, say n_0, at the maximum point of ψ_{n_0} both u_1 and u_2 satisfy the HJB equation. Then we can use the usual comparison principle available in the literature. We first list some standard estimates needed later in the proof.

LEMMA 5.2. *Let Φ and (x_0, y_0) be as above. Then*

(i) $\frac{|x_0-y_0|^2}{\epsilon} \le C$ *for some C independent of κ and ε;*

(ii) $\sqrt{\kappa}|x_0|, \sqrt{\kappa}|y_0| \le \hat{C}$ *for some $\hat{C}$ independent of κ and ε;*

(iii) $\frac{|x_0-y_0|^2}{\epsilon} \le \omega_\kappa^1(\sqrt{C\epsilon})$, *where ω_κ^1 is the local modulus of continuity of both u_1 and u_2 in the ball of radius R, dependent on κ but independent of ε, $R = R(\kappa) = \hat{C}/\sqrt{\kappa}$ in Ω_1.*

Proof. By our assumption

$$2\Phi^1(x_0, y_0) \ge \Phi^1(x_0, x_0) + \Phi^1(y_0, y_0). \tag{5.4}$$

Hence

$$\frac{2}{\epsilon}|x_0 - y_0|^2 \le u_1(x_0) - u_1(y_0) + u_2(x_0) - u_2(y_0). \tag{5.5}$$

Since u_1 and u_2 are bounded,

$$\frac{|x_0 - y_0|^2}{\epsilon} \le C,$$

which proves (i). This also implies

$$|x_0 - y_0| \le \sqrt{C\epsilon}.$$

To prove (ii), fix some $z \in \Omega_1$ such that $|z| = 1$; then $\Phi^1(x_0, y_0) \ge \Phi^1(z, z)$, which implies

$$\begin{aligned}\kappa\big(|x_0|^2 + |y_0|^2\big) &\le u_1(x_0) - u_1(z) - u_2(y_0) + u_2(z) - \frac{1}{\epsilon}|x_0 - y_0|^2 + 2\kappa|z|^2 \\ &\le C + 2\kappa \le C + 2.\end{aligned}$$

Hence $\sqrt{\kappa}|x_0| \le \hat{C}$, where $\hat{C}$ is independent of κ and ε. Similarly, $\sqrt{\kappa}|y_0| \le \hat{C}$. This proves (ii). Hence x_0 and y_0 lie in some ball B_R of radius $R = R(\kappa)$.

Now using the estimate in (i) and the modulus of continuity of u_1 and u_2 in the compact set $\bar{B}_{R(\kappa)}$ in Ω_1, we get

$$\frac{|x_0 - y_0|^2}{\epsilon} \le \omega_\kappa^1(\sqrt{C\epsilon}).$$

This proves (iii). □

Now we consider the different cases listed earlier.

Case 1. $(x_0, y_0) \in A \times C$ or $C \times A$.

Claim. This case does not occur.

Without loss of generality let $(x_0, y_0) \in A \times C$. Since $d(A, C) > \beta$,

$$\Rightarrow |x_0 - y_0| > \beta.$$

On the other hand by Lemma 5.2(i),

$$|x_0 - y_0| < \sqrt{C\epsilon}.$$

So choosing ϵ such that $\sqrt{C\epsilon} < \frac{\beta}{2}$,

$$|x_0 - y_0| < \frac{\beta}{2},$$

which is a contradiction. Hence Case 1 does not occur, for small ε.

Case 2. $(x_0, y_0) \in \Omega \setminus (A \cup C) \times \Omega \setminus (A \cup C)$.

In this case at $(x_0, y_0) \in \Omega_1 \times \Omega_1$, u_1, u_2 both satisfy the HJB equation. Hence we do all the calculations in Ω_1. Let us define the test functions ϕ_1 and ϕ_2 on Ω_1 as follows:

$$\phi_1(x) = u_2(y_0) + \frac{1}{\epsilon}|x - y_0|^2 + \kappa\big(|x|^2 + |y_0|^2\big), \tag{5.6}$$

$$\phi_2(y) = u_1(x_0) - \frac{1}{\epsilon}|x_0 - y|^2 - \kappa\big(|x_0|^2 + |y|^2\big). \tag{5.7}$$

Then, since (x_0, y_0) is point of supremum for Φ^1, $u_1 - \phi_1$ attains its maximum at x_0 and $u_2 - \phi_2$ attains its minimum at y_0. Also observe

$$D\phi_1(x_0) = \frac{2}{\epsilon}(x_0 - y_0) + 2\kappa x_0, \tag{5.8}$$

$$D\phi_2(y_0) = \frac{2}{\epsilon}(x_0 - y_0) - 2\kappa y_0, \tag{5.9}$$

and by Lemma 5.2

$$|D\phi_2(y_0)| \le \frac{2}{\epsilon}|x_0 - y_0| + \sqrt{\kappa}\hat{C}. \tag{5.10}$$

Now by definition of the viscosity sub- and supersolutions, and using u_1 as the subsolution and u_2 as the supersolution,

$$\begin{aligned} &u_1(x_0) + H(x_0, D\phi_1(x_0)) \le 0 \le u_2(y_0) + H(y_0, D\phi_2(y_0)) \\ &\quad \Rightarrow u_1(x_0) - u_2(y_0) \le H(y_0, D\phi_2(y_0)) - H(x_0, D\phi_1(x_0)). \end{aligned}$$

By our assumptions (A1)–(A7) and the definition of Hamiltonian H, one can easily prove that H satisfies the structural condition

$$|H(x,p) - H(y,q)| \le F|p - q| + L|q||x - y| + K_1|x - y|, \tag{5.11}$$

where K_1 is the Lipschitz constant for the running cost k. Using (5.11) we get

$$\begin{aligned} u_1(x_0) - u_2(y_0) \le{}& L\,|D\phi_2(y_0)|\,|x_0 - y_0| + K_1|x_0 - y_0| \\ &+ F\,|D\phi_2(y_0) - D\phi_1(x_0)|. \end{aligned}$$

Substituting from (5.8), (5.9), and (5.10),

$$u_1(x_0) - u_2(y_0) \le \frac{2L}{\epsilon}|x_0 - y_0|^2 + \sqrt{\kappa}L\hat{C}|x_0 - y_0| + K_1|x_0 - y_0| + 2\kappa F|x_0 + y_0|.$$

By Lemma 5.2 we then get

$$u_1(x_0) - u_2(y_0) \le 2L\omega^1_\kappa(\sqrt{C\epsilon}) + L\hat{C}\sqrt{C\kappa\varepsilon} + K_1(\sqrt{C\epsilon}) + 4F\hat{C}\sqrt{\kappa}. \tag{5.12}$$

Also observe that by (5.2)

$$\begin{aligned}\frac{C}{2} < C - \kappa &< \Phi^1(x_\kappa, x_\kappa) \\ &\le \Phi^1(x_0, y_0) \\ &\le u_1(x_0) - u_2(y_0) \\ &\le 2L\omega_\kappa^1\big(\sqrt{C\epsilon}\big) + 2L\hat{C}\sqrt{C\kappa\varepsilon} + K_1\big(\sqrt{C\epsilon}\big) + 4F\hat{C}\sqrt{\kappa}.\end{aligned}$$

Now fixing κ and sending ε to 0 and then choosing κ such that $4F\hat{C}\sqrt{\kappa} < \frac{C}{4}$ we will have

$$\frac{C}{2} < \frac{C}{4}.$$

This is a contradiction. Hence,

$$\sup_i \sup_{\Omega_i \times \Omega_i} \Phi^i(x, y) \le 0.$$

Case 3. $x_0, y_0 \notin A$, and one of $x_0, y_0 \in C$. Without loss of generality let $y_0 \in C$. $x_0 \notin A$ and u_1 is a subsolution of the QVI implies

$$u_1(x_0) + H(x_0, Du_1(x_0)) \le 0,$$

$$y_0 \in C \Rightarrow \max\{u_2(y_0) + H(y_0, Du_2(y_0)), u_2(y_0) - Nu_2(y_0)\} = 0,$$

and u_2 is a solution of the QVI, in particular it is a supersolution. Hence either $u_2 + H \ge 0$ or $u_2 - Nu_2 \ge 0$ at y_0.

If $u_2(y_0) + H(y_0, Du_2(y_0)) \ge 0$, we can proceed as in Case 2 and get a contradiction. Otherwise assume $u_2(y_0) - Nu_2(y_0) \ge 0$. Since u_2 is also a subsolution

$$u_2(x) \le Nu_2(x) \quad \forall x \in C.$$

Therefore,

$$u_2(y_0) = Nu_2(y_0) = \inf_{y' \in D} u_2(y') + c_c(y_0, y') = \inf_i \inf_{D_i} u_2(y') + c_c(y_0, y').$$

As each D_i is compact, the infimum is attained on each D_i. If the infimum over i is not attained, then we can choose y_0' in, say, D_2 such that

$$u_2(y_0) = Nu_2(y_0) > u_2(y_0') + c_c(y_0, y_0') - \kappa, \quad y_0' \in D_2.$$

Also $y_0' \notin A$. We estimate the difference $\Phi^1(x_0, y_0)$ and $\Phi^2(y_0', y_0')$ in the following lemma, which we will use to define another auxiliary function ψ_1, and consider the maximum point (x_1, y_1) of ψ_1, in the same spirit as in the earlier work on the impulse control problem (see [2], [7], [9]). We will show that after at most a finite number of such auxiliary functions, we necessarily arrive at Case 2.

Recall that y_0' lies in D, hence by (A2), $|y_0'| < R$. We will also need that x_0 and y_0 are not too close to y_0' in case $y_0' \in \Omega_1$. The following lemma proves this fact. More generally we prove here that whenever $u(x) = Nu(x)$ or $u(x) = Mu(x)$ the destination point is at a certain positive distance away from the point of supremum.

LEMMA 5.3. *Let $u \in BC(\Omega)$ be a solution of* (QVI). *If x, x', and $g(x,v')$ belong to $D_1 \subseteq \Omega_1$ and if*

$$\begin{aligned} &u(x) = Nu(x) > u(x') + c_c(x,x') - \kappa \\ \text{or} \quad &u(x) = Mu(x) = u(g(x,v')) + c_a(x,v'), \end{aligned}$$

then there exists an $\alpha_1 > 0$ depending only on the uniform continuity of u on $D_1 \subseteq \Omega_1$ but independent of ε and κ such that

$$|x - x'| > \alpha_1 \tag{5.13}$$

$$\text{or} \quad |x - g(x,v')| > \alpha_1, \tag{5.14}$$

depending on which equation $u(x)$ satisfies.

Proof. We claim that there exists $\alpha_1 > 0$ such that $|x - x'| > \alpha_1$. Suppose the contrary. That is, there exists sequence $x_n,\ x_n' \in \Omega_1$ such that

$$u(x_n) > u\big(x_n'\big) + c_c\big(x_n, x_n'\big) - \kappa \ \text{ and } \ |x_n - x_n'| \to 0.$$

Then by continuity of u, $|u(x_n) - u(x_n')| \to 0$. But

$$|u(x_n) - u\big(x_n'\big)| = c_c\big(x_n, x_n'\big) - \kappa > C' - \kappa > \frac{C'}{2} > 0,$$

which is a contradiction. Hence given $\frac{C'}{4}$ choose the corresponding α_1 given by uniform continuity of u on $D_1 \subseteq \Omega_1$ such that $|y - z| < \alpha_1 \Rightarrow |u(y) - u(z)| < \frac{C'}{4}$. Then

$$|x - x'| > \alpha_1.$$

This proves (5.13).

To prove that $|x - g(x,v')| > \alpha_1$, we proceed with arguments similar to those above and choose α_1 corresponding to the $\frac{C'}{4}$ in the definition of uniform continuity of u on D_1. ∎

In the next lemma we estimate the difference $\Phi^1(x_0,y_0)$ and $\Phi^2(y_0',y_0')$, which we are going to use to define new auxiliary function ψ_1.

LEMMA 5.4. *Let Φ be as defined in* (5.1) *and let $(x_0,y_0) \in \Omega_1 \times \Omega_1$ be as in* (5.3), *the point where Φ^1 attains supremum. Let $y_0' \in D_2$ be such that*

$$u_2(y_0) = Nu_2(y_0) > u_2\big(y_0'\big) + c_c\big(y_0, y_0'\big) - \kappa. \tag{5.15}$$

Then

$$\Phi^1(x_0,y_0) - \Phi^2\big(y_0',y_0'\big) \le \kappa K$$

for some constant $K > 1$ depending only on the constants of the problem and independent of ϵ and κ.

Proof.

$$\begin{aligned} \Phi^1(x_0,y_0) - \Phi^2\big(y_0',y_0'\big) = u_1(x_0) - u_2(y_0) - \frac{1}{\epsilon}|x_0 - y_0|^2 - \kappa\big(|x_0|^2 + |y_0|^2\big) \\ - u_1\big(y_0'\big) + u_2\big(y_0'\big) + 2\kappa|y_0'|^2. \end{aligned}$$

Using (5.15) we get

$$\Phi^1(x_0,y_0)-\Phi^2\big(y_0',y_0'\big)<u_1(x_0)-c_c\big(y_0,y_0'\big)-\frac{1}{\epsilon}|x_0-y_0|^2-\kappa\big(|x_0|^2+|y_0|^2\big)\\ -u_1\big(y_0'\big)+2\kappa|y_0'|^2+\kappa.$$

Also $u_1(y_0)\le Nu_1(y_0)\le u_1(y_0')+c_c(y_0,y_0')$. Hence,

$$\begin{aligned}\Phi^1(x_0,y_0)-\Phi^2\big(y_0',y_0'\big)&\le u_1(x_0)-u_1(y_0)-\frac{1}{\epsilon}|x_0-y_0|^2-\kappa\big(|x_0|^2+|y_0|^2\big)+2\kappa|y_0'|^2\\ &\quad+\kappa\le u_1(x_0)-u_1(y_0)+2\kappa|y_0'|^2+\kappa\\ &\le u_1(x_0)-u_1(y_0)+2\kappa R^2+\kappa\\ &\le \omega_\kappa^1\big(\sqrt{C\epsilon}\big)+2\kappa R^2+\kappa.\end{aligned}$$

Using the modulus of continuity of u_1, on $\bar{B_R}$ in Ω_1 for a given $\kappa>0$ choose $\epsilon>0$ such that

$$\omega_\kappa^1\big(\sqrt{C\epsilon}\big)<\kappa\Rightarrow\Phi^1(x_0,y_0)-\Phi^2\big(y_0',y_0'\big)\le\kappa K.$$

This proves the lemma. □

We use the above difference to define another auxiliary function ψ_1. We further restrict α_2 given by Lemma 5.3, if necessary, so that $\alpha_2<\frac{\beta}{2}$. Define

$$\psi_1^2(x,y)=\Phi^2(x,y)+2\kappa K\,\zeta_1(x,y),$$

$$\psi_1^i(x,y)=\Phi^i(x,y)\quad\forall i\ne 2,$$

where, $\zeta_1(x,y)\in C_0^\infty(\Omega_2\times\Omega_2)$, such that

$$\zeta_1\big(y_0',y_0'\big)=1;\quad 0\le\zeta_1\le 1;\quad |D\zeta_1|\le\frac{2}{\alpha_2};$$

$$\zeta_1(x,y)<1\text{ if }(x,y)\ne\big(y_0',y_0'\big);$$

$$\text{and}\ \ \zeta_1(x,y)=0\quad\forall(x,y)\ \text{ such that }|x-y_0'|^2+|y-y_0'|^2>\alpha_1,$$

i.e., ζ_1 has support in the α_1 ball around $(y_0',y_0')\in\Omega_2\times\Omega_2$, having maximum at (y_0',y_0') and it vanishes on all $\Omega_i\times\Omega_i$ other than $i=2$.

Observe that by the definition of ψ_1^i,

$$\begin{aligned}\psi_1^2\big(y_0',y_0'\big)&=\Phi^2\big(y_0',y_0'\big)+2\kappa K\\ &\ge\Phi^1(x_0,y_0)-K\kappa+2\kappa K\\ &\ge\sup_i\sup_{\Omega_i\times\Omega_i}\Phi^i(x,y)+\kappa K-\kappa\\ &\ge\psi_1^2(x,y)-2\kappa K\,\zeta_1(x,y)+\kappa(K-1).\end{aligned}$$

As ζ_1 is 0 for all $(x,y)\in\Omega_i\times\Omega_i$, $i\ne 2$, and for (x,y) outside the α_1 ball around (y_0',y_0') in $\Omega_2\times\Omega_2$, we have for all such (x,y)

$$\psi_1^2\big(y_0',y_0'\big)>\psi_1^2(x,y).$$

Hence ψ_1^2 has the supremum over $\Omega_2 \times \Omega_2$ in the α_1 ball around (y_0', y_0'). Let (x_1, y_1) be such that

$$\sup_{\Omega_2 \times \Omega_2} \psi_1^2 = \psi_1^2(x_1, y_1).$$

Then

$$\psi_1^2(x_1, y_1) \geq \psi_1^1(x_0, y_0) = \Phi^1(x_0, y_0) > C - \kappa. \tag{5.16}$$

Since $\alpha_1 < \frac{\beta}{2}$, $x_1, y_1 \notin A$. We remark here that by using the technique of Lemma 5.2, we can prove that

$$\frac{|x_1 - y_1|^2}{\epsilon} \leq \omega_\kappa^2\big(\sqrt{C\epsilon}\big) + 2K\kappa \quad \text{and} \quad |x_1|, |y_1| < \hat{C}\sqrt{\kappa}.$$

Thus either $x_1, y_1 \notin C$ or one of them is in C. If $x_1, y_1 \notin C$, we are in Case 2 or Case 4. If we are in Case 2, we can get the comparison by working with ψ_1 instead of Φ as in Case 2. We will show in the next step of the proof how to handle Case 4. Now if one of $x_1, y_1 \in C$, we are again in Case 3. So without loss of generality let $y_1 \in C$ and y_1 be such that $u_2(y_1) - Nu_2(y_1) \geq 0$. Then, as earlier, the approximate infimum will be attained at some point, say, $y_1' \in D$, some D_i which we call D_3. That is

$$u_2(y_1) = Nu_2(y_1) > u_2\big(y_1'\big) + c_c\big(y_1, y_1'\big) - \kappa.$$

We define ψ_2 on $\bigcup \Omega_i \times \Omega_i$, that is, ψ_2^i on $\Omega_i \times \Omega_i$, by

$$\psi_2^i(x, y) = \Phi^i(x, y) + 2\kappa K \sum_{j=1}^{2} \zeta_j(x, y),$$

where $\zeta_2(y_1', y_1') = 1$ and ζ_2 has support in the α_3 ball around (y_1', y_1') in $\Omega_3 \times \Omega_3$ with the properties $\zeta_2 \in C_0^\infty(\Omega \times \Omega)$, $0 \leq \zeta_2 \leq 1$, $|D\zeta_2| \leq \frac{2}{\alpha_3}$, $\zeta_2(x, y) < 1$ if $(x, y) \neq (y_1', y_1')$. Hence as before we can show that the supremum of ψ_2 is attained in the α_3 ball around (y_1', y_1'). Also we can show that ψ_2^3 satisfies the inequality similar to (5.16), namely,

$$\psi_2^2(x_1, y_1) \geq \psi_1^2(x_1, y_1) = \Phi^1(x_0, y_0) > C - \kappa.$$

Thus we can proceed to define $\psi_3, \psi_4, \ldots, \psi_n$ and so on, in case $u_2(y_i) = Nu_2(y_i)$. We now claim that this process has to terminate in finitely many steps, which is the content of the following lemma.

LEMMA 5.5. *Suppose $(x_n, y_n) \in \Omega_{n+1} \times \Omega_{n+1}$, $y_n' \in D_{n+2}$ are sequences such that*

$$u_2(y_n) = Nu_2(y_n) > u_2(y_n') + c_c(y_n, y_n') - \kappa, \quad y_n \in B(y_{n-1}', \alpha_{n+1});$$
$$\psi_n(x, y) = \psi_{n-1}(x, y) + 2\kappa K\, \zeta_n(x, y); \quad \psi_n(x_n, y_n) = \sup_{\Omega_{n+1} \times \Omega_{n+1}} \psi_n(x, y);$$

where ζ_n is such that $\zeta_n \in C_0^\infty(\Omega \times \Omega)$; actually ζ_n has support in the α_{n+1} ball around $(y_n', y_n') \in \Omega_{n+2} \times \Omega_{n+2}$. $0 \leq \zeta_n \leq 1$; $|D\zeta_n| < \frac{2}{\alpha_{n+1}}$; $\zeta_n(y_{n-1}', y_{n-1}') = 1$, $n = 1, 2, \ldots$. Then $n < n_0 = [\frac{8\hat{C}}{C'}]$, where $\hat{C}$ is a bound on u_1 and u_2 and C' is the lower bound on c_c.

Proof. Observe that y_i', $y_{i+1} \in D_{i+2}$. By uniform continuity of u_2 on $D_{i+2} \subseteq \Omega_{i+2}$, for all i,

$$|y_{i+1} - y_i'| < \alpha_{i+1} \Rightarrow |u_2(y_{i+1}) - u_2(y_i')| < \frac{C'}{4}.$$

By assumption,

$$\begin{aligned} u_2(y_0) &> u_2(y_0') + c_c(y_0, y_0') - \kappa \\ &> u_2(y_0') + C' - \kappa\ ; \quad \text{because } c_c \geq C' > 0 \\ &> u_2(y_1) - \frac{C'}{4} + C' - \kappa = u_2(y_1) + \frac{3}{4}C' - \kappa \\ &> u_2(y_1') + c_c(y_1, y_1') + \frac{3}{4}C' - 2\kappa \ > \ u_2(y_1') + C' + \frac{3}{4}C' - 2\kappa \\ &> u_2(y_2) - \frac{C'}{4} + C' + \frac{3}{4}C' - 2\kappa = u_2(y_2) + \frac{6}{4}C' - 2\kappa. \end{aligned}$$

Therefore, at the nth stage we will get

$$\hat{C} \geq u_2(y_0) > u_2(y_n) + \frac{3}{4}nC' - n\kappa.$$

By using $\kappa < \frac{C'}{2}$, if $n > n_0 = [\frac{8\hat{C}}{C'}]$, then $u_2(y_0) > \hat{C}$, which is a contradiction, because $|u_2| < \hat{C}$. □

Thus we have only a finite sequence of $\{y_n\}$ such that $u_2(y_n) = Nu_2(y_n)$. So, for $n > n_0 = [\frac{8\hat{C}}{C'}]$ necessarily $u_2(y_n) < Nu_2(y_n)$ and hence

$$u_2(y_n) + H(y_n, Du_2(y_n)) \geq 0.$$

Hence both u_1 and u_2 satisfy the HJB at the supremum point of auxiliary function ψ_n. Now we proceed as in Case 2 taking care of the extra terms.

In this case we define test functions ϕ_1 and ϕ_2 by

$$\phi_1(x) = u_2(y_n) + \frac{1}{\epsilon}|x - y_n|^2 + \kappa\big(|x|^2 + |y_n|^2\big) - 2\kappa K \sum_{j=1}^{n} \zeta_j(x, y_n), \tag{5.17}$$

$$\phi_2(y) = u_1(x_n) - \frac{1}{\epsilon}|x_n - y|^2 - \kappa\big(|x_n|^2 + |y|^2\big) + 2\kappa K \sum_{j=1}^{n} \zeta_j(x_n, y). \tag{5.18}$$

Then by the definition of (x_n, y_n), $u_1 - \phi_1$ has maximum at x_n and $u_2 - \phi_2$ has minimum at y_n. Using u_1 as the viscosity subsolution and u_2 as the viscosity supersolution, we get

$$u_1(x_n) - u_2(y_n) \leq H(y_n, D\phi_2(y_n)) - H(x_n, D\phi_1(x_n)).$$

Let $\alpha = \min\{\alpha_1, \ldots, \alpha_{n+1}\}$. Also, whenever $(x_n, y_n) \in \Omega_{j+1} \times \Omega_{j+1}$ we can write

$$D\phi_1(x_n) = \frac{2}{\epsilon}(x_n - y_n) + 2\kappa x_n - 2K\kappa \sum_{j=1}^{n} D\zeta_j(x_n, y_n), \tag{5.19}$$

$$D\phi_2(y_n) = \frac{2}{\epsilon}(x_n - y_n) - 2\kappa y_n + 2K\kappa \sum_{j=1}^{n} D\zeta_i(x_n, y_n), \tag{5.20}$$

$$|D\phi_1(y_n)| \leq \frac{2}{\epsilon}(x_n - y_n) + 2\kappa|y_n| + \frac{4nK\kappa}{\alpha}. \tag{5.21}$$

Hence by structural condition on H given by (5.11),

(5.22)
$$u_1(x_n) - u_2(y_n) \le L|D\phi_2(y_n)|\,|x_n - y_n| + K_1|x_n - y_n| + F|D\phi_2(y_n) - D\phi_2(x_n)|.$$

By using (5.19), (5.20), (5.21) in the above we get

$$u_1(x_n) - u_2(y_n) \le \frac{2L}{\epsilon}|x_n - y_n|^2 + 2\kappa L|y_n|\,|x_n - y_n| + \left(\frac{4K\kappa n}{\alpha}\right)|x_n - y_n| + K_1|x_n - y_n| + 4F\kappa(|x_n| + |y_n|) + \frac{8\kappa K n}{\alpha}. \tag{5.23}$$

Now by using the technique of Lemma 5.2 for ψ_n, we can prove that

$$\begin{aligned} |x_n - y_n| &< \sqrt{C\epsilon}, \\ \frac{|x_n - y_n|^2}{\epsilon} &\le \omega_\kappa^n\big(\sqrt{C\epsilon}\big) + 2\kappa K, \\ |x_n|\,|y_n| &\le \sqrt{\kappa}\hat{C}, \end{aligned}$$

where $\hat{C}, K$, and C are independent of ε and κ. Using these estimates in (5.23) we will get

$$u_1(x_n) - u_2(y_n) \le 2L\omega_\kappa^n\big(\sqrt{C\epsilon}\big) + 4L\kappa K + 2L\hat{C}\sqrt{C\kappa\varepsilon} + \left(\frac{4K\kappa n}{\alpha}\right)\sqrt{C\epsilon} + K_1\big(\sqrt{C\epsilon}\big) + 8F\hat{C}\sqrt{\kappa} + \frac{8\kappa K n}{\alpha}. \tag{5.24}$$

Also observe that from (5.3),

$$\frac{C}{2} < C - \kappa < \Phi^1(x_0, y_0) \le \psi_n^{n+1}(x_n, y_n).$$

Hence

$$\begin{aligned} \frac{C}{2} < C - \kappa &\le u_1(x_n) - u_2(y_n) - \frac{|x_n - y_n|^2}{\varepsilon} - \big(|x_n|^2 + |y_n|^2\big) + 2\kappa K\sum_{j=1}^{n}\zeta_j(x^n, y^n) \\ &\le u_1(x_n) - u_2(x_n) + 2\kappa K n. \end{aligned}$$

By using (5.24) in the above, with $n \le n_0$ given by Lemma 5.5, we get

$$\frac{C}{2} \le 2L\omega_\kappa^n\big(\sqrt{C\epsilon}\big) + 4L\kappa K + 2L\hat{C}\sqrt{C\kappa\varepsilon} + \left(\frac{4K\kappa n_0}{\alpha}\right)\sqrt{C\epsilon} + K_1(\sqrt{C\epsilon}) + 8F\hat{C}\sqrt{\kappa} + \frac{8\kappa K n_0}{\alpha} + 2\kappa K n_0.$$

Now first fixing κ and sending ε to 0 we get

$$\frac{C}{2} \le 8F\hat{C}\sqrt{\kappa} + 4L\kappa K + \frac{8\kappa K n_0}{\alpha} + 2\kappa K n_0.$$

Now we can choose κ so that the RHS of the above expression is strictly less than $\frac{C}{4}$ and hence we will get $\frac{C}{2} \le \frac{C}{4}$. This is a contradiction; hence, $\sup_i \sup_{\Omega_i\times\Omega_i} \psi_n^i(x, y) \le 0$. This implies that

$$\sup_i \sup_{\Omega_i\times\Omega_i} \Phi^i(x, y) \le \sup_i \sup_{\Omega_i\times\Omega_i} \psi_n^i(x, y) \le 0.$$

Thus in this case also we have $\sup_i \sup_{\Omega_i \times \Omega_i} \Phi^i(x,y) \le 0$.

Case 4. Now consider the last case where one of the x_0 or y_0 is in A. Without loss of generality we assume that $y_0 \in A$.

LEMMA 5.6. *Let Φ be as defined by* (5.1) *and let (x_0, y_0) be as in* (5.24)*, that is, $\Phi^1(x_0,y_0) = \sup_{\Omega_1 \times \Omega_1} \Phi^1$. Moreover, let y_0 be such that $u_2(y_0) = Mu_2(y_0) = u_2(g(y_0,v_0)) + c_a(y_0,v_0)$, where $g(y_0,v_0) \in \Omega_2$. Then*

$$\Phi^1(x_0,y_0) - \Phi^2(g(y_0,v_0), g(y_0,v_0)) < \kappa K$$

for some constant $K > 1$ depending only on the constants of the problem and independent of ϵ and κ.

Proof.

$$\begin{aligned}\Phi^1(x_0,y_0) - \Phi^2(g(y_0,v_0),(g(y_0,v_0)) &= u_1(x_0) - u_2(y_0) - \frac{1}{\epsilon}|x_0 - y_0|^2 - \kappa(|x_0|^2 + |y_0|^2) \\ &\quad - u_1(g(y_0,v_0)) + u_2(g(y_0,v_0)) + 2\kappa|g(y_0,v_0)|^2 \\ &= u_1(x_0) - c_a(y_0,v_0) - \frac{1}{\epsilon}|x_0 - y_0|^2 \\ &\quad - \kappa(|x_0|^2 + |y_0|^2) - u_1(g(y_0,v_0)) + 2\kappa|g(y_0,v_0)|^2.\end{aligned}$$

We add and subtract $u_1(y_0)$ in the above, and observing that $u_1(y_0) \le Mu_1(y_0) \le u_1(g(y_0,v_0)) + c_a(y_0,v_0)$, we get

$$\begin{aligned}\Phi^1(x_0,y_0) - \Phi^2(g(y_0,v_0), g(y_0,v_0)) &\le u_1(x_0) - u_1(y_0) - c_a(y_0,v_0) \\ &\quad - u_1(g(y_0,v_0)) + u_1(y_0) + 2\kappa|g(y_0,v_0)|^2 \\ &\le u_1(x_0) - u_1(y_0) + 2\kappa|g(y_0,v_0)|^2 \\ &\le \omega_\kappa^1(|x_0 - y_0|) + 2\kappa R^2.\end{aligned}$$

We can choose ϵ such that $\omega_\kappa^1(\sqrt{C\epsilon}) < \kappa$. Then by the Lemma 5.2,

$$\begin{aligned}&\omega_\kappa^1(|x_0 - y_0|) \le \omega_\kappa^1\big(\sqrt{C\epsilon}\big) < \kappa \\ &\quad \Rightarrow \Phi^1(x_0,y_0) - \Phi^2(g(y_0,v_0), g(y_0,v_0)) \le K\kappa,\end{aligned}$$

where K depends on the modulus of continuity of u_1 and R. This proves the lemma. □

To proceed, if necessary, we restrict $\alpha_2 < \frac{\beta}{2}$, where α_2 is as in Lemma 5.3 and define a C_0^∞ function ζ_1 on $\Omega \times \Omega$ by

$$\zeta_1(g(y_0,v_0), g(y_0,v_0)) = 1; \quad 0 \le \zeta_1 \le 1; \quad |D\zeta_1| < \frac{2}{\alpha_2};$$

$$\zeta_1(x,y) < 1 \text{ if } (x,y) \ne (g(y_0,v_0), g(y_0,v_0));$$

$$\text{and} \quad \operatorname{supp} \zeta_1 \subseteq B((g(y_0,v_0), g(y_0,v_0)), \alpha_2).$$

Note that ζ_1 is nonzero only on $\Omega_2 \times \Omega_2$ and it vanishes on all other $\Omega_i \times \Omega_i$. Define a new auxiliary function ψ_1 on $\Omega \times \Omega$ denoted by ψ_1^i on $\Omega_i \times \Omega_i$ such that

$$\begin{aligned}\psi_1^2(x,y) &= \Phi^i(x,y) + 2K\kappa\zeta_1(x,y), \\ \psi_1^i(x,y) &= \Phi^i(x,y) \quad \text{for } i \ne 2.\end{aligned}$$

Then arguing as in Case 3 we can conclude that ψ_1^2 attains its maximum in the α_2 ball around $(g(y_0, v_0), g(y_0, v_0))$. Let (x_1, y_1) be such that $\psi_1^2(x_1, y_1) = \sup_{\Omega_2 \times \Omega_2} \psi_1^2$. Since $\alpha_2 < \frac{\beta}{2}$, $x_1, y_1 \notin A$. Using techniques similar to those of Lemma 5.2 we can prove that

$$\frac{|x_1 - y_1|^2}{\epsilon} \leq \omega_\kappa^2\big(\sqrt{C\epsilon}\big) + 2K\kappa,$$
$$|x_1|,\ |y_1| < \hat{C}\sqrt{\kappa}.$$

Now either $(x_1, y_1) \in \Omega \setminus (A \cup C) \times \Omega \setminus (A \cup C)$ or one of x_1 or $y_1 \in C$. In both cases, we are either in Case 2 or in Case 3. Thus in any case, after finitely many steps, we will arrive at Case 2 and get that $\sup_i \sup_{\Omega_i \times \Omega_i} \Phi^i(x, y) \leq 0$. This proves the claim in Step 1.

Step 2. In Step 2 we show the uniqueness. For any $x \in \Omega$,

$$u_1(x) - u_2(x) \leq \Phi(x, x) + 2\kappa|x|^2.$$

Sending κ to 0, we get

$$\begin{aligned} u_1(x) - u_2(x) &\leq \Phi(x, x) \\ &\leq \sup_i \sup_{\Omega_i \times \Omega_i} \Phi^i(x, y) \\ &\leq 0, \end{aligned}$$

where the last inequality follows by Step 1. Now interchanging the roles of u_1 and u_2, we get other way inequality, which proves that $u_1 = u_2$ for all $x \in \Omega$, and hence the uniqueness. □

Acknowledgments. The authors would like to thank Prof. M. K. Ghosh for introducing them to the problem and Prof. L. C. Evans for useful references and discussions on switching problems. The authors would also like to thank referees for suggesting useful corrections.

REFERENCES

[1] A. Back, J. Gukenheimer, and M. Myers, *A dynamical simulation facility for hybrid systems,* in Hybrid Systems, Lecture Notes in Comput. Sci. 736, R. L. Grossman, A. Nerode, A. P. Rava, and H. Rischel, eds., Springer, New York, 1993.

[2] G. Barles, *Quasi-variational inequalities and first order Hamilton–Jacobi equations,* Nonlinear Anal., 9 (1985), pp. 131–148.

[3] M. Bardi and C. Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton–Jacobi–Bellman Equations,* Birkhäuser, Boston, 1997.

[4] A. Bensoussan and J. L. Menaldi, *Dynamics of hybrid control and dynamic programming,* Dyn. Contin. Discrete Impuls. Syst., 3 (1997), pp. 395–442.

[5] M. S. Branicky, V. S. Borkar, and S. K. Mitter, *A unified framework for hybrid control problem,* IEEE Trans. Automat. Control, 43 (1998), pp. 31–45.

[6] M. S. Branicky, *Studies in Hybrid Systems: Modelling, Analysis and Control,* Ph.D. dissertation, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, 1995.

[7] I. Capuzzo-Dolcetta and L. C. Evans, *Optimal switching for ordinary differential equations,* SIAM J. Control Optim., 22 (1984), pp. 143–161.

[8] P. P. Varaiya, *Smart cars on smart roads: Problems of control,* IEEE Trans. Automat. Control, 38 (1993), pp. 195–207.

[9] J. Yong, *Systems governed by ordinary differential equations with continuous, switching and impulse controls,* Appl. Math. Optim., 20 (1989), pp. 223–235.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 4, pp. 1289–1321

BOUNDED VARIATION SINGULAR STOCHASTIC CONTROL AND DYNKIN GAME*

FREDERIK BOETIUS†

Abstract. We consider a bounded variation singular stochastic control problem with value V in a general situation with control of a diffusion and nonlinear cost functional defined as solution to a backward stochastic differential equation (BSDE). Associated with this is a Dynkin game with value u. We establish the well-known relation $\frac{\partial}{\partial x}V = u$ for this general situation. A saddle point for the Dynkin game is given by the pair of first action times of an optimal control.

The methods are from stochastic analysis and include a priori estimates, pathwise construction, and comparison theorems for forward stochastic differential equations (FSDE) and BSDE.

Key words. backward stochastic differential equation, singular stochastic control, optimal stopping, Dynkin game, nonlinear cost functional, comparison theorem for SDE, pathwise construction

AMS subject classifications. Primary, 93E20; Secondary, 49J30, 60G40, 60H10, 60K30, 90B05, 90B50, 91A15

DOI. 10.1137/S0363012903429049

1. Introduction.

Recent years have seen a steeply increasing interest in mathematical finance. Among other theories, the theory of irreversible investment under uncertainty has benefited from this development and received treatment in both mathematical and economical journals (see the references below). Many examples for different kinds of real-life investment problems are collected in the monograph by Dixit and Pindyck [25]. Furthermore, the subject attracts attention because it is intimately related to the behavior and treatment of options. Whereas the option is typical for a single irreversible decision of a small investor, the irreversible investment problem is a large-scale analogue requiring continuous decision making. Both aspects, and the implications for economic equilibrium, are discussed at length by Baldursson and Karatzas [5]. The mathematical counterpart of this "duality" between irreversible investment and options is found in the relation between singular stochastic control and optimal stopping, which has been studied extensively for simple models and, recently, also in some complex situations.

One of the key elements in modern mathematical finance is the use of backward stochastic differential equations (BSDE) for pricing and hedging contingent claims, both in perfect and imperfect markets. With the help of BSDE these problems can be solved in the original probability spaces. Furthermore, they provide the means to overcome some shortcomings in the standard expected utility framework. An overview of the broad range of applications is given by El Karoui, Peng, and Quenez [31].

This paper tries to further extend the mathematical results on singular control and optimal stopping to more general situations. Allowing for partly reversible investment, the related stopping problem becomes a game of optimal stopping or Dynkin game. Cost functionals are defined as solutions to BSDE, so the optimization problem is posed in the context of g-semisolutions, in the terminology of Peng [61].

*Received by the editors June 1, 2003; accepted for publication (in revised form) January 15, 2005; published electronically October 7, 2005. This work is based on the author's Ph.D. thesis, which was supported by Bayerische Treuhandgesellschaft AG, Munich, a member of KPMG International, and by the Centre of Finance and Econometrics at the University of Constance, Germany.

http://www.siam.org/journals/sicon/44-4/42904.html

†KPMG, Ganghoferstr. 29, 80339 Munich, Germany (fboetius@KPMG.com).

There are two major approaches to the problem based on analytical or on probabilistic tools. To arrive at our results we make use of results from stochastic analysis, in particular, comparison theorems and a priori estimates for SDE. Keeping in mind the analytical characterizations in optimal control and optimal stopping, the main theorem, Theorem 1.3 below, also deals with the old relationship of smooth fit in control and the smooth pasting condition in optimal stopping.

1.1. Problem formulation. Assume as given a probability space $(\Omega, \mathcal{F}, \boldsymbol{P})$, a time interval $\bar{\imath} := [0, T]$, and a filtration $\{\mathcal{F}_t\}_{t\in\bar{\imath}}$ satisfying the usual conditions, generated by a standard d-dimensional Brownian motion W_t. For a random starting point η with values in $\mathbb{R}^n$ and initial time $t_0 \in \bar{\imath}$ we consider the *controlled forward backward stochastic differential equations (FBSDE)* on $[t_0, T]$:

$$\begin{aligned} dX_u &= b(u, X_u)\,du - dC_u + \sigma(u, X_u)^\top\,dW_u, \\ dY_u &= -g(u, X_u, Y_u, Z_u)\,du - a_u^\top\,dC_u + Z_u^\top\,dW_u, \\ X_{t_0} &= \eta, \\ Y_T &= \xi = h(X_T). \end{aligned} \tag{1.1}$$

We use the notation $a^\top\,dC := {a^U}^\top\,dC^U - {a^L}^\top\,dC^L$. C is called the control, and a class of admissible controls will be introduced in Definition 2.1. Further assumptions on the data (b, σ, C, h, g, a) will be fixed in Definition 2.2 below, which states conditions for existence and uniqueness of solutions to (1.1) for an n-dimensional state process X. We will focus on $n = 1$ and $\eta = x \in \mathbb{R}$.

Control problem. The solution to (1.1) is denoted by $(X_t^{t_0,x,C}, Y_t^{t_0,x,C}, Z_t^{t_0,x,C})$. Then $X^{t_0,x,C}$ is called the state process, and $Y^{t_0,x,C}$ is the cost associated with control C. To justify terminology, write the backward state component $Y = Y_t^{t_0,x,C}$ in the standard representation

$$Y_t = E\left[h(X_T) + \int_t^T g(s, X_s, Y_s, Z_s)\,ds + \int_{[t,T)} a_s^U\,dC_s^U - \int_{[t,T)} a_s^L\,dC_s^L \middle| \mathcal{F}_t\right].$$

Viewing h, g, and a as the terminal, running, and control cost, the process Y can be seen as the dynamic version of some cost functional dependent on the state process X and the control C exercised. If g is linear in y, then Y_t is obtained by integrating over appropriately discounted costs after time t.

To ease notation we set $t_0 = 0$ and omit this index henceforth.

Definition 1.1. *The control problem comprises determining the value function V and finding an optimal control $C^\star \in \mathcal{A}$ with the property*

$$V(t_0, x) := V(x) := \operatorname*{ess\,inf}_{C\in\mathcal{A}} Y_0^{x,C} = Y_0^{x,C^\star} \qquad [\boldsymbol{P}]. \tag{1.2}$$

The problem is said to be well-posed if $|V(x)| < \infty$.

While $V(x) < \infty$ is obvious, we assume $-\infty < V(x) < \infty$ for all x.

Thus in the control problem one faces the problem of minimizing the cost functional Y, which is defined in an unusual way as solution to a BSDE. Taking a look at its representation above we see that its components are more or less standard in a control problem, namely, a terminal cost h dependent on the state X_T when reaching the time horizon, proportional cost a^U for decreasing X, and proportional recovery of cost a^L for increasing X, and a running cost g. The unusual fact is that the running

cost depends not only on the state of the system but also on the current cost level and the process Z from the martingale representation of Y. Observe that if the running cost is of the form $g(t,x,y,z) = g^1(t,x) + g^2(t)y$, then the effect of g^2 is discounting and Y_0 can be written as a standard expected value without reference to Y_t.

A similar situation in a control problem was considered by Peng [60]. Including Y in the running cost is a step toward using his g-expectation concept [61]. In economic terms, a nonlinear cost g corresponds to a nonadditive stochastic differential utility as introduced by Duffie and Epstein [26]. Its main feature is that, in contrast to standard additive utility, preferences with respect to timing differences need not be induced by discounting. As an example in applications one may consider a situation where the credit rating and therefore the financing cost of an investor depend on the prospects of his business, measured by the expected payoff Y. In financial applications Z has an interpretation as information process or hedging portfolio, whence its inclusion may turn out to be useful.

As the data need not be Markov, b, σ, a, g, h, and therefore V also may depend on ω in an $\mathcal{F}_{t_0}$-measurable way; we omit this dependence in the notation. One might wish to consider also the behavior of the value V as the system evolves, and we actually need to do so when we introduce the associated Dynkin game. Then one has to take the past of ω into account. Observe that an extension of (1.2) by using $\operatorname{ess\,inf}_{C\in\mathcal{A}} Y_t^{x,C}$ for $t>0$ leads to senseless results. To avoid a cumbersome definition we set $V_t(x) := Y_t^{x,C^*}$ for an optimal control C^*, as we will consider such V_t only in the case that there is a solution to the control problem.

Associated stochastic game of optimal stopping. Now we turn to a two-player stochastic game of optimal stopping or Dynkin game associated with the above control problem. It will turn out that its value is the derivative of the value in the control problem and that a pair of optimal stopping times is determined by the first action time of an optimal control.

So denote by $\mathcal{T}_t$ the class of $\mathcal{F}_s$-stopping times with values a.s. in $[t,T]$, and $\mathcal{T} := \mathcal{T}_0$. Assume for a while that b, h, and g are partially differentiable with respect to (x,y,z), σ is linear in x, and g is linear in z (hence Dg is independent of z). Γ^x denotes a deflator process (see (2.2)). For initial condition (t_0,x) (again, we omit t_0) and $\sigma,\tau\in\mathcal{T}_0$ define the payoff $R_t^{t_0,x}(\sigma,\tau) = R_t^x = R_t$, where (R,Q) is the solution of the BSDE

$$\begin{aligned} dR_u &= -\langle Dg(t,X_t^{x,0,0},V_t(x)),(\Gamma_t^x,R_t,Q_t)\rangle \boldsymbol{\chi}_{t\le\sigma\wedge\tau}\,dt + Q_t^\top\,dW_t,\\ R_T &= \Big(h_x(X_T^{x,0,0})\boldsymbol{\chi}_{\sigma\wedge\tau=T} + a_\tau^L \boldsymbol{\chi}_{\substack{\tau\le\sigma\\ \tau<T}} + a_\sigma^U \boldsymbol{\chi}_{\substack{\sigma\le\tau\\ \sigma<T}}\Big)\Gamma^x_{\sigma\wedge\tau\wedge T}. \end{aligned} \tag{1.3}$$

Interpretation. R^x defined through (1.3) has the integral representation

$$\begin{aligned} R_t(\sigma,\tau) = E\bigg[\int_t^{\sigma\wedge\tau} &\langle Dg(s,X_s^{x,0,0},V_s(x)),(\Gamma_s^x,R_s,Q_s)\rangle\,ds\\ &+ h_x(X_T^{x,0,0})\Gamma_T^x\boldsymbol{\chi}_{\sigma\wedge\tau=T} + a_\sigma^U\Gamma_\sigma^x\boldsymbol{\chi}_{\substack{\sigma\le\tau\\ \sigma<T}} + a_\tau^L\Gamma_\tau^x\boldsymbol{\chi}_{\substack{\tau\le\sigma\\ \tau<T}}\,\bigg|\,\mathcal{F}_t\bigg]. \end{aligned} \tag{1.4}$$

We interpret this as a game for two players MIN and MAX: MIN pays MAX at rate $\langle Dg,\cdot\rangle$ as long as the game continues and the amount $h_x\Gamma_T$ upon reaching the time horizon T. If one of the players chooses to terminate early at times σ or τ, MIN pays MAX $a_\sigma^U\Gamma_\sigma$ or $a_\tau^L\Gamma_\tau$, depending on whether MIN or MAX stopped the game.

So MIN seeks to minimize $R^x(\sigma,\tau)$ by choice of σ, whereas MAX wishes to maximize the payoff by choice of τ.

Formally we define the upper and lower values

$$u^+(x) := \operatorname*{ess\,inf}_{\sigma\in\mathcal{T}} \operatorname*{ess\,sup}_{\tau\in\mathcal{T}} R_0^x(\sigma,\tau), \tag{1.5a}$$

$$u^-(x) := \operatorname*{ess\,sup}_{\tau\in\mathcal{T}} \operatorname*{ess\,inf}_{\sigma\in\mathcal{T}} R_0^x(\sigma,\tau). \tag{1.5b}$$

We assume $-\infty < u^-(x)$. By the definition of u^+ and u^-, $u^- \le u^+$ holds $\boldsymbol{P}$-a.s.

A solution of the game consists of a pair of stopping times such that neither of the two players has an incentive to deviate from his strategy. The optimization of the opponent is anticipated, and the right of choosing one's strategy first gives no advantage, i.e., has no additional value, which is the meaning of the Isaac's equation (1.6) below.

DEFINITION 1.2. *The associated stochastic game of optimal stopping or Dynkin game consists of determining the upper and lower values u^+ and u^- and the value function u as solution to the Isaac's equation*

$$u(x) := u^+(x) = u^-(x), \tag{1.6}$$

and of finding a saddle point $(\sigma^\star,\tau^\star)\in\mathcal{T}\times\mathcal{T}$, i.e., stopping times such that

$$\begin{aligned} u(x) &= R_0^x(\sigma^\star,\tau^\star) \\ &= \operatorname*{ess\,inf}_{\sigma\in\mathcal{T}} R_0^x(\sigma,\tau^\star) = \operatorname*{ess\,sup}_{\tau\in\mathcal{T}} R_0^x(\sigma^\star,\tau) \qquad [\boldsymbol{P}]. \end{aligned} \tag{1.7}$$

To obtain the relation between the two problems we will make some convexity (2.8) and smoothness (3.15) assumptions. These could be relaxed up to some point.

Our main result is the following theorem.

THEOREM 1.3. *Let the dimension of the state process be $n = 1$ and assume that* (2.8) *and* (3.15) *hold and that there exists an optimal control $C = (C^U, -C^L)$ for the control problem introduced in Definition* 1.1 *starting at x. Then the value function of the control problem V is partially differentiable at x with respect to the space variable. The associated Dynkin game introduced in Definition* 1.2 *has a solution, and its value is the partial derivative of the value of the control problem:*

$$\frac{\partial}{\partial x} V(x) = u(x). \tag{1.8}$$

Let $\sigma^0 := \inf\{t \ge 0 \,|\, C_t^U > C_0^U\}$ and $\tau^0 := \inf\{t \ge 0 \,|\, C_t^L > C_0^L\}$ denote the first action times of the controls C^U and C^L. They form a saddle point for the Dynkin game which has value

$$u(x) = R_0^x(\sigma^0,\tau^0). \tag{1.9}$$

1.2. Remarks and references. Singular stochastic control problems, their connection with a problem of optimal stopping, and the relation $\frac{\partial}{\partial x}V = u$ date back to an investigation of spaceship control by Bather and Chernoff [7]. The optimization problem received much interest after the work of Beneš, Shepp, and Witsenhausen [11], who were able to obtain some explicit solutions. The methods employed in their work and by several authors thereafter include characterizations of the value as a

solution to a Hamilton–Jacobi–Bellman PDE restricted to a bounded domain, with terminal condition and gradient constraints at the boundaries. If systems of this type are solved directly, additional information about the behavior of the value at the boundary is required. This usually takes the form of the *principle of smooth fit* stating that the value should be twice continuously differentiable across the boundary. Alternatively, one tries to derive the value from solutions to a related free boundary problem; thus in essence one establishes a relation with a stopping problem of the form $\frac{\partial}{\partial x}V = u$, where it is known as the *principle of smooth pasting* that the value u of the stopping problem is once continuously differentiable across the boundary. This was made rigorous by Karatzas [42].

In contrast to this analytical approach, probabilistic methods, namely, a pathwise construction, were introduced by Karatzas and Shreve [44] for the link between monotone follower and optimal stopping of Brownian motion. The rule for deriving an optimal stopping time as first action time of an optimal control is also due to [44]. These results were extended by the same authors in [45], [46] and also by Karatzas [43], Baldursson [4], El Karoui and Karatzas [29], [30], Baldursson and Karatzas [5], and Boetius and Kohlmann [18], among others. In the latter paper, instead of a monotone follower problem, the control of a diffusion is considered. Therefore exercise of control feeds back into the dynamics of the system, requiring the use of a comparison theorem.

The relation between singular stochastic control and Dynkin games was, to our knowledge, first noted by Taksar [67]. A result similar to ours for control of Brownian motion and very general cost data was obtained by Karatzas and Wang [49]. An extension to infinite time horizon and additional absolutely continuous control ("mixed control" and "mixed game" problems) is treated by Hamadene, Lepeltier, and Wu [39]. Theorem 1.3 focuses on the relation of the value functions and optimal strategies, respectively, saddle point, for the general case of bounded variation control of a diffusion with stochastic dynamics and cost structure and with a cost functional defined as solution to a nonlinear BSDE. Again, the basic ideas are those of Karatzas and Shreve [44], but the general formulation especially for the cost functional in the problem poses some technical difficulties and requires the frequent use of comparison theorems for BSDE.

It should be noted that the pathwise approach of [44] breaks down if control may be exercised in more than one dimension, as there is no extension of the powerful tool of comparison theorems in that case. This is due to the phenomenon of coalescing stochastic flows; see, e.g., Baxendale [8] and Darling [22]. Consequently, singular control problems in higher dimensions are inherently difficult. Results for that case, including existence of optimal controls, were obtained by Menaldi and Taksar [59], Soner and Shreve [66], and Kruk [52], [53], to mention but a few.

We should note that existence of an optimal control is an essential element in this paper. A simple example for nonexistence of an optimal control in a finite horizon problem by Lalley is presented in Karatzas and Shreve [44, p. 863], who also provide a nonconstructive general existence result for monotone control problems using weak-$*$ compactness arguments. Compactness arguments relying on strong growth assumptions for the data are used by Haussmann and Suo [40], and Ma and Yong [55] provide a proof for the case when cost data grow less than linearly. A general existence result for bounded variation control by Karatzas and Wang [49] makes use of a result of Komlós [51] on L^1-convergence of the Cesàro sequences of ϵ-optimal processes. Hamadene and Lepeltier [37], [36] obtain existence of a solution to both, the control problem and the Dynkin game, using BSDE if the Isaac's condition is satisfied.

Properties of optimal controls and state processes can be found in, e.g., Fleming and Soner [33, p. 360] and have been subject to a number of articles by several authors.

Besides the problems of irreversible or partially reversible investment mentioned already, applications extend from the original problem of spaceship control in [7] to queueing systems, control of storage or manufacturing systems, mathematical biology, optimal dividend distribution and risk control (in particular, in the insurance industry), optimal consumption and investment, portfolio optimization under transaction costs, and hedging and pricing under constraints. Some of the more recent articles in this context are those by Højgaard and Taksar [41], Cadenillas [19], Alvarez [2], Kushner [54], Benth, Karlsen, and Reikvam [13], and Schmock, Shreve, and Wystup [64], to mention but a few. We should point to Davis and Norman [24] for their treatment of a control problem in two dimensions. Further references are given, e.g., in [17].

The valuation of American contingent claims is one of the most prominent examples of optimal stopping problems in finance; the problem was first investigated in Samuelson [63] and McKean [58]. An account of this valuation problem is given, e.g., in Karatzas and Shreve [48], which we also cite for further references.

Further applications are given in the field of real options, where they usually exhibit the phenomenon of hysteresis or the "value of waiting to invest"; see, e.g., McDonald and Siegel [57] and especially Dixit and Pindyck [25] for an account of investment decisions where real options play an important role. Recent applications for technology adoption or credit default are given in, e.g., Alvarez [1] and Alvarez and Stenbacka [3].

Applications of Dynkin games in financial markets are considered in Ma and Yong [56] and Kifer [50]. Games and their applications in economics are widely discussed by Fudenberg and Tirole [35]. More recently it was discovered that the value of an optimal stopping problem or a Dynkin game can be represented as the solution to an appropriately reflected BSDE; see the articles by Cvitanić and Karatzas [21] and El Karoui, Kapoudjian, Pardoux, Peng, and Quenez [28]. Hamadene and Lepeltier [38] obtained the saddle point strategy for the mixed game problem when the Isaac's condition is satisfied.

There is a vast general literature on problems of optimal stopping, and one may refer to, e.g., Friedman [34, Chap. 16], Shiryaev [65], and Bensoussan and Lions [12] for an analytical approach and to Davis and Karatzas [23] for a probabilistic approach. The game of stopping under consideration was introduced by Dynkin [27] and studied further by Bismut [14], [15] and others. Further references are given in [17].

This paper is organized as follows. In section 1 we introduce the optimization problems, state the main result, and note some applications. Section 2 is devoted to notation and the basic tools from stochastic analysis. It also contains some consequences of the convex structure in the control problem. Section 3 contains the proof of Theorem 1.3. Here we perform the pathwise analysis and obtain the crucial estimates for the difference quotient of cost functionals. Finally, section 4 takes a look at several possible extensions.

2. Preliminaries. We now introduce in more detail terminology and basic assumptions (subsection 2.1), recall a priori estimates and comparison theorems (subsection 2.2) that are essential to many arguments, and discuss some consequences of the convex structure (subsection 2.3) that we will investigate.

2.1. Basic notation. We use the following notation as defined in Ma and Yong [56] for the different spaces of measurable random variables:

- For a σ-algebra, $\mathcal{G} \subset \mathcal{F}_T$ $L^p_{\mathcal{G}}(\Omega;\mathbb{R}^k)$ denotes the set of $\mathcal{G}$-measurable $\mathbb{R}^k$-valued random variables X such that $E\big[|X|^p\,dt\big] < \infty$.
- $L^p_{\mathcal{F}}(\Omega; L^p(0,T;\mathbb{R}^k))$ denotes the set of $\mathcal{F}_t$-progressively measurable $\mathbb{R}^k$-valued processes X_t such that $E\big[\int_0^T |X_t|^p\,dt\big] < \infty$; we write $L^p_{\mathcal{F}}(0,T;\mathbb{R}^k)$ if there is no danger of confusion.
- $L^p_{\mathcal{F}}(\Omega; C(\bar{\imath};\mathbb{R}^k))$ denotes the set of $\mathcal{F}_t$-progressively measurable continuous $\mathbb{R}^k$-valued processes X_t such that $E\big[\sup_{\bar{\imath}}|X_t|^p\big] < \infty$.
- $W^{1,\infty}(M,N)$ for Euclidean spaces M, N denotes the set of functions $f : M \to N$ that are Lipschitz continuous w.r.t. the metrices derived from the scalar products.
- $L^p_{\mathcal{F}}(0,T;W^{1,\infty}(M,N))$ for Euclidean spaces M, N denotes the set of functions $f : \bar{\imath} \times M \times \Omega \to N$ such that (a) for fixed $m \in M$, $(t,\omega) \mapsto f(t,m,\omega)$ is $\mathcal{F}_t$-progressively measurable, (b) $f(t,0,\omega) \in L^p_{\mathcal{F}}(0,T;N)$, and (c) there exists a constant $L \in \mathbb{R}_{>0}$ such that

$$\big|f(t,m,\omega) - f(t,m',\omega)\big| \le L|m-m'| \qquad \forall m,m' \in M,\ \text{a.e. } t \in \bar{\imath},\ \boldsymbol{P}\text{-a.s.}$$

- $L^p_{\mathcal{F}_T}(\Omega; W^{1,\infty}(M,N))$ for Euclidean spaces M, N denotes the set of functions $f : M \times \Omega \to N$ such that for any $m \in M$, $\omega \mapsto f(m,\omega)$ is $\mathcal{F}_T$-measurable, $m \mapsto f(m,\omega)$ is uniformly Lipschitz, and $f(0,\omega) \in L^p_{\mathcal{F}}(\Omega;N)$.

$\|\cdot\|_p$ denotes the usual p-norm in the spaces defined above.

For an $\mathbb{R}^k$-valued process $B = (B^j)^\top_{1\le j\le k}$ of bounded variation its absolute variation is denoted by $|B| := \sum_{j=1}^k |B^j|$; i.e., if $\mu_j = \mu_j^+ - \mu_j^-$ is the (ω-dependent) signed measure (with Hahn-decomposition (μ_j^+, μ_j^-)) defined by

$$\mu_j(A) := \int \boldsymbol{\chi}_A\,dB^j, \qquad \text{then} \qquad \int \boldsymbol{\chi}_A\,d|B| = \sum_{j=1}^k (\mu_j^+(A) + \mu_j^-(A)).$$

For an $\mathbb{R}^k$-valued progressively measurable process $a = (a^j)_{1\le j\le k}$ and an $\mathbb{R}^k$-valued bounded variation process $C = (C^j)_{1\le j\le k}$ we use—depending on whether we focus on a norm for a or C—the notation

$$\|a\|^2_{[C,\bar{\imath}]} := E\left[\sum_{j=1}^k \left(\int_{\bar{\imath}} |a^j_s|\,d|C^j|_s\right)^2\right] =: |C|^2_{[a,\bar{\imath}]}. \tag{2.1}$$

Using the Hölder inequality we obtain the estimates

$$E\left[\left(\int_{\bar{\imath}} a_s^\top\,dC_s\right)^2\right] \le k\|a\|^2_{[C,\bar{\imath}]} = k|C|^2_{[a,\bar{\imath}]} \le kE\left[\sum_{j=1}^k \left(|C^j|_T \int_{\bar{\imath}} |a^j(s)|^2\,d|C^j|_s\right)\right].$$

For a function $f : \mathbb{R} \to \mathbb{R}$ we define the Dini derivatives

$$\Delta^\pm f(x) := \limsup_{\delta \searrow 0} \frac{f(x\pm\delta) - f(x)}{\pm\delta}, \qquad \Delta_\pm f(x) := \liminf_{\delta \searrow 0} \frac{f(x\pm\delta) - f(x)}{\pm\delta}$$

and the one-sided differentials $D^+f := \Delta^+ f = \Delta_+ f$ and $D^-f := \Delta^- f = \Delta_- f$ if they exist. Note that D^+f and D^-f exist for convex functions f, are left continuous and right continuous, respectively, and satisfy $D^-f(x) \le D^+f(x) \le D^-f(y)$ for $x < y$ (cf. problem 3.6.19 of Karatzas and Shreve [47]).

DEFINITION 2.1. *Denote by $\mathcal{A}$ the class of admissible controls. It consists of all $\mathbb{R}$-valued, $\{\mathcal{F}_t\}$-adapted processes C with paths $\boldsymbol{P}$-a.s. left continuous with right-hand limits (LCRL) and of bounded variation satisfying $|C|^2_{[\mathbf{1},\bar{\imath}]} = E\big[|C|^2_T\big] < \infty$. We write C decomposed in $C = C^U - C^L$, where C^U, C^L are nonnegative, increasing, and LCRL, and the decomposition is minimal. We also use the notation $C = (C^U, -C^L)$.*

DEFINITION 2.2. *Data (b, σ, C) for the forward and (h, g, a, C) for the backward equation in (1.1) are called standard data if they satisfy*

- *$b \in L^2_{\mathcal{F}}(0, T; W^{1,\infty}(\mathbb{R}, \mathbb{R}))$ and $\sigma \in L^2_{\mathcal{F}}(0, T; W^{1,\infty}(\mathbb{R}, \mathbb{R}^d))$;*
- *$C \in \mathcal{A}$ and $a = (a^U, a^L) : \bar{\imath} \times \Omega \to \mathbb{R}^2$ progressively measurable such that $a^U \geq a^L$ and for all $\tilde{C} \in \mathcal{A}$ the process*

$$\int_{[0,t)} a_s^\top \, d\tilde{C}_s := \int_{[0,t)} a_s^U \, d\tilde{C}_s^U - \int_{[0,t)} a_s^L \, d\tilde{C}_s^L$$

  is in $L^2_{\mathcal{F}}(0, T; \mathbb{R})$ and has LCRL paths $\boldsymbol{P}$-a.s.;
- *$g \in L^2_{\mathcal{F}}(0, T; W^{1,\infty}(\mathbb{R} \times \mathbb{R} \times \mathbb{R}^d, \mathbb{R}))$ and $h \in L^2_{\mathcal{F}_T}(\Omega; W^{1,\infty}(\mathbb{R}, \mathbb{R}))$.*

So apart from sufficient measurability we require uniform Lipschitz continuity of the data (b, σ, g, h) and square integrability and we consider bounded variation control. The processes involved need not be Markov.

Note that (1.1) is well defined even for decompositions of C that are not minimal. At times, we will work with such decompositions as well. From the requirements on a it is clear that $Y^{x,C^U,C^L} \leq Y^{x,\bar{C}^U,\bar{C}^L}$ if (C^U, C^L) and $(\bar{C}^U, \bar{C}^L)$ are nonnegative, increasing, LCRL decompositions of the same process $C \in \mathcal{A}$ and (C^U, C^L) is minimal.

Using a transformation as in Peng [61], which shifts the control C from the equation into the data, it is easy to see that (1.1) has a unique solution $(X, Y, Z) \in L^2_{\mathcal{F}}(0, T; \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^d)$. See also the construction in subsection 3.2, where this transformation is applied expressis verbis. Dependence on data and parameters η, b, σ, C, ξ, h, a, g may be expressed by a superscript, and we also refer to the components of C, in particular when using a certain decomposition. Results on existence and uniqueness of solutions to FSDE can be found in, e.g., Karatzas and Shreve [47]; for BSDE one may consider, e.g., El Karoui, Peng, and Quenez [31] or the monographs by Yong and Zhou [68, Chap. 7] or Ma and Yong [56].

Sometimes we consider the backward equation without reference to a particular forward process. We then omit dependence of g on x and write ξ for the terminal condition, with suitably adapted meaning of standard data.

We further introduce *deflator processes* to refer to some flow properties of solutions to the forward equation in (1.1), under the additional convexity assumptions in (2.8) below. In effect we will require σ to be affine and b to be convex in x; thus $\sigma_x(t)$ and $D_x^+ b$ are well defined. The biased deflator process Γ^{x,C^U,C^L} for a control $(C^U, -C^L) \in \mathcal{A}$ is given as solution to the FSDE

$$\begin{aligned} d\Gamma_t^{x,C^U,C^L} &= \Gamma_t^{x,C^U,C^L} \big(D_x b(t, X_t^{x,C^U,C^L}) \, dt + \sigma_x(t)^\top \, dW_t \big), \\ \Gamma_0^{x,C^U,C^L} &= 1. \end{aligned} \tag{2.2}$$

Its uncontrolled version is abbreviated as $\Gamma^x := \Gamma^{x,0,0}$, which serves as deflator for the payoff in the Dynkin game (1.3). Properties of the deflator processes are discussed in more detail in subsection 2.3.

2.2. Comparison theorems and a priori estimates. Comparison theorems lie at the heart of the pathwise approach taken in this paper, and in some instances

we need a priori estimates for BSDE to obtain convergence results. Due to the impact of controls they appear in a slightly different form from those known commonly in the literature.

LEMMA 2.3 (a priori estimates for FSDE). *For $i \in \{1,2\}$ let (b^i, σ^i, C^i) denote standard data, let $x^i \in \mathbb{R}$, and let X^i be the corresponding solutions of the FSDE. Define*

$$\begin{aligned} &\boldsymbol{\delta} X_t := X_t^1 - X_t^2, && \boldsymbol{\delta} x := x^1 - x^2, && \boldsymbol{\delta} C_t = C_t^1 - C_t^2, \\ &\boldsymbol{\delta} b_t := b^1(t, X_t^2) - b^2(t, X_t^2), && \boldsymbol{\delta}\sigma_t := \sigma^1(t, X_t^2) - \sigma^2(t, X_t^2). \end{aligned}$$

Then there is a constant K_4 depending only on T and L such that

$$E\big[\sup_{\bar\imath} |\boldsymbol{\delta} X_t|^2\big] \le K_4\big(|\boldsymbol{\delta} x|^2 + \|\boldsymbol{\delta} b\|_2^2 + \|\boldsymbol{\delta}\sigma\|_2^2 + E\big[|\boldsymbol{\delta} C|_T^2\big]\big). \tag{2.3}$$

LEMMA 2.4 (a priori estimates for BSDE). *Let, for $i \in \{1,2\}$, (Y^i, Z^i) be the solutions of the BSDE* (1.1) *corresponding to standard data (g^i, ξ^i, a^i, C^i), and define*

$$\begin{aligned} &\boldsymbol{\delta} Y_t := Y_t^1 - Y_t^2, && \boldsymbol{\delta} Z_t := Z_t^1 - Z_t^2, && \boldsymbol{\delta}\xi := \xi^1 - \xi^2, \\ &\boldsymbol{\delta} g_t := g^1(t, Y_t^2, Z_t^2) - g^2(t, Y_t^2, Z_t^2), && \boldsymbol{\delta} a_t := a_t^1 - a_t^2, && \boldsymbol{\delta} C_t := C_t^1 - C_t^2. \end{aligned}$$

There are constants depending only on L and T such that for all $t \le T$

$$E\left[\int_t^T |\boldsymbol{\delta} Z_s|^2\, ds\right] \le K_2\big(\|\boldsymbol{\delta}\xi\|_2^2 + \|\boldsymbol{\delta} g\|_2^2 + \|\boldsymbol{\delta} a\|_{[C^1,\bar\imath]}^2 + |\boldsymbol{\delta} C|_{[a^2,\bar\imath]}^2\big)\big(1 + e^{K_3(T-t)}\big) \tag{2.4}$$

and the running maximum of differences satisfies

$$E\big[\sup_{\bar\imath} |\boldsymbol{\delta} Y_t|^2\big] \le K_4\big(\|\boldsymbol{\delta}\xi\|_2^2 + \|\boldsymbol{\delta} g\|_2^2 + \|\boldsymbol{\delta} a\|_{[C^1,\bar\imath]}^2 + |\boldsymbol{\delta} C|_{[a^2,\bar\imath]}^2\big). \tag{2.5}$$

The proof for both lemmata is along the lines of El Karoui, Peng, and Quenez [31] or Peng [61], the main tools being the Itô formula, and the Burkholder–Davis–Gundy, Hölder, and Gronwall inequalities. It may be worthwhile noting that an application of the latter requires a.e. continuity of the functions involved. We do not encounter difficulties in using the Gronwall inequality as the maps $t \mapsto E\big[|\boldsymbol{\delta} X_t|^2\big]$ and $t \mapsto E\big[|C|_t\big]$ are continuous a.e. in $\bar\imath$. Details are given in [17]. An even stronger result on jumps of right continous with left-hand limits (RCLL) processes can be found in Proposition 1.2.26 of Karatzas and Shreve [47].

THEOREM 2.5 (comparison theorem for FSDE). *Consider, for $i \in \{1,2\}$, the solutions X^i of the FSDE corresponding to standard data (b^i, σ^i, C^i) and initial condition $x^i \in \mathbb{R}$. Assume further that*

$$\begin{aligned} &b^1(\omega,t,x) \ge b^2(\omega,t,x), && \sigma^1(\omega,t,x) = \sigma^2(\omega,t,x), \\ &-d(\boldsymbol{\delta} C_t) \ge 0 \quad \textit{(i.e., } \boldsymbol{\delta} C \textit{ decreasing)}, && \boldsymbol{\delta} x \ge 0. \end{aligned}$$

Then the difference $\boldsymbol{\delta} X_t$ is positive a.s.:

$$X_t^1 \ge X_t^2 \quad \forall t \in \bar\imath \quad [\boldsymbol{P}].$$

COROLLARY 2.6. *It is easy to see that instead of $b^1 \ge b^2$ it is actually sufficient if one of the conditions*

$$b^1(\omega,t,X_t^1) \ge b^2(\omega,t,X_t^1), \qquad b^1(\omega,t,X_t^2) \ge b^2(\omega,t,X_t^2) \tag{2.6}$$

holds. If, e.g., b^1 is independent of x, this reduces to $b^1(\omega,t) \ge b^2(\omega,t,X_t^1)$.

The comparison theorem for the backward equation in (1.1) is quite similar.

THEOREM 2.7 (comparison theorem for BSDE). *Let, for $i \in \{1,2\}$, (g^i, ξ^i, a^i, C^i) denote standard data and (Y^i, Z^i) denote the corresponding solutions of the BSDE. a^i may depend on Y. Assume further that $\boldsymbol{P}$-a.s.*

$$g^1(t,y,z) - g^2(t,y,z) \geq 0, \qquad \xi^1 - \xi^2 \geq 0, \tag{2.7a}$$

$$a^1(t)^\top d(C^1_t - C^2_t) \geq 0, \qquad (a^1(t) - a^2(t))^\top dC^2_t \geq 0. \tag{2.7b}$$

Then the difference $\boldsymbol{\delta} Y_t := Y^1_t - Y^2_t$ is nonnegative $\boldsymbol{P}$-a.s.

Both theorems can be proved using the linearization arguments of El Karoui, Peng, and Quenez [31] or El Karoui and Quenez [32], which only need to be adapted to include the control process C. Note that the equality $\sigma^1 = \sigma^2$ is vital for this argument in the forward case.

It follows from the arguments in [31], [32] that, instead of the first inequality in (2.7a), it is sufficient if one of the conditions

$$g^1(t, Y^1_t, Z^1_t) - g^2(t, Y^1_t, Z^1_t) \geq 0, \quad g^1(t, Y^2_t, Z^2_t) - g^2(t, Y^2_t, Z^2_t) \geq 0 \tag{2.7c}$$

holds. This will prove useful in situations where we make use of the convexity of the driver $g(t,x,y,z)$. In situations where, e.g., g^1 does not depend on (Y^1, Z^1), we conclude that $g^1_t \geq g^2(t, Y^1_t, Z^1_t)$ is also sufficient.

Instead of (2.7b) it is actually sufficient if we have

$$\int_{[0,t)} a^1(s)^\top dC^1_s - \int_{[0,t)} a^2(s)^\top dC^2_s \quad \text{is increasing in } t. \tag{2.7d}$$

Hence the theorem remains true if (2.7b) is replaced by (2.7d) or

$$a^2(t)^\top d(C^1_t - C^2_t) \geq 0, \qquad (a^1(t) - a^2(t))^\top dC^1_t \geq 0. \tag{2.7e}$$

Often we can use properties of Y^i when verifying (2.7a), e.g., $Y^i \geq 0$, because then we can restrict the estimate to $y \in \mathbb{R}_{\geq 0}$. On the contrary, we usually know very little about the processes Z^i. Hence, if one of the drivers is linear in z, (2.7a) in effect requires that the partial derivatives g^1_z, g^2_z agree. This will force us to assume that g_z is independent of x, y, z, as in the associated Dynkin game the driver takes the form $\langle (g_x, g_y, g_z), (\Gamma, R, Q) \rangle$.

Note that all of the above results still hold if we use $C^i = (C^{i,U}, C^{i,L})$ where the decomposition is not minimal.

2.3. Convex structure of the control problem. An important prerequisite for our proof of the relation $\frac{\partial}{\partial x} V = u$ is convexity of the value V, which will be discussed in Theorem 2.10 below. We also include some useful estimates and comparison results for controlled forward processes and biased deflators in section 2.4 below.

DEFINITION 2.8. *Let the dimension of the state space be $n = 1$. The control problem* (1.2) *has convex structure if the data satisfy*

$$b \text{ is convex in the space variable } x, \tag{2.8a}$$

$$\sigma \text{ is affine in the space variable } x, \tag{2.8b}$$

$$h \text{ is convex and increasing in the space variable } x, \tag{2.8c}$$

$$g \text{ is convex in the variables } (x,y,z) \text{ and increasing in } x. \tag{2.8d}$$

It should be noted that in the case of affine dynamics b, σ in the forward equation, monotonicity of h and g in x is not required. In fact, most arguments in the discussion below remain unchanged or could even be simplified in that case.

Let, for $i \in \{1,2\}$, (b,σ,C^i,g,h,a) denote standard data for the controlled forward-backward system (1.1), and let $x^i \in \mathbb{R}$ and $\lambda \in [0,1]$. Denote, for $j \in \{1,2,\lambda\}$,

$$x^\lambda = \lambda x^1 + (1-\lambda)x^2, \qquad C^\lambda = \lambda C^1 + (1-\lambda)C^2,$$
$$X^j = X^{x^j,b,C^j,\sigma}, \qquad (Y^j,Z^j) = (Y,Z)^{g,h(T,X_T^j),a,C^j}.$$

Thus X^λ starts at a convex combination of x^1, x^2 and is controlled by a convex combination of controls C^1, C^2.

PROPOSITION 2.9. *Assume* (2.8). *Then the following inequalities hold* $\boldsymbol{P}$*-a.s. for all* $t \in \bar{\imath}$*:*

$$\bar{X}_t^\lambda := \lambda X_t^1 + (1-\lambda)X_t^2 \geq X_t^\lambda, \tag{2.9a}$$
$$\bar{Y}_t^\lambda := \lambda Y_t^1 + (1-\lambda)Y_t^2 \geq Y_t^\lambda. \tag{2.9b}$$

The proof can be done for the forward and backward cases separately: Starting with the forward equation, let

$$\bar{b}_t^\lambda := \lambda b(t,X_t^1) + (1-\lambda)b(t,X_t^2)$$

denote the driver of $\bar{X}^\lambda$. Then $\bar{b}_t^\lambda \geq b(t,\bar{X}_t^\lambda)$ as b is convex, which suffices according to Corollary 2.6. Now observe that we just need to prove (2.9b) with Y^λ defined by the (not necessarily minimal) decomposition $\lambda(C^{1,U},C^{1,L}) + (1-\lambda)(C^{2,U},C^{2,L})$; here $\boldsymbol{\delta}C = 0$. Defining the driver $\bar{g}^\lambda$ for $\bar{Y}^\lambda$ analogously and using (2.9a), (2.8c), and (2.8d) we obtain (2.9b).

The linearity assumption on σ may not be dropped easily, as the following example shows. Let $b = C = 0$ and $\sigma(t,x) = |x|$. Then the FSDE has a unique strong solution for each initial value $x \in \mathbb{R}$. Define X^i by

$$X_t^1 = 1 + \int_0^t |X_s^1|\,dW_s, \qquad X_t^2 = -1 + \int_0^t |X_s^2|\,dW_s;$$

hence

$$X_t^1 = \exp(W_t - \tfrac{1}{2}t), \qquad X_t^2 = -\exp(-W_t - \tfrac{1}{2}t).$$

Let $\lambda = \frac{1}{2}$; then $X_t^\lambda = 0$ is the solution of $X_t^\lambda = \int_0^t |X_s^\lambda|\,dW_s$. But

$$2\bar{X}_t^\lambda = \exp\left(-\frac{1}{2}t\right)\big(\exp(W_t) - \exp(-W_t)\big),$$

so $\boldsymbol{P}\{\bar{X}_t^\lambda > X_t^\lambda\} = \boldsymbol{P}\{\bar{X}_t^\lambda < X_t^\lambda\} = \frac{1}{2}$.

We then have as an easy consequence of (2.9) the following theorem.

THEOREM 2.10 (convexity of the value). *Assume that* (2.8) *holds. Then* V *is convex with respect to the starting point; i.e., if* $x^1, x^2 \in \mathbb{R}$ *and* $\lambda \in [0,1]$*, then*

$$V(\lambda x^1 + (1-\lambda)x^2) \leq \lambda V(x^1) + (1-\lambda)V(x^2).$$

The proof is analogous to the arguments in Karatzas and Shreve [44].

2.4. Properties of controlled forward processes. In this section we use the comparison theorem to work out some properties of the state process under a convex structure. We analyze differences of state processes and recall a result from the theory of stochastic flows on differentiability with respect to the initial condition.

Recall that the biased deflator process Γ^{x,C^U,C^L} for control $(C^U,-C^L)$ (with the uncontrolled version abbreviated as $\Gamma^x := \Gamma^{x,0,0}$) of (2.2) is the solution to the FSDE

$$\begin{aligned} d\Gamma_t^{x,C^U,C^L} &= \Gamma_t^{x,C^U,C^L}\big(D_x^+ b(t,X_t^{x,C^U,C^L})\,dt + \sigma_x(t)^\top\,dW_t\big),\\ \Gamma_0^{x,C^U,C^L} &= 1. \end{aligned}$$

When there is no ambiguity we also write $\Gamma^{x,C} := \Gamma^{x,C^U,C^L}$, as we do for controlled processes $X^{x,C}$. We further define geometric Brownian motions Γ^{up} and Γ^{lo}:

$$d\Gamma_t^{up} = \Gamma_t^{up} L\,dt + \Gamma_t^{up}\sigma_x(t)^\top\,dW_t, \qquad \Gamma_0^{up} = 1, \tag{2.10}$$

$$d\Gamma_t^{lo} = -\Gamma_t^{lo} L\,dt + \Gamma_t^{lo}\sigma_x(t)^\top\,dW_t, \qquad \Gamma_0^{lo} = 1. \tag{2.11}$$

Here L is a Lipschitz constant for b. These processes form universal bounds for "difference quotients" and differentials of the forward process.

LEMMA 2.11. *Let (b,σ,C) denote standard data of the FSDE satisfying* (2.8a) *and* (2.8b). *Then the following estimates hold for all $t\in\bar{\imath}$ a.s.:*

$$X_t^{x+\delta,C^U,0} - X_t^{x,C^U,0} \le X_t^{x+\delta,0,0} - X_t^{x,0,0} \le X_t^{x+\delta,0,C^L} - X_t^{x,0,C^L}, \tag{2.12}$$

$$\delta\Gamma_t^{lo} \le \delta\Gamma_t^{x,C^U,C^L} \le X_t^{x+\delta,C^U,C^L} - X_t^{x,C^U,C^L} \le \delta\Gamma_t^{x+\delta,C^U,C^L} \le \delta\Gamma_t^{up}, \tag{2.13}$$

$X_t^{x+\delta,C^U,0} - X_t^{x,C^U,0} \ge 0$, *and* $\Gamma_t^{lo} > 0$. *Furthermore, we have*

$$\lim_{\delta\searrow 0}\frac{1}{\delta}\big(X_t^{x+\delta,C} - X_t^{x,C}\big) = \Gamma_t^{x,C}. \tag{2.14}$$

Proof. Nonnegativity of the first term in (2.12) follows from the comparison theorem, Theorem 2.5, as both processes have the same driver and controls.

For the first inequality in (2.12), define the stopping time

$$\rho := \inf\{t\in\bar{\imath}\,|\,X_t^{x+\delta,C^U,0} \le X_t^{x,0,0}\}$$

and observe that, due to the left continuity of both processes and the comparison theorem (Theorem 2.5), ρ separates the intervals where the difference of these processes is positive and negative, respectively:

$$\begin{aligned} X_t^{x+\delta,C^U,0} &\ge X_t^{x,0,0} \qquad \text{if } t\le\rho,\\ X_t^{x+\delta,C^U,0} &\le X_t^{x,0,0} \qquad \text{if } t>\rho. \end{aligned}$$

Now set $\hat{X}^1 := X^{x+\delta,0,0} - X^{x,0,0}$ and $\hat{X}^2 := X^{x+\delta,C^U,0} - X^{x,C^U,0}$. For $t\le\rho$ consider $\tilde{X}^1 := X^{x+\delta,0,0} - X^{x+\delta,C^U,0}$ and $\tilde{X}^2 := X^{x,0,0} - X^{x,C^U,0}$. As $\hat{X}^1 - \hat{X}^2 = \tilde{X}^1 - \tilde{X}^2$ it suffices to prove $\tilde{X}^1 \ge \tilde{X}^2$. Both processes are nonnegative and have the same starting point 0 and control process 0. Their diffusion matrices are $\sigma(s,\tilde{X}_s^1)$ and $\sigma(s,\tilde{X}_s^2)$. Their drivers $\tilde{b}^1$ and $\tilde{b}^2$ satisfy

$$\begin{aligned} \tilde{b}_s^1 &= b(s,X_s^{x+\delta,0,0}) - b(s,X_s^{x+\delta,C^U,0}) \ge D_x^+ b(s,X_s^{x+\delta,C^U,0})\tilde{X}_s^1,\\ \tilde{b}_s^2 &= b(s,X_s^{x,0,0}) - b(s,X_s^{x,C^U,0}) \le D_x^- b(s,X_s^{x,0,0})\tilde{X}_s^2, \end{aligned}$$

due to the convexity of b in the space variable. As $D_x^+ b(s, X_s^{x+\delta,C^U,0}) \geq D_x^- b(s, X_s^{x,0,0})$ for $s \leq \rho$, the first inequality in (2.12) for $t \leq \rho$ follows from the comparison theorem (Theorem 2.5) and Corollary 2.6.

For $t > \rho$ consider $\hat{X}^1$ and $\hat{X}^2$. They are nonnegative, satisfy $\hat{X}_\rho^1 - \hat{X}_\rho^2 \geq 0$ by the above argument, and have the same control process 0. Their diffusion matrices are $\sigma(s, \hat{X}_s^1)$ and $\sigma(s, \hat{X}_s^2)$. Their drivers $\hat{b}^1$ and $\hat{b}^2$ satisfy

$$\begin{aligned}
\hat{b}_s^1 &= b(s, X_s^{x+\delta,0,0}) - b(s, X_s^{x,0,0}) \geq D_x^+ b(s, X_s^{x,0,0}) \hat{X}_s^1, \\
\hat{b}_s^2 &= b(s, X_s^{x+\delta,C^U,0}) - b(s, X_s^{x,C^U,0}) \leq D_x^- b(s, X_s^{x+\delta,C^U,0}) \hat{X}_s^2,
\end{aligned}$$

due to the convexity of b in the space variable. Then $\hat{X}_t^1 \geq \hat{X}_t^2$ for $t > \rho$ follows from $D_x^+ b(s, X_s^{x,0,0}) \geq D_x^- b(s, X_s^{x+\delta,C^U,0})$ and Corollary 2.6. This completes the proof of the first inequality in (2.12). □

For the second inequality consider the stopping time

$$\varrho := \inf\{t \in \bar{\imath} \,|\, X_t^{x+\delta,0,0} \leq X_t^{x,0,C^L}\}$$

and define the processes $\bar{X}^1 := X^{x+\delta,0,C^L} - X^{x+\delta,0,0}$, $\bar{X}^2 := X^{x,0,C^L} - X^{x,0,0}$, $\check{X}^1 := X^{x+\delta,0,C^L} - X^{x,0,C^L}$, and $\check{X}^2 := X^{x+\delta,0,0} - X^{x,0,0}$. Analogous arguments as employed for the first inequality show that $\bar{X}^1 - \bar{X}^2 \geq 0$ for $t \leq \varrho$ and $\check{X}^1 - \check{X}^2 \geq 0$ for $t > \varrho$, which are equivalent to the second inequality in (2.12).

The proof of (2.13) rests on quite similar arguments. $\Gamma^{lo} > 0$ $\boldsymbol{P}$-a.s. is a standard property of geometric Brownian motion. The first and last inequalities follow from the Lipschitz property of b, linearity of σ, and the comparison theorem (Theorem 2.5). For the second and third inequalities, use convexity of b in x, which gives the estimate

$$\begin{aligned}
D_x^+ b(t, X_t^{x,C^U,C^L}) &\big(X_t^{x+\delta,C^U,C^L} - X_t^{x,C^U,C^L}\big) \\
&\leq b(t, X_t^{x+\delta,C^U,C^L}) - b(t, X_t^{x,C^U,C^L}) \\
&\leq D_x^+ b(t, X_t^{x+\delta,C^U,C^L}) \big(X_t^{x+\delta,C^U,C^L} - X_t^{x,C^U,C^L}\big),
\end{aligned}$$

and apply, again, the comparison theorem.

Equation (2.14) is a standard property of stochastic flows; see, e.g., section V.7 of Protter [62]. We give a short outline of the argument in our situation. In view of (2.13) it suffices to prove

$$\Gamma_t^{x+\delta,C} \quad \searrow \quad \Gamma_t^{x,C} \qquad \text{as } \delta \searrow 0. \tag{2.15}$$

It is obvious from Theorem 2.5 that $X_t^{x+\delta,C}$ and hence $\Gamma_t^{x+\delta,C}$ decrease as $\delta \searrow 0$. By the dominated convergence theorem, with

$$\|\boldsymbol{\delta} D_x^+ b\|_2^2 := E\left[\int_0^T (D_x^+ b(s, X_s^{x+\delta,C}) - D_x^+ b(s, X_s^{x,C}))^2 |\Gamma_s^{x,C}|^2 \, ds\right],$$

$\|\boldsymbol{\delta} D_x^+ b\|_2^2 \overset{\delta\searrow 0}{\longrightarrow} 0$. Hence $E\big[\sup_{0\leq s\leq T} \big|\Gamma_s^{x+\delta,C} - \Gamma_s^{x,C}\big|^2\big]$ converges to zero as the a priori estimates of Lemma 2.3 show. Hence $\Gamma_t^{x+\delta,C} - \Gamma_t^{x,C}$ also converges uniformly in t $\boldsymbol{P}$-a.s. to zero, so (2.14) follows. □

3. Bounded variation control and associated Dynkin game. The proof of Theorem 1.3 essentially rests on the four relations

$$u^-(x) \le u^+(x), \qquad \Delta_- V(x) \le \Delta^+ V(x),$$
$$\Delta^+ V(x) \le u^-(x), \qquad u^+(x) \le \Delta_- V(x).$$

Here $\Delta^+ V$ and $\Delta_- V$ denote the upper right and lower left Dini derivatives of the value function V. The first inequality follows from the definitions (1.5) and the second from Theorem 2.10. The third and fourth are the subject of this section and are proved in Lemmata 3.8 and 3.15.

For our investigation of the value function of the control problem we choose an optimal trajectory and a state process that tracks it from a nearby starting point, up to a pair of stopping times. The essentials of this construction are illustrated best in the context of monotone controls, and we therefore recommend consulting, e.g., [44], [5], or [18]. In what follows, properties of tracking processes in a monotone control problem for the left- and right-hand side Dini derivative, respectively, are combined, which makes the construction more intricate. The aim of this pathwise analysis is to construct controls such that the difference of costs, calculated for original and disturbed starting points, resembles the payoff of a stochastic game of optimal stopping. Thus we establish correspondence between "difference quotients" of the control cost process and the cost of stopping times.

The formulation of tracking and the analysis of cost behavior is carried out without reference to specific properties of the trajectory tracked, but to draw conclusions on the value process we will have to assume that an optimal policy exists. We start with a discussion of right-hand side difference quotients. The left-hand side difference quotient is almost parallel, and there is a brief outline in section 3.3 of the steps necessary. The section closes with the proof of the main theorem.

3.1. Upper right Dini derivative: Construction and properties of the tracking process. First we will introduce the construction of a tracking process and recall some properties of the state processes in this situation. Then we analyze characteristics of the control involved and of the cost functionals. To also have these results available when discussing the lower left Dini derivatives we restrict our assumptions for a while to

(3.1a) b is differentiable in x and either convex or concave;

(3.1b) σ is linear in the space variable x;

(3.1c) h is increasing in the space variable x;

(3.1d) g is increasing in the space variable x;

(3.1e) both components of a are nonnegative.

Making use of the convexity of the data and its consequences presented in Lemma 2.11, we then find estimates for the difference quotient of cost functionals and investigate its limit behavior. This enables us to find an estimate for the upper right Dini derivative of the value function V in terms of the value function of a Dynkin game.

Construction of the tracking process from above. We consider a controlled process $X^\star = X^{x,C^U,C^L}$ and a process X^δ starting in $x+\delta$ and tracking $X^\star$ from above.

Let $\sigma \in \mathcal{T}$ and define the crossing time

$$\tau^\delta := \inf\{t \geq 0 \,|\, X_t^{x+\delta,C^U,0} \leq X_t^{x,C^U,C^L}\}. \tag{3.2}$$

The tracking process X^δ for the starting point $x+\delta$ is given by

$$X_t^\delta := \begin{cases} X_t^{x+\delta,C^U,0}, & t \leq \sigma \wedge \tau^\delta, \\ X_t^{x,C^U,C^L}, & t > \sigma \wedge \tau^\delta. \end{cases} \tag{3.3}$$

Its upper control is parallel to that of the state traced, and it has no lower control until σ occurs or X^{x,C^U,C^L} crosses its trajectory.

We want to obtain X^δ as controlled process $X^{x+\delta,C^{\delta,U},C^{\delta,L}}$ to compare the costs associated with X^δ and $X^\star$. To achieve this, define

$$C_t^{\delta,U} := \begin{cases} C_t^U, & t \leq \sigma, \\ C_t^U + X_\sigma^{x+\delta,C^U,0} - X_\sigma^{x,C^U,C^L}, & t > \sigma, \sigma \leq \tau^\delta, \\ C_t^U, & t > \sigma, \sigma > \tau^\delta, \end{cases} \tag{3.4a}$$

$$C_t^{\delta,L} := \begin{cases} 0, & t \leq \sigma \wedge \tau^\delta, \\ C_t^L - C_\sigma^\lambda, & t > \sigma, \sigma \leq \tau^\delta, \\ C_t^L - C_{\tau^\delta}^L - (X_{\tau^\delta}^{x+\delta,C^U,0} - X_{\tau^\delta}^{x,C^U,C^L}), & t > \tau^\delta, \sigma > \tau^\delta. \end{cases} \tag{3.4b}$$

We have to check that $C^{\delta,L}$ is increasing at τ^δ: Consider $t > \tau^\delta$ on $\{\sigma > \tau^\delta\}$. Then by definition of τ^δ

$$\begin{aligned} 0 &\leq C_t^L - C_{\tau^\delta+}^L + X_{\tau^\delta+}^{x,C^U,C^L} - X_{\tau^\delta+}^{x+\delta,C^U,0} \\ &= C_t^L - C_{\tau^\delta+}^L + X_{\tau^\delta}^{x,C^U,C^L} - (C_{\tau^\delta+}^U - C_{\tau^\delta}^U) \\ &\quad + (C_{\tau^\delta+}^L - C_{\tau^\delta}^L) - \big(X_{\tau^\delta}^{x+\delta,C^U,0} - (C_{\tau^\delta+}^U - C_{\tau^\delta}^U)\big) \\ &= C_t^L - C_{\tau^\delta}^L - \big(X_{\tau^\delta}^{x+\delta,C^U,0} - X_{\tau^\delta}^{x,C^U,C^L}\big). \end{aligned} \tag{3.5}$$

Hence $C^\delta := (C^{\delta,U}, -C^{\delta,L})$ is in $\mathcal{A}$. The elements of this decomposition are nonnegative, increasing, and LCRL, but need not be minimal. We take up this question in Lemma 3.8.

The next lemma tells us that C^δ is really the control needed to obtain X^δ.

LEMMA 3.1. *Let $C^\delta \in \mathcal{A}$ as defined in* (3.4). *Then $X^{x+\delta,C^{\delta,U},C^{\delta,L}} = X^\delta$ and*

$$0 \leq X_t^{x+\delta,C^{\delta,U},C^{\delta,L}} - X_t^{x,C^U,C^L} \leq X_t^{x+\delta,C^U,0} - X_t^{x,C^U,0}. \tag{3.6}$$

Furthermore, $Y_t^{x+\delta,C^{\delta,U},C^{\delta,L}} = Y_t^{x,C^U,C^L}$ and $Z_t^{x+\delta,C^{\delta,U},C^{\delta,L}} = Z_t^{x,C^U,C^L}$ for $\sigma \wedge \tau^\delta < t \leq T$.

See Baras, Elliott, and Kohlmann [6] for a treatment of jumps between stochastic processes in the context of stochastic flows.

Proof. $X_t^\delta = X_t^{x+\delta,C^{\delta,U},C^{\delta,L}}$ for $t \leq \sigma \wedge \tau^\delta$ is obvious from the construction. Consider the right-hand side limit $X_{\sigma\wedge\tau^\delta+}^{x+\delta,C^{\delta,U},C^{\delta,L}}$, which equals $X_{\sigma\wedge\tau^\delta+}^{x,C^U,C^L}$ by definition of C^δ; hence $X_t^{x+\delta,C^{\delta,U},C^{\delta,L}} = X_t^\delta$ for $t > \sigma \wedge \tau^\delta$. The equalities for Y^{x,C^U,C^L} and Z^{x,C^U,C^L} on $(\sigma \wedge \tau^\delta, T]$ follow from the uniqueness of solutions to BSDE.

It remains to prove (3.6). The first inequality follows from (3.3) and the definition of τ^δ. Observe that $X^{x,C^U,C^L} \geq X^{x,C^U,0}$ from the comparison theorem (Theorem 2.5).

For $t \le \sigma \wedge \tau^\delta$ the second inequality then follows from the definition of C^δ. Furthermore, $X_t^{x+\delta,C^{\delta,U},C^{\delta,L}} = X_t^{x,C^U,C^L}$ for $t > \sigma \wedge \tau^\delta$; it follows from the comparison theorem (Theorem 2.5) that the right-hand side is nonnegative, which completes the proof. □

Thus we may introduce the notation

$$X_t^\star := X_t^{x,C^U,C^L}, \qquad Y_t^\star := Y_t^{x,C^U,C^L}, \qquad Z_t^\star := Z_t^{x,C^U,C^L},$$
$$Y_t^\delta := Y_t^{x+\delta,C^{\delta,U},C^{\delta,L}}, \qquad Z_t^\delta := Z_t^{x+\delta,C^{\delta,U},C^{\delta,L}}.$$

$\frac{1}{\delta}\big(Y^\delta - Y^\star\big)$ will be the difference quotient under observation later in this section.

We will also make use of the following estimate, which reformulates some results on controlled processes of Lemma 2.11 for our situation.

LEMMA 3.2. *Assume* (3.1b) *holds. Then for any $C \in \mathcal{A}$ and initial values $x, x+\delta$ the estimate*

$$0 < \delta\Gamma_t^{lo} \le X_t^{x+\delta,C^U,C^L} - X_t^{x,C^U,C^L} \le \delta\Gamma_t^{up} \tag{3.7}$$

holds, and $X_t^{x+\delta,C^U,C^L}$ decreases as δ decreases. Furthermore, X^δ decreases to $X^\star$.

These are consequences of the Lipschitz property of the data b and σ, linearity of σ, and the comparison theorem (Theorem 2.5).

Some characteristics of the tracking process. Next define the first action time τ^0 of C^L by

$$\tau^0 := \inf\{t \ge 0 \,|\, X_t^{x,C^U,0} < X_t^\star\} = \inf\{t \ge 0 \,|\, C_t^L > C_0^L\}. \tag{3.8}$$

Observe that $\tau^\delta \searrow \tau^0$ a.s. Due to their definition, this follows from the right-hand side in (3.7) and existence of right-hand limits for all processes involved.

We know further that $\boldsymbol{P}$-a.s. if $C^L_{\tau^0+}(\omega) > C^L_{\tau^0}(\omega)$, there is a $\delta_0(\omega)$ such that $\tau^\delta(\omega) = \tau^0(\omega)$ for all $\delta \le \delta_0$. This again follows from (3.7) if we set $\delta_0 := (C^L_{\tau^0+} - C^L_{\tau^0})/\Gamma^{up}_{\tau^0}$.

Next assume (2.8a) and (2.8b) hold. Then for $t \in \bar{\imath}$

$$\delta\Gamma_t^{x,C^U,0}\boldsymbol{\chi}_{t\le\sigma\wedge\tau^0} \le X_t^\delta - X_t^\star \le \delta\Gamma_t^{x+\delta,C^U,0}\boldsymbol{\chi}_{t\le\sigma\wedge\tau^\delta} \qquad \boldsymbol{P}\text{-a.s.} \tag{3.9}$$

This follows from the convexity of b and definition of τ^0, τ^δ, and X^δ as in Lemma 2.11.

At the end of this section we will assume that (C^U, C^L) is in fact optimal for the control problem, but this is of no importance in the development of the next results.

Bounds for C^L. We wish to give an estimate for C_t^L if $t \le \tau^\delta$. We derive this from the definition of τ^δ in the following way: Let $\tilde{\Theta}_t(\omega) \in [X_t^{x,C^U,0}(\omega), X_t^\star(\omega)] \subset \mathbb{R}$ such that

$$b_x(t,\tilde{\Theta}_t)\big(X_t^\star - X_t^{x,C^U,0}\big) = b(t,X_t^\star) - b(t,X_t^{x,C^U,0}) \qquad \boldsymbol{P}\text{-a.s.}$$

Observe that b_x is either nondecreasing or nonincreasing in x, so $\tilde{\Theta}_t$ can be chosen in an $\mathcal{F}_t$-measurable, LCRL way. Thus $\tilde{\Theta}$ is progressively measurable, and we let $\Gamma_t^{\tilde{\Theta}}$ be the solution of the linear SDE

$$d\Gamma_t^{\tilde{\Theta}} = b_x(t,\tilde{\Theta}_t)\Gamma_t^{\tilde{\Theta}}\,dt + \sigma_x(t,\Gamma_t^{\tilde{\Theta}})^\top\,dW_t, \qquad \Gamma_0^{\tilde{\Theta}} = 1.$$

It is obvious from the comparison theorem (Theorem 2.5) and Lipschitz continuity of b that $\Gamma_t^{lo} \le \Gamma_t^{\tilde\Theta} \le \Gamma_t^{up}$ a.s. Hence $\Gamma_t^{\tilde\Theta} \in \mathbb{R}_{>0}$ for $t \le \tau^\delta$ a.s. Now consider $X^\star - X^{x,C^U,0}$: Apply the Itô formula to $\big(\Gamma_t^{\tilde\Theta}\big)^{-1}\big(X_t^\star - X_t^{x,C^U,0}\big)$ and recall that this equals zero for $t \le \tau^0$ to get the representation

$$X_t^\star - X_t^{x,C^U,0} = \int_0^t b_x(s,\tilde\Theta_s)\big(X_s^\star - X_s^{x,C^U,0}\big)\,ds + C_t^L + \int_0^t \sigma(s, X_s^\star - X_s^{x,C^U,0})^\top\,dW_s = \int_{\tau^0}^{\tau^0\vee t} \Gamma_t^{\tilde\Theta}\big(\Gamma_s^{\tilde\Theta}\big)^{-1}\,dC_s^L.$$

This leads to the following estimate for $\tau^0 < t \le \tau^\delta$:

$$\begin{aligned}\Big(\sup_{\tau^0<s\le t}\Gamma_s^{\tilde\Theta}\Big)^{-1}\Gamma_t^{\tilde\Theta}C_t^L &= \Big(\inf_{\tau^0<s\le t}\big(\Gamma_s^{\tilde\Theta}\big)^{-1}\Big)\Gamma_t^{\tilde\Theta}C_t^L\\ &\le \int_{\tau^0}^t \Gamma_t^{\tilde\Theta}\big(\Gamma_s^{\tilde\Theta}\big)^{-1}\,dC_s^L = X_t^\star - X_t^{x,C^U,0}\\ &\le X_t^{x+\delta,C^U,0} - X_t^{x,C^U,0} \le \delta\Gamma_t^{up}.\end{aligned}$$

Similarly we obtain a lower bound on C_t^L for t with $C_t^L > 0$, i.e., $t > \tau^0$. The results are summarized in the following lemma.

LEMMA 3.3. *Let τ^0 and τ^δ be as defined in* (3.8) *and* (3.2)*, and assume* (3.1) *holds. Then C_t^L has $\boldsymbol{P}$-a.s. upper and lower bounds for $\tau^0 < t \le \tau^\delta$:*

$$\big(X_t^\star - X_t^{x,C^U,0}\big)\frac{1}{\Gamma_t^{\tilde\Theta}}\Big(\inf_{\tau^0<s\le t}\Gamma_s^{\tilde\Theta}\Big) \le C_t^L \le \big(X_t^\star - X_t^{x,C^U,0}\big)\frac{1}{\Gamma_t^{\tilde\Theta}}\Big(\sup_{\tau^0<s\le t}\Gamma_s^{\tilde\Theta}\Big).$$

Thus C^L satisfies

$$C_t^L \le \delta\Gamma_t^{up}\big(\Gamma_t^{lo}\big)^{-1}\Big(\sup_{\tau^0\le s\le t}\Gamma_s^{up}\Big) =: \delta\cdot\overline{C_t^{L,\delta}} \qquad \text{for } t \le \tau^\delta \ \boldsymbol{P}\text{-a.s.}$$

Note that if (2.8) and (3.15) hold, it follows from convexity of b and the comparison theorem (Theorem 2.5) that for $t \le \tau^\delta$

$$\Gamma_t^{x,C^U,0} \le \Gamma_t^{\tilde\Theta} \le \Gamma_t^{x+\delta,C^U,0}. \tag{3.10}$$

Limit behavior of cost functionals. Our plan is to find bounds for the process $\frac{1}{\delta}\big(Y^\delta - Y^\star\big)$, which will result in a first connection with a Dynkin game. But before we can investigate this "difference quotient" we need more information about the processes Y^δ and their behavior as $\delta \searrow 0$. We start with the following lemma.

LEMMA 3.4. *Assume that* (3.1) *holds. Let, for $\delta, \delta' \in \mathbb{R}_{>0}$, the processes Y^δ, $Y^{\delta'}$ be the cost functionals associated with $x+\delta$, $x+\delta'$ according to the construction* (3.3)*. Then*

$$Y_t^\star \le Y_t^\delta \le Y_t^{\delta'} \qquad \text{for } \delta \le \delta' \text{ for all } t\ \boldsymbol{P}\text{-a.s.}$$

Proof. Let us look at the inequality $Y_t^\star \le Y_t^\delta$ first. For $t > \sigma \wedge \tau^\delta$ equality holds by definition of $(C^{\delta,U}, C^{\delta,L})$.

Now let $t \le \sigma \wedge \tau^\delta$. We wish to apply the comparison theorem, Theorem 2.7. From Lemma 3.1 and properties (3.1c) and (3.1d) of h, g it follows that

$$h(X_T^\delta) \ge h(X_T^\star), \qquad g(t, X_t^\delta, y, z) \ge g(t, X_t^\star, y, z).$$

Consider the difference of control processes $\boldsymbol{\delta}C := \big(C^{\delta,U} - C^U, -(C^{\delta,L} - C^L)\big)$. The components have the representation

$$\begin{aligned}
C^{\delta,U}_t - C^U_t &= \big(X^\delta_\sigma - X^\star_\sigma\big)\boldsymbol{\chi}_{\substack{t>\sigma\\ \sigma\le\tau^\delta}},\\
C^{\delta,L}_t - C^L_t &= -C^L_t\boldsymbol{\chi}_{t\le\sigma\wedge\tau^\delta} - C^L_\sigma\boldsymbol{\chi}_{\substack{t>\sigma\wedge\tau^\delta\\ \sigma\le\tau^\delta}} - \big(C^L_{\tau^\delta} + (X^\delta_{\tau^\delta} - X^\star_{\tau^\delta})\big)\boldsymbol{\chi}_{\substack{t>\sigma\wedge\tau^\delta\\ \sigma>\tau^\delta}}.
\end{aligned}$$

Recall the argument in (3.5) to see that both components of $\boldsymbol{\delta}C$ are increasing and therefore $\int_{[0,t)} a_s^\top\, d(\boldsymbol{\delta}C_s)$ is increasing in t.

So the assumptions of the comparison theorem (Theorem 2.7) are satisfied, which completes the proof of the first inequality. □

For the second inequality observe that we are done if $t > \sigma\wedge\tau^\delta$, as then $Y^\delta_t = Y^\star_t$. For $t \le \sigma\wedge\tau^\delta$ the processes satisfy

$$\begin{aligned}
Y^{\delta'}_t - Y^\delta_t = Y^{\delta'}_{\sigma\wedge\tau^\delta} - Y^\delta_{\sigma\wedge\tau^\delta} &+ \int_t^{\sigma\wedge\tau^\delta} g(s,X^{\delta'}_s,Y^{\delta'}_s,Z^{\delta'}_s) - g(s,X^\delta_s,Y^\delta_s,Z^\delta_s)\,ds\\
&- \int_t^{\sigma\wedge\tau^\delta} Z^{\delta'}_s - Z^\delta_s\,dW_s.
\end{aligned}$$

Hence it suffices to prove $Y^{\delta'}_{\sigma\wedge\tau^\delta} \ge Y^\delta_{\sigma\wedge\tau^\delta}$, as $g(s,X^{\delta'}_s,y,z) \ge g(s,X^\delta_s,y,z)$ and an application of the comparison theorem (Theorem 2.7) will complete the argument.

To this end use a slightly different representation (cf. (3.11) below):

$$\begin{aligned}
Y^\delta_{\sigma\wedge\tau^\delta} - Y^\star_{\sigma\wedge\tau^\delta} &= \big(h(X^\delta_T) - h(X^\star_T)\big)\boldsymbol{\chi}_{\sigma\wedge\tau^\delta=T} + a^U_\sigma\big(X^\delta_\sigma - X^\star_\sigma\big)\boldsymbol{\chi}_{\substack{\sigma\le\tau^\delta\\ \sigma<T}}\\
&\quad + a^L_{\tau^\delta}\big(X^\delta_{\tau^\delta} - X^\star_{\tau^\delta}\big)\boldsymbol{\chi}_{\substack{\tau^\delta<\sigma\\ \tau^\delta<T}}\\
&\le \big(h(X^{\delta'}_T) - h(X^\star_T)\big)\boldsymbol{\chi}_{\sigma\wedge\tau^\delta=T} + a^U_\sigma\big(X^{\delta'}_\sigma - X^\star_\sigma\big)\boldsymbol{\chi}_{\substack{\sigma\le\tau^\delta\\ \sigma<T}}\\
&\quad + a^L_{\tau^\delta}\big(C^L_{\tau^\delta+} - C^L_{\tau^\delta}\big)\boldsymbol{\chi}_{\substack{\tau^\delta<\sigma\\ \tau^\delta<T}} + \big(Y^{\delta'}_{\sigma\wedge\tau^\delta+} - Y^\star_{\sigma\wedge\tau^\delta+}\big)\boldsymbol{\chi}_{\substack{\tau^\delta<\sigma\\ \tau^\delta<T}}\\
&= Y^{\delta'}_{\sigma\wedge\tau^\delta} - Y^\star_{\sigma\wedge\tau^\delta},
\end{aligned}$$

where the inequality is due to the definition of $C^{\delta,L}$ and (3.5), $Y^{\delta'} \ge Y^\star$ and (3.1c), and (3.1e) and the monotonicity of X^δ in δ. The last equality follows from $\tau^\delta \le \tau^{\delta'}$. This completes the proof. □

Observe that we can interpret $Y^\delta - Y^\star$ as solution to a BSDE where the terminal value is $\mathcal{F}_{\sigma\wedge\tau^\delta}$-measurable. In this form it resembles the payoff of a stochastic game of optimal stopping:

$$\begin{aligned}
Y^\delta_t - Y^\star_t = E\Bigg[&\int_t^{\sigma\wedge\tau^\delta} g(s,X^\delta_s,Y^\delta_s,Z^\delta_s) - g(s,X^\star_s,Y^\star_s,Z^\star_s)\,ds\\
&+ \big(h(X^\delta_T) - h(X^\star_T)\big)\boldsymbol{\chi}_{\sigma\wedge\tau^\delta=T} + a^U_\sigma\big(X^\delta_\sigma - X^\star_\sigma\big)\boldsymbol{\chi}_{\substack{\sigma\le\tau^\delta\\ \sigma<T}}\\
&+ a^L_{\tau^\delta}\big(X^\delta_{\tau^\delta} - X^\star_{\tau^\delta}\big)\boldsymbol{\chi}_{\substack{\tau^\delta<\sigma\\ \tau^\delta<T}} + \int_{[t,\sigma\wedge\tau^\delta)} a^L_s\,dC^L_s \,\Bigg|\, \mathcal{F}_t\Bigg].
\end{aligned} \tag{3.11}$$

We will use this representation in the discussion of limit behavior of $\frac{1}{\delta}\big(Y^\delta - Y^\star\big)$. As a first step we establish a result on the convergence of Y^δ to $Y^\star$.

LEMMA 3.5. *Assume that* (3.1) *holds. Then* Y_t^δ *converges to* $Y_t^\star$ $\boldsymbol{P}$*-a.s. uniformly in* t*, and further,*

$$\lim_{\delta \searrow 0} E\left[\sup_{t\le s\le T} \left|Y_s^\delta - Y_s^\star\right|^2\right] = 0, \tag{3.12}$$

$$\lim_{\delta \searrow 0} E\left[\int_t^T \left|Z_s^\delta - Z_s^\star\right|^2 ds\right] = 0. \tag{3.13}$$

The convergence in (3.12) *is monotone.*

Proof. We know from Lemma 3.4 that $\sup_{t\le s\le T}(Y_s^\delta - Y_s^\star)$ is nonnegative and decreases as δ decreases to zero. If we set $A := \{\omega \mid \lim_{\delta\searrow 0}(\sup_{t\le s\le T} Y_s^\delta - Y_s^\star) > 0\}$, then $\boldsymbol{P}(A) = 0$ follows from (3.12).

But (3.12) and (3.13) follow from Lemma 2.4 if we can show that—in the notation of that lemma—$\|\boldsymbol{\delta}\xi\|_2^2$, $\|\boldsymbol{\delta}g\|_2^2$, and $|\boldsymbol{\delta}C|_{[a,\bar{\imath}]}^2$ converge to zero. Hence it suffices to prove convergence of the data in the respective 2-norms.

We first consider the driver and write

$$\boldsymbol{\delta}g_s = g(s, X_s^\delta, Y_s^\star, Z_s^\star) - g(s, X_s^\star, Y_s^\star, Z_s^\star) \le L\cdot\delta\cdot\Gamma_s^{up},$$

where L denotes a Lipschitz constant of g. Hence $\|\boldsymbol{\delta}g_s\|_2^2 \le L^2\cdot\delta^2\cdot\|\Gamma_s^{up}\|_2^2$, which converges to zero as $\delta \searrow 0$. The same argument can be applied to $\boldsymbol{\delta}\xi = h(X_T^\delta) - h(X_T^\star)$.

From the definition of the processes X^δ, $X^\star$ we have

$$\begin{aligned}\left(X_{\tau^\delta}^\delta - X_{\tau^\delta}^\star\right) &= \left(X_{\tau^0}^\delta - X_{\tau^0}^\star\right) + \int_{\tau^0}^{\tau^\delta} \left(b(s, X_s^\delta) - b(s, X_s^\star)\right) ds \\ &\quad + \int_{\tau^0}^{\tau^\delta} \sigma\left(s, (X_s^\delta - X_s^\star)\right) dW_s - (C_{\tau^\delta}^L - C_{\tau^0}^L).\end{aligned} \tag{3.14}$$

Also recall that $C^{\delta,L}$ is constant on $[t, \tau^0]$. Therefore we can estimate the difference of controls in the following way:

$$\begin{aligned}|\boldsymbol{\delta}C|_{[a,\bar{\imath}]}^2 &= E\left[\left(\int_{\bar{\imath}} a_s^L\, d\left|C^{\delta,L} - C^L\right|_s\right)^2 + \left(\int_{\bar{\imath}} a_s^U\, d\left|C^{\delta,U} - C^U\right|_s\right)^2\right] \\ &= E\left[\left(\left(a_{\tau^\delta}^L(X_{\tau^\delta}^\delta - X_{\tau^\delta}^\star)\right)\boldsymbol{\chi}_{\substack{\tau^\delta<\sigma\\ \tau^\delta<T}} + \int_{[t,\sigma\wedge\tau^\delta)} a_s^L\, dC_s^L\right)^2\right. \\ &\qquad \left. + \left(a_\sigma^U(X_\sigma^\delta - X_\sigma^\star)\right)^2\boldsymbol{\chi}_{\substack{\sigma\le\tau^\delta\\ \sigma<T}}\right] \\ &\le E\left[\left(\left(a_{\tau^\delta}^L(X_{\tau^\delta}^\delta - X_{\tau^\delta}^\star)\right) + \int_{[\tau^0,\tau^\delta)} a_s^L\, dC_s^L\right)^2\boldsymbol{\chi}_{\substack{\tau^0<\sigma\\ \tau^0<T}}\right. \\ &\qquad \left. + \left(a_\sigma^U(X_\sigma^\delta - X_\sigma^\star)\right)^2\boldsymbol{\chi}_{\substack{\sigma\le\tau^\delta\\ \sigma<T}}\right] \\ &\le \delta^2 E\left[3\left(\max_{\tau^0\le s\le\tau^\delta} a_s^L\right)^2\left(|\Gamma_{\tau^0}^{up}|^2 + \left(\int_{\tau^0}^{\tau^\delta} L\Gamma_s^{up}\, ds\right)^2\right.\right. \\ &\qquad \left.\left. + \int_{\tau^0}^{\tau^\delta} |\sigma(s, \Gamma_s^{up})|^2\, ds\right)\right] + \delta^2 E\left[\left(a_\sigma^U\Gamma_\sigma^{up}\right)^2\right].\end{aligned}$$

The second inequality is due to (3.14), where we have used (3.7), (3.6), Lipschitz continuity of b, and linearity of σ. The last term obviously converges to zero as $\delta \searrow 0$, which completes the proof. □

3.2. Upper right Dini derivative: Estimates for the difference quotient. From now on we assume smoothness of the data of the FBSDE (1.1):

(3.15a) b is differentiable in x;

(3.15b) h is differentiable in x;

(3.15c) g is partially differentiable in x, y, and z;

(3.15d) g_y is increasing in x;

(3.15e) g_z is independent of x, y, and z;

(3.15f) both components of a are nonnegative and continuous.

Under (3.15e) we may drop dependence on z in the notation for g_x, g_y, and g_z.

We now define processes $\tilde{R}^{up,\delta}$ and $\tilde{R}^{lo,\delta}$, which converge to the same limiting process $\tilde{R}$ as $\delta \searrow 0$ and serve as majorants and minorants to the difference quotient $\frac{1}{\delta}(Y^\delta - Y^\star)$. Thus they help us investigate its limit behavior. To facilitate the argument we use the translation to a control-free FBSDE as in Peng [61]; we define

$$\tilde{Y}_t^\delta := \frac{1}{\delta}\left((Y_t^\delta - Y_t^\star) + \int_{[0,\sigma\wedge\tau^\delta\wedge t)} a_s^L \, dC_s^L\right) \qquad \text{and} \qquad \tilde{Z}_t^\delta := \frac{1}{\delta}(Z_t^\delta - Z_t^\star).$$

Then $(\tilde{Y}^\delta, \tilde{Z}^\delta)$ is solution to a BSDE with data

$$\begin{aligned}
\xi^{\tilde{Y}} &:= \frac{1}{\delta}\big(h(X_T^\delta) - h(X_T^\star)\big)\boldsymbol{\chi}_{\sigma\wedge\tau^\delta = T} + a_\sigma^U \frac{1}{\delta}\big(X_\sigma^\delta - X_\sigma^\star\big)\boldsymbol{\chi}_{\substack{\sigma\le\tau^\delta \\ \sigma<T}} \\
&\quad + a_{\tau^\delta}^L \frac{1}{\delta}\big(X_{\tau^\delta}^\delta - X_{\tau^\delta}^\star\big)\boldsymbol{\chi}_{\substack{\tau^\delta<\sigma \\ \tau^\delta<T}} + \frac{1}{\delta}\int_{[0,\sigma\wedge\tau^\delta)} a_s^L \, dC_s^L, \\
g^{\tilde{Y}}(t,y,z) &:= \Bigg(\Delta_x\tilde{g}_t \frac{1}{\delta}(X_t^\delta - X_t^\star) + \Delta_y\tilde{g}_t\boldsymbol{\chi}_{y\ge 0}\cdot y \\
&\qquad - \Delta_y\tilde{g}_t \frac{1}{\delta}\int_{[0,t)} a_s^L \, dC_s^L\Bigg)\boldsymbol{\chi}_{t\le\sigma\wedge\tau^\delta} + \Delta_z\tilde{g}_t \cdot z\boldsymbol{\chi}_{t\le\sigma}, \\
C_t^{\tilde{Y}} &:= (0,0).
\end{aligned}$$

Here $\Delta_x\tilde{g}_t$, $\Delta_y\tilde{g}_t$, and $\Delta_z\tilde{g}_t$ are defined as

$$\begin{aligned}
\Delta_x\tilde{g}_t &:= \big(g(t,X_t^\delta,Y_t^\star,Z_t^\star) - g(t,X_t^\star,Y_t^\star,Z_t^\star)\big)\frac{1}{X_t^\delta - X_t^\star}\boldsymbol{\chi}_{X_t^\delta\neq X_t^\star}, \\
\Delta_y\tilde{g}_t &:= \big(g(t,X_t^\delta,Y_t^\delta,Z_t^\star) - g(t,X_t^\delta,Y_t^\star,Z_t^\star)\big)\frac{1}{Y_t^\delta - Y_t^\star}\boldsymbol{\chi}_{Y_t^\delta\neq Y_t^\star}, \\
\Delta_z\tilde{g}_t &:= \big(g(t,X_t^\delta,Y_t^\delta,Z_t^\delta) - g(t,X_t^\delta,Y_t^\delta,Z_t^\star)\big)\frac{1}{Z_t^\delta - Z_t^\star}\boldsymbol{\chi}_{Z_t^\delta\neq Z_t^\star}.
\end{aligned}$$

Observe that $Y_t^\delta - Y_t^\star = 0$ for $t > \sigma\wedge\tau^\delta$; hence $Y_t^\delta - Y_t^\star$ is $\mathcal{F}_{\sigma\wedge\tau^\delta}$-measurable. Therefore $Z_t^\delta - Z_t^\star = 0$ for $t > \sigma\wedge\tau^\delta$ and we can in effect extend the $\Delta_z\tilde{g}$-component in the definition of $g^{\tilde{Y}}$ from $[0,\sigma\wedge\tau^\delta]$ to $[0,\sigma]$. Observe also that $\tilde{Y}^\delta \ge 0$ by (3.15f)

and Lemma 3.4, and that $\tilde{Y}^\delta_t = \frac{1}{\delta}(Y^\delta_t - Y^\star_t)$ for $t \le \sigma \wedge \tau^0$. The representation of $(\tilde{Y}^\delta, \tilde{Z}^\delta)$ as solution of a BSDE with the data $(\xi^{\tilde{Y}}, g^{\tilde{Y}}, C^{\tilde{Y}})$ can be verified directly.

From now on we will assume that (3.15) holds. Therefore we drop the z-variable in the notation of partial derivatives of g. Observe also that $\Delta_x\tilde{g}$, $\Delta_y\tilde{g}$, and $\Delta_z\tilde{g}$ do not depend on $Z^\star$ and Z^δ in this case.

Now assume g to be partially differentiable and define $(\tilde{R}^{up,\delta}, \tilde{Q}^{up,\delta})$ as solution to the BSDE with data

$$\begin{aligned}
\xi^{up,\delta} &:= h_x(X_T^{x+\delta,C^U,0})\Gamma_T^{x+\delta,C^U,0}\boldsymbol{\chi}_{\sigma\wedge\tau^\delta=T} + a^U_\sigma\Gamma_\sigma^{x+\delta,C^U,0}\boldsymbol{\chi}_{\substack{\sigma\le\tau^\delta\\ \sigma<T}} \\
&\quad + \big(\max_{\tau^0\le s\le\tau^\delta} a^L_s\big)\big(\sup_{\tau^0\le s\le\tau^\delta}\Gamma_s^{x+\delta,C^U,0}\big)^2\big(\inf_{\tau^0\le s\le\tau^\delta}\Gamma_s^{x,C^U,0}\big)^{-1}\boldsymbol{\chi}_{\substack{\tau^0<\sigma\\ \tau^0<T}}, \\
g^{up,\delta}(t,y,z) &:= \Big(g_x(t,X^\delta_t,Y^\star_t)\Gamma_t^{x+\delta,C^U,0} + g_y(t,X^\delta_t,Y^\delta_t)\boldsymbol{\chi}_{y\ge0}\boldsymbol{\chi}_{A^+_{g,\delta}}\cdot y \\
&\quad + L\big(\max_{\tau^0\le s\le\tau^\delta\wedge t} a^L_s\big)\overline{C^{L,\delta}_t}\Big)\boldsymbol{\chi}_{t\le\sigma\wedge\tau^\delta} + g_z(t)\cdot z\boldsymbol{\chi}_{t\le\sigma}, \\
C^{up,\delta}_t &:= (0,0).
\end{aligned}$$

Similarly we let $(\tilde{R}^{lo,\delta}, \tilde{Q}^{lo,\delta})$ be the solution to the BSDE with data

$$\begin{aligned}
\xi^{lo,\delta} &:= h_x(X_T^{x,C^U,0})\Gamma_T^{x,C^U,0}\boldsymbol{\chi}_{\sigma\wedge\tau^0=T} + a^U_\sigma\Gamma_\sigma^{x,C^U,0}\boldsymbol{\chi}_{\substack{\sigma\le\tau^0\\ \sigma<T}} \\
&\quad + \big(\min_{\tau^0\le s\le\tau^\delta} a^L_s\big)\big(\inf_{\tau^0\le s\le\tau^\delta}\Gamma_s^{x,C^U,0}\big)^2\big(\sup_{\tau^0\le s\le\tau^\delta}\Gamma_s^{x+\delta,C^U,0}\big)^{-1}\boldsymbol{\chi}_{\substack{\tau^\delta<\sigma\\ \tau^\delta<T}}, \\
g^{lo,\delta}(t,y,z) &:= g_x(t,X^\star_t,Y^\star_t)\Gamma_t^{x,C^U,0}\boldsymbol{\chi}_{t\le\sigma\wedge\tau^0} + \Big(g_y(t,X^\star_t,Y^\star_t)\boldsymbol{\chi}_{y\ge0}\boldsymbol{\chi}_{A^-_{g,\delta}}\cdot y \\
&\quad - L\big(\max_{\tau^0\le s\le\tau^\delta\wedge t} a^L_s\big)\overline{C^{L,\delta}_t}\Big)\boldsymbol{\chi}_{t\le\sigma\wedge\tau^\delta} + g_z(t)\cdot z\boldsymbol{\chi}_{t\le\sigma}, \\
C^{lo,\delta}_t &:= (0,0).
\end{aligned}$$

Recall the definition of $\overline{C^{L,\delta}_t}$ in Lemma 3.3. L again denotes a Lipschitz constant of g.

We need $A^+_{g,\delta}$ and $A^-_{g,\delta}$ to eliminate the negative and positive parts in the $g_y\cdot y$ -terms for $t > \sigma\wedge\tau^0$. The reason is that $\boldsymbol{\chi}_{t\le\tau^\delta}$ decreases to $\boldsymbol{\chi}_{t\le\tau^0}$ in δ, so we have to add a condition that ensures the required monotonicity of the data. More specifically, we define

$$\begin{aligned}
A^+_{g,\delta} &:= \{t \le \sigma\wedge\tau^0\} \cup \{g_y(t,X^\delta_t,Y^\delta_t) \ge 0\}, \\
A^-_{g,\delta} &:= \{t \le \sigma\wedge\tau^0\} \cup \{g_y(t,X^\star_t,Y^\star_t) \le 0\}.
\end{aligned}$$

We also use the nonstandard convention $\max_{\tau^0\le s\le\tau^\delta\wedge t} a^L_s = 0$ for $t < \tau^0$.

We are now in a position to formulate the estimate.

LEMMA 3.6. *Assume that* (2.8) *and* (3.15) *hold. Then* $\tilde{R}^{up,\delta}, \tilde{Y}^\delta, \tilde{R}^{lo,\delta}$ *satisfy*

$$\tilde{R}^{lo,\delta}_t \le \tilde{Y}^\delta_t \le \tilde{R}^{up,\delta}_t \qquad \forall t\in\bar{\imath}\ \boldsymbol{P}\text{-a.s.} \tag{3.16}$$

Further, $\tilde{R}^{up,\delta}$ *decreases and* $\tilde{R}^{lo,\delta}$ *increases as* $\delta \searrow 0$ *to the same limiting process* $\tilde{R}$ *defined below, in* $L^2_{\mathcal{F}}(0,T;\mathbb{R})$ *and for all* $t\in\bar{\imath}$ $\boldsymbol{P}$*-a.s.*

Proof. To apply the comparison theorem (Theorem 2.7) with supplementary condition (2.7c) to $\tilde{R}^{up,\delta}$, $\tilde{Y}^\delta$, $\tilde{R}^{lo,\delta}$, we have to estimate the drivers and terminal conditions of the three processes.

Consider the drivers first. Their components satisfy

$$\begin{aligned} g_x(t,X_t^\star,Y_t^\star)\Gamma_t^{x,C^U,0}\boldsymbol{\chi}_{t\le\sigma\wedge\tau^0} &\le \Delta_x\tilde g_t\frac{1}{\delta}\big(X_t^\delta - X_t^\star\big)\boldsymbol{\chi}_{t\le\sigma\wedge\tau^\delta} \\ &\le g_x(t,X_t^\delta,Y_t^\star)\Gamma_t^{x+\delta,C^U,0}\boldsymbol{\chi}_{t\le\sigma\wedge\tau^\delta}, \\ g_y(t,X_t^\star,Y_t^\star)\frac{1}{\delta}\big(Y_t^\delta - Y_t^\star\big)\boldsymbol{\chi}_{A^-_{g,\delta}} &\le \Delta_y\tilde g_t\frac{1}{\delta}\big(Y_t^\delta - Y_t^\star\big) \\ &\le g_y(t,X_t^\delta,Y_t^\delta)\frac{1}{\delta}\big(Y_t^\delta - Y_t^\star\big)\boldsymbol{\chi}_{A^+_{g,\delta}}, \end{aligned}$$

$$\begin{aligned} &g_y(t,X_t^\star,Y_t^\star)\frac{1}{\delta}\int_{[\tau^0,\tau^\delta\wedge t)} a_s^L\,dC_s^L - L\big(\max_{\tau^0\le s\le\tau^\delta\wedge t} a_s^L\big)\overline{C_t^{L,\delta}} \\ &\qquad\le \Delta_y\tilde g_t\frac{1}{\delta}\int_{[\tau^0,\tau^\delta\wedge t)} a_s^L\,dC_s^L - \Delta_y\tilde g_t\frac{1}{\delta}\int_{[\tau^0,\tau^\delta\wedge t)} a_s^L\,dC_s^L \\ &\text{(which equals zero)} \\ &\qquad\le g_y(t,X_t^\delta,Y_t^\delta)\frac{1}{\delta}\int_{[\tau^0,\tau^\delta\wedge t)} a_s^L\,dC_s^L + L\big(\max_{\tau^0\le s\le\tau^\delta\wedge t} a_s^L\big)\overline{C_t^{L,\delta}}, \\ &\qquad\qquad g_z(t)\frac{1}{\delta}\big(Z_t^\delta - Z_t^\star\big) = \Delta_z\tilde g_t\frac{1}{\delta}\big(Z_t^\delta - Z_t^\star\big). \end{aligned}$$

The estimates for $\Delta_x\tilde g$ follow from (3.15c), (2.8d), and (3.9). For the estimates of $\Delta_y\tilde g$ use (2.8d), (3.15c), Lemma 3.4, and the definition of $A^\pm_{g,\delta}$.

For the third line estimate of the La^L-term recall Lipschitz continuity of g and Lemma 3.3. As g_z is independent of (x,y,z) and $\tau^\delta\ge\tau^0$, we can summarize this as

$$g^{lo,\delta}(t,\tilde Y^\delta,\tilde Z^\delta) \le g^{\tilde Y}(t,\tilde Y^\delta,\tilde Z^\delta) \le g^{up,\delta}(t,\tilde Y^\delta,\tilde Z^\delta). \tag{3.17}$$

In a similar way we can estimate the terms in h_x, a^U, and h, respectively, in the definitions of $\xi^{lo,\delta}$, $\xi^{\tilde Y}$, and $\xi^{up,\delta}$ by (2.8c), (3.15b), and (3.9). For the terms in a^L observe that by Lemma 3.3, (2.13), and (3.10),

$$\begin{aligned} &a^L_{\tau^\delta}\frac{1}{\delta}\big(X^\delta_{\tau^\delta} - X^\star_{\tau^\delta}\big)\boldsymbol{\chi}_{\substack{\tau^\delta<\sigma\\ \tau^\delta<T}} + \frac{1}{\delta}\int_{[\sigma\wedge\tau^0,\sigma\wedge\tau^\delta)} a_s^L\,dC_s^L \\ &\quad\le \frac{1}{\delta}\big(\max_{\tau^0\le s\le\tau^\delta} a_s^L\big)\Big(X^\delta_{\tau^\delta} - X^\star_{\tau^\delta} + \big(X^\star_{\tau^\delta} - X^{x,C^U,0}_{\tau^\delta}\big)(\Gamma^{\tilde\Theta}_{\tau^\delta})^{-1}\big(\sup_{\tau^0\le s\le\tau^\delta}\Gamma^{\tilde\Theta}_s\big)\Big)\boldsymbol{\chi}_{\substack{\tau^0<\sigma\\ \tau^0<T}} \\ &\qquad\le \big(\max_{\tau^0\le s\le\tau^\delta} a_s^L\big)\big(\sup_{\tau^0\le s\le\tau^\delta}\Gamma^{x+\delta,C^U,0}_s\big)^2\big(\inf_{\tau^0\le s\le\tau^\delta}\Gamma^{x,C^U,0}_s\big)^{-1}\boldsymbol{\chi}_{\substack{\tau^0<\sigma\\ \tau^0<T}}. \end{aligned}$$

Similarly we can estimate the a^L-term in $\xi^{lo,\delta}$:

$$\begin{aligned} &\big(\min_{\tau^0\le s\le\tau^\delta} a_s^L\big)\big(\inf_{\tau^0\le s\le\tau^\delta}\Gamma^{x,C^U,0}_s\big)^2\big(\sup_{\tau^0\le s\le\tau^\delta}\Gamma^{x+\delta,C^U,0}_s\big)^{-1}\boldsymbol{\chi}_{\substack{\tau^\delta<\sigma\\ \tau^\delta<T}} \\ &\qquad\le a^L_{\tau^\delta}\frac{1}{\delta}\big(X^\delta_{\tau^\delta} - X^\star_{\tau^\delta}\big)\boldsymbol{\chi}_{\substack{\tau^\delta<\sigma\\ \tau^\delta<T}} + \frac{1}{\delta}\int_{[\sigma\wedge\tau^0,\sigma\wedge\tau^\delta)} a_s^L\,dC_s^L. \end{aligned}$$

Hence $\xi^{lo,\delta}\le\xi^{\tilde Y^\delta}\le\xi^{up,\delta}$. An application of the comparison theorem (Theorem 2.7) and (2.7c) completes the proof of (3.16). □

Let us now prove monotony of $\tilde{R}^{lo,\delta}$ and $\tilde{R}^{up,\delta}$ in δ. We show that the data are monotone in δ so that we can again apply the comparison theorem (Theorem 2.7).

Consider the terminal values first. The summands in $\xi^{up,\delta}$ are nonnegative. The terms in h_x and a^U decrease with δ by (3.15b), (3.15f), (2.15), and $\tau^\delta \searrow \tau^0$. The third term decreases to $a^L_{\tau^0}\Gamma^{x,C^U,0}_{\tau^0}$, as a^L and Γ are continuous.

The terms in h_x and a^U of $\xi^{lo,\delta}$ are independent of δ. The a^L-term increases as $\delta \searrow 0$, as $\sup_{\tau^0\le s\le\tau^\delta}\Gamma^{x+\delta,C^U,C^L}_s$ decreases and $\boldsymbol{\chi}_{\tau^\delta<\sigma,\tau^\delta<T}$ increases. Hence $\xi^{up,\delta}$ and $\xi^{lo,\delta}$ decrease and increase, respectively, to $\tilde{\xi}$, where

$$\tilde{\xi} := h_x(X^{x,C^U,0}_T)\Gamma^{x,C^U,0}_T\boldsymbol{\chi}_{\sigma\wedge\tau^0=T} + a^U_\sigma\Gamma^{x,C^U,0}_\sigma\boldsymbol{\chi}_{\substack{\sigma\le\tau^0\\ \sigma<T}} + a^L_{\tau^0}\Gamma^{x,C^U,0}_{\tau^0}\boldsymbol{\chi}_{\substack{\tau^0<\sigma\\ \tau^0<T}}.$$

Next we show that the drivers $g^{up,\delta}$ and $g^{lo,\delta}$ decrease and increase, respectively, and for $t\neq\sigma\wedge\tau^0$ converge to $\tilde{g}$ defined by

$$\begin{aligned}\tilde{g}(t,y,z) := {}& \Big(g_x(t,X^\star_t,Y^\star_t)\Gamma^{x,C^U,0}_t + g_y(t,X^\star_t,Y^\star_t)\boldsymbol{\chi}_{y\ge0}\cdot y\Big)\boldsymbol{\chi}_{t\le\sigma\wedge\tau^0}\\ &+ g_z(t)\cdot z\boldsymbol{\chi}_{t\le\sigma}.\end{aligned}$$

Convergence and monotony of the g_x-term of $g^{up,\delta}$ is obvious; for $g^{lo,\delta}$ it is trivial. The terms in $\pm L(\max_{\tau^0\le s\le\tau^\delta\wedge t}a^L_s)$ equal zero for $t<\sigma\wedge\tau^0$; they are decreasing in absolute value. As $\boldsymbol{\chi}_{\tau^0\le t\le\sigma\wedge\tau^\delta}$ converges to zero, both terms decrease and increase, respectively, to zero for $t\neq\sigma\wedge\tau^0$.

For $t\le\sigma\wedge\tau^0$ the $g_y\cdot y$ -term in $g^{up,\delta}$ decreases by (3.15d), monotony of X^δ in δ and Lemma 3.4. For $t>\sigma\wedge\tau^0$ it decreases and is nonnegative thanks to the definition of $A^+_{g,\delta}$. As $\boldsymbol{\chi}_{t\le\sigma\wedge\tau^\delta}$ decreases to zero, the full $g_y\cdot y$ -component also decreases to zero.

Convergence of the $g_y\cdot y$ -term in $g^{lo,\delta}$ for $t\le\sigma\wedge\tau^0$ is trivial. For $t>\sigma\wedge\tau^0$ we are restricted to the negative values by definition of $A^-_{g,\delta}$, and as $\boldsymbol{\chi}_{t\le\sigma\wedge\tau^\delta}$ decreases to zero, the $g_y\cdot y$ -term increases to zero.

The g_z-term remains unaffected by (3.15e). So $g^{lo,\delta}\nearrow\tilde{g}$ and $g^{up,\delta}\searrow\tilde{g}$ as $\delta\searrow0$ for $t\neq\sigma\wedge\tau^0$.

Let $(\tilde{R},\tilde{Q})$ be the solution to the BSDE with data $(\tilde{\xi},\tilde{g},0)$. $\tilde{R}$ will serve as limiting process for $\tilde{R}^{up,\delta}$ and $\tilde{R}^{lo,\delta}$. By the comparison theorem, left continuity of the processes involved, and the above discussion of the data $(\xi^{up,\delta},g^{up,\delta})$ and $(\xi^{lo,\delta},g^{lo,\delta})$, we have

$$\tilde{R}^{lo,\delta}_t\le\tilde{R}_t\le\tilde{R}^{up,\delta}_t\qquad\forall t\in\bar{\imath}\ \boldsymbol{P}\text{-a.s.},$$

and $\tilde{R}_t-\tilde{R}^{lo,\delta}_t$ and $\tilde{R}^{up,\delta}_t-\tilde{R}_t$ decrease as $\delta\searrow0$.

Furthermore, the differences of the data and hence the processes $\tilde{R}_t-\tilde{R}^{lo,\delta}_t$ and $\tilde{R}^{up,\delta}_t-\tilde{R}_t$ decrease to zero in $L^2_{\mathcal{F}}(0,T;\mathbb{R})$. To be precise, set

$$\begin{aligned}\|\boldsymbol{\delta}g^{up,\delta}\|^2_2 &:= E\left[\int_0^{\sigma\wedge\tau^\delta}\left|\tilde{g}(s,\tilde{R}_s,\tilde{Q}_s)-g^{up,\delta}(s,\tilde{R}_s,\tilde{Q}_s)\right|^2ds\right],\\ \|\boldsymbol{\delta}g^{lo,\delta}\|^2_2 &:= E\left[\int_0^{\sigma\wedge\tau^\delta}\left|\tilde{g}(s,\tilde{R}_s,\tilde{Q}_s)-g^{lo,\delta}(s,\tilde{R}_s,\tilde{Q}_s)\right|^2ds\right],\\ \|\boldsymbol{\delta}\xi^{up,\delta}\|^2_2 &:= E\big[\big|\tilde{\xi}-\xi^{up,\delta}\big|^2\big],\\ \|\boldsymbol{\delta}\xi^{lo,\delta}\|^2_2 &:= E\big[\big|\tilde{\xi}-\xi^{lo,\delta}\big|^2\big],\end{aligned}$$

which converge to zero as $\delta \searrow 0$ by the dominated convergence theorem. The a priori estimates in Lemma 2.4 show that

$$\lim_{\delta\searrow 0} E\Big[\sup_{0\le s\le T}\big|\tilde R^{up,\delta}_s - \tilde R_s\big|^2\Big] = \lim_{\delta\searrow 0} E\Big[\sup_{0\le s\le T}\big|\tilde R^{lo,\delta}_s - \tilde R_s\big|^2\Big] = 0.$$

By monotony of $\tilde R^{lo,\delta}$ and $\tilde R^{up,\delta}$, these converge also $\boldsymbol{P}$-a.s. to $\tilde R$ uniformly in t. As $T \in \mathbb{R}_{>0}$ is bounded, $\tilde R^{lo,\delta}$ and $\tilde R^{up,\delta}$ converge to $\tilde R$ in $L^2_{\mathcal{F}}(0,T;\mathbb{R})$. □

Recall the payoff R of the associated Dynkin game in Definition 1.2. Now define the data

$$\begin{aligned}\xi^R &:= \Big(h_x(X^{x,0,0}_T)\boldsymbol{\chi}_{\sigma\wedge\tau^0=T} + a^U_\sigma \boldsymbol{\chi}_{\substack{\sigma\le\tau^0\\ \sigma<T}} + a^L_{\tau^0}\boldsymbol{\chi}_{\substack{\tau^0<\sigma\\ \tau^0<T}}\Big)\Gamma^x_{\sigma\wedge\tau^0\wedge T},\\ g^R(t,y,z) &:= \Big(g_x(t,X^{x,0,0}_t, Y^\star_t)\Gamma^x_t + g_y(t,X^{x,0,0}_t,Y^\star_t)\cdot y\Big)\boldsymbol{\chi}_{t\le\sigma\wedge\tau^0}\\ &\quad + g_z(t)\cdot z\boldsymbol{\chi}_{t\le\sigma},\end{aligned}$$

and $C^R := 0$. Then the solution (R,Q) of the BSDE with data $(\xi^R, g^R, a, 0)$ is the payoff defined in (1.3) and has the representation

$$\begin{aligned}R^x_t(\sigma,\tau^0) = E\Bigg[&\int_t^{\sigma\wedge\tau^0}\langle Dg(s,X^{x,0,0}_s,Y^\star_s),(\Gamma^x_s,R_s,Q_s)\rangle\,ds\\ &+\ h_x(X^{x,0,0}_T)\Gamma^x_T\boldsymbol{\chi}_{\sigma\wedge\tau^0=T} + a^U_\sigma\Gamma^x_\sigma\boldsymbol{\chi}_{\substack{\sigma\le\tau^0\\ \sigma<T}} + a^L_{\tau^0}\Gamma^x_{\tau^0}\boldsymbol{\chi}_{\substack{\tau^0<\sigma\\ \tau^0<T}}\,\Big|\,\mathcal{F}_t\Bigg].\end{aligned} \tag{3.18}$$

Lemma 3.7. *Assume that* (2.8) *and* (3.15) *hold. Then*

$$\lim_{\delta\searrow 0}\frac{1}{\delta}\big(Y^\delta_t - Y^\star_t\big) \le R^x_t(\sigma,\tau^0) \tag{3.19}$$

holds $\boldsymbol{P}$*-a.s. for all* $t \le \sigma\wedge\tau^0$.

Proof. We use the notation of Lemma 3.6. From the construction of $\tilde Y^\delta$ we have $\tilde Y^\delta_t = \frac{1}{\delta}\big(Y^\delta_t - Y^\star_t\big)$ for $t \le \sigma\wedge\tau^0$. Hence by (3.16)

$$\tilde R^{lo,\delta}_t \le \frac{1}{\delta}\big(Y^\delta_t - Y^\star_t\big) \le \tilde R^{up,\delta}_t \qquad \text{for } t\le\sigma\wedge\tau^0 \ \boldsymbol{P}\text{-a.s.}$$

As $\tilde R^{lo,\delta}$ and $\tilde R^{up,\delta}$ converge to $\tilde R$, so does $\frac{1}{\delta}\big(Y^\delta - Y^\star\big)$. Hence the limit in (3.19) exists, and it suffices to prove

$$\tilde R_t \le R^x_t(\sigma,\tau^0) \qquad \text{for } t\le\sigma\wedge\tau^0 \ \boldsymbol{P}\text{-a.s.}$$

First observe that $\tilde R$ remains unchanged if we restrict the g_z-term in the definition of $\tilde g$ to $[0,\sigma\wedge\tau^0]$. This follows from the fact that $\tilde R_t$ is $\mathcal{F}_{\sigma\wedge\tau^0}$-measurable for $t > \sigma\wedge\tau^0$; hence $\tilde Q_t = 0$ on $(\sigma\wedge\tau^0, T]$.

Recall that $X^\star_t = X^{x,C^U,0}_t$ for $t \le \tau^0$ by definition and that $X^{x,C^U,0} \le X^{x,0,0}$ and $\Gamma^{x,C^U,0} \le \Gamma^{x,0,0}$ by (2.8a) and the comparison theorem (Theorem 2.5).

Hence $\tilde g(t,y,z) \le g^R(t,y,z)$ is a consequence of (2.8d), (3.15c). Using (2.8c), (3.15b), and (3.15f) we deduce $\tilde\xi \le \xi^R$. Hence the assumptions of the comparison theorem are satisfied, which completes the proof. □

Let us now consider the case where $C = (C^U, -C^L)$ is optimal for the problem starting in x; i.e., the value $V_t(x)$ is the state process of a controlled BSDE:

$$V_t(x) = Y_t^\star = Y_t^{x,C} \qquad \forall t \geq 0.$$

LEMMA 3.8. *Assume that* (2.8) *and* (3.15) *hold and that there exists an optimal control* (C^U, C^L) *for the control problem* 1.1 *in* x. *Let* $V(x)$ *be its value. Then the upper right Dini derivative of* V *with respect to the initial condition satisfies*

$$\Delta^+ V(x) \leq u^-(x).$$

Proof. By optimality of C in x and Lemma 3.7 we have

$$\limsup_{\delta \searrow 0} \frac{1}{\delta}\big(V(x+\delta) - V(x)\big) \leq \limsup_{\delta \searrow 0} \frac{1}{\delta}\big(Y_0^\delta - Y_0^\star\big) \leq R_0^x(\sigma, \tau^0).$$

This holds even if the decomposition (3.4) is not minimal. As $\sigma \in \mathcal{T}$ is arbitrary we obtain

$$\Delta^+ V(x) \leq \operatorname*{ess\,inf}_{\sigma \in \mathcal{T}} R_0(\sigma, \tau^0) \leq u^-(x). \tag{3.20}$$ □

3.3. Lower left Dini derivative. The construction of an estimate for the lower left Dini derivative of V is essentially the same as that in section 3.1. This is intuitively obvious if we consider $-X$ instead of X as the controlled forward process, and all proofs relying on the set of conditions in (3.1) translate one to one to the new situation. For example, C^U and C^L just change their roles. We therefore only outline the argument by listing the statements.

However, for our investigation of the difference quotient the transformation $X \mapsto -X$ cannot be applied as it does not preserve convexity of the data. So we give the definitions of upper and lower bounds and limiting process for the difference quotient in detail.

Construction of the tracking process from below. We construct a process $X^{-\delta}$ that tracks a given controlled process $X^\star = X^{x,C^U,C^L}$, starting at a distance δ below x and jumping on $X^\star$ some time afterwards.

Let $\tau \in \mathcal{T}$ arbitrary and set the crossing time

$$\sigma^\delta := \inf\{t \geq 0 \,|\, X_t^{x-\delta,0,C^L} \geq X_t^{x,C^U,C^L}\}. \tag{3.21}$$

The tracking process $X^{-\delta}$ for the starting point $x - \delta$ is defined as

$$X_t^{-\delta} := \begin{cases} X_t^{x-\delta,0,C^L}, & t \leq \sigma^\delta \wedge \tau, \\ X_t^{x,C^U,C^L}, & t > \sigma^\delta \wedge \tau. \end{cases} \tag{3.22}$$

It has parallel lower and no upper control as long as either τ occurs or $X^\star$ crosses its path.

To obtain $X^{-\delta}$ as a controlled process define $C^{-\delta} = (C^{-\delta,U}, -C^{-\delta,L})$ by

$$C_t^{-\delta,U} := \begin{cases} 0, & t \leq \sigma^\delta \wedge \tau, \\ C_t^U - C_{\sigma^\delta}^U - \big(X_{\sigma^\delta}^{x,C^U,C^L} - X_{\sigma^\delta}^{x-\delta,0,C^L}\big), & t > \sigma^\delta, \sigma^\delta \leq \tau, \\ C_t^U - C_\tau^U, & t > \tau, \sigma^\delta > \tau, \end{cases} \tag{3.23a}$$

$$C_t^{-\delta,L} := \begin{cases} C_t^L, & t \leq \tau, \\ C_t^L, & t > \tau, \sigma^\delta \leq \tau, \\ C_t^L + (X_\tau^{x,C^U,C^L} - X_\tau^{x-\delta,0,C^L}), & t > \tau, \sigma^\delta > \tau. \end{cases} \tag{3.23b}$$

From the definition of σ^δ we have for $t > \sigma^\delta$, $\sigma^\delta \le \tau$

$$\begin{aligned}0 \le C^U_t - C^U_{\sigma^\delta+} + X^{x-\delta,0,C^L}_{\sigma^\delta+} - X^{x,C^U,C^L}_{\sigma^\delta+}&\\ = C^U_t - C^U_{\sigma^\delta+} + X^{x-\delta,0,C^L}_{\sigma^\delta} + (C^L_{\sigma^\delta+} - C^L_{\sigma^\delta})&\\ - \big(X^{x,C^U,C^L}_{\sigma^\delta} - (C^U_{\sigma^\delta+} - C^U_{\sigma^\delta}) + (C^L_{\sigma^\delta+} - C^L_{\sigma^\delta})\big)&\\ = C^U_t - C^U_{\sigma^\delta} - \big(X^{x,C^U,C^L}_{\sigma^\delta} - X^{x-\delta,0,C^L}_{\sigma^\delta}\big)&\end{aligned}$$

so $C^{-\delta,U}$ is increasing; hence $C^{-\delta} \in \mathcal{A}$. Again, this decomposition need not be minimal. In addition to the notation of $X^\star$, $Y^\star$, and $Z^\star$ in section 3.1 we set

$$Y^{-\delta}_t := Y^{x-\delta,C^{-\delta,U},C^{-\delta,L}}_t, \qquad Z^{-\delta}_t := Z^{x-\delta,C^{-\delta,U},C^{-\delta,L}}_t.$$

The following statements and their proofs are parallel to Lemmata 3.1, 3.2, 3.3, 3.4, and 3.5. As noted above, instead of a line-by-line imitation we could use a transformation

$$\begin{aligned}\check b(t,x) &:= -b(t,-x), & \check\sigma(t,x) &:= \sigma(t,x),\\ \check h(x) &:= -h(-x), & \check g(t,x,y,z) &:= -g(t,-x,-y,-z)\end{aligned}$$

by which

$$\begin{aligned}(X^{x,C^U,C^L}, Y^{x,C^U,C^L}, Z^{x,C^U,C^L}) \mapsto &(\check X^{-x,C^L,C^U}, \check Y^{-x,C^L,C^U}, \check Z^{-x,C^L,C^U})\\ &= (-X^{x,C^U.C^L}, -Y^{x,C^U,C^L}, -Z^{x,C^U,C^L})\end{aligned}$$

and apply the results of section 3.1.

LEMMA 3.9. *Let $C^{-\delta}$ be as defined in* (3.23) *and assume* (3.1) *holds. Then we have $X^{x-\delta,C^{-\delta,U},C^{-\delta,L}} = X^{-\delta}$ and*

$$0 \le X^{x,C^U,C^L}_t - X^{x-\delta,C^{-\delta,U},C^{-\delta,L}}_t \le X^{x,0,C^L}_t - X^{x-\delta,0,C^L}_t. \tag{3.24}$$

Furthermore, $Y^{x-\delta,C^{-\delta,U},C^{-\delta,L}}_t = Y^{x,C^U,C^L}_t$ and $Z^{x-\delta,C^{-\delta,U},C^{-\delta,L}}_t = Z^{x,C^U,C^L}_t$ for $\sigma^\delta \wedge \tau < t \le T$. As δ decreases, $X^{x-\delta,C^U,C^L}_t$ increases and $X^{-\delta}$ increases to $X^\star$.

A bound for C^U. Define the first action time σ^0 of C^U by

$$\sigma^0 := \inf\{t \ge 0 \,|\, X^{x,0,C^L}_t > X^\star_t\} = \inf\{t \ge 0 \,|\, C^U_t > C^U_0\} \tag{3.25}$$

and observe that $\sigma^\delta \searrow \sigma^0$ $\boldsymbol{P}$-a.s. With $\hat\Theta$ defined suitably and $\Gamma^{\hat\Theta}$ the solution of the linear SDE

$$d\Gamma^{\hat\Theta}_t = b_x(t,\hat\Theta_t)\Gamma^{\hat\Theta}_t\,dt + \sigma_x(t,\Gamma^{\hat\Theta}_t)\,dW_t, \qquad \hat X_0 = 1,$$

the following representation leads to the analogue of Lemma 3.3:

$$\begin{aligned}X^{x,0,C^L}_t - X^\star_t = \int_0^t b_x(s,\hat\Theta_s)\big(X^{x,0,C^L}_s - X^\star_s\big)\,ds + C^U_t&\\ + \int_0^t \sigma(s, X^{x,0,C^L}_s - X^\star_s)\,dW_s = \int_{\sigma^0}^{\sigma^0\vee t} \Gamma^{\hat\Theta}_t\big(\Gamma^{\hat\Theta}_s\big)^{-1}\,dC^U_s.&\end{aligned}$$

LEMMA 3.10. *Let σ^0 and σ^δ be as defined in* (3.25) *and* (3.21)*, and assume* (3.1) *holds. Then C_t^U has $\boldsymbol{P}$-a.s. upper and lower bounds for $\sigma^0 < t \le \sigma^\delta$:*

$$\big(X_t^{x,0,C^L} - X_t^\star\big)\frac{1}{\Gamma_t^{\hat\Theta}}\Big(\inf_{\sigma^0<s\le t}\Gamma_s^{\hat\Theta}\Big) \le C_t^U \le \big(X_t^{x,0,C^L} - X_t^\star\big)\frac{1}{\Gamma_t^{\hat\Theta}}\Big(\sup_{\sigma^0<s\le t}\Gamma_s^{\hat\Theta}\Big).$$

Thus C^U satisfies

$$C_t^U \le \delta\Gamma_t^{up}\big(\Gamma_t^{lo}\big)^{-1}\Big(\sup_{\sigma^0\le s\le t}\Gamma_s^{up}\Big) =: \delta\cdot\overline{C_t^{U,-\delta}} \qquad \text{for } t\le\sigma^\delta \ \boldsymbol{P}\text{-a.s.}$$

Limit behavior of cost functionals. The difference of tracked and tracking cost functional has the representation

$$\begin{aligned} Y_t^\star - Y_t^{-\delta} = E\Big[&\int_t^{\sigma^\delta\wedge\tau} g(s,X_s^\star,Y_s^\star,Z_s^\star) - g(s,X_s^{-\delta},Y_s^{-\delta},Z_s^{-\delta})\,ds \\ &+ \big(h(X_T^\star) - h(X_T^{-\delta})\big)\boldsymbol{\chi}_{\sigma^\delta\wedge\tau=T} + a_\tau^L\big(X_\tau^\star - X_\tau^{-\delta}\big)\boldsymbol{\chi}_{\substack{\tau<\sigma^\delta\\ \tau<T}} \\ &+ a_{\sigma^\delta}^U\big(X_{\sigma^\delta}^\star - X_{\sigma^\delta}^{-\delta}\big)\boldsymbol{\chi}_{\substack{\sigma^\delta\le\tau\\ \sigma^\delta<T}} + \int_{[t,\sigma^\delta\wedge\tau)} a_s^U\,dC_s^U \,\Big|\, \mathcal{F}_t\Big]. \end{aligned} \tag{3.26}$$

Again the value is monotone in δ, and $(Y^{-\delta}, Z^{-\delta})$ converges to $(Y^\star, Z^\star)$.

LEMMA 3.11. *Assume that* (3.1) *holds. Let, for $\delta,\delta'\in\mathbb{R}_{>0}$, the processes $Y^{-\delta}, Y^{-\delta'}$ be as defined in Lemma* 3.9*. Then*

$$Y_t^{-\delta'} \le Y_t^{-\delta} \le Y_t^\star \qquad \text{for } \delta\le\delta' \ \forall\ t \ \boldsymbol{P}\text{-a.s.}$$

LEMMA 3.12. *Assume that* (3.1) *holds. Then $Y_t^{-\delta}$ converges to $Y_t^\star$ $\boldsymbol{P}$-a.s. uniformly in t, and further,*

$$\lim_{\delta\searrow 0} E\Big[\sup_{t\le s\le T}\big|Y_s^\star - Y_s^{-\delta}\big|^2\Big] = 0, \tag{3.27}$$

$$\lim_{\delta\searrow 0} E\bigg[\int_t^T \big|Z_s^\star - Z_s^{-\delta}\big|^2\,ds\bigg] = 0. \tag{3.28}$$

The convergence in (3.27) *is monotone.*

Estimates for the difference quotient $\frac{1}{\delta}\big(Y^\star - Y^{-\delta}\big)$. The structure of our approach is the same as in section 3.2, but we have to make slight modifications in the definitions of upper and lower bounds $\hat R^{up,-\delta}$ and $\hat R^{lo,-\delta}$ for the transformed process $\hat Y^{-\delta}$. Again, the bounds converge to a common limit $\hat R$ that gives an estimate for the payoff of the associated Dynkin game.

We rewrite $\frac{1}{\delta}\big(Y^\star - Y^{-\delta}\big)$ as an uncontrolled process similar to the construction in Lemma 3.6 and consider

$$\hat Y_t^{-\delta} := \frac{1}{\delta}\left(\big(Y_t^\star - Y_t^{-\delta}\big) + \int_{[0,\sigma^\delta\wedge\tau\wedge t)} a_s^U\,dC_s^U\right) \quad\text{and}\quad \hat Z_t^{-\delta} := \frac{1}{\delta}\big(Z_t^\star - Z_t^{-\delta}\big).$$

It can be verified directly that $(\hat{Y}^{-\delta}, \hat{Z}^{-\delta})$ is the solution of a BSDE with data

$$\begin{aligned}
\xi^{\hat{Y}} &:= \frac{1}{\delta}\big(h(X_T^\star) - h(X_T^{-\delta})\big)\boldsymbol{\chi}_{\sigma^\delta\wedge\tau=T} + a_\tau^L\frac{1}{\delta}\big(X_\tau^\star - X_\tau^{-\delta}\big)\boldsymbol{\chi}_{\substack{\tau<\sigma^\delta\\ \tau<T}}\\
&\quad + a_{\sigma^\delta}^U\frac{1}{\delta}\big(X_{\sigma^\delta}^\star - X_{\sigma^\delta}^{-\delta}\big)\boldsymbol{\chi}_{\substack{\sigma^\delta\le\tau\\ \sigma^\delta<T}} + \frac{1}{\delta}\int_{[0,\sigma^\delta\wedge\tau)} a_s^U\,dC_s^U,\\
g^{\hat{Y}}(t,y,z) &:= \Big(\Delta_x\hat{g}_t\frac{1}{\delta}(X_t^\star - X_t^{-\delta}) + \Delta_y\hat{g}_t\boldsymbol{\chi}_{y\ge0}\cdot y\\
&\quad - \Delta_y\hat{g}_t\frac{1}{\delta}\int_{[0,t)} a_s^U\,dC_s^U\Big)\boldsymbol{\chi}_{t\le\sigma^\delta\wedge\tau} + \Delta_z\hat{g}_t\cdot z\boldsymbol{\chi}_{t\le\tau},\\
C_t^{\hat{Y}} &:= (0,0).
\end{aligned}$$

Here $\Delta_x\hat{g}_t$, $\Delta_y\hat{g}_t$, and $\Delta_z\hat{g}_t$ are defined as

$$\begin{aligned}
\Delta_x\hat{g}_t &:= \big(g(t,X_t^\star,Y_t^\star,Z_t^\star) - g(t,X_t^{-\delta},Y_t^\star,Z_t^\star)\big)\frac{1}{X_t^\star - X_t^{-\delta}}\boldsymbol{\chi}_{X_t^{-\delta}\neq X_t^\star},\\
\Delta_y\hat{g}_t &:= \big(g(t,X_t^{-\delta},Y_t^\star,Z_t^\star) - g(t,X_t^{-\delta},Y_t^{-\delta},Z_t^\star)\big)\frac{1}{Y_t^\star - Y_t^{-\delta}}\boldsymbol{\chi}_{Y_t^{-\delta}\neq Y_t^\star},\\
\Delta_z\hat{g}_t &:= \big(g(t,X_t^{-\delta},Y_t^{-\delta},Z_t^\star) - g(t,X_t^{-\delta},Y_t^{-\delta},Z_t^{-\delta})\big)\frac{1}{Z_t^\star - Z_t^{-\delta}}\boldsymbol{\chi}_{Z_t^{-\delta}\neq Z_t^\star}.
\end{aligned}$$

Observe that $\hat{Y}^{-\delta}\ge 0$ by (3.15f) and Lemma 3.11, and that for $t\le\sigma^0\wedge\tau$, $\hat{Y}_t^{-\delta} = \frac{1}{\delta}(Y_t^\star - Y_t^{-\delta})$.

We will assume that (3.15) holds. Again we define processes $(\hat{R}^{up,-\delta},\hat{Q}^{up,-\delta})$ and $(\hat{R}^{lo,-\delta},\hat{Q}^{lo,-\delta})$ as solutions to the BSDE with data $(\xi^{up,-\delta},g^{up,-\delta},C^{up,-\delta})$ and $(\xi^{lo,-\delta},g^{lo,-\delta},C^{lo,-\delta})$, where

$$\begin{aligned}
\xi^{up,-\delta} &:= h_x(X_T^{x,0,C^L})\Gamma_T^{x,0,C^L}\boldsymbol{\chi}_{\sigma^\delta\wedge\tau=T} + a_\tau^L\Gamma_\tau^{x,0,C^L}\boldsymbol{\chi}_{\substack{\tau<\sigma^\delta\\ \tau<T}}\\
&\quad + \big(\max_{\sigma^0\le s\le\sigma^\delta} a_s^U\big)\big(\sup_{\sigma^0\le s\le\sigma^\delta}\Gamma_s^{x,0,C^L}\big)^2\big(\inf_{\sigma^0\le s\le\sigma^\delta}\Gamma_s^{x-\delta,0,C^L}\big)^{-1}\boldsymbol{\chi}_{\substack{\sigma^0\le\tau\\ \sigma^0<T}},\\
\xi^{lo,-\delta} &:= h_x(X_T^{x-\delta,0,C^L})\Gamma_T^{x-\delta,0,C^L}\boldsymbol{\chi}_{\sigma^0\wedge\tau=T} + a_\tau^L\Gamma_\tau^{x-\delta,0,C^L}\boldsymbol{\chi}_{\substack{\tau<\sigma^0\\ \tau<T}}\\
&\quad + \big(\min_{\sigma^0\le s\le\sigma^\delta} a_s^U\big)\big(\inf_{\sigma^0\le s\le\sigma^\delta}\Gamma_s^{x-\delta,0,C^L}\big)^2\big(\sup_{\sigma^0\le s\le\sigma^\delta}\Gamma_s^{x,0,C^L}\big)^{-1}\boldsymbol{\chi}_{\substack{\sigma^\delta\le\tau\\ \sigma^\delta<T}},\\
g^{up,-\delta}(t,y,z) &:= \Big(g_x(t,X_t^\star,Y_t^\star)\Gamma_t^{x,0,C^L} + g_y(t,X_t^\star,Y_t^\star)\boldsymbol{\chi}_{y\ge0}\boldsymbol{\chi}_{A^+_{g,-\delta}}\cdot y\\
&\quad + L\big(\max_{\sigma^0\le s\le\sigma^\delta\wedge t} a_s^U\big)\overline{C_t^{U,-\delta}}\Big)\boldsymbol{\chi}_{t\le\sigma^\delta\wedge\tau} + g_z(t)\cdot z\boldsymbol{\chi}_{t\le\tau},\\
g^{lo,-\delta}(t,y,z) &:= g_x(t,X_t^{-\delta},Y_t^\star)\Gamma_t^{x-\delta,0,C^L}\boldsymbol{\chi}_{t\le\sigma^0\wedge\tau}\\
&\quad + \Big(g_y(t,X_t^{-\delta},Y_t^{-\delta})\boldsymbol{\chi}_{y\ge0}\boldsymbol{\chi}_{A^-_{g,-\delta}}\cdot y\\
&\quad - L\big(\max_{\sigma^0\le s\le\sigma^\delta\wedge t} a_s^U\big)\overline{C_t^{U,-\delta}}\Big)\boldsymbol{\chi}_{t\le\sigma^\delta\wedge\tau} + g_z(t)\cdot z\boldsymbol{\chi}_{t\le\tau},\\
C_t^{up,-\delta} &:= C_t^{lo,-\delta} := (0,0).
\end{aligned}$$

Here $A^+_{g,-\delta}$ and $A^-_{g,-\delta}$ are defined as

$$A^+_{g,-\delta} := \{t \le \sigma^0 \wedge \tau\} \cup \{g_y(t, X^\star_t, Y^\star_t) \ge 0\},$$
$$A^-_{g,-\delta} := \{t \le \sigma^0 \wedge \tau\} \cup \{g_y(t, X^{-\delta}_t, Y^{-\delta}_t) \le 0\}.$$

We use the convention $\sup_{\sigma^0 \le s \le \sigma^\delta \wedge t} a^U_s = 0$ for $t < \sigma^0$. L denotes a Lipschitz constant for g.

The limiting process $\hat{R}$ is defined as the state process of a BSDE with data

$$\hat{\xi} := h_x(X^{x,0,C^L}_T)\Gamma^{x,0,C^L}_T \boldsymbol{\chi}_{\sigma^0 \wedge \tau = T} + a^L_\tau \Gamma^{x,0,C^L}_\tau \boldsymbol{\chi}_{\substack{\tau < \sigma^0 \\ \tau < T}} + a^U_{\sigma^0}\Gamma^{x,0,C^L}_{\sigma^0} \boldsymbol{\chi}_{\substack{\sigma^0 \le \tau \\ \sigma^0 < T}},$$

$$\hat{g}(t,y,z) := \Big(g_x(t, X^\star_t, Y^\star_t)\Gamma^{x,0,C^L}_t + g_y(t, X^\star_t, Y^\star_t)\boldsymbol{\chi}_{y \ge 0} \cdot y\Big)\boldsymbol{\chi}_{t \le \sigma^0 \wedge \tau} + g_z(t) \cdot z \boldsymbol{\chi}_{t \le \tau}$$

and control zero. Its property as limit of the difference quotient is the subject of the following lemma, which is the analogue to Lemma 3.6. The proof is straightforward.

LEMMA 3.13. *Assume that* (2.8) *and* (3.15) *hold. The processes* $\hat{R}^{up,-\delta}$, $\hat{Y}^{-\delta}$, $\hat{R}^{lo,-\delta}$ *satisfy*

$$\hat{R}^{lo,-\delta}_t \le \hat{Y}^{-\delta}_t \le \hat{R}^{up,-\delta}_t \qquad \forall t \in \bar{\imath}\ \boldsymbol{P}\text{-a.s.} \tag{3.29}$$

Further, $\hat{R}^{up,-\delta}_t$ *decreases and* $\hat{R}^{lo,-\delta}_t$ *increases as* $\delta \searrow 0$ *to the same limiting process* $\hat{R}$ *defined above, in* $L^2_{\mathcal{F}}(0,T;\mathbb{R})$ *and for all* $t \in \bar{\imath}$ $\boldsymbol{P}$*-a.s.*

We now can formulate the estimate in terms of the Dynkin game 1.2.

LEMMA 3.14. *Assume that* (2.8) *and* (3.15) *hold. Then*

$$\lim_{\delta \searrow 0} \frac{1}{\delta}\big(Y^\star_t - Y^{-\delta}_t\big) \ge R^x_t(\sigma^0, \tau) \tag{3.30}$$

holds $\boldsymbol{P}$*-a.s. for all* $t \le \sigma^0 \wedge \tau$.

Proof. By Lemma 3.13 it suffices to prove

$$\hat{R}_t \ge R^x_t(\sigma^0, \tau) \qquad \boldsymbol{P}\text{-a.s. for } t \le \sigma^0 \wedge \tau.$$

This follows from the comparison theorem (Theorem 2.7), if the data satisfy $\hat{g}(t,y,z) \ge g^R(t,y,z)$ and $\hat{\xi} \ge \xi^R$. But this is a consequence of the convexity assumptions. Especially observe that $\Gamma^{x,0,C^L} \ge \Gamma^{x,0,0}$ holds for the deflator processes, as b is convex and $X^{x,0,C^L} \ge X^{x,0,0}$. □

We now conclude this discussion with the following lemma.

LEMMA 3.15. *Assume that* (2.8) *and* (3.15) *hold and that there exists an optimal control* (C^U, C^L) *for the control problem* 1.1 *in* x. *Let* $V(x)$ *be its value. Then the lower left Dini derivative of* V *with respect to the initial condition satisfies*

$$\Delta_- V(x) \ge u^+(x).$$

Proof. By the optimality of $(C^U, -C^L)$ in x and Lemma 3.7 we have

$$\liminf_{\delta \searrow 0} \frac{1}{\delta}\big(V(x) - V(x-\delta)\big) \ge \liminf_{\delta \searrow 0} \frac{1}{\delta}\big(Y^\star_0 - Y^{-\delta}_0\big) \ge R^x_0(\sigma^0, \tau).$$

As $\tau \in \mathcal{T}$ is arbitrary this gives

$$\Delta_- V(x) \ge \operatorname*{ess\,sup}_{\tau \in \mathcal{T}} R^x_0(\sigma^0, \tau) \ge u^+(x) \qquad \boldsymbol{P}\text{-a.s} \tag{3.31}$$ □

3.4. Proof of Theorem 1.3. By Lemma 3.15, the definitions in (1.5), Lemma 3.8, and Theorem 2.10 we have the relations

$$\Delta^+ V(x) \le u^-(x) \le u^+(x) \le \Delta_- V(x) \le \Delta^+ V(x). \tag{3.32}$$

Consequently, equality holds in (3.32). So V is differentiable at x with partial derivative equal to the solution of the Isaac's equation $u^-(x) = u^+(x) = u(x)$.

Let us suppress dependence on x. By (3.20), (3.31), (3.32), and the Isaac's equation

$$\operatorname*{ess\,inf}_{\sigma\in\mathcal{T}} R_0(\sigma,\tau^0) = \operatorname*{ess\,sup}_{\tau\in\mathcal{T}} \operatorname*{ess\,inf}_{\sigma\in\mathcal{T}} R_0(\sigma,\tau)$$
$$= u(x) = \operatorname*{ess\,inf}_{\sigma\in\mathcal{T}} \operatorname*{ess\,sup}_{\tau\in\mathcal{T}} R_0(\sigma,\tau) = \operatorname*{ess\,sup}_{\tau\in\mathcal{T}} R_0(\sigma^0,\tau).$$

Hence (σ^0,τ^0) is an optimal pair or saddle point for the associated Dynkin game.

4. Extensions. In this section we briefly discuss some variations, such as weaker assumptions on the data and certain subspaces of controls.

Relaxing requirements on the data. Although the results of Theorem 1.3 are presented in a fairly general situation, some of the conditions in (2.8) and (3.15) can be found to be very restrictive. So we note that the removal of global Lipschitz conditions on the cost data g, h w.r.t. the state X and extension to a stochastic, possibly infinite time horizon can be achieved through similar localization arguments as in Boetius and Kohlmann [18]. To ensure validity of comparison theorems and a priori estimates one may consult, e.g., El Karoui and Quenez [32].

Removing the monotonicity of the cost data in x will be possible in exchange for stronger assumptions on the forward equation, notably the requirement that b is affine, and integrability conditions to ensure existence of the cost functionals of the Dynkin game. In this way quadratic costs can be included in our considerations.

To lift the assumption of Lipschitz continuity for the running cost g w.r.t. y and z is less easy because then the BSDE may no longer have a solution; see Bender and Kohlmann [10] or Bender [9] for solvability of BSDE under weak conditions.

Relaxation of convexity assumptions without losing convexity of the value is considered by Alvarez [2].

Monotone control as limiting case. From Definitions 1.1 and 1.2 we obtain a monotone control problem and a problem of optimal stopping if we require $C^L = 0$ and $\tau = T$. Let V^m and u^m denote the respective values. The monotone control problem models an irreversible investment problem, so it is natural to ask whether this is obtainable as a limit in some sense of partially reversible investment problems. In fact, when we consider the control problem and Dynkin game under the convex structure (2.8) and choose $a^L \le 0$, we obtain

$$V^m(x) = V(x), \qquad u^m(x) = u(x).$$

For the control problem this becomes intuitively clear if we recall that the cost data are increasing in x; thus exercising control C^L leads to a worse cost situation in the running and terminal costs for any future time, and, in addition, we incur the cost $-a^L_t\, dC^L_t$. Thus, letting $a^L = 0$ will allow us to treat the monotone case as a limiting situation. This is even more obvious in the Dynkin game: if $a^L \le 0$, the player seeking to maximize the cost functional R will never terminate the game, because he would forego the chance of profiting from the running cost g_x and the terminal cost h_x.

Modifications on the space of controls. Monotone follower problems with finite fuel, a limitation on the amount of control to be exercised, form a particularly interesting class of control problems. In many situations it is possible to express the value of the control problem with finite fuel in terms of the value without such additional condition and the risk with zero control available; see e.g., Chow, Menaldi, and Robin [20], or Karatzas [43]. The most interesting element is that for the monotone follower there is no time value of control or additional hysteresis effect, so a control optimal under a finite fuel restriction does not save resources compared to optimal behavior without restriction. The reason for this is that the "displacement" caused by a certain amount of controlling activity is constant over time, and so is its influence on the cost functional.

This result carries over to the situation of Theorem 1.3 if one considers a generalized finite fuel condition defined as follows: Take the starting values from an interval $[a,b] \subset \mathbb{R}$ and call controls $C \in \mathcal{A}$ admissible, or elements of $\mathcal{A}_{[a,b]}$, if the controlled forward process $X^{0,x,C}$ stays in the moving interval $[X^{0,a,0}, X^{0,b,0}]$. Denoting the value by $V_{[a,b]}$ it is easy to see that Theorem 1.3 still holds.

The same is true if we consider subclasses of $\mathcal{A}$ or $\mathcal{A}_{[a,b]}$ whose elements have a finite *partition of monotone control* $\boldsymbol{P}$-a.s., which means that there exists a sequence of $\mathcal{F}_t$-stopping times $(\theta_n)_{\mathbb{N}}$ such that $\theta_0 = 0$,

$$\theta_n \xrightarrow{=} T, \qquad C_t \text{ is monotone in } [\theta_{n-1}, \theta_n).$$

Details can be found in [17] or Boetius and Kohlmann [18].

These modifications are useful when one tries to construct solutions to the partially irreversible investment problem from optimal strategies of small investors. The latter take the form of optimal stopping or "entry-exit" sequential stopping problems; see [16] or Baldursson and Karatzas [5]. In the case of monotone control one finds that the control problem and a family of stopping problems are equivalent. In the case of bounded variation control this leads to a representation of the Dynkin game in terms of two closely related entry-exit problems; see [17].

Acknowledgments. I am very much indebted to Michael Kohlmann for many helpful and constructive discussions and to two unknown referees for their useful and precise remarks and recommendations that made this paper more readable.

REFERENCES

[1] L. H. R. Alvarez, *Optimal exit and valuation under demand uncertainty: A real options approach*, European J. Oper. Res., 114 (1999), pp. 320–329.

[2] L. H. R. Alvarez, *Singular stochastic control, linear diffusions, and optimal stopping: A class of solvable problems*, SIAM J. Control Optim., 39 (2001), pp. 1697–1710.

[3] L. H. R. Alvarez and Rune Stenbacka, *Adoption of uncertain multi-stage technology projects: A real options approach*, J. Math. Econom., 35 (2001), pp. 71–97.

[4] F. M. Baldursson, *Singular stochastic control and optimal stopping*, Stochastics, 21 (1987), pp. 1–40.

[5] F. M. Baldursson and I. Karatzas, *Irreversible investment and industry equilibrium*, Finance Stoch., 1 (1997), pp. 69–89.

[6] J. S. Baras, R. J. Elliot, and M. Kohlmann, *The partially observed stochastic minimum principle*, SIAM J. Control Optim., 27 (1989), pp. 1279–1292.

[7] J. A. Bather and H. Chernoff, *Sequential decisions in the control of a spaceship*, in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 3, (1966), University of California Press, Berkeley, CA, 1967, pp. 181–207.

[8] P. Baxendale, *Stability and equilibrium properties of stochastic flows of diffeomorphisms*, in Diffusion Processes and Related Problems in Analysis, Vol. 2, Progr. Probab. 27, M. A. Pinsky and V. Wihsturz, eds., Birkhäuser Boston, Boston, 1992, pp. 3–35.

[9] C. Bender, *Rückwärtsstochastische Differentialgleichungen und Anwendungen bei der Bewertung von Finanzderivaten*, Diplomarbeit, Universität Konstanz, Konstanz, Germany, 2002.
[10] C. Bender and M. Kohlmann, *BSDEs with stochastic Lipschitz condition*, CoFE Discussion Paper 00-08, http://cofe.uni-konstanz.de (2000).
[11] V. E. Beneš, L. A. Shepp, and H. S. Witsenhausen, *Some solvable stochastic control problems*, Stochastics, 4 (1980), pp. 39–83.
[12] A. Bensoussan and J. L. Lions, *Applications of Variational Inequalities in Stochastic Control*, North–Holland, Amsterdam, New York, 1982.
[13] F. E. Benth, K. H. Karlsen, and K. Reikvam, *Optimal portfolio management rules in a non-Gaussian market with durability and intertemporal substitution*, Finance Stoch., 5 (2001), pp. 447–467.
[14] J. M. Bismut, *Sur un problème de Dynkin*, Wahrscheinlichkeitstheorie und Verw. Giebiete, 39 (1977), pp. 31–53.
[15] J. M. Bismut, *Temps d'arrêt optimal, quasi-temps d'arrêt et retournement du temps*, Ann. Probab., 7 (1979), pp. 933–963.
[16] F. Boetius, *Entry-exit problems and bounded variation singular stochastic control*, working paper, 2003.
[17] F. Boetius, *Singular Stochastic Control and Its Relations to Dynkin Game and Entry-Exit Problems*, Ph.D. thesis, Universität Konstanz, Konstanz, Germany, 2003.
[18] F. Boetius and M. Kohlmann, *Connections between optimal stopping and singular stochastic control*, Stochastic Process Appl., 77 (1998), pp. 253–281.
[19] A. Cadenillas, *Consumption-investment problems with transaction costs: Survey and open problems*, Math. Methods Oper. Res., 51 (2000), pp. 43–68.
[20] P.-L. Chow, J.-L. Menaldi, and M. Robin, *Additive control of stochastic linear systems with finite horizon*, SIAM J. Control Optim., 23 (1985), pp. 858–899.
[21] J. Cvitanić and I. Karatzas, *Backward stochastic differential equations with reflection and Dynkin games*, Ann. Probab., 24 (1996), pp. 2024–2056.
[22] R. W. R. Darling, *Isotropic stochastic flows*, in Diffusion Processes and Related Problems in Analysis, Vol. 2, Progr. Probab. 27, M. A. Pinsky and V. Wihsturz, eds., Birkhäuser Boston, Boston, 1992, pp. 75–94.
[23] M. H. A. Davis and I. Karatzas, *A deterministic approach to optimal stopping*, in Probability, Statistics, and Optimization, F. P. Kelly, ed., Wiley, New York, 1994, pp. 455–466.
[24] M. H. A. Davis and A. Norman, *Portfolio selection with transaction costs*, Math. Oper. Res., 15 (1990), pp. 676–713.
[25] A. K. Dixit and R. S. Pindyck, *Investment under Uncertainty*, Princeton University Press, Princeton, NJ, 1994.
[26] D. Duffie and L. G. Epstein, *Stochastic differential utility*, Econometrica, 60 (1992), pp. 353–394.
[27] E. B. Dynkin, *Game variant of a problem on optimal stopping*, Soviet Math. Dokl., 10 (1969), pp. 270–274.
[28] N. El Karoui, C. Kapoudjian, E. Pardoux, S. Peng, and M. C. Quenez, *Reflected solutions of backward SDE's, and related obstacle problems for PDE's*, Ann. Probab., 25 (1997), pp. 702–737.
[29] N. El Karoui and I. Karatzas, *Probabilistic aspects of finite-fuel, reflected follower problems*, Acta Appl. Math., 11 (1988), pp. 223–258.
[30] N. El Karoui and I. Karatzas, *A new approach to the Skorohod problem, and its applications*, Stochastics Stochastics Rep., 34 (1991), pp. 57–82.
[31] N. El Karoui, S. Peng, and M. C. Quenez, *Backward stochastic differential equations in finance*, Math. Finance, 7 (1997), pp. 1–71.
[32] N. El Karoui and M. C. Quenez, *Non-linear pricing theory and backward stochastic differential equations*, in Financial Mathematics (Bressanone, Italy, 1996), Lecture Notes in Math. 1656, W. Runggaldier, ed., Springer-Verlag, Berlin, 1997, pp. 191–246.
[33] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Appl. Math. 25, Springer-Verlag, New York, 1993.
[34] A. Friedman, *Stochastic Differential Equations and Applications*, Vol. 2, Academic Press, New York, London, 1976.
[35] D. Fudenberg and J. Tirole, *Game Theory*, MIT Press, Cambridge, MA, 1991.
[36] S. Hamadene and J.-P. Lepeltier, *Backward equations, stochastic control and zero-sum stochastic differential games*, Stochastics Stochastics Rep., 54 (1995), pp. 221–231.
[37] S. Hamadene and J.-P. Lepeltier, *Zero-sum stochastic differential games and backward equations*, Systems Control Lett., 24 (1995), pp. 259–263.
[38] S. Hamadene and J.-P. Lepeltier, *Reflected BSDEs and mixed game problem*, Stochastic Process Appl., 85 (2000), pp. 177–188.

[39] S. Hamadene, J.-P. Lepeltier, and Z. Wu, *Infinite horizon reflected backward stochastic differential equations and applications in mixed control and game problems*, Probab. Math. Statist., 19 (1999), pp. 211–234.
[40] U. G. Haussmann and W. Suo, *Singular optimal stochastic controls* I: *Existence*, SIAM J. Control Optim., 33 (1995), pp. 916–936.
[41] B. Højgaard and M. Taksar, *Optimal risk control for a large corporation in the presence of returns on investments*, Finance Stoch., 5 (2001), pp. 527–547.
[42] I. Karatzas, *A class of singular stochastic control problems*, Adv. in Appl. Probab., 15 (1983), pp. 225–254.
[43] I. Karatzas, *Probabilistic aspects of finite fuel stochastic control*, Proc. Natl. Acad. Sci. USA, 82 (1985), pp. 5579–5581.
[44] I. Karatzas and S. E. Shreve, *Connections between optimal stopping and singular stochastic control* I. *Monotone follower problems*, SIAM J. Control Optim., 22 (1984), pp. 856–877.
[45] I. Karatzas and S. E. Shreve, *Connections between optimal stopping and singular stochastic control* II. *Reflected follower problems*, SIAM J. Control Optim., 23 (1985), pp. 433–451.
[46] I. Karatzas and S. E. Shreve, *Equivalent models for finite-fuel stochastic control*, Stochastics, 18 (1986), pp. 245–276.
[47] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, Grad. Texts in Math. 113, 3rd ed., Springer-Verlag, New York, 1994.
[48] I. Karatzas and S. E. Shreve, *Methods of Mathematical Finance*, Appl. Math. 39, Springer-Verlag, New York, 1998.
[49] I. Karatzas and H. Wang, *Connections between bounded-variation control and Dynkin games*, in Optimal Control and Partial Differential Equations, IOS Press, Amsterdam, 2001, pp. 353–362.
[50] Y. Kifer, *Game options*, Finance Stoch., 4 (2000), pp. 443–463.
[51] J. Komlós, *A generalization of a problem of Steinhaus*, Acta Math. Acad. Sci. Hungar., 18 (1967), pp. 217–229.
[52] L. Kruk, *Optimal policies for n-dimensional singular stochastic control problems. Part* I: *The Skorohod problem*, SIAM J. Control Optim., 38 (2000), pp. 1603–1622.
[53] L. Kruk, *Optimal policies for n-dimensional singular stochastic control problems. Part* II: *The radially symmetric case. Ergodic control*, SIAM J. Control Optim., 39 (2000), pp. 635–659.
[54] H. J. Kushner, *Control and optimal control of assemble to order manufacturing systems under heavy traffic*, Stochastics Stochastics Rep., 66 (1999), pp. 233–27.
[55] J. Ma and J. Yong, *Dynamic programming for multidimensional stochastic control problems*, Acta Math. Sin. (Engl. Ser.), 15 (1999), pp. 485–506.
[56] J. Ma and J. Yong, *Forward-backward Stochastic Differential Equations and Their Applications*, Lecture Notes in Math. 1702, Springer-Verlag, Berlin, 1999.
[57] R. McDonald and D. Siegel, *The value of waiting to invest*, Quart. J. Econom., 101 (1986), pp. 707–727.
[58] H. P. McKean, Jr., *Appendix: A free boundary problem for the heat equation arising from a problem in mathematical economics*, Ind. Management Rev., 6 (1965), pp. 32–39.
[59] J. L. Menaldi and M. I. Taksar, *Optimal correction problem of a multidimensional stochastic system*, Automatica J. IFAC, 25 (1989), pp. 223–232.
[60] S. Peng, *A generalized dynamic programming principle and Hamilton–Jacobi–Bellman equation*, Stochastics Stochastics Rep., 38 (1992), pp. 119–134.
[61] S. Peng, *Monotonic limit theorem of BSDE and nonlinear decomposition theorem of Doob–Meyer's type*, Probab. Theory Related Fields, 113 (1999), pp. 473–499.
[62] P. Protter, *Stochastic Integration and Differential Equations*, Appl. Math. 21, Springer-Verlag, Berlin, 1990.
[63] P. A. Samuelson, *Rational theory of warrant pricing*, Ind. Mangement Rev., 6 (1965), pp. 13–31.
[64] U. Schmock, S. E. Shreve, and U. Wystup, *Valuation of exotic options under shortselling constraints*, Finance Stoch., 6 (2002), pp. 143–172.
[65] A. N. Shiryaev, *Optimal Stopping Rules*, Springer-Verlag, Berlin, Heidelberg, New York, 1978.
[66] H. M. Soner and S. E. Shreve, *Regularity of the value function for a two-dimensional singular stochastic control problem*, SIAM J. Control Optim., 27 (1989), pp. 876–907.
[67] M. Taksar, *Average optimal singular control and a related stopping problem*, Math. Oper. Res., 10 (1985), pp. 63–81.
[68] J. Yong and X. Y. Zhou, *Stochastic Controls: Hamiltonian Systems and HJB Equations*, Springer-Verlag, New York, 1999.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 4, pp. 1322–1344

OPTIMAL CONSUMPTION-INVESTMENT PROBLEMS IN INCOMPLETE MARKETS WITH STOCHASTIC COEFFICIENTS*

NETZAHUALCÓYOTL CASTAÑEDA-LEYVA† AND DANIEL HERNÁNDEZ-HERNÁNDEZ‡

Abstract. The goal of this paper is to solve an optimal consumption-investment problem in the context of an incomplete financial market. The model is a generalization of the Black and Scholes diffusion model, where the coefficients of the diffusion modelling the stock's price depend on some stochastic economic factors. Based on the martingale approach, a basic methodology to get the optimal solution is presented. Combining this procedure with stochastic control techniques, explicit solutions for HARA and logarithmic utility functions are obtained.

Key words. optimal investment and consumption, incomplete markets, stochastic volatility, martingale method, optimal control, Black–Scholes model

AMS subject classifications. 91B28, 49L20

DOI. 10.1137/S0363012904440885

1. Introduction.

Since the fundamental work of Black and Scholes to value European options, their model has become a cornerstone in the development and study of many problems in mathematical finance. In recent years, different generalizations of this classical model have been studied, trying to model more precisely the dynamics of the asset prices. In this sense, it is natural to consider a model with stochastic coefficients (interest rate, return rate, and volatility), depending on economic external factors. In fact, several contributions show empirical arguments justifying these kinds of models. For example, Fouque, Papanicolaou, and Sircar [FPS00] present a detailed analysis when the external factor is a mean reverting Ornstein–Uhlenbeck (O–U) process, which can be, for instance, a *leader* interest rate. On the other hand, recently, Barndorff–Nielsen and Shephard [BaSp02] propose a model for volatility based on the O–U process with background subordinator (nonnegative Levy process). They also give a detailed statistical analysis, identifying important volatility effects in the asset prices: heavy tailed of returns, volatility clustering, and right skewness in some cases.

The relevance of the diffusion models presented in this paper is not limited only to economical or empirical qualities, they also have proved to be tractable. For example, we can find explicit solutions of problems in the context of optimal investment (Zariphopoulou [Za01]), optimal consumption process (Fleming and Hernández-Hernández [FlHe02]), and valuation (Davis [Da00]). This feature contrasts with technical constraints or difficulties in the implementation of other affine approaches. For instance, in the model proposed by Barndorff–Nielsen and Shephard [BaSp02] their background subordinator induces a constraint for the trading portfolio proportion, which should belong to the interval $[0, 1]$. This fact was mentioned by Benth, Karlsen,

*Received by the editors February 9, 2004; accepted for publication (in revised form) December 28, 2004; published electronically October 21, 2005.
http://www.siam.org/journals/sicon/44-4/44088.html

†Universidad Autónoma de Aguascalientes, Departamento de Estadística, Centro de Ciencias Básicas, Av. Universidad 940, Ciudad Universitaria, Aguascalientes, Ags., C.P. 20100, México (ncastane@uaa.mx). The research of this author was supported by grants UAA-PIEST-04-01 and PROMEP-UAAGS-EXB-69.

‡Corresponding author. Centro de Investigación en Matemáticas Apartado Postal 402, Guanajuato, Gto., C.P. 36000, México (dher@cimat.mx). The research of this author was supported by Conacyt grant 37643-E.

and Reikvam [BKR03] in their study of some investment problems.

The goal of this work is to solve the problem of maximizing the expected utility of terminal wealth and/or consumption in some finite time interval $[0, T]$, as well as to find an *optimal* trading strategy.

The investor's financial market is composed of a bank account, a risky asset, and an economic external correlated factor. The dynamics of the risky asset price and the external factor are diffusion processes where, as it was already mentioned, this last one affects the coefficients of the model. On the other hand, since the external factor is not traded, this problem drops into the family of incomplete markets. We deal with two particular utility functions: hyperbolic absolute risk aversion (HARA) and logarithmic. However, since the arguments are similar, most of the effort is concentrated in solving the optimization problem when the utility function is HARA, which is also the most complex.

We will use the martingale method to solve this problem, formulating the investor's problem as a convex optimization one, referred to as the *primal problem*. In this context, the primal problem has an associated *dual* problem, which turns out to be an stochastic optimal control problem, where the control processes belong to the set of equivalent local martingale measures.

The martingale method goes back to the fundamental contribution by Harrison and Pliska [HaPl81], and it has now become a popular approach to study optimal terminal wealth and/or consumption problems. This method is particularly powerful when the financial market is incomplete. For instance, in Karatzas and Shreve [KaSr98] and some references therein, a wide class of optimization problems for incomplete markets are studied using this approach. However, explicit optimal solutions are not presented in general, except for logarithmic utility or when the coefficients are deterministic. In Kramkov and Schachermayer [KrSc99] optimal investment problems for incomplete markets are analyzed when the stock prices are driven by semimartingales, for a wide class of utility functions. In both references, under suitable conditions, some characterizations of the optimization problem are presented. In particular, they show that there is no duality gap between the primal and dual problems.

We shall solve the investor's problem using a composition of the martingale method and stochastic control techniques. With this goal in mind, we pose the primal and dual problems and state the existence of their solutions, which shall imply the absence of duality gap. When the utility function is HARA, the solution to the dual problem relies on stochastic control techniques, while in the logarithmic case the solution is straightforward.

The paper is organized as follows. In section 2 the model and the investor's problem as well as its primal representation is established. Next, in section 3 we write down the associated dual problem. The martingale method is also explained, and a practical condition for absence of duality gap is given. Furthermore, a relevant relationship between optimal solutions of both problems is obtained. Also, we present the basic steps of the martingale method. Finally, in section 4 closed form solutions, when the utility function is HARA and logarithmic, are presented.

Throughout the paper all inequalities involving random variables are understood to be satisfied almost surely.

2. The model and primal problem. In this section the underlying model of the financial market is introduced, and a stochastic dynamic model for optimal consumption and investment on a finite horizon is considered.

Let $(\Omega, \mathcal{F}_T, P)$ be a complete probability space, where a two-dimensional standard

Brownian motion (BM) $\{(W_{1t}, W_{2t}), \mathcal{F}_t\}_{0\leq t\leq T}$ is defined, with $\{\mathcal{F}_t\}_{0\leq t\leq T}$ being the augmentation of the filtration $\left\{\mathcal{F}_t^{(W_{1t},W_{2t})}\right\}_{0\leq t\leq T}$. The securities market is modelled using this BM and involves a risky asset, a bond, and an external economic factor such that, for $0 \leq t \leq T$:

1. The bond price process is given by $S_t^0 \triangleq \exp\int_0^t r(Y_s)\,ds$, where the interest rate $r(\cdot)$ is a real function in $C_b^2(\mathbf{R})$, where $C_b^2(\mathbf{R})$ is the class of $C^2(\mathbf{R})$ functions which are bounded together with their first and second derivatives.
2. The asset price process S_t satisfies the stochastic differential equation (SDE)

$$dS_t = S_t\left[\mu(Y_t)\,dt + \sigma(Y_t)\,dW_{1t}\right], \quad \text{with} \quad S_0 = 1. \tag{2.1}$$

   It is assumed that the functions $\mu(\cdot)$ and $\sigma(\cdot)$ belong to $C_b^2(\mathbf{R})$, with $\sigma(\cdot) \geq \sigma_0$, for some constant $\sigma_0 > 0$.
3. The dynamics of the external factor Y_t is modelled as a diffusion process solving the SDE

$$dY_t = g(Y_t)\,dt + \beta\left(\rho dW_{1t} + \varepsilon dW_{2t}\right), \quad \text{with} \quad Y_0 = y \in \mathbf{R}, \tag{2.2}$$

   where $|\rho| \leq 1$, $\varepsilon \triangleq \sqrt{1-\rho^2}$, $\beta \neq 0$, and $g(\cdot)$ belongs to $C^1(\mathbf{R})$, with $g'(\cdot)$ bounded.

The parameter ρ is the correlation coefficient between the BM W_1 driving the asset price and the BM from the external factor $\breve{W} \triangleq \rho W_1 + \varepsilon W_2$. Except when $\rho = \pm 1$, the securities market is *incomplete*, since the external factor Y cannot be traded. Finally, without loss of generality, we fix $\beta = 1$.

Let π_t be the net amount of capital allocated in the risk asset, and c_t the rate at which capital is consumed at time t. Then, the investor's wealth process evolves as

$$dX_t = -c_t dt + \frac{X_t - \pi_t}{S_t^0} dS_t^0 + \frac{\pi_t}{S_t} dS_t,$$

with initial capital $X_0 = x > 0$. Formally, $\{\pi_t, \mathcal{F}_t\}_{0\leq t\leq T}$ is a trading **portfolio** process if it is progressively measurable and $\int_0^T \pi_u^2 du < \infty$, whereas $\{c_t, \mathcal{F}_t\}_{0\leq t\leq T}$ is a **consumption** process if it is nonnegative and progressively measurable with $\int_0^T c_t dt < \infty$. Their associated **wealth** process, denoted by $X^{\pi,c} \triangleq X^{x,y,\pi,c}$, is the solution to the integral equation

$$X_t^{\pi,c} + \int_0^t c_s ds \triangleq x + \int_0^t \left[r(Y_s)\,X_s^{\pi,c} + \left[\mu(Y_s) - r(Y_s)\right]\pi_s\right] ds + \int_0^t \pi_s \sigma(Y_s)\,dW_{1s}.$$

We say that a trading *strategy* (π, c) is **admissible** if $X^{\pi,c} \geq 0$; the set of such strategies is denoted as $\mathcal{A}(x, y)$.

The investor's problem consists of

$$\text{maximizing} \quad E\left\{U_1(X_T^{\pi,c}) + \int_0^T U_2(c_t)\,dt\right\}, \quad \text{over} \quad (\pi, c) \in \mathcal{A}(x, y), \tag{2.3}$$

as well as to find an *optimal* trading strategy $(\hat{\pi}, \hat{c})$. The utility functions U_1, $U_2 : \mathbf{R}_+ \to \mathbf{R}$ are differentiable, strictly increasing, and concave. In addition to these fundamental properties, it is assumed that $U_i'(\infty) \triangleq \lim_{b\to\infty} U_i'(b) = 0$ and $U_i'(0+) \triangleq \lim_{b\downarrow 0} U_i'(b) = \infty$, for $i = 1, 2$.

The solution of this optimization problem will be obtained using the martingale approach (see [HaPl81]), and the first step in this direction will be to characterize the family $\mathcal{A}(x,y)$ and get the *primal* representation of the investor's problem (2.3); see Lemma 2.2 and expression (P) below.

Define the function $\theta : \mathbf{R} \to \mathbf{R}$ as

$$\theta(y) \triangleq \frac{\mu(y) - r(y)}{\sigma(y)}, \quad y \in \mathbf{R}.$$

Now, denote by $\mathcal{M}$ the set of progressively measurable processes $\{\nu_t, \mathcal{F}_t\}_{t\in[0,T]}$ such that $E\int_0^T \nu_u^2 du < \infty$ and the local martingale, given by

$$Z_t^\nu \triangleq \exp\left(-\int_0^t \left[\theta(Y_u)\, dW_{1u} + \nu_u dW_{2u}\right] - \frac{1}{2}\int_0^t \left[\theta^2(Y_u) + \nu_u^2\right] du\right), \tag{2.4}$$

is a martingale for all $y \in \mathbf{R}$. Note that all bounded processes belong to $\mathcal{M}$, since $\theta(\cdot)$ is bounded. Then, for each $\nu \in \mathcal{M}$ we can define a probability measure P^ν on $(\Omega, \mathcal{F}_T)$ as

$$dP^\nu \triangleq Z_T^\nu dP. \tag{2.5}$$

Observe that $P \ll P^\nu \ll P$ and $Z_t^\nu \triangleq dP^\nu/dP|_{\mathcal{F}_t}$, for $t \in [0,T]$. Under the measure P^ν, the two-dimensional process $\{(W_{1t}^\nu, W_{2t}^\nu), \mathcal{F}_t\}_{0\le t\le T}$

$$W_{1t}^\nu \triangleq W_{1t} + \int_0^t \theta(Y_u)\, du \quad \text{and} \quad W_{2t}^\nu \triangleq W_{2t} + \int_0^t \nu_u du, \tag{2.6}$$

is a BM, and the dynamics of the processes Y_t and Z_t can be written as

$$dY_t = \left[g(Y_t) - \rho\theta(Y_t) - \varepsilon\nu_t\right] dt + \rho dW_{1t}^\nu + \varepsilon dW_{2t}^\nu, \tag{2.7}$$

$$dZ_t^\nu = Z_t^\nu\left(\left[\theta^2(Y_t) + \nu_t^2\right] dt - \theta(Y_t)\, dW_{1t}^\nu - \nu_t dW_{2t}^\nu\right). \tag{2.8}$$

Furthermore, the discounted price and wealth processes satisfy

$$d\left[\frac{S_t}{S_t^0}\right] = \frac{S_t}{S_t^0}\sigma(Y_t)\, dW_{1t}^\nu, \tag{2.9}$$

$$d\left[\frac{X_t^{\pi,c}}{S_t^0}\right] + \frac{c_t}{S_t^0}dt = \frac{\pi_t}{S_t^0}\sigma(Y_t)\, dW_{1t}^\nu, \quad (\pi, c) \in \mathcal{A}(x,y). \tag{2.10}$$

Remark 2.1. The above displayed equations imply the following:

(i) The process S_t/S_t^0 is a continuous P^ν-martingale, since $\sigma(\cdot)$ is bounded. Hence, $\mathcal{M} \subset \mathcal{P} \triangleq \{Q : P \ll Q \ll P$ and S/S^0 is a Q-local martingale for all $y \in \mathbf{R}\}$, in the sense that $P^\nu \in \mathcal{P}$ for each $\nu \in \mathcal{M}$.

(ii) The discounted process $X_t^{\pi,c}/S_t^0 + \int_0^t \left(c_s/S_s^0\right) ds$ is a nonnegative continuous P^ν-local martingale and, by Fatou's lemma, is also a P^ν-supermartingale.

The following lemma allows us to characterize the set of admissible trading strategies $\mathcal{A}(x,y)$, and it will be useful to write down the primal problem. It is analogous to Theorem 1 in [Cu97] and Theorem 5.6.2 in [KaSr98], and even though it is a known result its proof is crucial to put into firm ground the main contributions of this paper. Some parts of the proof are quoted from the above references. Condition (2.11) below is referred hereafter as the *budget constraint*.

LEMMA 2.2. *Let B be a nonnegative $\mathcal{F}_T$-measurable random variable and c_t a consumption rate process such that*

$$\sup_{\nu\in\mathcal{M}} E^{\nu}\left\{\frac{B}{S_T^0}+\int_0^T \frac{c_t}{S_t^0}dt\right\}\leq x. \tag{2.11}$$

Then, there exists a trading portfolio π such that $(\pi,c)\in\mathcal{A}(x,y)$ and $X_T^{\pi,c}\geq B$. Conversely, if $(\pi,c)\in\mathcal{A}(x,y)$, then $B\stackrel{\circ}{=}X_T^{\pi,c}$ satisfies the budget constraint (2.11).

Proof. The last part of the lemma is straightforward, since, from Remark 2.1, when $\pi\in\mathcal{A}(x,y)$ and $\nu\in\mathcal{M}$ the discounted process $X_t^{\pi,c}/S_t^0+\int_0^t\left(c_s/S_s^0\right)ds$ is a P^{ν}-supermartingale. Hence

$$E^{\nu}\left[\frac{X_T^{\pi,c}}{S_T^0}+\int_0^T\frac{c_t}{S_t^0}dt\right]\leq E^{\nu}X_0^{\pi,c}=x,$$

and, since ν was chosen arbitrarily, the budget constraint (2.11) follows.

Now, to establish the first part, we define the following discounted process:

$$\frac{\check{X}_t}{S_t^0}\stackrel{\circ}{=}\operatorname{ess\,sup}_{\nu\in\mathcal{M}}E^{\nu}\left[\frac{B}{S_T^0}+\int_t^T\frac{c_u}{S_u^0}du\mid\mathcal{F}_t\right],\quad t\in[0,T]\,. \tag{2.12}$$

Note that, by hypothesis,

$$\check{X}_0=\sup_{\nu\in\mathcal{M}}E^{\nu}\left\{\frac{B}{S_T^0}+\int_0^T\frac{c_t}{S_t^0}dt\right\}\leq x\quad\text{and}\quad\check{X}_T\equiv B. \tag{2.13}$$

It will be shown that the process $\check{X}$ induces an admissible trading strategy (π,c) such that $X_T^{\pi,c}\geq B$. First, it will be verified that $\check{X}$ satisfies the dynamic programming equation (DPE)

$$\frac{\check{X}_s}{S_s^0}=\operatorname{ess}\sup_{\nu\in\mathcal{M}}E^{\nu}\left[\frac{\check{X}_t}{S_t^0}+\int_s^t\frac{c_u}{S_u^0}du\mid\mathcal{F}_s\right],\quad 0\leq s\leq t\leq T. \tag{2.14}$$

From the definition of $\check{X}_t$ it follows that

$$E^{\nu}\left[\frac{B}{S_T^0}+\int_t^T\frac{c_u}{S_u^0}du\mid\mathcal{F}_s\right]=E^{\nu}\left[E^{\nu}\left(\frac{B}{S_T^0}+\int_t^T\frac{c_u}{S_u^0}du\mid\mathcal{F}_t\right)\mid\mathcal{F}_s\right]\leq E^{\nu}\left[\frac{\check{X}_t}{S_t^0}\mid\mathcal{F}_s\right],$$

and then

$$\begin{aligned}\frac{\check{X}_s}{S_s^0}&=\operatorname{ess}\sup_{\nu\in\mathcal{M}}E^{\nu}\left[\frac{B}{S_T^0}+\int_s^T\frac{c_u}{S_u^0}du\mid\mathcal{F}_s\right]\\&=\operatorname{ess}\sup_{\nu\in\mathcal{M}}E^{\nu}\left[\frac{B}{S_T^0}+\int_t^T\frac{c_u}{S_u^0}du+\int_s^t\frac{c_u}{S_u^0}du\mid\mathcal{F}_s\right]\\&\leq\operatorname{ess}\sup_{\nu\in\mathcal{M}}E^{\nu}\left[\frac{\check{X}_t}{S_t^0}+\int_s^t\frac{c_u}{S_u^0}du\mid\mathcal{F}_s\right].\end{aligned}$$

Next, the reverse inequality shall be verified, namely

$$\frac{\check{X}_s}{S^0_s} \geq E^\nu\left[\frac{\check{X}_t}{S^0_t} + \int_s^t \frac{c_u}{S^0_u} du \mid \mathcal{F}_s\right], \quad \text{for each } \nu \in \mathcal{M}. \tag{2.15}$$

Given $\nu \in \mathcal{M}$ and $t \in [0,T]$ fixed, define $\mathcal{M}^\nu(t) \triangleq \{\eta \in \mathcal{M} : \eta \equiv \nu \text{ in } [0,t]\}$ and

$$J^\eta_t \triangleq E^\eta\left[\frac{B}{S^0_T} + \int_t^T \frac{c_u}{S^0_u} du \mid \mathcal{F}_t\right] = E\left[\frac{Z^\eta_T}{Z^\eta_t}\left(\frac{B}{S^0_T} + \int_t^T \frac{c_u}{S^0_u} du\right) \mid \mathcal{F}_t\right], \quad \eta \in \mathcal{M}^\nu(t).$$

The second equality in the last expression is due to Bayes' formula for conditional expectations; see III.3.9 in [JaSh87] or Lemma 3.5.3 in [KaSr91]. On the other hand, note that Z^η_T/Z^η_t depends only on the values of ν in $[t,T]$. Then, $\check{X}_t/S^0_t = \sup_{\eta\in\mathcal{M}^\nu(t)} J^\eta_t$. In fact,

$$\frac{\check{X}_t}{S^0_t} = \lim_{n\to\infty} J^{\eta_n}_t, \tag{2.16}$$

for some increasing sequence $\{J^{\eta_n}_t\}_{n\geq 1}$, with $\eta_n \in \mathcal{M}^\nu(t)$. This is true since the set $\{J^\eta_t\}_{\eta\in\mathcal{M}^\nu(t)}$ is a closed family by pair maximization (see Theorem A.3 in [KaSr98]). From (2.16) and the conditional monotone convergence theorem, inequality (2.15) holds if for all $n \geq 1$

$$\frac{\check{X}_s}{S^0_s} \geq E^\nu\left[J^{\eta_n}_t + \int_s^t \frac{c_u}{S^0_u} du \mid \mathcal{F}_s\right].$$

This inequality is established observing that

$$\begin{aligned}
\frac{\check{X}_s}{S^0_s} &\geq E^{\eta_n}\left[\frac{B}{S^0_T} + \int_s^T \frac{c_u}{S^0_u} du \mid \mathcal{F}_s\right] = E^{\eta_n}\left[\frac{B}{S^0_T} + \int_s^t \frac{c_u}{S^0_u} du + \int_t^T \frac{c_u}{S^0_u} du \mid \mathcal{F}_s\right] \\
&= E^{\eta_n}\left[E^{\eta_n}\left(\frac{B}{S^0_T} + \int_t^T \frac{c_u}{S^0_u} du \mid \mathcal{F}_t\right) + \int_s^t \frac{c_u}{S^0_u} du \mid \mathcal{F}_s\right] \\
&= E^\nu\left[J^{\eta_n}_t + \int_s^t \frac{c_u}{S^0_u} du \mid \mathcal{F}_s\right].
\end{aligned}$$

Now, from the DPE (2.14), it follows that the process $\check{X}_\cdot/S^0_\cdot + \int_0^\cdot \left(c_t/S^0_t\right) dt$ is a P^ν-supermartingale with a right continuous with left limits (RCLL) modification, for each $\nu \in \mathcal{M}$. Using the Doob–Meyer supermartingale decomposition theorem and the local martingale representation theorem, this process can be written as

$$\frac{\check{X}_t}{S^0_t} + \int_0^t \frac{c_s}{S^0_s} ds =: \check{X}_0 + \int_0^t (\psi^\nu_{1s} dW^\nu_{1s} + \psi^\nu_{2s} dW^\nu_{2s}) - A^\nu_t, \tag{2.17}$$

where ψ^ν_1, ψ^ν_2, are progressively measurable processes with $\int_0^T ([\psi^\nu_{1s}]^2 + [\psi^\nu_{2s}]^2) ds < \infty$, and A^ν is a predictable integrable increasing process with $A^\nu_0 \equiv 0$; see Theorem 3.3.9 in [LiSh01] and Problem 3.4.16 in [KaSr91]. Thus, denoting by 0 the null process $\nu \equiv 0$, and using (2.17), the following identity holds:

$$\int_0^t (\psi^\nu_{1s} dW^\nu_{1s} + \psi^\nu_{2s} dW^\nu_{2s}) - A^\nu_t = \int_0^t \left(\psi^0_{1s} dW^0_{1s} + \psi^0_{2s} dW^0_{2s}\right) - A^0_t.$$

According to expression (2.6), we obtain that

$$0 = \int_0^t \left[\left(\psi_{1s}^{\nu} - \psi_{1s}^{0}\right) dW_{1s} + \left(\psi_{2s}^{\nu} - \psi_{2s}^{0}\right) dW_{2s}\right] + A_t^0 - A_t^{\nu} + \int_0^t \left[\left(\psi_{1s}^{\nu} - \psi_{1s}^{0}\right) \theta\left(Y_s\right) + \psi_{2s}^{\nu}\nu_s\right] ds.$$

This equation has the form $L + V + \phi \equiv 0$, where L is a continuous P-local martingale, V is a predictable finite variation process, and ϕ is a continuous process with zero quadratic variation, such that $L_0 = V_0 = \phi_0 = 0$. The above suggests that all those terms should be the zero process. In fact, from Proposition I.4.49.d in [JaSh87], the *covariation* $\langle L, V\rangle$ is identically zero. Thus, $0 = \langle \phi, \phi\rangle = \langle L + V, L + V\rangle = \langle L\rangle + \langle V\rangle + 2\langle L, V\rangle$. Hence $\langle L\rangle = \langle V\rangle = 0$, i.e.,

$$\psi_1^{\nu} \equiv \psi_1^0, \quad \psi_2^{\nu} \equiv \psi_2^0, \quad \text{and} \quad A_{\cdot}^{\nu} \equiv A_{\cdot}^0 + \int_0^{\cdot} \psi_{2s}^0 \nu_s ds \geq 0, \quad \nu \in \mathcal{M}.$$

This, together with the fact that $\left\{\int_0^t \psi_{2s}^0 ds < 0\right\} \cup \left\{\int_0^t \psi_{2s}^0 ds > 0\right\}$ is a null event, implies that $A_t^{\nu} = A_t^0$ for $t \in [0, T]$. Hence

$$\psi \stackrel{\circ}{=} \psi_1^0 \equiv \psi_1^{\nu}, \quad \psi_2^{\nu} \equiv 0, \quad \text{and} \quad A^{\nu} \equiv A^0, \quad \nu \in \mathcal{M}.$$

Thus, (2.17) can be written as

$$\frac{\check{X}_t}{S_t^0} + \int_0^t \frac{c_s}{S_s^0} ds = \check{X}_0 + \int_0^t \psi_s dW_{1s}^{\nu} - A_t^0, \quad \nu \in \mathcal{M}. \tag{2.18}$$

Now, assume for a moment that the budget constraint (2.11) holds with equality

$$\check{X}_0 = \sup_{\nu \in \mathcal{M}} E^{\nu}\left\{\frac{B}{S_T^0} + \int_0^T \frac{c_t}{S_t^0} dt\right\} = x. \tag{2.19}$$

Next, define the trading portfolio $\pi_t \stackrel{\circ}{=} S_t^0 \psi_t / \sigma\left(Y_t\right)$, for $t \in [0, T]$. Then, using (2.10), (2.18), and (2.19), it follows that $X^{x,y,\pi,c}$ satisfies

$$\begin{aligned} \frac{X_t^{x,y,\pi,c}}{S_t^0} &= x - \int_0^t \frac{c_s}{S_s^0} ds + \int_0^t \frac{\pi_s}{S_s^0} \sigma\left(Y_s\right) dW_{1s}^{\nu} = x - \int_0^t \frac{c_s}{S_s^0} ds + \int_0^t \psi_s dW_{1s}^{\nu} \\ &= \frac{\check{X}_t}{S_t^0} + A_t^0 \geq \frac{\check{X}_t}{S_t^0} \geq 0. \end{aligned}$$

In particular, $X_T^{x,y,\pi,c} \geq \check{X}_T \equiv B$. On the other hand, when $\check{X}_0 < x$, substituting $\check{X}_0$ by x and applying the above arguments to the trading strategy (π, c), but also investing in the bank account the exceeding initial capital $x - \check{X}_0$, we get

$$X^{x,y,\pi,c} \geq X^{\check{X}_0,y,\pi,c} \geq 0 \quad \text{and} \quad X_T^{x,y,\pi,c} \geq B. \qquad \square$$

Thanks to Lemma 2.2, the original investor's problem can be written as a convex optimization problem, referred to as *primal problem*, consisting of

$$\text{(P)} \qquad \text{maximizing} \quad E\left\{U_1\left(B\right) + \int_0^T U_2\left(c_t\right) dt\right\} \quad \text{over} \quad (B, c) \in \mathcal{B}\left(x, y\right),$$

where $\mathcal{B}(x,y)$ is the set of pairs (B,c) such that B is a nonnegative $\mathcal{F}_T$-measurable random variable and c is a consumption rate process satisfying the budget constraint (2.11).

The next theorem suggests the relationship between the trading portfolio π and the final wealth B. Its proof is based on arguments given in the proof of the previous lemma. This result is analogous to Theorem 5.8.9 in [KaSr98].

THEOREM 2.3. *Let c be a consumption rate process and $\check{\nu} \in \mathcal{M}$. Then, the following statements are equivalent:*
(i)

$$(B,c) \in \mathcal{B}(x,y) \quad \text{and} \quad E^{\check{\nu}}\left[\frac{B}{S_T^0} + \int_0^T \frac{c_t}{S_t^0} dt\right] = x,$$

(ii) $(\pi,c) \in \mathcal{A}(x,y)$, $X_T^{\pi,c} \equiv B$, *and* $X^{\pi,c}/S^0 + \int_0^{\cdot} \left(c_s/S_s^0\right) ds$ *is a* $P^{\check{\nu}}$*-martingale with representation*

$$\frac{X_t^{\pi,c}}{S_t^0} + \int_0^t \frac{c_s}{S_s^0} ds = x + \int_0^t \psi_s dW_{1s}^{\check{\nu}}, \quad t \in [0,T], \tag{2.20}$$

where ψ is a progressively measurable process with $\int_0^T \psi_u^2 du < \infty$.

Proof. (i) implies (ii): We shall verify that $X_t^{\pi,c} \equiv \check{X}_t$, where

$$\pi_t = \frac{S_t^0}{\sigma(Y_t)} \psi_t \tag{2.21}$$

is the trading portfolio and $\check{X}_t$ is defined in (2.12). From (2.13) and (i), we have

$$E^{\check{\nu}}\left[\frac{\check{X}_T}{S_T^0} + \int_0^T \frac{c_t}{S_t^0} dt\right] = E^{\check{\nu}}\left[\frac{B}{S_T^0} + \int_0^T \frac{c_t}{S_t^0} dt\right] = x = \sup_{\nu \in \mathcal{M}} E^{\nu}\left[\frac{B}{S_T^0} + \int_0^T \frac{c_t}{S_t^0} dt\right] = \check{X}_0.$$

Then $\check{X}_{\cdot}/S_{\cdot}^0 + \int_0^{\cdot} \left(c_s/S_s^0\right) ds$ is a $P^{\check{\nu}}$-martingale, since it is a $P^{\check{\nu}}$-supermartingale with constant mean. Thus, from (2.18), $A^0 \equiv 0$ and $X^{\pi,c} \equiv \check{X}$. The rest follows from the martingale representation theorem.

(ii) implies (i): From Remark 2.1, for each $\nu \in \mathcal{M}$ the discounted process $X_{\cdot}^{\pi,c}/S_{\cdot}^0 + \int_0^{\cdot} \left(c_s/S_s^0\right) ds$ is a P^{ν}-supermartingale; in particular, it is a $P^{\check{\nu}}$-martingale. Hence, $(B,c) \in \mathcal{B}(x,y)$ and $E^{\check{\nu}}\left[B/S_T^0 + \int_0^T \left(c_s/S_s^0\right) ds\right] = x$, where $B \equiv X_T^{\pi,c}$. ∎

Remark 2.4. The following identities will be used later; see [Cu97]. For any consumption process c, $\nu \in \mathcal{M}$ and $t \in [0,T]$,

$$E^{\nu}\left[\int_t^T \frac{c_u}{S_u^0} du \mid \mathcal{F}_t\right] = E\left[\int_t^T \frac{Z_u^{\nu}}{Z_t^{\nu}} \frac{c_u}{S_u^0} du \mid \mathcal{F}_t\right], \tag{2.22}$$

$$E^{\nu} \int_0^T \frac{c_u}{S_u^0} du = E \int_0^T \frac{Z_u^{\nu}}{S_u^0} c_u du. \tag{2.23}$$

3. Dual problem. In this section the dual problem is posed using techniques from convex optimization, and the relationship between the optimal values of the primal and dual problems is studied.

The Legendre–Frechel transform $\tilde{U}_i$ corresponding to the utility function U_i is defined as

$$\tilde{U}_i(z) \triangleq \max_{b\geq 0}\{U_i(b) - zb\}, \quad z > 0. \tag{3.1}$$

From the definition of $\tilde{U}_i(\cdot)$ and elementary calculus, it follows that, for $z > 0$,

$$\tilde{U}_i(z) =: U_i(I_i(z)) - zI_i(z), \tag{3.2}$$

where $I(\cdot)$ is the inverse function of $U_i'(\cdot)$. The associated *dual functional* to the primal problem (P) is defined, for $\nu \in \mathcal{M}$ and $\lambda \geq 0$, as follows:

$$\begin{aligned} L(\nu,\lambda) &\triangleq L(\nu,\lambda;x,y) \\ &\triangleq \sup_{B\geq 0, c\geq 0}\left\{E\left[U_1(B) + \int_0^T U_2(c_t)\,dt\right] - \lambda E^{\nu}\left[\frac{B}{S_T^0} + \int_0^T \frac{c_t}{S_t^0}dt\right]\right\} + \lambda x. \end{aligned}$$

Here the argument "$B \geq 0$, $c \geq 0$" means that B is a nonnegative $\mathcal{F}_T$-measurable random variable and c is a consumption rate process. We point out that the present definition is a variant of the classical dual functional given in (8.6.2) in [Lu69]. The *dual* problem consists of

$$\text{(D)} \qquad \text{minimizing} \quad L(\nu,\lambda), \quad \text{over} \quad \nu \in \mathcal{M} \ \text{ and } \ \lambda > 0.$$

It is not difficult to verify that the following inequality, relating the optimal values of the primal and dual problems, holds:

$$\sup_{(B,c)\in\mathcal{B}(x,y)} E\left\{U_1(B) + \int_0^T U_2(c_t)\,dt\right\} \leq \inf_{\nu\in\mathcal{M},\lambda>0} L(\nu,\lambda). \tag{3.3}$$

When equality holds in (3.3), we say that there is no duality *gap*. In the next section we will establish this property for HARA utility functions. On the other hand, note that using (2.23), (3.1), and (3.2) the dual functional can be written as

$$L(\nu,\lambda) = E\left[\tilde{U}_1\left(\lambda\frac{Z_T^{\nu}}{S_T^0}\right) + \int_0^T \tilde{U}_2\left(\lambda\frac{Z_t^{\nu}}{S_t^0}\right)dt\right] + \lambda x, \quad \nu \in \mathcal{M}, \quad \lambda > 0. \tag{3.4}$$

This representation of $L(\nu,\lambda)$ is inspired as a natural extension, from complete to incomplete markets, of the results presented in section 3.6 in [KaSr98].

The next proposition shows, under suitable conditions, the relationship between the optimal solutions of the primal (P) and dual (D) problems.

PROPOSITION 3.1. *Assume that for some* $(\hat{\nu},\hat{\lambda}) \in \mathcal{M}\times\mathbf{R}_+$ *the pair* $(\hat{B},\hat{c})$, *defined as*

$$\hat{B} \triangleq I_1\left(\hat{\lambda}\frac{Z_T^{\hat{\nu}}}{S_T^0}\right) \quad \textit{and} \quad \hat{c}_t \triangleq I_2\left(\hat{\lambda}\frac{Z_t^{\hat{\nu}}}{S_t^0}\right), \quad t \in [0,T], \tag{3.5}$$

belongs to $\mathcal{B}(x,y)$ *and*

$$E^{\hat{\nu}}\left[\frac{\hat{B}}{S_T^0}+\int_0^T\frac{\hat{c}_t}{S_t^0}dt\right]=x. \tag{3.6}$$

Then, $(\hat{B},\hat{c})$ *is the optimal solution to the primal problem* (P)*, whereas* $(\hat{\nu},\hat{\lambda})$ *is the optimal solution to the dual problem* (D)*. In particular, there is no duality gap.*

Proof. From (3.4) and (3.2), it follows that

$$\begin{aligned}
\inf_{\nu\in\mathcal{M},\lambda>0}L(\nu,\lambda)&=\inf_{\nu\in\mathcal{M},\lambda>0}E\left\{\tilde{U}_1\left(\lambda\frac{Z_T^\nu}{S_T^0}\right)+\int_0^T\tilde{U}_2\left(\lambda\frac{Z_t^\nu}{S_t^0}\right)dt+\lambda x\right\}\\
&\le E\left[\tilde{U}_1\left(\hat{\lambda}\frac{Z_T^{\hat{\nu}}}{S_T^0}\right)+\int_0^T\tilde{U}_2\left(\hat{\lambda}\frac{Z_t^{\hat{\nu}}}{S_t^0}\right)dt\right]+\hat{\lambda}x\\
&=E\left[U_1(\hat{B})+\int_0^T U_2(\hat{c}_t)\,dt\right]-\hat{\lambda}E\left[\frac{Z_T^{\hat{\nu}}}{S_T^0}\frac{\hat{B}}{S_T^0}+\int_0^T\frac{Z_t^{\hat{\nu}}}{S_t^0}\hat{c}_t dt\right]+\hat{\lambda}x\\
&=E\left[U_1(\hat{B})+\int_0^T U_2(\hat{c}_t)\,dt\right]-\hat{\lambda}E^{\hat{\nu}}\left[\frac{\hat{B}}{S_T^0}+\int_0^T\frac{\hat{c}_t}{S_t^0}dt\right]+\hat{\lambda}x\\
&=E\left[U_1(\hat{B})+\int_0^T U_2(\hat{c}_t)\,dt\right]\\
&\le\sup_{(B,c)\in\mathcal{B}(x,y)}E\left\{U_1(B)+\int_0^T U_2(c_t)\,dt\right\}.
\end{aligned}$$

Using (3.3), we conclude that there is no duality gap and, furthermore, $(\hat{B},\hat{c})$ is the optimal solution to the primal problem (P), whereas $(\hat{\nu},\hat{\lambda})$ is the optimal solution to the dual problem (D). □

Remark 3.2. The *optimal pair* $(\hat{B},\hat{c})$ given in (3.5) is similar to the one given in (6.3.16) and (6.3.17) in [KaSr98], with $\hat{\lambda}=\mathcal{Y}_{\hat{\nu}}(x)$ and $\mathcal{Y}_{\hat{\nu}}(\cdot)$ being the inverse function of $\mathcal{X}_{\hat{\nu}}(\cdot)$. An existence result and a characterization of the optimal solution to the dual problem (D) are also given in section 6.5 in [KaSr98]. However, except for deterministic coefficients (section 6.6) and logarithmic case (Example 6.7.2), the optimal process $\hat{\nu}$ is not obtained explicitly.

Remark 3.3. Based on the previous results, the following steps can be formulated to solve the investor's problem using the martingale method:

1. Given the utility function $U_1(\cdot)$ and $U_2(\cdot)$, write down the dual problem and obtain its optimal solution $(\hat{\nu},\hat{\lambda})\in\mathcal{M}\times\mathbf{R}_+$.
2. Verify that the pair $(\hat{B},\hat{c})$ given by (3.5) belongs to $\mathcal{B}(x,y)$ and satisfies (3.6), and then apply Proposition 3.1 and Theorem 2.3.
3. Finally, from (2.21) and (2.20), the optimal trading portfolio $\hat{\pi}$ is obtained.

These steps will be applied successfully for HARA and logarithmic utility functions in the next section.

4. Results for HARA utility function. In this section we solve explicitly the optimal consumption-investment problem for HARA utility functions $U_1(b)=U_2(b)=\frac{1}{\gamma}b^\gamma$ with parameter $\gamma\neq 0$ and $\gamma<1$.

The corresponding Legendre–Frechel transform $\tilde{U}_i:\mathbf{R}^+\to\mathbf{R}$ is given by $\tilde{U}(z)=U(I(z))-zI(z)=-z^\alpha/\alpha$, where $I(z)=z^{\alpha-1}$ and $\alpha\triangleq-\gamma/(1-\gamma)$. Notice that

$\alpha < 1$, $\alpha \neq 0$ and $\gamma = -\alpha/(1-\alpha)$. Hence, the dual functional (3.4) can be written as

$$L(\nu,\lambda) =: \lambda x - \frac{1}{\alpha}\lambda^{\alpha}\Lambda_{\nu}, \tag{4.1}$$

where $\Lambda_{\nu} \stackrel{\circ}{=} E\left(\left[Z_T^{\nu}/S_T^0\right]^{\alpha} + \int_0^T \left[Z_t^{\nu}/S_t^0\right]^{\alpha} dt\right)$. Fixing $\nu \in \mathcal{M}$, and using elementary calculus, the optimal value of the parameter λ can be derived and is given by

$$\hat{\lambda}(\nu) = (\Lambda_{\nu}/x)^{1/(1-\alpha)};$$

see definition of the dual problem (D). Substituting this value in (4.1), we obtain

$$L(\nu, \hat{\lambda}(\nu)) = \frac{1}{\gamma}x^{\gamma}\Lambda_{\nu}^{1-\gamma}, \qquad \nu \in \mathcal{M}.$$

Depending on the sign of the HARA parameter, $0 < \gamma < 1$ [$\gamma < 0$], the dual problem consisting of minimizing $L(\nu, \hat{\lambda}(\nu))$ over the set of processes ν in $\mathcal{M}$, is equivalent to

$$\text{minimize [maximize]} \quad J(T,y,\nu) \stackrel{\circ}{=} \Lambda_{\nu}, \quad \text{over} \quad \nu \in \mathcal{M}. \tag{4.2}$$

This is a stochastic control problem, referred to as the *auxiliary* problem, and will be solved using dynamic programming techniques as well as analytic arguments. Note that, in this case, the martingale method reduces the original investor's problem with two control variables to only one control process $\nu \in \mathcal{M}$.

On the other hand, observe that

$$\begin{aligned}[Z_t^{\nu}]^{\alpha} &= e^{-\alpha\int_0^t [\theta(Y_s)dW_{1s} + \nu_s dW_{2s}] - \frac{1}{2}\alpha^2\int_0^t \left[\theta^2(Y_s)+\nu_s^2\right]ds - \frac{1}{2}\alpha(1-\alpha)\int_0^t \left[\theta^2(Y_s)+\nu_s^2\right]ds} \\ &= Z_t^{\alpha,\nu} e^{-\frac{1}{2}\alpha(1-\alpha)\int_0^t \left[\theta^2(Y_s)+\nu_s^2\right]ds}, \quad t \in [0,T], \end{aligned} \tag{4.3}$$

with $Z^{\alpha,\nu}$ (see (2.4)) given by

$$Z_t^{\alpha,\nu} \stackrel{\circ}{=} \exp\left(-\alpha\int_0^t \left[\theta(Y_u)\,dW_{1u} + \nu_u dW_{2u}\right] - \frac{1}{2}\alpha^2\int_0^t \left[\theta^2(Y_u) + \nu_u^2\right]du\right).$$

Proceeding as in (2.5) and (2.6), we can define a new measure $P^{\alpha,\nu}$ in $\mathcal{F}_T$ and a BM $(W_1^{\alpha,\nu}, W_2^{\alpha,\nu})$, respectively. Under this measure, the dynamics of the external factor Y evolves as

$$dY_t = \left[g(Y_t) - \alpha\rho\theta(Y_t) - \alpha\varepsilon\nu_t\right]dt + \rho dW_{1t}^{\alpha,\nu} + \varepsilon dW_{2t}^{\alpha,\nu}, \quad \text{with} \quad Y_0 = y \in \mathbf{R}. \tag{4.4}$$

Now, from (4.3) and (2.23), we obtain the following representation for the functional $J(T,y,\nu)$:

$$\begin{aligned} J(T,y,\nu) &= E\left[\left(\frac{Z_T^{\nu}}{S_T^0}\right)^{\alpha} + \int_0^T \left(\frac{Z_t^{\nu}}{S_t^0}\right)^{\alpha} dt\right] \\ &= E^{\alpha,\nu}\left[e^{-\alpha\int_0^T \left[r(Y_t)+\frac{1}{2}(1-\alpha)\left(\theta^2(Y_t)+\nu_t^2\right)\right]dt}\right. \\ &\quad \left. + \int_0^T e^{-\alpha\int_0^t \left[r(Y_s)+\frac{1}{2}(1-\alpha)\left(\theta^2(Y_s)+\nu_s^2\right)\right]ds} dt\right] \\ &= E^{\alpha,\nu}\left[e^{\int_0^T q(Y_t,\nu_t)dt} + \int_0^T e^{\int_0^t q(Y_s,\nu_s)ds} dt\right], \quad (T,y,\nu) \in \mathbf{R}_+ \times \mathbf{R} \times \mathcal{M}, \end{aligned}$$

where $q(y,v) \triangleq -\alpha\left[r(y)+\frac{1}{2}(1-\alpha)\left(\theta^2(y)+v^2\right)\right]$, for $(y,v)\in\mathbf{R}^2$.

Case $\gamma<0$. The *value function* associated with the auxiliary problem (4.2) is defined as

$$W(T,y) \triangleq \sup_{\nu\in\mathcal{M}} J(T,y,\nu), \quad (T,y)\in\mathbf{R}_+\times\mathbf{R}, \tag{4.5}$$

with $W(0,y)=1$. Some basic properties of the value function will be established next, in order to solve this optimal control problem.

First, note that, denoting by $|\cdot|_\infty$ the supremum norm, the following estimates hold:

$$\begin{aligned} &E\left[e^{-\alpha\int_0^T\left(|r|_\infty+\frac{1}{2}(1-\alpha)\left(|\theta|_\infty^2+\nu_u^2\right)\right)du}+\int_0^T e^{-\alpha\int_0^t\left(|r|_\infty+\frac{1}{2}(1-\alpha)\left(|\theta|_\infty^2+\nu_u^2\right)\right)du}dt\right] \\ &\quad\le J(T,y,\nu)\le 1+T, \end{aligned} \tag{4.6}$$

and they imply that

$$0<K_1\le W(T,y)\le 1+T, \tag{4.7}$$

where

$$K_1 \triangleq (1+T)\,e^{-\alpha\left(|r|_\infty+\frac{1}{2}(1-\alpha)|\theta|_\infty^2\right)T}. \tag{4.8}$$

Note that the constant K_1 does not depend on the initial condition y.

Now, it will be verified that the function $W(T,\cdot)$ is Lipschitz. Using the fact that $q(y,v)$ is bounded above, the dominated convergence theorem can be applied to establish the y-differentiability of $J(T,y,\nu)$ (see Theorem 5.5.5 in [Fr75]). In fact,

$$\begin{aligned} J_y(T,y,\nu)=E^{\alpha,\nu}&\left[e^{\int_0^T q(Y_t,\nu_t)dt}\int_0^T q_y(Y_t,\nu_t)\frac{\partial}{\partial y}Y_t dt\right. \\ &\left.+\int_0^T e^{\int_0^t q(Y_s,\nu_s)ds}\int_0^t q_y(Y_s,\nu_s)\frac{\partial}{\partial y}Y_s ds dt\right], \end{aligned}$$

where $\frac{\partial}{\partial y}Y_s=\exp\left(\int_0^s\left[g'(Y_u)-\alpha\rho\theta'(Y_u)\right]du\right)$. Furthermore,

$$\left|\int_0^T q_y(Y_s,\nu_s)\frac{\partial}{\partial y}Y_s ds\right| \le K_2 \triangleq \alpha\left(|r'|_\infty+(1-\alpha)|\theta|_\infty|\theta'|_\infty\right)Te^{\left(|g'|_\infty+\alpha|\theta'|_\infty\right)T},$$

which implies that

$$|J_y(T,y,\nu)|\le K_2(1+T).$$

Observe that K_2 does not depend on ν and y, and hence $W(T,\cdot)$ is Lipschitz with constant $K_2(1+T)$. This estimate together with (4.7) yields that, when W_y is well defined,

$$\frac{|W_y(T,y)|}{W(T,y)}\le K \triangleq \frac{K_2}{K_1}(1+T). \tag{4.9}$$

The above estimate will be used later in this section.

It is convenient to restrict for a moment the set of control processes to those in $\mathcal{M}$ taking values in $[-M, M]$ for a fixed constant $M > 0$. This set will be denoted by $\mathcal{M}^M$, and the corresponding *constrained value function* by $W^M(T, y)$. This restriction will be removed later.

The verification theorem below states that

$$w(T, y) = W^M(T, y),$$

where $w(T, y)$ is the unique smooth function (see Theorem IV.4.3 and Remark IV.4.1 in [FlSo93]) in $C^{1,2}\left(\bar{\mathbf{R}}_+ \times \mathbf{R}\right) \cap C_p\left(\bar{\mathbf{R}}_+ \times \mathbf{R}\right)$ satisfying the associated Hamilton–Jacobi–Bellman (HJB) equation:

$$\begin{aligned} w_T = 1 + \frac{1}{2} w_{yy} + (g - \alpha\rho\theta) w_y - \alpha \left[r + \frac{1}{2}(1-\alpha)\theta^2 \right] w \\ + \alpha \sup_{v \in [-M, M]} \left\{ -\varepsilon w_y v - \frac{1}{2}(1-\alpha) w v^2 \right\} \quad \text{with} \quad w(0, y) = 1. \end{aligned} \tag{4.10}$$

The supremum selector on the r.h.s. of the above equation induces a Markov *policy* defined, for $(t, y) \in [0, T] \times \mathbf{R}$, as follows:

$$\begin{aligned} \nu^*(t, y) &\triangleq \arg \max_{v \in [-M, M]} \left\{ -\varepsilon w_y(t, y) v - \frac{1}{2}(1-\alpha) w(t, y) v^2 \right\} \\ &= \begin{cases} -\dfrac{\varepsilon}{1-\alpha} \dfrac{w_y(t, y)}{w(t, y)}, & \text{if} \quad \dfrac{\varepsilon}{1-\alpha} \dfrac{|w_y(t, y)|}{w(t, y)} \le M \quad \text{and} \quad w(t, y) \neq 0 \\ -M \operatorname{sgn} w_y(t, y), & \text{otherwise.} \end{cases} \end{aligned} \tag{4.11}$$

THEOREM 4.1 (verification). *Given $T > 0$, let $w(T, y)$ be the unique solution to* (4.10). *Then,*

(i) $w(T, y) \ge J(T, y, \nu)$, *for* $(T, y) \in \bar{\mathbf{R}}_+ \times \mathbf{R}$ *and* $\nu \in \mathcal{M}^M$,

(ii) $w(T, y) = W^M(T, y) = J(T, y, \hat{\nu})$, *where* $\hat{\nu} \in \mathcal{M}^M$ *is the Markov policy given by*

$$\hat{\nu}_t \triangleq \nu^*(T - t, Y_t), \quad t \in [0, T]. \tag{4.12}$$

In particular, $\hat{\nu}$ is the optimal process for the constrained auxiliary problem associated with (4.5).

Proof. (i) Given $v \in [-M, M]$, define the functional $\mathcal{L}^v$ as

$$\mathcal{L}^v f \triangleq f_t + \frac{1}{2} f_{yy} + (g - \alpha\rho\theta - \alpha\varepsilon v) f_y, \quad \text{with} \quad f \in C^{1,2}([0, T] \times \mathbf{R}).$$

In particular, when $f(t, y) = w(T - t, y)$, we get

$$\begin{aligned} [\mathcal{L}^v + q(y, v)] w(T - t, y) = -w_t + \frac{1}{2} w_{yy} + (g - \alpha\rho\theta) w_y - \alpha \left[r + \frac{1}{2}(1-\alpha)\theta^2 \right] w \\ + \alpha \left[-\varepsilon w_y v - \frac{1}{2}(1-\alpha) w v^2 \right]. \end{aligned}$$

Hence, from (4.10), we have

$$[\mathcal{L}^{\nu_t} + q(Y_t, \nu_t)] w(T - t, Y_t) \le -1; \quad t \in [0, T] \quad \nu \in \mathcal{M}^M. \tag{4.13}$$

This inequality together with the Feynman-Kac formula imply, using a change of variable in the time parameter, that

$$
\begin{aligned}
w(T,y) &= E^{\alpha,\nu}\left[e^{\int_0^T q(Y_u,\nu_u)du} w\left(0,Y_T\right)\right. \\
&\qquad \left. - \int_0^T e^{\int_0^t q(Y_u,\nu_u)du}\left[\mathcal{L}^{\nu_t} + q(Y_t,\nu_t)\right] w(T-t,Y_t)dt\right] \\
&\geq E^{\alpha,\nu}\left[e^{\int_0^T q(Y_u,\nu_u)du} + \int_0^T e^{\int_0^t q(Y_u,\nu_u)du}dt\right] \qquad (4.14)\\
&= J\left(T,y,\nu\right).
\end{aligned}
$$

See, for instance, (D.13) in [FlSo93] and Corollary 4.4.5 in [KaSr91].

(ii) Since $w_y(t,y)$ is continuous, the Markov policy $\nu^*(t,y)$ defined in (4.11) is bounded, continuous, and y-locally Lipschitz. Hence, the Markov control process $\hat{\nu}$ defined in (4.12) belongs to $\mathcal{M}^M$ and, from the definition of $\nu^*(t,y)$, for $\nu \triangleq \hat{\nu}$ inequalities (4.13) and (4.14) become equalities. Therefore, $w(T,y) = W^M(T,y) = J(T,y,\hat{\nu})$. □

COROLLARY 4.2. *Let $w(T,y)$ be the unique solution to* (4.10), *with $M > \varepsilon K/(1-\alpha)$. Then,*

$$w(T,y) = W(T,y),$$

where $W(T,y)$ is the unconstrained value function defined in (4.5), *and*

$$\hat{\nu}_t \triangleq \nu^*(T-t,Y_t) = -\frac{\varepsilon}{1-\alpha}\frac{W_y(T-t,Y_t)}{W(T-t,Y_t)}, \quad t\in[0,T], \qquad (4.15)$$

is the optimal control process. Furthermore, $W(T,y) \in C^{1,2}\left(\bar{\mathbf{R}}_+ \times \mathbf{R}\right) \cap C_b^{0,1}\left(\bar{\mathbf{R}}_+ \times \mathbf{R}\right)$ and solves the following partial differential equation (PDE):

$$W_T = 1 + \frac{1}{2}W_{yy} + (g - \alpha\rho\theta)W_y - \alpha\left[r + \frac{1}{2}(1-\alpha)\theta^2\right]W - \frac{1}{2}\gamma\varepsilon^2\frac{W_y^2}{W}, \qquad (4.16)$$

with initial data $W(0,y) = 1$.

Proof. Using (4.9) it follows that, for $M > \varepsilon K/(1-\alpha)$, we have

$$\frac{\varepsilon}{1-\alpha}\frac{|W_y(T,y)|}{W(T,y)} < M,$$

and hence $w(T,y) = W^M(T,y) = W(T,y)$. Thus, using (4.12) and (4.11), the optimal process is given by (4.15). Finally, substituting the Markov policy (4.11) in the HJB equation (4.10), we obtain (4.16). □

Case $0 < \gamma < 1$. In this case the *value function* associated with the auxiliary problem (4.2) is defined as

$$W(T,y) \triangleq \inf_{\nu\in\mathcal{M}} J(T,y,\nu) = \inf_{\nu\in\mathcal{M}} E^{\alpha,\nu}\left\{e^{\int_0^T q(Y_t,\nu_t)dt} + \int_0^T e^{\int_0^t q(Y_s,\nu_s)ds}dt\right\}. \qquad (4.17)$$

As in the previous case, we expect similar conclusions if a suitable bound for the ratio $W_y(T,y)/W(T,y)$ can be obtained. However, in this case we cannot apply the dominated convergence theorem to find an estimate of $W_y(T,y)$ independent of M, because $q(y,v)$ is not bounded above. Instead, a more involved technique will be used, exploring qualitative properties of the corresponding HJB equation.

First, note that it is not difficult to prove that $W(T,y)$ is increasing with respect to T, and also that estimate (4.7) holds in the reverse sense, with K_1 given by (4.8), i.e.,

$$1+T \leq W(T,y) \leq K_1. \tag{4.18}$$

Now, let us restrict the control processes to the set $\mathcal{M}^M$, for some positive constant $M>0$. Hence, taking $[-M,M]$ as the control space, the following verification theorem holds. Its proof is based on the same arguments used in the case when γ is negative and will be omitted. In this case, it can also be guaranteed the existence and uniqueness of $w(T,y) \in C^{1,2}\left(\bar{\mathbf{R}}_+ \times \mathbf{R}\right) \cap C_p\left(\bar{\mathbf{R}}_+ \times \mathbf{R}\right)$, satisfying the associated HJB equation (4.10), with $\alpha < 0$.

THEOREM 4.3 (verification). *Given $M>0$, let $w(T,y)$ be the unique solution to* (4.10). *Then,*

(i) $w(T,y) \leq J(T,y,\nu)$, *for* $(T,y) \in \mathbf{R}_+ \times \mathbf{R}$ *and* $\nu \in \mathcal{M}^M$,

(ii) $w(T,y) = W^M(T,y) = J(T,y,\hat{\nu})$, *where $\hat{\nu}$ is the Markov control process in $\mathcal{M}^M$ given by* (4.12), *with $\nu^*(t,y)$ as in* (4.11). *In particular, $\hat{\nu}$ is the optimal control process for the constrained auxiliary problem related with* (4.17).

The following is the main result of this paper.

THEOREM 4.4. *There exists a constant $\tilde{K}>0$ such that for $M > \varepsilon\tilde{K}/(1-\alpha)$,*

$$w(T,y) = W(T,y)$$

and $\hat{\nu}$, given by (4.15), *is an optimal control process, where $W(T,y)$ is the unconstrained value function* (4.17). *Furthermore, $W \in C^{1,2}\left(\bar{\mathbf{R}}_+ \times \mathbf{R}\right) \cap C_b^{0,1}\left(\bar{\mathbf{R}}_+ \times \mathbf{R}\right)$ and solves the PDE* (4.16), *with initial data $W(0,y)=1$.*

Proof. To estimate $W_y(T,y)$ we shall obtain first an upper bound for $W_T^M(T,y)$ and then extract qualitative properties of $W^M(T,y)$ from the HJB equation (4.10), where $w(T,y) = W^M(T,y)$. To get such an upper bound for $W_T^M(T,y)$, we extend, and denote by the same symbol, the optimal process $\hat{\nu}$ from the constrained problem in $[0,T]$ to the interval $[0,T+\Delta]$, in such a way that it vanishes in $(T,T+\Delta]$; see (4.12) and (4.11). This extended process belongs to $\mathcal{M}(T+\Delta)$, since $\theta(\cdot)$ is bounded.

Therefore,

$$\begin{aligned}
W^M(T+\Delta,y)-W^M(T,y) &\le J(T+\Delta,y,\hat{\nu})-J(T,y,\hat{\nu})\\
&= E^{\alpha,\hat{\nu}}_{T+\Delta}\left[e^{\int_0^{T+\Delta}q(Y_t,\hat{\nu}_t)dt}+\int_0^{T+\Delta}e^{\int_0^t q(Y_s,\hat{\nu}_s)ds}dt\right]\\
&\quad -E^{\alpha,\hat{\nu}}_{T}\left[e^{\int_0^{T}q(Y_t,\hat{\nu}_t)dt}+\int_0^{T}e^{\int_0^t q(Y_s,\hat{\nu}_s)ds}dt\right]\\
&= E^{\alpha,\hat{\nu}}_{T+\Delta}\left[e^{\int_0^{T}q(Y_t,\hat{\nu}_t)dt}\left(e^{\int_T^{T+\Delta}q(Y_t,0)dt}-1\right)\right.\\
&\quad \left.+e^{\int_0^{T}q(Y_t,\hat{\nu}_t)dt}\int_T^{T+\Delta}e^{\int_T^t q(Y_s,0)ds}dt\right]\\
&\le E^{\alpha,\hat{\nu}}_{T+\Delta}e^{\int_0^{T}q(Y_t,\hat{\nu}_t)dt}\left[e^{-\alpha\left(|r|_\infty+\frac{1}{2}(1-\alpha)|\theta|^2_\infty\right)\Delta}-1\right.\\
&\quad \left.+\int_T^{T+\Delta}e^{-\alpha\int_T^t\left(|r|_\infty+\frac{1}{2}(1-\alpha)|\theta|^2_\infty\right)ds}dt\right]\\
&= \left[e^{K_3\Delta}-1+\int_0^{\Delta}e^{K_3 t}dt\right]E^{\alpha,\hat{\nu}}_{T+\Delta}e^{\int_0^{T}q(Y_t,\hat{\nu}_t)dt}.
\end{aligned}$$

The last inequality is due to the fact that $q(y,0)\le K_3 \stackrel{\circ}{=} -\alpha\left[|r|_\infty+\frac{1}{2}(1-\alpha)|\theta|^2_\infty\right]$. Observe that K_3 does not depend on M and y. Then, from the dominated convergence theorem, we have

$$0\le W^M_T(T,y)\le (1+K_3)\,W^M(T,y). \tag{4.19}$$

The next step in the proof shall be to get upper and lower bounds for $W^M_y(T,y)$. This will be done only when y is positive; the same arguments can be adapted when y is negative. Define

$$\Phi(p)\stackrel{\circ}{=}\sup_{v\in[-M,M]}\left\{-\varepsilon pv-\frac{1}{2}(1-\alpha)v^2\right\},\quad p\in\mathbf{R}. \tag{4.20}$$

Then, the HJB equation (4.10) can be written as

$$W^M_T=1+\frac{1}{2}W^M_{yy}+(g-\alpha\rho\theta)W^M_y-\alpha\left[r+\frac{1}{2}(1-\alpha)\theta^2\right]W^M+\alpha W^M\Phi\left(\frac{W^M_y}{W^M}\right). \tag{4.21}$$

On the other hand, the mean value theorem and (4.18) enables us to find a sequence $\{y_n\}$, with $y_n\in[n,2n]$, such that

$$\left|W^M_y(T,y_n)\right|=\frac{1}{n}\left|W^M(T,2n)-W^M(T,n)\right|\le\frac{2}{n}K_1.$$

This inequality allows us to reduce the analysis of W^M_y to only the critical points of $W^M_y(T,\cdot)$, *i.e.,* the points $\tilde{y}>0$ where $W^M_{yy}(T,\tilde{y})=0$. We study three different cases, based on the coefficient of W^M_y in the PDE (4.21):

(a) $g(\tilde{y}) - \alpha\rho\theta(\tilde{y}) \leq -1$ and $W_y^M(T,\tilde{y}) > 0$ or $g(\tilde{y}) - \alpha\rho\theta(\tilde{y}) \geq 1$ and $W_y^M(T,\tilde{y}) < 0$,

(b) $g(\tilde{y}) - \alpha\rho\theta(\tilde{y}) \geq -1$ and $W_y^M(T,\tilde{y}) > 0$, and

(c) $g(\tilde{y}) - \alpha\rho\theta(\tilde{y}) \leq 1$ and $W_y^M(T,\tilde{y}) < 0$.

Case (a): Condition (a) is equivalent to $|g(\tilde{y}) - \alpha\rho\theta(\tilde{y})| \geq 1$ and $[g(\tilde{y}) - \alpha\rho\theta(\tilde{y})] W_y^M(T,\tilde{y}) < 0$. Thus, from (4.21), (4.19), and (4.18), we get

$$\begin{aligned} 0 &< -[g(\tilde{y}) - \alpha\rho\theta(\tilde{y})] W_y^M \\ &= -W_T^M + 1 - \alpha\left[r(\tilde{y}) + \frac{1}{2}(1-\alpha)\theta^2(\tilde{y})\right] W^M + \alpha W^M \Phi\left(\frac{W_y^M}{W^M}\right) \\ &\leq 1 - \alpha\left[r(\tilde{y}) + \frac{1}{2}(1-\alpha)\theta^2(\tilde{y})\right] W^M \\ &\leq 1 + K_1 K_3. \end{aligned}$$

That is, $\left|(g(\tilde{y}) - \alpha\rho\theta(\tilde{y})) W_y^M(T,\tilde{y})\right| \leq \tilde{K}_1 \doteq 1 + K_1 K_3$, where $\tilde{K}_1$ does not depend on $\tilde{y}$ and M. Hence

$$\left|W_y^M(T,\tilde{y})\right| \leq \tilde{K}_1. \tag{4.22}$$

To study cases (b) and (c) we use the logarithmic transformation $V(T,y) \doteq \log W^M(T,y)$. Noting that $W_T^M = W^M V_T$, $W_y^M = W^M V_y$, and $W_{yy}^M = W^M\left(V_y^2 + V_{yy}\right)$, the PDE (4.21) can be written in terms of V as

$$\frac{1}{2}V_y^2 + [g - \alpha\rho\theta] V_y + \frac{1}{2}V_{yy} = V_T - \frac{1}{W^M} + \alpha\left[r + \frac{1}{2}(1-\alpha)\theta^2\right] - \alpha\Phi(V_y). \tag{4.23}$$

Note that from (4.19) and (4.20),

$$V_T \leq 1 + K_3 \quad \text{and} \quad 0 \leq -\alpha\Phi(V_y) \leq \frac{1}{2}\gamma\varepsilon^2 V_y^2.$$

These estimates, together with (4.23), imply that

$$\frac{1}{2}\left(1 - \gamma\varepsilon^2\right) V_y^2 + [g - \alpha\rho\theta] V_y + \frac{1}{2}V_{yy} \leq (K_3 + 1) + \alpha\left[r + \frac{1}{2}(1-\alpha)\theta^2\right] \leq 1 + 2K_3.$$

Then, $V_y^2 + 2\tilde{g}V_y + \frac{1}{1-\gamma\varepsilon^2}V_{yy} \leq \tilde{\alpha}$, with $\tilde{\alpha} \doteq 2(1 + 2K_3)/\left(1 - \gamma\varepsilon^2\right) > 0$ and

$$\tilde{g}(y) \doteq (g(y) - \alpha\rho\theta(y))/\left(1 - \gamma\varepsilon^2\right).$$

If $\tilde{y} > 0$ is a critical point of $V_y(T,\cdot)$, it follows that

$$V_y^2(T,\tilde{y}) + 2\tilde{g}(\tilde{y}) V_y(T,\tilde{y}) \leq \tilde{\alpha},$$

which is equivalent to $[V_y(T,\tilde{y}) + \tilde{g}(\tilde{y})]^2 \leq \tilde{\alpha} + \tilde{g}^2(\tilde{y})$. Thus,

$$-\tilde{g}(\tilde{y}) - \sqrt{\tilde{\alpha} + \tilde{g}^2(\tilde{y})} \leq V_y(T,\tilde{y}) \leq -\tilde{g}(\tilde{y}) + \sqrt{\tilde{\alpha} + \tilde{g}^2(\tilde{y})}. \tag{4.24}$$

Case (b): When $\tilde{g}(\tilde{y}) \geq -1$ and $V_y(T,\tilde{y}) > 0$. Since the function $\tilde{h}(u) \doteq -u + \sqrt{\tilde{\alpha} + u^2}$ is bounded when $u \geq -1$, then the r.h.s. of (4.24) is bounded. Hence,

$$0 < V_y(T,\tilde{y}) \leq \tilde{K}_2 \doteq 1 + \sqrt{1 + \tilde{\alpha}}. \tag{4.25}$$

Note that the constant $\tilde{K}_2$ does not depend on $\tilde{y}$ and M.

Case (c): When $\tilde{g}(\tilde{y}) \leq 1$ and $V_y(T,\tilde{y}) < 0$. Similarly, since the function $\tilde{h}(u) \doteq -u - \sqrt{\tilde{\alpha} + u^2}$ is bounded when $u \leq 1$, then the l.h.s. of (4.24) is bounded, and hence

$$-\tilde{K}_2 \leq V_y(T,\tilde{y}) < 0. \tag{4.26}$$

Putting together (4.22), (4.25), and (4.26), we get

$$|V_y(T,\tilde{y})| \leq \tilde{K} \doteq \tilde{K}_1 \vee \tilde{K}_2, \tag{4.27}$$

for any positive critical point $\tilde{y}$ of $V_y(T,\cdot)$. Now, if $M > [\varepsilon/(1-\alpha)]\tilde{K}$, using (4.27) we conclude that

$$\frac{\varepsilon}{1-\alpha}\frac{\left|W_y^M\right|}{W^M} < M \quad \text{and} \quad w(T,y) = W^M(T,y) = W(T,y).$$

This implies, using (4.12) and (4.11), that the control process $\hat{\nu}$ defined in (4.15) is optimal. Finally, substituting the Markov policy (4.11) in the HJB equation (4.10) yields (4.16). □

Solution to the investor's problem. Now we shall obtain explicitly an optimal trading strategy for the investor's problem when the utility function is HARA. The main argument is based on a Dynkin's result, which turns out to be the key idea to go back from the dual optimization problem solved above to the investor's one; see Proposition 5.4.2 in [KaSr98].

The optimal process $\hat{\nu}$ for the associated dual problem is given by (4.15), with $W(T,y)$ being the smooth solution of the HJB equation (4.16). According to the steps described at the end of the last section, we consider as candidates for being the optimal final wealth and consumption process to

$$\hat{B} \doteq I\left(\hat{\lambda}\frac{Z_T^{\hat{\nu}}}{S_T^0}\right) = \frac{x}{\Lambda_{\hat{\nu}}}\left(\frac{Z_T^{\hat{\nu}}}{S_T^0}\right)^{\alpha-1} \quad \text{and} \quad \hat{c}_t \doteq I\left(\hat{\lambda}\frac{Z_t^{\hat{\nu}}}{S_t^0}\right) = \frac{x}{\Lambda_{\hat{\nu}}}\left(\frac{Z_t^{\hat{\nu}}}{S_t^0}\right)^{\alpha-1}, \quad t \in [0,T],$$

where

$$\hat{\lambda} = \left(\frac{\Lambda_{\hat{\nu}}}{x}\right)^{\frac{1}{1-\alpha}} \quad \text{and} \quad \Lambda_{\hat{\nu}} = E\left[\left(\frac{Z_T^{\hat{\nu}}}{S_T^0}\right)^{\alpha} + \int_0^T \left(\frac{Z_t^{\hat{\nu}}}{S_t^0}\right)^{\alpha} dt\right] = W(T,y).$$

To derive the expression of the optimal trading portfolio $\hat{\pi}$, we will prove first that the process defined as

$$M_t \doteq \left(\frac{Z_t^{\hat{\nu}}}{S_t^0}\right)^{\alpha} W(T-t, Y_t) + \int_0^t \left(\frac{Z_u^{\hat{\nu}}}{S_u^0}\right)^{\alpha} du, \quad t \in [0,T],$$

is a martingale, and then find its stochastic integral representation. Note first that the initial and terminal expectation of this process coincides, i.e., $EM_T = M_0$, since $M_0 = W(T,y) = \Lambda_{\hat{\nu}}$ and $M_T = \left(Z_T^{\hat{\nu}}/S_T^0\right)^{\alpha} + \int_0^T \left(Z_t^{\hat{\nu}}/S_t^0\right)^{\alpha} dt$. Under the original measure P, we write down the system of SDE in $[0,T]$,

$$\begin{aligned} dY_t &= g(Y_t)\,dt + \rho dW_{1t} + \varepsilon dW_{2t}, & Y_0 &= y, \\ d\left[\frac{Z_t^{\hat{\nu}}}{S_t^0}\right] &= -\frac{Z_t^{\hat{\nu}}}{S_t^0}\left[r(Y_t)\,dt + \theta(Y_t)\,dW_{1t} + \hat{\nu}_t dW_{2t}\right], & \frac{Z_0^{\hat{\nu}}}{S_0^0} &= z = 1. \end{aligned}$$

This system has associated a differential operator, defined as

$$\mathcal{L}f \triangleq f_t + gf_y - rzf_z - z\left(\rho\theta + \varepsilon\nu^*\right) f_{yz} + \frac{1}{2}f_{yy} + \frac{1}{2}z^2\left(\theta^2 + \nu^{*2}\right) f_{zz},$$

for $f \in C^{1,2,2}\left([0,T] \times \mathbf{R} \times \mathbf{R}_+\right)$. In particular, when $f\left(t,y,z\right) \triangleq W\left(T-t,y\right) z^\alpha$, and using (4.15) and (4.16), it is not difficult to verify that

$$\mathcal{L}f\left(t,y,z\right) = -z^\alpha.$$

Therefore,

$$M_t = f\left(t, Y_t, \frac{Z_t^{\hat{\nu}}}{S_t^0}\right) - \int_0^t \mathcal{L}f\left(s, Y_s, \frac{Z_s^{\hat{\nu}}}{S_s^0}\right) ds.$$

Thus, from Proposition 5.4.2 in [KaSr91], M is a nonnegative local martingale. Furthermore, it is also a martingale since it is a supermartingale with constant mean.

Now define the nonnegative process

$$\frac{\hat{X}_t}{S_t^0} \triangleq E^{\hat{\nu}}\left[\frac{\hat{B}}{S_T^0} + \int_t^T \frac{\hat{c}_u}{S_u^0} du \mid \mathcal{F}_t\right], \quad t \in [0,T].$$

From (2.4), we obtain

$$\begin{aligned}
\frac{\hat{X}_t}{S_t^0} &= E^{\hat{\nu}}\left[\frac{\hat{B}}{S_T^0} + \int_t^T \frac{\hat{c}_u}{S_u^0} du \mid \mathcal{F}_t\right] \\
&= \frac{x}{W\left(T,y\right) Z_t^{\hat{\nu}}} E\left[\left(\frac{Z_T^{\hat{\nu}}}{S_T^0}\right)^\alpha + \int_t^T \left(\frac{Z_u^{\hat{\nu}}}{S_u^0}\right)^\alpha du \mid \mathcal{F}_t\right] \\
&= \frac{x}{W\left(T,y\right) Z_t^{\hat{\nu}}} E\left[M_T - \int_0^t \left(\frac{Z_u^{\hat{\nu}}}{S_u^0}\right)^\alpha du \mid \mathcal{F}_t\right] \\
&= x\frac{\left[Z_t^{\hat{\nu}}\right]^{\alpha-1}}{\left[S_t^0\right]^\alpha} \frac{W\left(T-t,Y_t\right)}{W\left(T,y\right)}.
\end{aligned}$$

However, by Ito's formula

$$\begin{aligned}
d\left[Z^{\hat{\nu}}\right]^{\alpha-1} &= \left(1-\alpha\right)\left[Z^{\hat{\nu}}\right]^{\alpha-1}\left[\left(1-\frac{\alpha}{2}\right)\left(\theta^2+\hat{\nu}^2\right)dt + \theta dW_1 + \hat{\nu}dW_2\right], \\
dW &= \left(-W_t + \frac{1}{2}W_{yy} + gW_y\right)dt + W_y\left(\rho dW_1 + \varepsilon dW_2\right), \\
d\left[S^0\right]^{-\alpha} &= -\alpha r\left[S^0\right]^{-\alpha} dt.
\end{aligned}$$

Here W represents the process $W\left(T-t,Y_t\right)$, for $t \in [0,T]$. Then, using (4.15) and

(4.16), we get

$$
\begin{aligned}
&d\left(\frac{\left[Z^{\hat{\nu}}\right]^{\alpha-1}}{\left[S^0\right]^{\alpha}}\,W\right)\\
&\quad=\frac{\left[Z^{\hat{\nu}}\right]^{\alpha-1}}{\left[S^0\right]^{\alpha}}\left(-\alpha r W dt+(1-\alpha)\,W\left[\left(1-\frac{\alpha}{2}\right)\left(\theta^2+\hat{\nu}^2\right)dt+\theta dW_1+\hat{\nu}dW_2\right]\right.\\
&\qquad\left.+\left[-W_t+\frac{1}{2}W_{yy}+gW_y\right]dt\right.\\
&\qquad\left.+W_y\left(\rho dW_1+\varepsilon dW_2\right)+(1-\alpha)\left(\rho\theta+\varepsilon\hat{\nu}\right)W_y dt\right)\\
&\quad=\frac{\left[Z^{\hat{\nu}}\right]^{\alpha-1}}{\left[S^0\right]^{\alpha}}\left(-dt+W\left[(1-\alpha)\,\theta+\rho\frac{W_y}{W}\right]\left[\theta dt+dW_1\right]\right)\\
&\quad=\frac{\left[Z^{\hat{\nu}}\right]^{\alpha-1}}{\left[S^0\right]^{\alpha}}\left(-dt+W\left[(1-\alpha)\,\theta+\rho\frac{W_y}{W}\right]dW_1^{\hat{\nu}}\right).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
d\frac{\hat{X}}{S^0}+\frac{\hat{c}}{S^0}dt&=\frac{x}{W(T,y)}d\left(\frac{\left[Z^{\hat{\nu}}\right]^{\alpha-1}}{\left[S^0\right]^{\alpha}}W\right)+\frac{x}{W(T,y)}\left(\frac{Z^{\hat{\nu}}}{S^0}\right)^{\alpha-1}\frac{1}{S^0}dt\\
&=\frac{x}{W(T,y)}\frac{\left[Z^{\hat{\nu}}\right]^{\alpha-1}}{\left[S^0\right]^{\alpha}}W\left[(1-\alpha)\,\theta+\rho\frac{W_y}{W}\right]dW_1^{\hat{\nu}}\\
&=:\frac{\hat{\pi}}{S^0}\sigma dW_1^{\hat{\nu}},
\end{aligned}
$$

where

$$
\hat{\pi}_t\doteq\frac{\hat{X}_t}{\sigma(Y_t)}\left[(1-\alpha)\,\theta(Y_t)+\rho\frac{W_y(T-t,Y_t)}{W(T-t,Y_t)}\right]=:\pi^*(T-t,\hat{X}_t,Y_t),
$$

with

$$
\pi^*(t,x,y)\doteq\frac{x}{\sigma(y)}\left[(1-\alpha)\,\theta(y)+\rho\frac{W_y(t,y)}{W(t,y)}\right],\quad(t,x,y)\in[0,T]\times\mathbf{R}_+\times\mathbf{R}.
$$

Moreover, the optimal consumption process $\hat{c}$ can be written in a feedback form as

$$
\begin{aligned}
\hat{c}_t&=\frac{x}{\Lambda_{\hat{\nu}}}\left(\frac{Z_t^{\hat{\nu}}}{S_t^0}\right)^{\alpha-1}=x\left(\frac{Z_t^{\hat{\nu}}}{S_t^0}\right)^{\alpha-1}\frac{W(T-t,Y_t)}{W(T,y)}\frac{1}{W(T-t,Y_t)}\\
&=:c^*\left(T-t,\hat{X}_t,Y_t\right),
\end{aligned}
$$

with

$$
c^*(t,x,y)\doteq\frac{x}{W(t,y)},\quad(t,x,y)\in[0,T]\times\mathbf{R}_+\times\mathbf{R}.
$$

From the formula of $\hat{X}$ together with (2.10), we conclude that $(\hat{\pi},\hat{c})\in\mathcal{A}(x,y)$ and $X^{\hat{\pi},\hat{c}}\equiv\hat{X}$. Moreover, $X_T^{\hat{\pi},\hat{c}}\equiv\hat{B}$ is the optimal terminal wealth.

The optimal investment and consumption policies obtained above depend on the solution of the PDE (4.16). Performing a power transformation, it is possible to

get a functional representation, which might be useful to approximate its solution through a fixed point algorithm. Suppose that $W(T,y) =: [h(T,y)]^{\delta}$, for some $\delta > 0$. Then, $W_T = \delta h^{\delta-1} h_T$, $W_y = \delta h^{\delta-1} h_y$, $W_y^2/W = \delta^2 h^{\delta-2} h_y^2$, and $W_{yy} = \delta h^{\delta-1} h_{yy} + \delta(\delta-1) h^{\delta-2} h_y^2$. Using (4.16), it follows that $h(T,y)$ solves the PDE

$$h_T = \frac{1}{\delta} h^{1-\delta} + \frac{1}{2} h_{yy} + (g - \alpha\rho\theta) h_y - \frac{\alpha}{\delta}\left[r + \frac{1}{2}(1-\alpha)\theta^2\right] h + \frac{1}{2}\left(\delta - 1 - \gamma\delta\varepsilon^2\right)\frac{h_y^2}{h},$$

with initial condition $h(0,y) = 1$. However, observe that when $\delta \stackrel{\circ}{=} 1/\left(1-\gamma\varepsilon^2\right)$ the last nonlinear term in the previous PDE vanishes. Hence, for this value of δ, the PDE (4.16) is equivalent to

(4.28)
$$h_T = \frac{1}{\delta} h^{1-\delta} + \frac{1}{2} h_{yy} + (g - \alpha\rho\theta) h_y - \frac{\alpha}{\delta}\left[r + \frac{1}{2}(1-\alpha)\theta^2\right] h, \quad \text{with} \quad h(0,y) = 1.$$

The previous power transformation was introduced by Zariphopoulou in [Za01] for an optimal investment problem.

On the other hand, since the function $h(T,y)$ belongs to the set of functions $C^{1,2}\left(\bar{\mathbf{R}}_+ \times \mathbf{R}\right) \cap C_b^{0,1}\left(\bar{\mathbf{R}}_+ \times \mathbf{R}\right)$, the Feynman–Kac formula can be used to get the following representation:

$$h(T,y) = E\left(e^{-\frac{\alpha}{\delta}\int_0^T \left[r(\check{Y}_u) + \frac{1}{2}(1-\alpha)\theta^2(\check{Y}_u)\right] du}\right.$$
$$\left. + \quad \frac{1}{\delta}\int_0^T h^{1-\delta}\left(T-u, \check{Y}_u\right) e^{-\frac{\alpha}{\delta}\int_0^u \left[r(\check{Y}_s) + \frac{1}{2}(1-\alpha)\theta^2(\check{Y}_s)\right] ds} du\right), \tag{4.29}$$

for $(T,y) \in \bar{\mathbf{R}}_+ \times \mathbf{R}$. Here $\{\check{Y}_t\}_{t\in[0,T]}$ is the solution of the SDE

$$d\check{Y}_t = \left[g(\check{Y}_t) - \alpha\rho\theta(\check{Y}_t)\right] dt + d\check{W}_t, \tag{4.30}$$

with initial condition $\check{Y}_0 = y$ and $\check{W}_t$ being a Brownian motion; see Theorem 5.7.6 and Corollary 4.4.5 in [KaSr91].

Remark 4.5 (logarithmic utility). Analogous results can be derived when $U_1(b) = U_1(b) = \log b$. In this case, the dual optimization problem is equivalent to

$$\text{minimize} \quad E\left\{\int_0^T \nu_t^2 dt + \int_0^T \int_0^t \nu_s^2 ds dt\right\} \quad \text{over} \quad \nu \in \mathcal{M}.$$

The optimal solution can be easily obtained and is given by $(\hat{\nu}, \hat{\lambda}) \equiv (0, (1+T)/x)$. Further, the optimal trading strategy $(\hat{\pi}, \hat{c})$ is given by $\hat{\pi}_t = \pi^*(t, X_t^{\hat{\pi},\hat{c}}, Y_t)$ and $\hat{c}_t = c^*(t, X_t^{\hat{\pi},\hat{c}}, Y_t)$, where

$$\pi^*(t,x,y) \stackrel{\circ}{=} x\frac{\mu(y) - r(y)}{\sigma^2(y)} \quad \text{and} \quad c^*(t,x,y) \stackrel{\circ}{=} \frac{x}{1+T-t},$$

$$\text{for} \qquad (t,x,y) \in [0,T] \times \mathbf{R}_+ \times \mathbf{R}.$$

Remark 4.6 (investment or consumption). Modifying slightly the above arguments it is also possible to get explicit optimal solutions for optimal investment or consumption problems. The optimal investment problem is defined as the investor's problem (2.3) with $U_2(\cdot) = 0$, whereas when $U_1(\cdot) = 0$ it corresponds to the optimal consumption problem. These results are summarized below.

1. *Investment problem with logarithmic utility function.* The dual functional is

$$L(\nu, \lambda) = E \int_0^T r(Y_u)\, du - 1 + \lambda x - \log \lambda - E \log Z_T^{\nu},$$

whereas the optimal values for both the dual and primal problems are, respectively, $\hat{\lambda} = 1/x$, $\hat{\nu} = 0$,

$$X_t^{\hat{\pi}} = x \frac{S_t^0}{Z_t^0}, \quad \text{and} \quad \hat{\pi}_t = \frac{\mu(Y_t) - r(Y_t)}{\sigma^2(Y_t)} X_t^{\hat{\pi}}, \quad t \in [0, T].$$

2. *Investment problem with HARA utility function.* The dual functional is

$$L(\nu, \lambda) = \lambda x - \frac{1}{\alpha} \lambda^{\alpha} \Lambda_{\nu},$$

where $\Lambda_{\nu} \triangleq E\left(Z_T^{\nu}/S_T^0\right)^{\alpha}$; the optimal control process $\hat{\nu}$ is given by (4.15). Here the corresponding value function $W(T, y)$ is the solution of the PDE

$$W_T = \frac{1}{2} W_{yy} + (g - \alpha\rho\theta) W_y - \alpha \left[r + \frac{1}{2}(1 - \alpha)\theta^2 \right] W - \frac{1}{2} \gamma \varepsilon^2 \frac{W_y^2}{W},$$

with $W(0, y) = 1$. The optimal wealth and trading portfolio processes are

$$X_t^{\hat{\pi}} = x \left(\frac{S_t^0}{Z_t^0} \right)^{\alpha} \frac{W(T - t, Y_t)}{W(T, y)}, \quad \hat{\pi}_t = \pi^* \left(T - t, X_t^{\hat{\pi}}, Y_t \right), \quad t \in [0, T],$$

where $\pi^*(t, x, y) = (x/\sigma(y)) [(1 - \alpha)\theta(y) + \rho W_y(t, y)/W(t, y)]$; for details see [CaHe03].

3. *Consumption problem with logarithmic utility function.* The dual functional is

$$L(\nu, \lambda) = E \int_0^T \int_0^t r(Y_u)\, du dt + \lambda x - T(1 + \log \lambda) - E \int_0^T \log Z_t^{\nu} dt,$$

and the optimal values are $\hat{\lambda} = T/x$, $\hat{\nu} = 0$. Furthermore,

$$X_t^{\hat{\pi}, \hat{c}} = x \frac{T - t}{T} \frac{S_t^0}{Z_t^0}, \quad \hat{\pi}_t = \frac{\mu(Y_t) - r(Y_t)}{\sigma^2(Y_t)} X_t^{\hat{\pi}, \hat{c}}, \quad \text{and} \quad \hat{c}_t = \frac{1}{T - t} X_t^{\hat{\pi}, \hat{c}}, \quad t \in [0, T].$$

4. *Consumption problem with HARA utility function.* The dual functional is

$$L(\nu, \lambda) = \lambda x - \frac{1}{\alpha} \lambda^{\alpha} \Lambda_{\nu}, \quad \text{where} \quad \Lambda_{\nu} \triangleq E \int_0^T \left(\frac{Z_t^{\hat{\nu}}}{S_t^0} \right)^{\alpha} dt.$$

The optimal solution to the dual problem is the process $\hat{\nu}$ given by (4.15), where the corresponding value function $W(T,y)$ solves the PDE

$$W_T = 1 + \frac{1}{2}W_{yy} + (g - \alpha\rho\theta)\,W_y - \alpha\left[r + \frac{1}{2}(1-\alpha)\,\theta^2\right]W - \frac{1}{2}\gamma\varepsilon^2\frac{W_y^2}{W},$$

with $W(0,y) = 0$. The optimal processes are

$$X_t^{\hat{\pi},\hat{c}} = x\left(\frac{S_t^0}{Z_t^0}\right)^{\alpha}\frac{W(T-t,Y_t)}{W(T,y)}, \quad \hat{\pi}_t = \pi^*(T-t, X_t^{\hat{\pi},\hat{c}}, Y_t),$$

$$\text{and} \quad \hat{c}_t = c^*(T-t, X_t^{\hat{\pi},\hat{c}}, Y_t),$$

where $\pi^*(t,x,y) \triangleq (x/\sigma(y))\left[(1-\alpha)\,\theta(y) + \rho W_y(t,y)/W(t,y)\right]$ and $c^*(t,x,y) \triangleq x/W(t,y)$.

REFERENCES

[BaSp02] O. Barndorff-Nielsen and N. Shephard, *Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics*, J. R. Stat. Soc., B Stat. Methodol. 63 (2001), pp. 167–241.

[BKR03] F. E. Benth, K. H. Karlsen, and K. Reikvam, *Merton's trading portfolio optimization problem in a Black and Scholes market with non-Gaussian stochastic volatility of Ornstein-Uhlenbeck type*, Math. Finance, 13 (2003), pp. 215–244.

[BiPl99] T. R. Bielecki and S. R. Pliska, *Risk-sensitive dynamic asset management*, Appl. Math. Optim., 39 (1999), pp. 337–360.

[CaHe03] N. Castañeda-Leyva and D. Hernández-Hernández, *Optimal investment in incomplete financial markets with stochastic volatility*, Contemp. Math., 336 (2003), pp. 119–136.

[Cu97] D. Cuoco, *Optimal consumption and equilibrium prices with portfolio constraints and stochastic income*, J. Econom. Theory, 72 (1997), pp. 33–73.

[Da00] M. Davis, *Optimal hedging with basis risk*, preprint, 2000.

[FlHe02] W. H. Fleming and D. Hernández-Hernández, *An optimal consumption model with stochastic volatility*, Finance Stoch., 7 (2003), pp. 245–262.

[FlSo93] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[FPS00] J.-P. Fouque, G. Papanicolaou, and K. R. Sircar, *Derivatives in Financial Markets with Stochastic Volatility*, Cambridge University Press, Cambridge, UK, 2000.

[Fr75] A. Friedman, *Stochastic Differential Equations and Applications*, Academic Press, New York, 1975.

[HaPl81] J. M. Harrison and S. R. Pliska, *Martingales and stochastic integrals in the theory of continuous trading*, Stochastic Process. Appl., 11 (1981), pp. 215–260.

[JaSh87] J. Jacod and A. N. Shiryaev, *Limit Theorems for Stochastic Processes*, Springer-Verlag, Berlin, 1987.

[KaSr91] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd ed. Springer-Verlag, New York, 1991.

[KaSr98] I. Karatzas and S. E. Shreve, *Methods of Mathematical Finance*, Springer-Verlag, New York, 1998.

[KrSc99] D. Kramkov and W. Schachermayer, *The asymptotic elasticity of utility functions and optimal investments in incomplete markets*, Ann. Appl. Probab., 9 (1999), pp. 904–950.

[LiSh01] R. S. Lipster and A. N. Shiryaev, *Statistics of Random Processes*, Springer-Verlag, Berlin, 2001.

[Lu69] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley & Sons, New York, 1969.

[Za01] T. Zariphopoulou, *A solution approach to valuation with unhedgeable risks*, Finance Stoch., 5 (2001), pp. 61–82.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 4, pp. 1345–1369

GENERALIZED SAMPLED-DATA STABILIZATION OF WELL-POSED LINEAR INFINITE-DIMENSIONAL SYSTEMS*

HARTMUT LOGEMANN†, RICHARD REBARBER‡, AND STUART TOWNLEY§

Abstract. We consider well-posed linear infinite-dimensional systems, the outputs of which are sampled in a generalized sense using a suitable weighting function. Under certain natural assumptions on the system, the weighting function, and the sampling period, we show that there exists a generalized hold function such that unity sampled-data feedback renders the closed-loop system exponentially stable (in the state-space sense) as well as L^2-stable (in the input-output sense). To illustrate our main result, we describe an application to a structurally damped Euler–Bernoulli beam.

Key words. generalized hold, generalized sampling, infinite-dimensional systems, sampled-data control, stabilization

AMS subject classifications. 34G10, 47D06, 93C25, 93C57, 93D15

DOI. 10.1137/S0363012903434340

1. Introduction.

The design of sampled-data controllers is important both for applications, because of digital implementation issues, and for theoretical development. Sampled-data control for infinite-dimensional systems has been considered in a number of papers; see [12, 13, 14, 15, 18, 19, 30]. In this paper we develop generalized sampled-data control for well-posed linear continuous-time infinite-dimensional systems. Generalized sampled-data control has been frequently studied for finite-dimensional systems (see, for instance, [2, 10]) and for infinite-dimensional systems in Tarn et al. [28] and Tarn, Zavgren, and Zeng [29]. A well-posed system Σ has generating operators (A, B, C), where A is the generator of a strongly continuous semigroup $\mathbf{T} = (\mathbf{T}_t)_{t\geq 0}$ governing the state evolution of the uncontrolled system, B is the control operator, and C is the observation operator; see, for example, [5, 23, 25, 27, 31]. Denote by u and y the input and output of Σ. For a given sampling period $\tau > 0$, a generalized sampled-data feedback control will have the form

$$u(t) = v(t) - H(t - k\tau)y_k, \qquad t \in [k\tau, (k+1)\tau), \quad k = 0, 1, 2, \dots . \tag{1.1}$$

In (1.1), $H(\cdot)$ represents a generalized hold element in the feedback, $v(\cdot)$ denotes an external input to the closed-loop sampled-data feedback system, and y_k is the kth sample of the output y. In the most general setting, y_k is obtained via generalized sampling (i.e., weighted averaging):

$$y_k := \int_0^{\tau-\delta} w(s)y((k-1)\tau + \delta + s)\, ds,$$

*Received by the editors September 5, 2003; accepted for publication (in revised form) January 5, 2005; published electronically October 21, 2005. This work was supported in part by EPSRC (grant GR/S01580/01) and the National Science Foundation (grant DMS-0206951).
http://www.siam.org/journals/sicon/44-4/43434.html
†Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK (hl@maths.bath.ac.uk).
‡Department of Mathematics, University of Nebraska-Lincoln, Lincoln, NE 68588-0130 (rrebarbe@math.unl.edu).
§Corresponding author. School of Engineering, Computer Science and Mathematics, University of Exeter, Exeter EX4 4QE, UK (townley@maths.ex.ac.uk).

where $\delta \in (0, \tau)$ and w is a suitable scalar-valued weighting function defined on $[(k-1)\tau + \delta, k\tau]$. This kind of generalized sampling is natural for well-posed systems where the output typically is in L^2_{loc} but is not necessarily continuous. The feedback element $H(\cdot)$ in (1.1) is also referred to as a periodic gain, as in [28, 29] and Chammas and Leondes [2].

Control objective. Choose a generalized hold function H defined on $[0, \tau]$, such that the unity sampled-data feedback given by (1.1), when applied to the well-posed system Σ, yields an exponentially stable closed-loop system.

Our main result is Theorem 4.4. Loosely speaking, Theorem 4.4, part (1), states that for a given well-posed system Σ, we can choose H to meet the control objective if

(i) the unstable portion of the spectrum of A consists of at most finitely many eigenvalues with finite algebraic multiplicities,
(ii) the semigroup generated by the stable part of A is exponentially stable,
(iii) the unstable (finite-dimensional) part of the observed discrete-time system $(C, \mathbf{T}_\tau)$ is observable,
(iv) $\int_0^{\tau-\delta} w(s)e^{\lambda s}ds \neq 0$ for all unstable eigenvalues λ of A,
(v) the unstable subspace of Σ is contained in the closure of its reachable subspace.

In Proposition 4.6 we show that conditions (i)–(iv) above are in fact necessary, and in Remark 4.3 it is noted that condition (iv) is in fact satisfied "generically." Furthermore, if the semigroup generated by A is analytic, then (v) is also necessary. In [19] we showed, however, that in general (v) is not necessary for stabilization by idealized sampling and generalized hold sampled-data control. This necessity issue is also discussed in [18, 19, 30].

In Theorem 4.4, part (2), we show that the resulting closed-loop system with external input v is L^2-stable in an input-output sense. In part (3) we show that if the square-integrable input v is such that $\dot{v}$ is also square-integrable, and if the initial state satisfies a certain natural smoothness condition, then the output $y(t)$ of the sampled-data feedback system converges to 0 as $t \to \infty$.

Our main result extends, generalizes, and improves the basic result in [29] in a number of ways. First, the results in [29] are proved for systems with bounded operators B and C and then stated without proof for a class of systems with unbounded B and C satisfying the conditions of the set-up developed in [4]. The unboundedness in this class of systems is quite limited and allows only a few systems described by partial differential equations with boundary control and observation. The results in [29] were further developed in [28] to encompass a class of neutral systems. In our paper, we work in the context of the theory of well-posed systems, the largest class of infinite-dimensional systems for which there exists a well-developed state-space and frequency-domain theory; see, for example, [5, 22, 23, 25, 26, 27, 31, 32]. Well-posed systems allow for considerable unboundedness of the control and observation operators B and C, and they encompass many of the most commonly studied partial differential equations with boundary control and observation and all functional differential equations of retarded and neutral type with delays in the inputs and outputs. Second, in contrast to [28, 29], not only do we prove results on exponential stability but we also obtain results on input-output stability.

The paper is organized as follows: In section 2 we describe in detail various results relevant to the sampled-data control of well-posed systems. In section 3 we discuss issues relating to sampled-data feedback stabilization. In section 4 we present our

main result. In section 5 we illustrate our results by applying them to a structurally damped Euler–Bernoulli beam.

Notation. $\mathbb{N}$ denotes the set of positive integers; $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$; $\mathbb{R}_+ := [0, \infty)$; for $\alpha \in \mathbb{R}$, set $\mathbb{C}_\alpha := \{s \in \mathbb{C} \mid \operatorname{Re} s > \alpha\}$; for a real or complex Banach space Z, $\alpha \in \mathbb{R}$ and $0 < p \leq \infty$, we define the exponentially weighted spaces $L^p_\alpha(\mathbb{R}_+, Z) := \{f \in L^p_{\mathrm{loc}}(\mathbb{R}_+, Z) : f(\cdot)\exp(-\alpha\,\cdot) \in L^p(\mathbb{R}_+, Z)\}$ and $W^{1,p}_\alpha(\mathbb{R}_+, Z) := \{f \in L^p_{\mathrm{loc}}(\mathbb{R}_+, Z) : f(\cdot)\exp(-\alpha\,\cdot) \in W^{1,p}(\mathbb{R}_+, Z)\}$; we endow $L^p_\alpha(\mathbb{R}_+, Z)$ with the norm $\|f\|_{L^p_\alpha} := \|e^{-\alpha\,\cdot} f(\cdot)\|_{L^p}$; $W^{1,p}_c([a, b], Z)$ denotes the subspace of all functions in $W^{1,p}([a, b], Z)$ with support contained in the open interval (a, b); $\mathcal{B}(Z_1, Z_2)$ denotes the space of bounded linear operators from a Banach space Z_1 to a Banach space Z_2; we write $\mathcal{B}(Z)$ for $\mathcal{B}(Z, Z)$; let $A : \operatorname{dom}(A) \subset Z \to Z$ be a linear operator, where $\operatorname{dom}(A)$ denotes the domain of A; the resolvent set of A and the spectrum of A are denoted by $\varrho(A)$and $\sigma(A)$, respectively; if $A \in \mathcal{B}(Z)$, then $r(A)$ denotes the spectral radius of A.

2. Preliminaries on well-posed systems. Before developing our main results for generalized sampled-data control of well-posed linear systems we first need to cover some basic background material on well-posed linear systems. We cover only those basic properties we need and some specific results relevant in a context of sampled-data control. There are a number of equivalent definitions of well-posed systems; see [5, 22, 23, 25, 26, 27, 31, 32]. We will be brief in the following and refer the reader to [22, 23] for the original definition of a well-posed system, to [31] for issues related especially to admissibility, and to [25] for a more comprehensive treatment. Throughout this section, we will consider a well-posed system Σ with state-space X, input space $\mathbb{R}^m$, and output space $\mathbb{R}^p$, generating operators (A, B, C), input-output operator G, and transfer function $\mathbf{G}$. Here X is a real Hilbert space with norm denoted by $\|\cdot\|$, A is the generator of a strongly continuous semigroup $\mathbf{T} = (\mathbf{T}_t)_{t\geq 0}$ on X, $B \in \mathcal{B}(\mathbb{R}^m, X_{-1})$, and $C \in \mathcal{B}(X_1, \mathbb{R}^p)$, where X_1 denotes the space $\operatorname{dom}(A)$ endowed with the norm $\|z\|_1 := \|(s_0 I - A)z\|$, while X_{-1} denotes the completion of X with respect to the norm $\|z\|_{-1} = \|(s_0 I - A)^{-1} z\|$, where $s_0 \in \varrho(A)$ (different choices of s_0 lead to equivalent norms). Clearly, the norm $\|\cdot\|_1$ is equivalent to the graph norm of A. Moreover, $X_1 \subset X \subset X_{-1}$ and the canonical injections are bounded and dense. The semigroup $\mathbf{T}$ restricts to a strongly continuous semigroup on X_1 and extends to a strongly continuous semigroup on X_{-1} with the exponential growth constant being the same on all three spaces; the generator of the restriction (extension) of $\mathbf{T}$ is a restriction (extension) of A; we shall use the same symbol $\mathbf{T}$ (respectively, A) for the original semigroup (respectively, generator) and the associated restrictions and extensions: with this convention, we may write $A \in \mathcal{B}(X, X_{-1})$ (considered as a generator on X_{-1}, the domain of A is X). The spectra of A and its extension coincide. For $s_0 \in \varrho(A)$, $s_0 I - A$, considered as an operator in $\mathcal{B}(X, X_{-1})$, provides an isometric isomorphism from X to X_{-1} (we refer the reader to [7] for more details on the extrapolation space X_{-1}). The operator B is an *admissible control operator* for $\mathbf{T}$; i.e., for each $t \in \mathbb{R}_+$ there exists $\beta_t \geq 0$ such that

$$\left\| \int_0^t \mathbf{T}_{t-s} B u(s)\, ds \right\| \leq \beta_t \|u\|_{L^2([0,t],\mathbb{R}^m)} \quad \forall\, u \in L^2([0, t], \mathbb{R}^m).$$

The operator C is an *admissible observation operator* for $\mathbf{T}$; i.e., for each $t \in \mathbb{R}_+$ there exists $\gamma_t \geq 0$ such that

$$\left(\int_0^t \|C\mathbf{T}_s z\|^2 ds \right)^{1/2} \leq \gamma_t \|z\| \quad \forall\, z \in X_1.$$

The control operator B is said to be *bounded* if it is so as a map from the input space $\mathbb{R}^m$ to the state space X; otherwise it is said to be *unbounded*. The observation operator C is said to be *bounded* if it can be extended continuously to X; otherwise C is said to be *unbounded*.

The so-called Λ-*extension* C_Λ of C is defined by

$$C_\Lambda z = \lim_{s\to\infty,\, s\in\mathbb{R}} Cs(sI-A)^{-1}z,$$

with $\operatorname{dom}(C_\Lambda)$ consisting of all $z \in X$ for which the above limit exists. For every $z \in X$, $\mathbf{T}_t z \in \operatorname{dom}(C_\Lambda)$ for almost all (a.a.) $t \in \mathbb{R}_+$, and if $\alpha > \omega(\mathbf{T})$, then $C_\Lambda \mathbf{T} z \in L^2_\alpha(\mathbb{R}_+, \mathbb{R}^p)$, where

$$\omega(\mathbf{T}) := \lim_{t\to\infty} \frac{1}{t} \ln \|\mathbf{T}_t\|$$

denotes the exponential growth constant of $\mathbf{T}$. The transfer function $\mathbf{G}$ satisfies

$$\frac{1}{s-s_0}\left(\mathbf{G}(s)-\mathbf{G}(s_0)\right) = -C(sI-A)^{-1}(s_0 I-A)^{-1}B \quad \forall\, s, s_0 \in \mathbb{C}_{\omega(\mathbf{T})},\ s \neq s_0, \tag{2.1}$$

and for every $\alpha > \omega(\mathbf{T})$, $\mathbf{G}$ is analytic and bounded on $\mathbb{C}_\alpha$. Moreover, the input-output operator $G : L^2_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^m) \to L^2_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^p)$ is continuous and right-shift invariant; for every $\alpha > \omega(\mathbf{T})$, $G \in \mathcal{B}(L^2_\alpha(\mathbb{R}_+, \mathbb{R}^m), L^2_\alpha(\mathbb{R}_+, \mathbb{R}^p))$ and

$$(\mathfrak{L}(Gu))(s) = \mathbf{G}(s)(\mathfrak{L}(u))(s) \quad \forall\, s \in \mathbb{C}_\alpha,\ \forall\, u \in L^2_\alpha(\mathbb{R}_+, \mathbb{R}^m),$$

where $\mathfrak{L}$ denotes the Laplace transform. It follows from (2.1) that if two well-posed systems have the same generating operators, then the difference of their transfer functions is constant: roughly speaking, the generating operators determine the input-output behavior of a well-posed system up to a constant.

In the following, let $s_0 \in \mathbb{C}_{\omega(\mathbf{T})}$ be fixed but arbitrary. For $x^0 \in X$ and $u \in L^2_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^m)$, let x and y denote the state and output functions of Σ, respectively, corresponding to the initial condition $x(0) = x^0 \in X$ and the input function u. Then $x(t) = \mathbf{T}_t x^0 + \int_0^t \mathbf{T}_{t-s} B u(s)\, ds$ for all $t \in \mathbb{R}_+$, and $y(t) = C_\Lambda \mathbf{T}_t x^0 + (Gu)(t)$ for a.a. $t \in \mathbb{R}$. Moreover, $x(t) - (s_0 I - A)^{-1} B u(t) \in \operatorname{dom}(C_\Lambda)$ for a.a. $t \in \mathbb{R}_+$ and

$$\dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = x^0, \quad \text{for a.a. } t \in \mathbb{R}_+, \tag{2.2a}$$

$$y(t) = C_\Lambda\left(x(t) - (s_0 I - A)^{-1} B u(t)\right) + \mathbf{G}(s_0)u(t) \quad \text{for a.a. } t \geq 0. \tag{2.2b}$$

Of course, the differential equation (2.2a) has to be interpreted in X_{-1}. In the following, we identify Σ and (2.2) and refer to (2.2) as a well-posed system. We say that the well-posed system (2.2) is *exponentially stable* if $\omega(\mathbf{T}) < 0$. If the well-posed system (2.2) is *regular*, i.e., the limit

$$\lim_{s\to\infty,\, s\in\mathbb{R}} \mathbf{G}(s) = D$$

exists, then $x(t) \in \operatorname{dom}(C_\Lambda)$ for a.a. $t \in \mathbb{R}_+$ and the output equation (2.2b) simplifies to

$$y(t) = C_\Lambda x(t) + Du(t) \quad \text{for a.a. } t \geq 0.$$

Moreover, in the regular case, we have that $(sI - A)^{-1} B\mathbb{R}^m \subset \operatorname{dom}(C_\Lambda)$ for all $s \in \varrho(A)$ and

$$\mathbf{G}(s) = C_\Lambda (sI - A)^{-1} B + D \quad \forall\, s \in \mathbb{C}_{\omega(\mathbf{T})}.$$

The matrix $D \in \mathbb{R}^{p\times m}$ is called the *feedthrough matrix* of (2.2). We mention that if the control operator B or the observation operator C is bounded, then (2.2) is regular.

The following result relates to the asymptotic behavior of the output y of the well-posed system (2.2) under the assumption that x^0 and u satisfy certain "smoothness" conditions.

PROPOSITION 2.1. *Let $\alpha > \omega(\mathbf{T})$, $x^0 \in X$, and $u \in W^{1,2}_\alpha(\mathbb{R}_+, \mathbb{R}^m)$. If there exists $t_0 \in \mathbb{R}_+$ such that $\mathbf{T}_{t_0}(Ax^0 + Bu(0)) \in X$, then the output y of the well-posed system (2.2) is continuous on $[t_0, \infty)$ [1] and satisfies*

$$\lim_{t\to\infty} y(t)e^{-\alpha t} = 0.$$

Proof. Let $x^0 \in X$, $t_0 \in \mathbb{R}_+$, and $u \in W^{1,2}_\alpha(\mathbb{R}_+, \mathbb{R}^m)$ be such that $\mathbf{T}_{t_0}(Ax^0 + Bu(0)) \in X$. The output y of the well-posed system (2.2) is given by

$$y(t) = C_\Lambda \mathbf{T}_t x^0 + (Gu)(t) \quad \text{for a.a. } t \in \mathbb{R}_+. \tag{2.3}$$

Let us first assume that $\alpha = 0$. Then, by hypothesis, $0 = \alpha > \omega(\mathbf{T})$; that is, the well-posed system (2.2) is exponentially stable. Define a right-shift-invariant operator $F : L^2_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^m) \to L^2_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^p)$ by setting

$$(Ff)(t) := \int_0^t ((Gf)(\zeta) - \mathbf{G}(0)f(\zeta))\, d\zeta \quad \forall\, f \in L^2_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^m),\ \forall\, t \in \mathbb{R}_+.$$

The transfer function $\mathbf{F}$ of F is given by $\mathbf{F}(s) = (\mathbf{G}(s) - \mathbf{G}(0))/s$. Clearly, $\mathbf{F}$ is analytic and bounded on $\mathbb{C}_0$ and so, $F \in \mathcal{B}(L^2(\mathbb{R}_+, \mathbb{R}^m), L^2(\mathbb{R}_+, \mathbb{R}^p))$. Using that G commutes with the integration operator (by right-shift invariance), a routine calculation gives

$$Gu = F\dot{u} + \mathbf{G}(0)u + G(u(0)\theta) - \mathbf{G}(0)u(0),$$

where θ denotes the unit-step function. Setting

$$y_1 := F\dot{u} + \mathbf{G}(0)u \qquad \text{and} \qquad y_2 := C_\Lambda \mathbf{T}x^0 + G(u(0)\theta) - \mathbf{G}(0)u(0),$$

it follows from (2.3) that

$$y(t) = y_1(t) + y_2(t) \quad \text{for a.a. } t \in \mathbb{R}_+. \tag{2.4}$$

It is clear that y_1 is continuous. Since $u, \dot{u} \in L^2(\mathbb{R}_+, \mathbb{R}^m)$, we may conclude that $\lim_{t\to\infty} u(t) = 0$. Using again that $\dot{u} \in L^2(\mathbb{R}_+, \mathbb{R}^m)$, it follows from the boundedness of F and G that $F\dot{u}$ and $(d/dt)(F\dot{u})$ are in $L^2(\mathbb{R}_+, \mathbb{R}^p)$, showing that $\lim_{t\to\infty}(F\dot{u})(t) = 0$. Thus, $\lim_{t\to\infty} y_1(t) = 0$. Taking the Laplace transform of y_2 gives

$$(\mathfrak{L}y_2)(s) = C(sI - A)^{-1}x^0 + \frac{1}{s}(\mathbf{G}(s) - \mathbf{G}(0))u(0) \quad \forall\, s \in \mathbb{C}_0.$$

Invoking (2.1) we obtain that for all $s \in \mathbb{C}_0$,

$$(\mathfrak{L}y_2)(s) = C(sI-A)^{-1}x^0 + C(sI-A)^{-1}A^{-1}Bu(0) = C(sI-A)^{-1}A^{-1}(Ax^0 + Bu(0)),$$

[1] The output y of the well-posed system (2.2) is an element in $L^2_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^m)$, and so, strictly speaking, y is not a function but an equivalence class of functions coinciding almost everywhere in $\mathbb{R}_+$. We say that y is continuous on $[t_0, \infty)$ if there exists a representative in the equivalence class which is continuous on $[t_0, \infty)$.

implying that $y_2(t) = C_\Lambda \mathbf{T}_t A^{-1}(Ax^0 + Bu(0))$ for a.a. $t \in \mathbb{R}_+$. Hence, since $\mathbf{T}_{t_0}(Ax^0 + Bu(0)) \in X$,

$$y_2(t) = C\mathbf{T}_{t-t_0} A^{-1} \mathbf{T}_{t_0}(Ax^0 + Bu(0)) \quad \text{for a.a. } t \in [t_0, \infty). \tag{2.5}$$

Obviously, the right-hand side of (2.5) is continuous on $[t_0, \infty)$ and converges to 0 as $t \to \infty$. The claim now follows from (2.4).

Let us now assume that $\alpha \neq 0$. Define the operator $G_\alpha : L^2_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^m) \to L^2_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^p)$ by setting $G_\alpha(u) := e^{-\alpha \cdot} G(e^{\alpha \cdot} u)$. It is trivial that there exists a well-posed system Σ_α with generating operators $(A - \alpha I, B, C)$ and input-output operator G_α (the exponentially weighted version of the well-posed system (2.2)). Since $\alpha > \omega(\mathbf{T})$, it is clear that Σ_α is exponentially stable. If y is the output of the well-posed system (2.2), then

$$y(t)e^{-\alpha t} = C_\Lambda \mathbf{T}_t e^{-\alpha t} x^0 + (G_\alpha(e^{-\alpha \cdot} u))(t) \quad \text{for a.a. } t \in \mathbb{R}_+. \tag{2.6}$$

The right-hand side of (2.6) is the output of the exponentially stable well-posed system Σ_α corresponding to the initial value x^0 and the control function $e^{-\alpha \cdot} u \in W^{1,2}(\mathbb{R}_+, \mathbb{R}^m)$. Moreover, since $\mathbf{T}_{t_0}(Ax^0 + Bu(0)) \in X$,

$$\mathbf{T}_{t_0} e^{-\alpha t_0}\big((A - \alpha I)x^0 + B(e^{-\alpha \cdot} u)(0)\big) = e^{-\alpha t_0} \mathbf{T}_{t_0}\big(Ax^0 + Bu(0) - \alpha x^0\big) \in X.$$

Thus, by what we have already proved, it follows that the right-hand side of (2.6), and hence the function $t \mapsto y(t)e^{-\alpha t}$, is continuous on $[t_0, \infty)$ and converges to 0 as $t \to \infty$. □

We close this section with a simple sufficient condition for a triple of operators (A, B, C) to be the generating operators of a well-posed system. Here $A : \mathrm{dom}(A) \subset X \to X$ generates a strongly continuous semigroup $\mathbf{T} = (\mathbf{T}_t)_{t \geq 0}$, and $B \in \mathcal{B}(\mathbb{R}^m, X_{-1})$ and $C \in \mathcal{B}(X_1, \mathbb{R}^p)$ are admissible control and observation operators for $\mathbf{T}$, respectively. Assume that the semigroup $\mathbf{T}$ is analytic; let $s_0 \in \varrho(A)$ and let $\alpha \geq 0$. Then the fractional powers $(s_0 I - A)^{-\alpha}$ and $(s_0 I - A)^\alpha$ are well-defined (where $(s_0 I - A)^0 := I$), $(s_0 I - A)^\alpha$ is closed, and $(s_0 I - A)^{-\alpha} \in \mathcal{B}(X)$. We endow the domain of $(s_0 I - A)^\alpha$ with the norm

$$\|z\|_\alpha := \|(s_0 I - A)^\alpha z\|$$

and denote the resulting Hilbert space by X_α. Let $X_{-\alpha}$ be the completion of X with respect to the norm

$$\|z\|_{-\alpha} := \|(s_0 I - A)^{-\alpha} z\|.$$

It is trivial that $X_0 = X$ and $(s_0 I - A)^{-\alpha} \in \mathcal{B}(X, X_\alpha)$. If $\alpha \in (0, 1)$, then X_α and $X_{-\alpha}$ can be interpreted as interpolation spaces: between X and X_1 in the case of the former and between X and X_{-1} in the case of the latter. The operator $(s_0 I - A)^\alpha$ extends to an operator in $\mathcal{B}(X, X_{-\alpha})$ and similarly, $(s_0 I - A)^{-\alpha}$ extends to an operator in $\mathcal{B}(X_{-\alpha}, X)$; we shall use the same symbol $(s_0 I - A)^\alpha$ (respectively, $(s_0 I - A)^{-\alpha}$) to denote the extensions.

PROPOSITION 2.2. *Assume that the semigroup $\mathbf{T}$ generated by A is analytic and that $B \in \mathcal{B}(\mathbb{R}^m, X_{-1})$ and $C \in \mathcal{B}(X_1, \mathbb{R}^p)$ are admissible control and observation operators for $\mathbf{T}$, respectively. If there exist $\alpha, \beta \in [0, 1]$ with $\alpha + \beta \leq 1$ and such that $B \in \mathcal{B}(\mathbb{R}^m, X_{-\alpha})$ and $C \in \mathcal{B}(X_\beta, \mathbb{R}^p)$, then there exists a regular well-posed system with generating operators (A, B, C).*

Proof. Fix $\lambda \in \varrho(A)$. It follows from the hypothesis that $\tilde{B} := (\lambda I - A)^{-\alpha} B \in \mathcal{B}(\mathbb{R}^m, X)$ and $\tilde{C} = C(\lambda I - A)^{-\beta} \in \mathcal{B}(X, \mathbb{R}^p)$. Since $\alpha + \beta \leq 1$, the operator $(\lambda I - A)^{\alpha+\beta}(sI - A)^{-1}$ is in $\mathcal{B}(X)$ for all $s \in \varrho(A)$. Consequently, the function $\mathbf{G}$ defined by

$$\mathbf{G}(s) := \tilde{C}(\lambda I - A)^{\alpha+\beta}(sI - A)^{-1}\tilde{B}$$

is analytic on $\varrho(A)$. Moreover,

$$\begin{aligned}(\lambda I - A)^{\alpha+\beta}(sI - A)^{-1} &= (\lambda I - A)^{\alpha+\beta-1}(\lambda I - A)(sI - A)^{-1} \\ &= (\lambda I - A)^{\alpha+\beta-1}[(\lambda + s)(sI - A)^{-1} + I] \ \forall\, s \in \varrho(A). \end{aligned} \tag{2.7}$$

Fix $\gamma > \omega(\mathbf{T})$. The fact that A generates an analytic semigroup guarantees the existence of a constant $M > 0$ such that $\|(sI - A)^{-1}\| \leq M/|s - \gamma|$ for all $s \in \mathbb{C}_\gamma$. Therefore we obtain from (2.7) that the $\mathcal{B}(X)$-valued function $s \mapsto (\lambda I - A)^{\alpha+\beta}(sI - A)^{-1}$ is bounded on $\mathbb{C}_\gamma$. Consequently, $\mathbf{G}$ is bounded on $\mathbb{C}_\gamma$. Moreover, since

$$(sI - A)^{-1}(\lambda I - A)^{\alpha} z = (\lambda I - A)^{\alpha}(sI - A)^{-1} z \in X \quad \forall\, z \in X,\ \forall\, s \in \varrho(A)$$

and

$$(\lambda I - A)^{\alpha}(\lambda I - A)^{\beta} z = (\lambda I - A)^{\alpha+\beta} z \in X \quad \forall\, z \in X_1,$$

an application of the resolvent identity yields for all $s, s_0 \in \varrho(A)$ with $s \neq s_0$

$$\begin{aligned}\frac{1}{s_0 - s}(\mathbf{G}(s) - \mathbf{G}(s_0)) &= \tilde{C}(\lambda I - A)^{\alpha+\beta}(sI - A)^{-1}(s_0 I - A)^{-1}\tilde{B} \\ &= C(sI - A)^{-1}(s_0 I - A)^{-1} B.\end{aligned}$$

Invoking a result in [5], we may now conclude that there exists a well-posed system with generating operators (A, B, C). To show that this system is regular, it suffices to prove that $(s_0 I - A)^{-1} B\mathbb{R}^m \subset \operatorname{dom} C_\Lambda$ for $s_0 \in \varrho(A)$; see [31]. But this follows trivially from the identity

$$C(sI - A)^{-1}(s_0 I - A)^{-1} B = \tilde{C}(\lambda I - A)^{\alpha+\beta}(s_0 I - A)^{-1}(sI - A)^{-1}\tilde{B}$$

and the facts that $\tilde{C} \in \mathcal{B}(X, \mathbb{R}^p)$, $(\lambda I - A)^{\alpha+\beta}(s_0 I - A)^{-1} \in \mathcal{B}(X)$, and $\tilde{B} \in \mathcal{B}(\mathbb{R}^m, X)$. ∎

3. The sampled-data system. Let $\tau > \delta > 0$, $H \in L^2([0, \delta], \mathbb{R}^{m\times p})$, $w \in L^2([0, \tau - \delta], \mathbb{R})$, and $v \in L^2_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^m)$. We apply the following sampled-data feedback control law to the well-posed system (2.2):

$$u(t) = \begin{cases} v(t) - H(t - k\tau)y_k, & t \in [k\tau, k\tau + \delta) \\ v(t), & t \in [k\tau + \delta, (k+1)\tau) \end{cases} \quad \forall\, k \in \mathbb{N}_0, \ \text{where} \tag{3.1a}$$

$$y_0 := 0 \qquad \text{and} \qquad y_k := \int_0^{\tau-\delta} w(s) y((k-1)\tau + \delta + s)\, ds \quad \forall\, k \in \mathbb{N}. \tag{3.1b}$$

The function v represents the input signal of the sampled-data feedback system and emphasises our input-output as well as state-space point of view.

Remark 3.1. Defining $H_\tau \in L^2([0, \tau], \mathbb{R}^{m\times p})$ by

$$H_\tau(t) := \begin{cases} H(t), & t \in [0, \delta], \\ 0, & t \in (\delta, \tau], \end{cases} \tag{3.2}$$

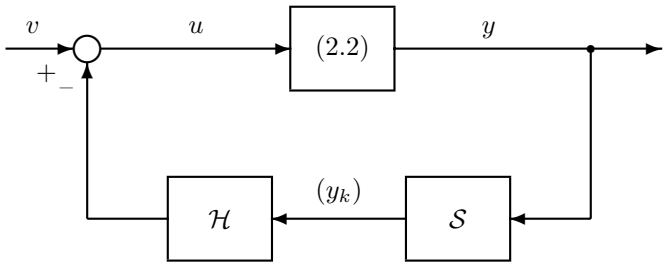


FIG. 1. *Feedback system with generalized sampling $\mathcal{S}$ and generalized hold $\mathcal{H}$.*

and setting

$$\big(\mathcal{H}((y_k))\big)(t) := H_\tau(t-k\tau)y_k \quad \forall\, t \in [k\tau,(k+1)\tau),\ \forall\, k \in \mathbb{N}_0,$$

(3.1a) can be written in the form $u = v - \mathcal{H}((y_k))$. The operator $\mathcal{H}$ represents a generalized hold operation with hold function H_τ (see, for example, [1]). Similarly, (3.1b) describes a generalized sampling operation (see [1]). The function w is called the weighting function of the sampler (3.1b). Note that instantaneous sampling of the form $y_k = y(k\tau)$ is in general not possible since typically the output y of a well-posed system (2.2) need not be continuous. Indeed, the state-space formula (2.2b) for the output does not hold for all $t \in \mathbb{R}_+$, but only for a.a. $t \in \mathbb{R}$: in particular, it might not hold at $t = k\tau$ for some $k \in \mathbb{N}_0$.

The sampled-data feedback system obtained by applying the control law (3.1) to the well-posed system (2.2) is illustrated in Figure 1, where $\mathcal{S}$ denotes the generalized sampling operation given by (3.1b).

It is clear that for given initial state $x^0 \in X$ and given input function $v \in L^2_{\mathrm{loc}}(\mathbb{R}_+,\mathbb{R}^m)$, the (unique) state trajectory $x(\cdot;x^0,v)$ of the sampled-data feedback system given by (2.2) and (3.1) can be obtained recursively from (2.2b), (3.1b), and

$$x(0;x^0,v) = x^0, \tag{3.3a}$$

$$x(k\tau+t;x^0,v) = \mathbf{T}_t x(k\tau;x^0,v) + \int_0^t \mathbf{T}_{t-s}B(v(k\tau+s) - H_\tau(s)y_k)\,ds$$

$$\forall\, t \in (0,\tau],\ \forall\, k \in \mathbb{N}_0. \tag{3.3b}$$

Note that $x(\cdot;x^0,v)$ is a continuous X-valued function defined on $\mathbb{R}_+$. For simplicity, in the following we shall occasionally use the abbreviation $x := x(\cdot;x^0,v)$. We define

$$x_k := x(k\tau), \quad x_{k,\delta} := x(k\tau+\delta) \qquad \forall\, k \in \mathbb{N}_0.$$

For $\sigma,\tau > 0$, we define the left-shift/truncation operator $\mathbf{L}^\tau_\sigma : L^2_{\mathrm{loc}}(\mathbb{R}_+,\mathbb{R}^m) \to L^2(\mathbb{R}_+,\mathbb{R}^m)$ by setting

$$(\mathbf{L}^\tau_\sigma f)(t) := \begin{cases} f(t+\sigma), & t \in [0,\tau], \\ 0, & t \in (\tau,\infty). \end{cases}$$

In the following lemma we establish the basic discrete-time equations (involving x_k, $x_{k,\delta}$, y_k, and $\mathbf{L}^\tau_{k\tau+\delta}v$) associated with the sampled-data feedback system given by (2.2) and (3.1).

LEMMA 3.2. *Let $\tau > \delta > 0$, $H \in L^2([0,\delta],\mathbb{R}^{m\times p})$, and $w \in L^2([0,\tau-\delta],\mathbb{R})$. We assume that*

$$\int_0^{\tau-\delta} w(s)\mathbf{T}_s z\,ds \in X_1 \quad \forall\, z \in X. \tag{3.4}$$

Then the following statements hold.

(1) *The operator*

$$L_w : X \to X_1,\ z \mapsto \int_0^{\tau-\delta} w(s)\mathbf{T}_s z\,ds \tag{3.5}$$

is in $\mathcal{B}(X, X_1)$.

(2) *The sequences (x_k), $(x_{k,\delta})$, and (y_k) satisfy, for all $k \in \mathbb{N}_0$,*

$$x_{k+1} = \mathbf{T}_{\tau-\delta}x_{k,\delta} + \int_0^{\tau-\delta} \mathbf{T}_{\tau-\delta-s}Bv(k\tau+\delta+s)\,ds, \tag{3.6}$$

$$y_{k+1} = CL_w x_{k,\delta} + M_w \mathbf{L}^\tau_{k\tau+\delta} v, \tag{3.7}$$

$$x_{k+1,\delta} = (\mathbf{T}_\tau + K_H C L_w)x_{k,\delta} + M_{H,w}\mathbf{L}^\tau_{k\tau+\delta} v, \tag{3.8}$$

where $K_H \in \mathcal{B}(\mathbb{R}^p, X)$, $M_w \in \mathcal{B}(L^2(\mathbb{R}_+,\mathbb{R}^m),\mathbb{R}^p)$, and $M_{H,w} \in \mathcal{B}(L^2(\mathbb{R}_+,\mathbb{R}^m), X)$ are defined by

$$K_H z = -\int_0^\delta \mathbf{T}_{\delta-s}BH(s)z\,ds \quad \forall\, z \in \mathbb{R}^p, \tag{3.9}$$

$$M_w f = \int_0^{\tau-\delta} w(s)(Gf)(s)\,ds \quad \forall\, f \in L^2(\mathbb{R}_+,\mathbb{R}^m), \tag{3.10}$$

$$M_{H,w} f = K_H M_w f + \int_0^\tau \mathbf{T}_{\tau-s}Bf(s)\,ds \quad \forall\, f \in L^2(\mathbb{R}_+,\mathbb{R}^m), \tag{3.11}$$

respectively.

Remark 3.3. It is easy to show, using integration by parts, that (3.4) holds for any $w \in L^2([0,\tau-\delta],\mathbb{R})$ for which there exist a partition $0 = t_0 < t_1 < \cdots < t_n = \tau-\delta$ and functions $w_j \in W^{1,1}([t_{j-1},t_j],\mathbb{R})$ such that $w(t) = w_j(t)$ for all $t \in (t_{j-1},t_j)$ and all $j = 1,2,\ldots,n$.

Proof of Lemma 3.2. Statement (1) follows from a routine application of the closed-graph theorem. To prove statement (2), note first that (3.6) follows immediately from the variation-of-parameters formula combined with the fact that the control u given by (3.1a) satisfies

$$u(t) = v(t) \quad \forall\, t \in [k\tau+\delta,(k+1)\tau). \tag{3.12}$$

To derive (3.7), we use (2.2b) and (3.12) to obtain

$$\begin{aligned} y(k\tau+\delta+s) = C_\Lambda\big(x(k\tau+\delta+s) - (s_0 I - A)^{-1}Bv(k\tau+\delta+s)\big) \\ + \mathbf{G}(s_0)v(k\tau+\delta+s) \quad \text{for a.a. } s \in [0,\tau-\delta]. \end{aligned} \tag{3.13}$$

It follows from the variation-of-parameters formula that the function $\tilde{x} : s \mapsto x(k\tau+\delta+s)$ is the state trajectory of (2.2) corresponding to the initial condition $\tilde{x}(0) = x(k\tau+\delta) = x_{k,\delta}$ and the control function $s \mapsto v(k\tau+\delta+s)$. By (3.13), the function $s \mapsto y(k\tau+\delta+s)$ is the corresponding output, and thus

$$y(k\tau+\delta+s) = C_\Lambda \mathbf{T}_s x_{k,\delta} + (G\mathbf{L}^\tau_{k\tau+\delta}v)(s) \quad \text{for a.a. } s \in [0,\tau-\delta].$$

Combining this with (3.1b) gives

$$y_{k+1} = \int_0^{\tau-\delta} w(s) \left(C_\Lambda \mathbf{T}_s x_{k,\delta} + (G\mathbf{L}^\tau_{k\tau+\delta} v)(s) \right) ds.$$

A standard argument involving the approximation of $x_{k,\delta}$ by elements in X_1, the admissibility of C and the boundedness of the operator L_w (see statement (1)) shows that

$$\int_0^{\tau-\delta} w(s) C_\Lambda \mathbf{T}_s x_{k,\delta}\, ds = C L_w x_{k,\delta}.$$

Hence, with M_w given by (3.10),

$$y_{k+1} = C L_w x_{k,\delta} + M_w \mathbf{L}^\tau_{k\tau+\delta} v,$$

which is (3.7). To prove (3.8), note that $k\tau + \delta + s \in [(k+1)\tau, (k+1)\tau + \delta]$ for all $s \in [\tau - \delta, \tau]$ and so, by (3.1a),

$$u(k\tau + \delta + s) = v(k\tau + \delta + s) - H(s + \delta - \tau) y_{k+1} \quad \forall\, s \in [\tau - \delta, \tau],\ \forall\, k \in \mathbb{N}_0.$$

Combining this with (3.12), we may conclude that

$$x_{k+1,\delta} = \mathbf{T}_\tau x_{k,\delta} + \int_0^\tau \mathbf{T}_{\tau-s} B v(k\tau + \delta + s)\, ds - \int_{\tau-\delta}^\tau \mathbf{T}_{\tau-s} B H(s + \delta - \tau) y_{k+1}\, ds.$$

Changing the integration variable s in the second integral to $\zeta = s + \delta - \tau$ gives

$$\begin{aligned} x_{k+1,\delta} &= \mathbf{T}_\tau x_{k,\delta} + \int_0^\tau \mathbf{T}_{\tau-s} B v(k\tau + \delta + s)\, ds - \int_0^\delta \mathbf{T}_{\delta-\zeta} B H(\zeta) y_{k+1}\, d\zeta \\ &= \mathbf{T}_\tau x_{k,\delta} + K_H y_{k+1} + \int_0^\tau \mathbf{T}_{\tau-s} B v(k\tau + \delta + s)\, ds \quad \forall\, k \in \mathbb{N}_0, \end{aligned}$$

where K_H is given by (3.9). Together with (3.7) and (3.11) this yields (3.8). □

The sampled-data feedback system given by (2.2) and (3.1) is called *exponentially bounded* if there exist constants $N \geq 1$ and $\nu \in \mathbb{R}$ such that

$$\|x(t; x^0, 0)\| \leq N e^{\nu t} \|x^0\| \quad \forall\, t \in \mathbb{R}_+,\ \forall\, x^0 \in X, \tag{3.14}$$

where $x(t; x^0, 0)$ is given by (3.3) (with $v = 0$). The number ν is called an *exponential bound* of the sampled-data feedback system. Obviously any bounded operator $\Delta \in \mathcal{B}(X)$ satisfies $\|\Delta^k\| \leq \|\Delta\|^k$; i.e., Δ is *power bounded*. If $q > 0$ is such that there exists $M \geq 1$ so that

$$\|\Delta^k\| \leq M q^k \quad \forall\, k \in \mathbb{N}_0, \tag{3.15}$$

then q is a *power bound* for Δ.

LEMMA 3.4. *Let $\tau > \delta > 0$, $H \in L^2([0, \delta], \mathbb{R}^{m \times p})$, and $w \in L^2([0, \tau - \delta], \mathbb{R})$. Let $L_w \in \mathcal{B}(X, X_1)$ and $K_H \in \mathcal{B}(\mathbb{R}^p, X)$ be given by* (3.5) *and* (3.9)*, respectively, and assume that* (3.4) *holds. Furthermore, let $\nu \in \mathbb{R}$. Then the following statements hold.*

(1) *If $e^{\nu\tau}$ is a power bound for the operator $\mathbf{T}_\tau + K_H C L_w$, then $\nu \in \mathbb{R}$ is an exponential bound for the sampled-data feedback system given by* (2.2) *and* (3.1).

(2) *Under the additional assumption that* $\mathbf{T}$ *is a group, the converse of statement* (1) *holds; that is, if* $\nu \in \mathbb{R}$ *is an exponential bound for the sampled-data feedback system given by* (2.2) *and* (3.1)*, then* $e^{\nu\tau}$ *is a power bound for* $\mathbf{T}_\tau + K_H C L_w$.

The lemma shows in particular that the sampled-data feedback system is exponentially bounded. We define the *exponential growth* $\omega_{\rm sd}$ of the sampled-data feedback system to be the infimum of all $\nu \in \mathbb{R}$ for which there exists $N \geq 1$ such that (3.14) holds. Note that $-\infty \leq \omega_{\rm sd} < \infty$. If $\omega_{\rm sd} < 0$, then we say that the sampled-data feedback system is *exponentially stable*. Similarly, the infimum of all $q > 0$ for which there exists $M \geq 1$ such that (3.15) holds is called the *power growth* of Δ. If the power growth is smaller than 1, we say that Δ is *power stable*. It follows from Gelfand's spectral radius formula

$$r(\Delta) = \lim_{k\to\infty} \|\Delta^k\|^{1/k}$$

that the power growth of Δ coincides with $r(\Delta)$. As a consequence, Lemma 3.4 has the following corollary.

COROLLARY 3.5. *Let* $\tau > \delta > 0$, $H \in L^2([0,\delta], \mathbb{R}^{m\times p})$, *and* $w \in L^2([0,\tau-\delta],\mathbb{R})$. *Let* $L_w \in \mathcal{B}(X, X_1)$ *and* $K_H \in \mathcal{B}(\mathbb{R}^p, X)$ *be given by* (3.5) *and* (3.9)*, respectively, and assume that* (3.4) *holds. Then* $r(\mathbf{T}_\tau + K_H C L_w) \geq e^{\omega_{\rm sd}\tau}$*; under the additional assumption that* $\mathbf{T}$ *is a group, we have* $r(\mathbf{T}_\tau + K_H C L_w) = e^{\omega_{\rm sd}\tau}$ (*we adopt the convention* $e^{-\infty\tau} := 0$).

Proof of Lemma 3.4. We define $\Delta \in \mathcal{B}(X)$ by setting

$$\Delta := \mathbf{T}_\tau + K_H C L_w.$$

To prove statement (1), let $\nu \in \mathbb{R}$ and assume that $e^{\nu\tau}$ is a power bound for Δ. By the variation-of-parameter formula we obtain for the state trajectory $x(\cdot; x^0, 0)$ of the sampled-data feedback system

$$x(k\tau + t; x^0, 0) = \mathbf{T}_t x_k - \int_0^t \mathbf{T}_{t-s} B H_\tau(s) y_k \, ds \quad \forall\, t \in [0,\tau),\ \forall\, k \in \mathbb{N}_0,$$

where H_τ is given by (3.2). Using (3.6) and (3.7), we obtain

$$x(k\tau+t; x^0, 0) = \mathbf{T}_{t+\tau-\delta} x_{k-1,\delta} - \int_0^t \mathbf{T}_{t-s} B H_\tau(s) C L_w x_{k-1,\delta}\, ds \quad \forall\, t \in [0,\tau),\ \forall\, k \in \mathbb{N}.$$

Invoking the admissibility of B, (3.8), and the hypothesis, we may conclude that there exist $N_1, N_2 \geq 0$ such that

$$\|x(k\tau + t; x^0, 0)\| \leq N_1 \|x_{k-1,\delta}\| \leq N_2 (e^{\nu\tau})^{k-1} \|x_{0,\delta}\| \quad \forall\, t \in [0,\tau),\ \forall\, k \in \mathbb{N}.$$

Noting that $x(t; x^0, 0) = \mathbf{T}_t x^0$ for all $t \in [0,\tau]$ and setting

$$N := \left(\max\{ \sup_{0\leq s\leq\tau} \|\mathbf{T}_s\|, N_2 \|\mathbf{T}_\delta\| e^{-\nu\tau} \} \right) \sup_{0\leq s\leq\tau} e^{-\nu s},$$

it follows that

$$\|x(k\tau + t; x^0, 0)\| \leq N e^{\nu(k\tau+t)} \|x^0\| \quad \forall\, t \in [0,\tau),\ \forall\, k \in \mathbb{N}_0.$$

This holds for all $x^0 \in X$, showing that ν is an exponential bound for the sampled-data feedback system.

To prove statement (2), assume that $\mathbf{T}$ is a group and let $\nu \in \mathbb{R}$ be an exponential bound for the sampled-data feedback system. Then there exists $N \geq 1$ such that (3.14) holds and therefore

$$\|x_{k,\delta}\| \leq Ne^{\nu(k\tau+\delta)}\|x^0\| = Ne^{\nu\delta}(e^{\nu\tau})^k\|x^0\| \quad \forall\, k \in \mathbb{N}_0.$$

Since $x_{0,\delta} = \mathbf{T}_\delta x^0$, it follows from (3.8) that $x_{k,\delta} = \Delta^k \mathbf{T}_\delta x^0$. Hence, using the group property of $\mathbf{T}$, we obtain

$$\|\Delta^k \mathbf{T}_\delta x^0\| \leq N\|\mathbf{T}_{-\delta}\|e^{\nu\delta}(e^{\nu\tau})^k\|\mathbf{T}_\delta x^0\| \quad \forall\, k \in \mathbb{N}_0.$$

Since this holds for all $x^0 \in X$, it follows that $e^{\nu\tau}$ is a power bound for Δ. ∎

4. Main result. We first state and prove a technical lemma.

LEMMA 4.1. *Let $S \in \mathbb{R}^{n\times n}$, $a > 0$, and $f \in L^1([0,a],\mathbb{R})$. The matrix $\int_0^a f(t)e^{St}\,dt$ is invertible if and only if $\int_0^a f(t)e^{\lambda t}\,dt \neq 0$ for all $\lambda \in \sigma(S)$.*

Proof. Using the Jordan form of S, it is easy to show that a complex number μ is an eigenvalue of the matrix $\int_0^a f(t)e^{St}dt$ if and only if $\mu = \int_0^a f(t)e^{\lambda t}\,dt$ for some $\lambda \in \sigma(S)$. ∎

In the following we shall impose a number of assumptions on the well-posed system (2.2), the weighting function w, and the sampling constants $\tau > \delta > 0$.

A1. There exists $\beta < 0$ such that $\sigma(A) \cap \overline{\mathbb{C}}_\beta$ consists of finitely many isolated eigenvalues of A with finite algebraic multiplicities.

If A1 holds, then there exists a simple closed curve Γ in the complex plane not intersecting $\sigma(A)$, enclosing $\sigma(A) \cap \overline{\mathbb{C}}_\beta$ in its interior and having $\sigma(A) \cap (\mathbb{C} \setminus \overline{\mathbb{C}}_\beta)$ in its exterior. The operator

$$\Pi := \frac{1}{2\pi i}\int_\Gamma (sI - A)^{-1}\,ds \tag{4.1}$$

is a projection operator, and we have

$$X = X^+ \oplus X^-, \qquad \text{where} \quad X^+ := \Pi X, \quad X^- := (I - \Pi)X. \tag{4.2}$$

It follows from a standard result (see, for example, Lemma 2.5.7 in [6]) that $\dim X^+ < \infty$, $X^+ \subset X_1$, X^+ and X^- are $\mathbf{T}_t$-invariant for all $t \geq 0$, and

$$\sigma(A|_{X^+}) = \sigma(A) \cap \overline{\mathbb{C}}_\beta, \qquad \sigma(A|_{X^-}) = \sigma(A) \cap (\mathbb{C} \setminus \overline{\mathbb{C}}_\beta).$$

It is useful to introduce the notation

$$A^+ := A|_{X^+}, \quad A^- := A|_{X_1 \cap X^-}, \qquad \mathbf{T}_t^+ := \mathbf{T}_t|_{X^+}, \quad \mathbf{T}_t^- := \mathbf{T}_t|_{X^-}. \tag{4.3}$$

Clearly, $\mathbf{T}_t^+$ is a semigroup on the finite-dimensional space X^+ with generator A^+, i.e., $\mathbf{T}_t^+ = e^{A^+ t}$, and $\mathbf{T}_t^-$ is a strongly continuous semigroup on X^- with generator A^-. Since the spectrum of A considered as an operator on X coincides with the spectrum of A considered as an operator on X_{-1}, the projection operator Π on X defined in (4.1) extends to a projection on X_{-1}. We will use the same symbol Π for the original projection and its associated extension. Obviously, the operator A^- extends to an operator in $\mathcal{B}(X^-, (X_{-1})^-)$, and the same symbol A^- will be used to denote this extension. The decomposition (4.2) induces decompositions of the control operator $B \in \mathcal{B}(\mathbb{R}^m, X_{-1})$ and the observation operator $C \in \mathcal{B}(X_1, \mathbb{R}^p)$:

$$B^+ := \Pi B, \quad B^- := (I - \Pi)B, \quad C^+ := C|_{X^+}, \quad C^- := C|_{X_1 \cap X^-}. \tag{4.4}$$

The following simple lemma will be useful in the proof of Theorem 4.4.

LEMMA 4.2. *Assume that* A1 *holds. There exists a well-posed system* Σ^- *with generating operators* (A^-, B^-, C^-) [2] *and input-output operator* $G^- := G - G^+$, *where* G^+ *denotes the input-output operator of the* (*finite-dimensional*) *system* (A^+, B^+, C^+), *that is,* $(G^+u)(t) = \int_0^t C^+ e^{A^+(t-s)} B^+ u(s)\, ds$ *for all* $t \in \mathbb{R}_+$ *and all* $u \in L^2_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^m)$. *Moreover, for any* $x^0 \in X$ *and* $u \in L^2_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^m)$, *the output* y *of the well-posed system* (2.2) *can be written in the form*

$$y(t) = (C^-)_\Lambda \mathbf{T}^-_t (I - \Pi) x^0 + (G^- u)(t) + C^+ \Pi x(t) \quad \text{for a.a. } t \in \mathbb{R}_+, \tag{4.5}$$

where $x(t) = \mathbf{T}_t x^0 + \int_0^t \mathbf{T}_{t-s} B u(s)\, ds$ *for all* $t \in \mathbb{R}_+$. *The* Λ*-extension of* C^- *satisfies*

$$(C^-)_\Lambda z = C_\Lambda z \quad \forall\, z \in \mathrm{dom}((C^-)_\Lambda) = \mathrm{dom}(C_\Lambda) \cap X^-. \tag{4.6}$$

Proof. It is trivial that the Λ-extension of C^- satisfies (4.6). The admissibility of B and C immediately implies that B^- and C^- are admissible control and observation operators for $\mathbf{T}^-$, respectively. Defining $\mathbf{G}^+(s) := C^+(sI - A^+)^{-1}B^+$, it follows from (2.1) that

$$\begin{aligned} &\frac{1}{s - s_0}(\mathbf{G}(s) - \mathbf{G}(s_0)) - \frac{1}{s - s_0}(\mathbf{G}^+(s) - \mathbf{G}^+(s_0)) = \\ &- C^-(sI - A^-)^{-1}(s_0 I - A^-)^{-1} B^- \quad \forall\, s, s_0 \in \mathbb{C}_{\omega(\mathbf{T})},\ s \neq s_0. \end{aligned}$$

Choosing $\alpha > \omega(\mathbf{T})$ and setting $\mathbf{G}^-(s) := \mathbf{G}(s) - \mathbf{G}^+(s)$ for all $s \in \mathbb{C}_\alpha$, it is clear that $\mathbf{G}^-$ is analytic and bounded on $\mathbb{C}_\alpha$ and $\mathbf{G}^-$ satisfies

$$\frac{1}{s - s_0}\big(\mathbf{G}^-(s) - \mathbf{G}^-(s_0)\big) = -C^-(sI - A^-)^{-1}(s_0 I - A^-)^{-1} B^- \ \ \forall\, s, s_0 \in \mathbb{C}_\alpha,\ s \neq s_0.$$

Invoking a result in [5], we may now conclude that there exists a well-posed system Σ^- with generating operators (A^-, B^-, C^-) and input-output operator G^- (or, equivalently, transfer function $\mathbf{G}^-$).[3] To prove (4.5), let $x^0 \in X$ and $u \in L^2_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{R}^m)$ and note that

$$\Pi \mathbf{T}_t x^0 \in X^+ \subset X_1 \subset \mathrm{dom}(C_\Lambda) \quad \forall\, t \in \mathbb{R}_+$$

and

$$(I - \Pi)\mathbf{T}_t x^0 = \mathbf{T}^-_t (I - \Pi) x^0 \in \mathrm{dom}(C_\Lambda) \cap X^- = \mathrm{dom}((C^-)_\Lambda) \quad \text{for a.a. } t \in \mathbb{R}_+.$$

Thus, by (4.6), we may write the output $y = C_\Lambda \mathbf{T} x^0 + Gu$ in the form

$$y = (C^-)_\Lambda \mathbf{T}^- (I - \Pi) x^0 + G^- u + C^+ \mathbf{T}^+ \Pi x^0 + G^+ u. \tag{4.7}$$

[2] For (A^-, B^-, C^-) to be the generating operators of a well-posed system it is of course necessary that B^- maps into $(X^-)_{-1} = ((I - \Pi)X)_{-1}$, the extrapolation space associated with A^-. Since, by definition, B^- maps into $(I - \Pi)X_{-1} =: (X_{-1})^-$, there seems to be a difficulty. However, it is clear that the spaces $(X^-)_{-1}$ and $(X_{-1})^-$ are both completions of X^- endowed with the norm $\|\cdot\|_{-1}$. Hence there exists an isometric isomorphism $(X^-)_{-1} \to (X_{-1})^-$ whose restriction to X^- is the identity, and so we can safely identify $(X^-)_{-1}$ and $(X_{-1})^-$.

[3] Alternatively, the claim that there exists a well-posed system Σ^- with generating operators (A^-, B^-, C^-) and input-output operator G^- can be proved by direct verification of the defining properties of a well-posed system as given in, for example, [25, 27, 31].

With x given by $x(t) = \mathbf{T}_t x^0 + \int_0^t \mathbf{T}_{t-s} B u(s)\, ds$, it is clear that Πx is the state trajectory of the finite-dimensional system given by (A^+, B^+, C^+) corresponding to the initial state Πx^0 and the input function u. Therefore, $C^+\mathbf{T}^+\Pi x^0 + G^+ u = C^+\Pi x$, and (4.5) follows from (4.7). □

We recall that the linear bounded map

$$R_{t_0} : L^2([0,t_0],\mathbb{R}^m) \to X, \;\; f \mapsto \int_0^{t_0} \mathbf{T}_{t_0-s} B f(s)\, ds \tag{4.8}$$

is called the reachability operator of the well-posed system (2.2) at time t_0.

We assume, in addition to A1, that the following conditions are satisfied. Let $t_0 > 0$ be fixed and assume that $\tau > \delta \geq t_0$.

A2. The semigroup $\mathbf{T}^-$ is exponentially stable; that is, $\omega(\mathbf{T}^-) < 0$.

A3. The pair $(C^+, \mathbf{T}^+_\tau)$ is observable.

A4. The constants τ and δ and the function $w \in L^2([0,\tau-\delta],\mathbb{R})$ are such that (3.4) holds and

$$\int_0^{\tau-\delta} w(s) e^{\lambda s}\, ds \neq 0 \quad \forall\, \lambda \in \sigma(A^+). \tag{4.9}$$

A5. $\overline{\operatorname{im}}\, R_{t_0} \supset X^+$.

Remark 4.3. Of course, A2 holds if the generator A^- satisfies the spectrum-determined-growth assumption. Trivially, for A5 to hold, it is sufficient that the well-posed system (2.2) is approximately controllable in time t_0. If the function w is a nonzero constant, then it is clear that (4.9) holds if and only if

$$(\tau-\delta)\lambda \neq 2\pi i k \quad \forall\, \lambda \in \sigma(A^+),\; \forall\, k \in \mathbb{Z}\setminus\{0\}.$$

The observability condition A3 is implied by observability of the pair (C^+, A^+) and the nonpathological sampling assumption

$$\tau(\lambda-\mu) \neq 2\pi i k \quad \forall\, \lambda,\mu \in \sigma(A^+),\; \forall\, k \in \mathbb{Z}\setminus\{0\}. \tag{4.10}$$

We do not want to focus here on the issue of pathological sampling and instead refer the reader to Proposition 6.2.11 in [24] for more on this. We note that conditions (4.10) and (4.9) are "generically" satisfied in the following sense: the set of all $\tau > t_0$ for which (4.10) holds is open and dense in (t_0,∞), and, for given $\tau > \delta \geq t_0$, the set of all $w \in L^2([0,\tau-\delta],\mathbb{R})$ for which (4.9) holds is open and dense in $L^2([0,\tau-\delta],\mathbb{R})$.

The control function u generated by the sampled-data control law (3.1) depends on the initial value $x^0 \in X$ and the input function $v \in L^2_{\mathrm{loc}}(\mathbb{R}_+,\mathbb{R}^m)$. We express this dependence by writing $u = u(\cdot; x^0, v)$. It is natural to define the output $y(\cdot; x^0, v)$ of the sampled-data feedback system given by (2.2) and (3.1) to be the output of (2.2) corresponding to the initial condition x^0 and the control $u(\cdot; x^0, v)$. We are now in the position to formulate the main result of this paper.

THEOREM 4.4. *Assume that* A1–A5 *are satisfied. For every* $\varepsilon \in (0, -\omega(\mathbf{T}^-))$ *there exists* $H \in L^2([0,\delta],\mathbb{R}^{m\times p})$ *such that the following statements hold.*

(1) *The sampled-data feedback system given by* (2.2) *and* (3.1) *is exponentially stable with exponential growth* $\omega_{\mathrm{sd}} < \omega(\mathbf{T}^-) + \varepsilon < 0$.

(2) *For every* $\alpha \in [\omega(\mathbf{T}^-)+\varepsilon, 0]$ *there exists* $N \geq 1$ *such that*

$$\|y(\cdot; x^0, v)\|_{L^2_\alpha} \leq N(\|x^0\| + \|v\|_{L^2_\alpha}) \quad \forall\, x^0 \in X,\;\; \forall\, v \in L^2_\alpha(\mathbb{R}_+,\mathbb{R}^m).$$

(3) *If* $\alpha \in [\omega(\mathbf{T}^-)+\varepsilon, 0]$, $x^0 \in X$, *and* $v \in W^{1,2}_\alpha(\mathbb{R}_+, \mathbb{R}^m)$ *and there exists* $t_1 \in \mathbb{R}_+$ *such that* $\mathbf{T}_{t_1}(Ax^0 + Bv(0)) \in X$, *then* $y(\cdot; x^0, v)$ *is continuous on* $[t_1, \infty)$ *and*

$$\lim_{t\to\infty} y(t; x^0, v)e^{-\alpha t} = 0.$$

Statement (2) shows in particular that there exists $H \in L^2([0,\delta], \mathbb{R}^{m\times p})$ such that the sampled-data feedback system given by (2.2) and (3.1) is L^2_α-input-output stable.

Proof of Theorem 4.4. We define $\Delta_H \in \mathcal{B}(X)$ by setting

$$\Delta_H := \mathbf{T}_\tau + K_H C L_w, \tag{4.11}$$

where the operators $L_w \in \mathcal{B}(X, X_1)$ and $K_H \in \mathcal{B}(\mathbb{R}^p, X)$ are given by (3.5) and (3.9), respectively. It is convenient to set $\omega^- := \omega(\mathbf{T}^-)$. Let $\varepsilon \in (0, -\omega^-)$.

(1) To prove that for a suitable hold function H, $\omega_{\mathrm{sd}} < \omega^- + \varepsilon$, we note that, by Corollary 3.5, it is sufficient to show the existence of a function $H \in L^2([0,\delta], \mathbb{R}^{m\times p})$ such that $r(\Delta_H) < e^{(\omega^-+\varepsilon)\tau}$. Defining the operators

$$K_H^+ := \Pi K_H, \quad K_H^- := (I-\Pi)K_H, \quad L_w^\pm := L_w|_{X^\pm} = \int_0^{\tau-\delta} w(s)\mathbf{T}_s^\pm \, ds, \tag{4.12}$$

we have $K_H^\pm \in \mathcal{B}(\mathbb{R}^p, X^\pm)$, $L_w^+ \in \mathcal{B}(X^+)$, and $L_w^- \in \mathcal{B}(X^-, X_1 \cap X^-)$, where $X_1 \cap X^-$ is endowed with the norm $\|\cdot\|_1$. The operator Δ_H can then be written in the form

$$\Delta_H = \begin{pmatrix} \mathbf{T}_\tau^+ + K_H^+ C^+ L_w^+ & K_H^+ C^- L_w^- \\ K_H^- C^+ L_w^+ & \mathbf{T}_\tau^- + K_H^- C^- L_w^- \end{pmatrix}. \tag{4.13}$$

By A4, $\int_0^{\tau-\delta} w(s)e^{\lambda s}\, ds \neq 0$ for all $\lambda \in \sigma(A^+)$ and hence an application of Lemma 4.1 shows that the matrix $L_w^+ = \int_0^{\tau-\delta} w(s)e^{A^+ s}\, ds$ is invertible. Since L_w^+ and $\mathbf{T}_\tau^+ = e^{A^+\tau}$ commute, we have that

$$(C^+L_w^+, (L_w^+)^{-1}\mathbf{T}_\tau^+ L_w^+) = (C^+L_w^+, \mathbf{T}_\tau^+). \tag{4.14}$$

Using A3, i.e., observability of the pair $(C^+, \mathbf{T}_\tau^+)$, it follows that the pair $(C^+L_w^+, \mathbf{T}_\tau^+)$ is observable. Hence, by the pole-placement theorem for finite-dimensional systems, there exists $Q \in \mathcal{B}(\mathbb{R}^p, X^+)$ such that

$$\sigma(\mathbf{T}_\tau^+ + QC^+L_w^+) = \{0\}. \tag{4.15}$$

Denoting the canonical basis of $\mathbb{R}^p$ by $(e_1, e_2, \dots, e_p)$, it follows from the fact that $\delta \geq t_0$ (see A4) combined with assumption A5 that for every $\eta > 0$, there exist $h_1, h_2, \dots, h_p \in L^2([0,\delta], \mathbb{R}^m)$ such that

$$\sum_{j=1}^p \|R_\delta h_j - Qe_j\|^2 \leq \eta^2. \tag{4.16}$$

Setting $H := -(h_1, h_2, \dots, h_p) \in L^2([0,\delta], \mathbb{R}^{m\times p})$, it follows that

$$R_\delta h_j = K_H e_j \quad \forall\, j \in \{1, 2, \dots, p\}.$$

Therefore, invoking (4.16), we obtain that for all $z = (z_1, z_2, \dots, z_p)^T \in \mathbb{R}^p$,

$$\|K_H z - Qz\| = \left\| \sum_{j=1}^{p} z_j (K_H e_j - Q e_j) \right\| \le \sum_{j=1}^{p} \|K_H e_j - Q e_j\| \, |z_j|$$
$$\le \left(\sum_{j=1}^{p} \|K_H e_j - Q e_j\|^2 \right)^{1/2} \|z\| \le \eta \|z\|.$$

Thus, $\|K_H - Q\| \le \eta$, and so, since Q maps into X^+,

$$\|K_H^+ - Q\| = \|\Pi(K_H - Q)\| \le \|\Pi\| \eta, \tag{4.17a}$$
$$\|K_H^-\| = \|(I - \Pi)(K_H - Q)\| \le \|I - \Pi\| \eta. \tag{4.17b}$$

Using (4.13), we may write

$$\Delta_H = \begin{pmatrix} \mathbf{T}_\tau^+ + QC^+L_w^+ & QC^-L_w^- \\ 0 & \mathbf{T}_\tau^- \end{pmatrix} + \begin{pmatrix} (K_H^+ - Q)C^+L_w^+ & (K_H^+ - Q)C^-L_w^- \\ K_H^- C^+ L_w^+ & K_H^- C^- L_w^- \end{pmatrix}. \tag{4.18}$$

We denote the first operator on the right-hand side of (4.18) by Δ and the second by P_H. Obviously, by (4.15), $r(\Delta) = e^{\omega^- \tau}$. By upper semicontinuity of the spectrum (see [11], pp. 208), there exists $\gamma > 0$ such that

$$r(\Delta_H) = r(\Delta + P_H) < e^{(\omega^- + \varepsilon)\tau}, \tag{4.19}$$

provided that $\|P_H\| \le \gamma$. It follows from (4.17) that the latter can be accomplished by choosing $\eta > 0$ sufficiently small.

(2) To prove statement (2) of the theorem, choose $H \in L^2([0, \delta], \mathbb{R}^{m \times p})$ such that (4.19) holds. Choose $\nu \in (\omega^-, \omega^- + \varepsilon)$ such that $e^{\nu\tau}$ is a power bound for Δ_H. Let $x^0 \in X$, $\alpha \in (\nu, 0]$, and $v \in L^2_\alpha(\mathbb{R}_+, \mathbb{R}^m)$. Recall that the feedback control produced by the sampled-data control law (3.1) is denoted by $u(\cdot; x^0, v)$. With H_τ defined by (3.2) we have

$$u(t; x^0, v)e^{-\alpha t} = e^{-\alpha t} v(t) - e^{-\alpha(t - k\tau)} H_\tau(t - k\tau) y_k e^{-\alpha k\tau} \quad \forall\, t \in [k\tau, (k+1)\tau),\ \forall\, k \in \mathbb{N}_0.$$

In the following, the numbers $N_i > 0$ are suitable constants, depending only on α but not on x^0 and v. It follows from the above identity that

$$\int_{k\tau}^{(k+1)\tau} \|u(t; x^0, v)e^{-\alpha t}\|^2 dt \le N_1 \left(\|y_k e^{-\alpha k\tau}\|^2 + \int_{k\tau}^{(k+1)\tau} \|v(t)e^{-\alpha t}\|^2 dt \right) \quad \forall\, k \in \mathbb{N}_0. \tag{4.20}$$

Using that $e^{\nu\tau}$ is a power bound for Δ_H and that $0 \ge \alpha > \nu$, we may conclude from (3.7), (3.8), and (4.11) that

$$\sum_{k=0}^{\infty} \|x_{k,\delta} e^{-\alpha k\tau}\|^2 \le N_2 \left(\|x_{0,\delta}\|^2 + \int_0^\infty \|v(t)e^{-\alpha t}\|^2 dt \right) \tag{4.21}$$

and

$$\sum_{k=0}^{\infty} \|y_k e^{-\alpha k\tau}\|^2 \le N_3 \left(\|x_{0,\delta}\|^2 + \int_0^{\infty} \|v(t)e^{-\alpha t}\|^2 dt \right). \tag{4.22}$$

Now $y_0 = 0$, and so $u(t) = v(t)$ for all $t \in [0, \tau)$. Hence,

$$x(t; x^0, v) = \mathbf{T}_t x^0 + \int_0^t \mathbf{T}_{t-s} B v(s)\, ds \quad \forall\, t \in [0, \tau), \tag{4.23}$$

showing that

$$\|x_{0,\delta}\| = \|x(\delta; x^0, v)\| \le N_4(\|x^0\| + \|v\|_{L^2}) \le N_4(\|x^0\| + \|v\|_{L^2_\alpha}). \tag{4.24}$$

Inserting this into (4.21) and (4.22) yields

$$\sum_{k=0}^{\infty} \|x_{k,\delta} e^{-\alpha k\tau}\|^2 \le N_5(\|x^0\|^2 + \|v\|^2_{L^2_\alpha}) \tag{4.25}$$

and

$$\sum_{k=0}^{\infty} \|y_k e^{-\alpha k\tau}\|^2 \le N_6(\|x^0\|^2 + \|v\|^2_{L^2_\alpha}). \tag{4.26}$$

It follows from (4.20) and (4.26) that

$$\|u(\cdot; x^0, v)\|_{L^2_\alpha} \le N_7(\|x^0\| + \|v\|_{L^2_\alpha}). \tag{4.27}$$

To derive a similar estimate for $x(\cdot; x^0, v)$, we note that by the variations-of-parameter formula we have, for $k \in \mathbb{N}$ and $t \in [0, \tau)$,

$$\begin{aligned} x(k\tau + t; x^0, v) = \mathbf{T}_{t+\tau-\delta} x_{k-1,\delta} &- \int_{k\tau}^{k\tau+t} \mathbf{T}_{k\tau+t-s} B H_\tau(s - k\tau) y_k\, ds \\ &+ \int_{(k-1)\tau+\delta}^{k\tau+t} \mathbf{T}_{k\tau+t-s} B v(s)\, ds, \end{aligned}$$

where H_τ is defined in (3.2). A change of variables leads to

$$x(k\tau + t; x^0, v) = \mathbf{T}_{t+\tau-\delta} x_{k-1,\delta} - \int_0^t \mathbf{T}_{t-s} B H_\tau(s) y_k\, ds + \int_{\delta-\tau}^t \mathbf{T}_{t-s} B v(k\tau + s)\, ds.$$

Hence,

$$\begin{aligned} \|x(k\tau + t; x^0, v) e^{-\alpha(k\tau+t)}\|^2 \le N_8 \Bigg(&\|x_{k-1,\delta} e^{-\alpha(k-1)\tau}\|^2 + \|y_k e^{-\alpha k\tau}\|^2 \\ &+ \int_{(k-1)\tau}^{(k+1)\tau} \|v(s)e^{-\alpha s}\|^2 ds \Bigg) \ \forall\, k \in \mathbb{N},\ \forall\, t \in [0, \tau), \end{aligned} \tag{4.28}$$

and so,

$$\int_{k\tau}^{(k+1)\tau} \|x(t;x^0,v)e^{-\alpha t}\|^2 dt \le N_9\Bigg(\|x_{k-1,\delta}e^{-\alpha(k-1)\tau}\|^2 + \|y_k e^{-\alpha k\tau}\|^2 + \int_{(k-1)\tau}^{(k+1)\tau} \|v(s)e^{-\alpha s}\|^2 ds\Bigg) \quad \forall\, k \in \mathbb{N}. \tag{4.29}$$

Combining this with (4.23), (4.25), and (4.26) shows that

$$\|x(\cdot;x^0,v)\|_{L^2_\alpha} \le N_{10}(\|x^0\| + \|v\|_{L^2_\alpha}). \tag{4.30}$$

Using that $\alpha > \nu > \omega^-$, we have that the weighted semigroup $t \mapsto \mathbf{T}^-_t e^{-\alpha t}$ is exponentially stable and $G^- \in \mathcal{B}(L^2_\alpha(\mathbb{R}_+,\mathbb{R}^m), L^2_\alpha(\mathbb{R}_+,\mathbb{R}^p))$. Combining this with (4.27) and (4.30), an application of (4.5) (with $u = u(\cdot;x^0,v)$, $x = x(\cdot;x^0,v)$, and $y = y(\cdot;x^0,v)$) yields the claim.

(3) Since the space of all $W^{1,2}_c([0,\delta],\mathbb{R}^{m\times p})$ is dense in $L^2([0,\delta],\mathbb{R}^{m\times p})$, an inspection of the proof of statement (1) shows that there exists $H \in W^{1,2}_c([0,\delta],\mathbb{R}^{m\times p})$ such that (4.19) holds. Choose $\nu \in (\omega^-, \omega^- + \varepsilon)$ such that $e^{\nu\tau}$ is a power bound for Δ_H. Fix $\alpha \in (\nu, 0]$. Let $x^0 \in X$ and $v \in W^{1,2}_\alpha(\mathbb{R}_+,\mathbb{R}^m)$ be such that $Ax^0 + Bv(0) \in X$. It follows from (3.1a) and (4.22) that $u(\cdot;x^0,v) \in W^{1,2}_\alpha(\mathbb{R}_+,\mathbb{R}^m)$. Denoting the output of the well-posed system Σ^- corresponding to the initial value $(I-\Pi)x^0$ and the control $u(\cdot;x^0,v)$ by y^-, we have that

$$y^- = (C^-)_\Lambda \mathbf{T}^-(I-\Pi)x^0 + G^- u(\cdot;x^0,v). \tag{4.31}$$

Since $u(0;x^0,v) = v(0)$, we may conclude that

$$\mathbf{T}^-_{t_1}\big(A^-(I-\Pi)x^0 + B^- u(0;x^0,v)\big) = (I-\Pi)\mathbf{T}_{t_1}\big(Ax^0 + Bv(0)\big) \in (I-\Pi)X = X^-.$$

An application of Proposition 2.1 to Σ^- now yields that y^- is continuous on $[t_1,\infty)$ and

$$\lim_{t\to\infty} \|y^-(t)e^{-\alpha t}\| = 0. \tag{4.32}$$

Since $v \in L^2_\alpha(\mathbb{R}_+,\mathbb{R}^m)$, it is clear that $\int_{(k-1)\tau}^{(k+1)\tau} \|v(s)e^{-\alpha s}\|^2 ds$ converges to 0 as $k\to\infty$. Furthermore, it follows from (4.25) and (4.26) that $x_{k,\delta}e^{-\alpha k\tau}$ and $y_k e^{-\alpha k\tau}$ converge to 0 as $k\to\infty$. Consequently, the right-hand side of (4.28) converges to 0 as $k\to\infty$ and therefore,

$$\lim_{t\to\infty} \|x(t)e^{-\alpha t}\| = 0. \tag{4.33}$$

Finally, by (4.31) and Lemma 4.2 (applied to the well-posed system (2.2) with control $u = u(\cdot;x^0,v)$),

$$y(\cdot;x^0,v) = y^-(t) + C^+\Pi x(\cdot;x^0,v).$$

Therefore, $y(\cdot;x^0,v)$ is continuous on $[t_1,\infty)$, and, furthermore, we may conclude from (4.32) and (4.33) that $\lim_{t\to\infty} y(t;x^0,v)e^{-\alpha t} = 0$. ∎

Remark 4.5. (1) If in Theorem 4.4 assumption A5 is replaced by the stronger assumption that $\operatorname{im} R_{t_0} \supset X^+$ (that is, every state in X^+ is reachable from 0 in time

t_0), then an inspection of the above proof shows that there exists $H \in L^2([0,\delta],\mathbb{R}^{m\times p})$ such that (i) $\omega_{\rm sd} \le \omega(\mathbf{T}^-)$ and (ii) the conclusions of statements (2) and (3) of Theorem 4.4 remain true for every $\alpha \in (\omega(\mathbf{T}^-),0]$.

(2) From a practical point of view, it is important that the "structure" of the stabilizing hold function H (the existence of which is guaranteed by Theorem 4.4) is as simple as possible. In this context, we define $S([0,\delta],\mathbb{R}^{m\times p})$ to be the space of $\mathbb{R}^{m\times p}$-valued step functions on $[0,\delta]$ and $CPL_c([0,\delta],\mathbb{R}^{m\times p})$ to be the space of $\mathbb{R}^{m\times p}$-valued continuous piecewise affine-linear functions with support contained in the open interval $(0,\delta)$. We recall that $S([0,\delta],\mathbb{R}^{m\times p})$ and $CPL_c([0,\delta],\mathbb{R}^{m\times p})$ are dense in $L^2([0,\delta],\mathbb{R}^{m\times p})$. Moreover, it is clear that $CPL_c([0,\delta],\mathbb{R}^{m\times p}) \subset W_c^{1,2}([0,\delta],\mathbb{R}^{m\times p})$. Therefore an inspection of the proof of Theorem 4.4 shows that, for every $\varepsilon \in (0,-\omega(\mathbf{T}^-)$, there exist

(i) $H \in S([0,\delta],\mathbb{R}^{m\times p})$ such that statements (1) and (2) of Theorem 4.4 hold;

(ii) $H \in CPL_c([0,\delta],\mathbb{R}^{m\times p})$ such that statements (1)–(3) of Theorem 4.4 hold.

It follows from [18, 30] that assumptions A1 and A2 are necessary conditions for the stabilization of (2.2) by any of the commonly used sampled-data feedback designs including the control law (3.1) (see [18, 30]). In this context the following proposition is of interest.

PROPOSITION 4.6. *Let $\tau > \delta > 0$, $H \in L^2([0,\delta],\mathbb{R}^{m\times p})$, and $w \in L^2([0,\tau-\delta],\mathbb{R})$. Assume that* (3.4) *holds. If the sampled-data feedback system given by* (2.2) *and* (3.1) *is exponentially stable, then conditions* A1–A4 *hold, and if the semigroup $\mathbf{T}$ is analytic, then* A5 *holds also.*

Proof. Assume that the sampled-data feedback system given by (2.2) and (3.1) is exponentially stable. It follows from [30] that A1 and A2 hold. We claim that the pair $(C^+L_w^+,\mathbf{T}_\tau^+)$ is observable. Suppose not; then we can find $z \in X^+$, $z \neq 0$, and $\zeta \in \mathbb{C}$ with $|\zeta| \ge 1$ so that

$$\mathbf{T}_\tau^+ z = \zeta z \qquad \text{and} \qquad C^+L_w^+ z = 0.$$

Now choose $z^0 \in X^+$ such that $z = \mathbf{T}_\delta^+ z^0$. We consider the state trajectory $x(\cdot;x^0,0)$ of the sampled-data feedback system corresponding to the initial state

$$x^0 := \begin{pmatrix} z^0 \\ 0 \end{pmatrix}$$

and the external input function $v = 0$. Then, using (4.13),

$$x(k\tau+\delta;x^0,0) = x_{k,\delta} = \Delta_H^k x_{0,\delta} = \Delta_H^k \mathbf{T}_\delta x^0 = \Delta_H^k \begin{pmatrix} \mathbf{T}_\delta^+ z^0 \\ 0 \end{pmatrix} = \Delta_H^k \begin{pmatrix} z \\ 0 \end{pmatrix} = \zeta^k \begin{pmatrix} z \\ 0 \end{pmatrix}.$$

Since $z \neq 0$, we may conclude that $x(k\tau+\delta;x^0,0)$ does not converge to 0 as $k \to \infty$, yielding a contradiction to the exponential stability of the sampled-data feedback system. Hence the pair $(C^+L_w^+,\mathbf{T}_\tau^+)$ is observable. To show that A3 and A4 hold, let $\mathcal{O}_L$ and $\mathcal{O}$ be the observability matrices for the pairs $(C^+L_w^+,\mathbf{T}_\tau^+)$ and $(C^+,\mathbf{T}_\tau^+)$, respectively. Since L_w^+ and $\mathbf{T}_\tau^+ = e^{A^+\tau}$ commute, it follows that

$$\mathcal{O}_L = \mathcal{O}\,L_w^+.$$

If (4.9) fails to hold, then, by Lemma 4.1, L_w^+ is singular, implying that $\mathcal{O}_L$ loses rank. If A3 fails to hold, then $(C^+,\mathbf{T}_\tau^+)$ is not observable and again $\mathcal{O}_L$ will lose rank. In both cases $(C^+L_w^+,\mathbf{T}_\tau^+)$ will not be observable, which is impossible. Therefore both A3 and A4 must hold.

To complete the proof we just need to show that A5 also holds if $\mathbf{T}$ is analytic. Define the operator $B_\tau^+ : \mathbb{R}^p \to X^+$ by

$$B_\tau^+ z = \int_0^\tau \mathbf{T}_{\tau-s}^+ B^+ H_\tau(s) z \, ds \quad \forall\, z \in \mathbb{R}^p,$$

where H_τ is defined in (3.2). It follows from [30] that the pair $(\mathbf{T}_\tau^+, B_\tau^+)$ is controllable. A routine argument based on the Hautus criterion for controllability then shows that the pair (A^+, B^+) is also controllable. Finally, an application of Proposition 1.2 in [19] yields that condition A5 is satisfied. □

5. Example. We will illustrate Theorem 4.4 with a standard model for an Euler–Bernoulli beam with structural damping (see Chen and Russell [3]). Let $z(\xi, t)$ be the lateral deflection of a beam, where $\xi \in [0, 1]$ and $t > 0$ denote space and time, respectively. We assume that the flexural rigidity EI and the mass density per unit length m are both constant. We normalize so that $EI/m = 1$. The Euler–Bernoulli beam with structural damping is described by the following fourth-order partial differential equation

$$z_{tt}(\xi, t) - 2\gamma z_{t\xi\xi}(\xi, t) + z_{\xi\xi\xi\xi}(\xi, t) = 0, \tag{5.1}$$

where $\gamma \in (0, 1)$ denotes the damping constant. We assume that the beam is hinged at $\xi = 0$ and has a freely sliding clamped end at $\xi = 1$, with shear (also known as lateral) force $u(t)$ at $\xi = 1$:

$$z(0, t) = 0, \qquad z_{\xi\xi}(0, t) = 0, \tag{5.2a}$$

$$z_\xi(1, t) = 0, - z_{\xi\xi\xi}(1, t) = u(t). \tag{5.2b}$$

For this system we consider a standard observation, the velocity at $\xi = 1$:

$$y(t) = z_t(1, t). \tag{5.3}$$

The applicability of our considerations below to other boundary conditions is briefly discussed in Remark 5.1 at the end of this section.

Our first aim is to represent the controlled and observed partial differential equation given by (5.1)–(5.3) as an abstract well-posed system of the form (2.2). We write $L^2(0, 1)$ and $W^{q,2}(0, 1)$, respectively, in place of the more cumbersome $L^2([0, 1], \mathbb{R})$ and $W^{q,2}([0, 1], \mathbb{R})$. Let $A_0 : \mathrm{dom}(A_0) \subset L^2(0, 1) \to L^2(0, 1)$ be given by

$$\begin{aligned} &A_0 f = d^4 f / d\xi^4, \\ &\mathrm{dom}(A_0) = \{f \in W^{4,2}(0, 1) : f(0) = 0,\ f''(0) = 0,\ f'(1) = 0,\ f'''(1) = 0\}. \end{aligned}$$

The operator A_0 is closed, bijective, self-adjoint, and coercive and has compact resolvent. The numbers $(-\pi/2 + \pi k)^4$, where $k \in \mathbb{N}$, are the eigenvalues of A_0 with associated eigenvectors e_k given by

$$e_k(\xi) = \sqrt{2} \sin((-\pi/2 + \pi k)\xi), \quad k \in \mathbb{N}.$$

The family $(e_k)_{k \in N}$ forms an orthonormal basis of $L^2(0, 1)$. Moreover,

$$A_0^{1/2} f = -f'', \quad \mathrm{dom}(A_0^{1/2}) = \{f \in W^{2,2}(0, 1) : f(0) = 0,\ f'(1) = 0\}.$$

Let $X := \mathrm{dom}(A_0^{1/2}) \times L^2(0,1)$. Endowed with the inner product

$$\langle (x_1, x_2)^T, (y_1, y_2)^T \rangle := \langle A_0^{1/2} x_1, A_0^{1/2} y_1 \rangle_{L^2} + \langle x_2, y_2 \rangle_{L^2},$$

X becomes a Hilbert space. Defining the operator

$$A = \begin{pmatrix} 0 & I \\ -A_0 & -2\gamma A_0^{1/2} \end{pmatrix}, \quad \mathrm{dom}(A) = \mathrm{dom}(A_0) \times \mathrm{dom}(A_0^{1/2}), \tag{5.4}$$

(5.1) and (5.2) (with $u = 0$) can be written in the form $\dot{x} = Ax$, where $x(t) = (z(\cdot, t), z_t(\cdot, t))^T$. The eigenvalues of A are given by

$$\lambda_{\pm k} = (-\gamma \pm i\sqrt{1-\gamma^2})(-\pi/2 + \pi k)^2, \quad k \in \mathbb{N}, \tag{5.5}$$

with associated eigenvectors

$$f_{\pm k} = \frac{\sqrt{2}}{1 - e^{\mp 2i\varphi}} \begin{pmatrix} e_k/\lambda_{\pm k} \\ e_k \end{pmatrix}, \quad k \in \mathbb{N},$$

where $\varphi := \arccos(-\gamma)$, so that $e^{i\varphi} = -\gamma + i\sqrt{1-\gamma^2}$. It is a routine exercise to check that $(f_{\pm k})_{k \in \mathbb{N}}$ is a Riesz basis for X. For $k \in \mathbb{N}$, the unit vectors

$$g_{\pm k} = \frac{1}{\sqrt{2}} \begin{pmatrix} -e_k/\lambda_{\mp k} \\ e_k \end{pmatrix} \in \mathrm{dom}(A^*)$$

are eigenvectors of A^* with associated eigenvalues $\bar{\lambda}_{\pm k} = \lambda_{\mp k}$. Furthermore, introducing the set $\mathbb{Z}^* := \mathbb{Z} \setminus \{0\}$, we have that

$$\langle f_j, g_l \rangle = \begin{cases} 0, & j \neq l, \\ 1, & j = l, \end{cases}$$

i.e., $(f_j)_{j \in \mathbb{Z}^*}$ and $(g_j)_{j \in \mathbb{Z}^*}$ are biorthogonal. Consequently, A is a Riesz spectral operator (as defined in [6]) and thus can be represented in the form

$$Ax = \sum_{j \in \mathbb{Z}^*} \lambda_j \langle x, g_j \rangle f_j \quad \forall\, x \in \mathrm{dom}(A) = \left\{ x \in X : \sum_{j \in \mathbb{Z}^*} |\lambda_j|^2 |\langle x, g_j \rangle|^2 < \infty \right\};$$

moreover, $\sigma(A) = \{\lambda_j : j \in \mathbb{Z}^*\}$ and A generates a strongly continuous semigroup $\mathbf{T}$ given by

$$\mathbf{T}_t x = \sum_{j \in \mathbb{Z}^*} e^{\lambda_j t} \langle x, g_j \rangle f_j \quad \forall\, x \in X;$$

see, e.g., Theorem 2.3.5 in [6]. It follows from the location of $\sigma(A)$ combined with a standard result in semigroup theory (see, e.g., Theorem 5.2 in [16, p. 61]) that the semigroup $\mathbf{T}$ is analytic.

To write the controlled partial differential equation given by (5.1) and (5.2) in the abstract form (2.2a), we need to determine the input operator B. Moreover, in order to prove admissibility of B, we need to expand B in terms of the functions f_j. To this end it is useful to recall that the inner product on X has a continuous extension to $X_{-1} \times \mathrm{dom}(A^*)$, where $\mathrm{dom}(A^*)$ is endowed with the graph norm of

A^*. More precisely, there exists a bounded nondegenerate sesquilinear form $[\cdot,\cdot]$ on $X_{-1}\times\mathrm{dom}(A^*)$ such that $[x_1,x_2]=\langle x_1,x_2\rangle$ for all $(x_1,x_2)\in X\times\mathrm{dom}(A^*)$. The space X_{-1} may be identified with the dual of $\mathrm{dom}(A^*)$. Following the procedure outlined in [8], we obtain that

$$B=(0,\delta_1)^T, \tag{5.6}$$

where δ_1 denotes the Dirac distribution (or unit mass) with support at $\xi=1$.[4] Consequently, the controlled partial differential equation given by (5.1) and (5.2) can be written in the form (2.2a) with $x(t)=(z(\cdot,t),z_t(\cdot,t))^T$ and the operators A and B given by (5.4) and (5.6), respectively.

In order to verify that B is admissible, we first note that $(f_j)_{j\in\mathbb{Z}^*}$ is a Schauder basis of X_{-1}. Indeed, for arbitrary $x\in X_{-1}$, we have that

$$x=AA^{-1}x=A\sum_{j\in\mathbb{Z}^*}\langle A^{-1}x,g_j\rangle f_j=\sum_{j\in\mathbb{Z}^*}\langle A^{-1}x,g_j\rangle\lambda_j f_j,$$

and it is clear that the coefficients $\langle A^{-1}x,g_j\rangle\lambda_j$ in the expansion on the right-hand side are unique. It is easy to see that $\langle A^{-1}x,g_j\rangle=[x,g_j]/\lambda_j$ for $x\in X_{-1}$ and $j\in\mathbb{Z}^*$. Thus, for arbitrary $x\in X_{-1}$,

$$x=\sum_{j\in\mathbb{Z}^*}[x,g_j]f_j.$$

Since $[B,g_j]=\sin(-\pi/2+\pi|j|)=(-1)^{|j|+1}$, we obtain the following expansion for B in X_{-1}:

$$B=\sum_{j\in\mathbb{Z}^*}(-1)^{|j|+1}f_j. \tag{5.7}$$

A standard application of the Carleson measure criterion (see [8, 33]) yields that B is an admissible control operator for the semigroup $\mathbf{T}$. Since the observation (5.3) is described by the operator $C:=B^*$, we conclude that C is an admissible observation operator. From (5.5) and (5.7), it is easy to see that for any $\varepsilon>0$, $B\in\mathcal{B}(\mathbb{R},X_{-(1/4+\varepsilon)})$ and $C\in\mathcal{B}(X_{1/4+\varepsilon},\mathbb{R})$. Hence we can apply Proposition 2.2 to conclude that (A,B,C) are the generating operators of a regular well-posed system.

The semigroup generated by A has exponential growth constant $-\gamma\pi^2/4$, the real part of the rightmost eigenvalue of A. Our aim is to construct a hold function H such that the sampled-data feedback control law (3.1) with weighting $w(s)\equiv 1$ achieves closed-loop exponential growth $\omega_{\mathrm{sd}}\le-9\gamma\pi^2/4$. To this end, fix $\beta\in(-9\gamma\pi^2/4,-\gamma\pi^2/4)$. Then assumption A1 holds, the subspace X^+ of X is spanned by $\{f_{-1},f_1\}$, and $\sigma(A^+)=\sigma(A)\cap\overline{\mathbb{C}}_\beta=\{\lambda_1,\overline{\lambda}_1\}$. It is clear that $\omega(\mathbf{T}^-)=-9\gamma\pi^2/4<0$, showing that A2 holds. It is straightforward to show that $(f_j)_{j\in\mathbb{Z}^*}$ is a Schauder basis of X_1, so that $(f_j)_{j\in\mathbb{Z}^*}$ is a Schauder basis of each of the three spaces X_1, X, and X_{-1}. With respect to this basis we have that

$$A=\mathrm{diag}_{j\in\mathbb{Z}^*}(\lambda_j),\quad \mathbf{T}_t=\mathrm{diag}_{j\in\mathbb{Z}^*}(e^{\lambda_j t}),\quad B=(((-1)^{|j|+1})_{j\in\mathbb{Z}^*})^T,\quad C=(c_j)_{j\in\mathbb{Z}^*},$$

where

$$c_k=2(-1)^{k+1}/(1-e^{-2i\varphi}),\quad c_{-k}=\overline{c}_k\qquad\forall\,k\in\mathbb{N}.$$

[4] Strictly speaking, B is the operator in $\mathcal{B}(\mathbb{R},X_{-1})$ given by $Bv=v(0,\delta_1)^T$, but it is convenient to identify B and $B1=(0,\delta_1)^T$.

Furthermore,

$$A^+ = \operatorname{diag}(\dots, 0, 0, \overline{\lambda}_1, \lambda_1, 0, 0, \dots), \quad \mathbf{T}_t^+ = \operatorname{diag}(\dots, 0, 0, e^{\overline{\lambda}_1 t}, e^{\lambda_1 t}, 0, 0, \dots),$$

$$B^+ = (\dots, 0, 0, 1, 1, 0, 0, \dots)^T, \quad C^+ = (\dots, 0, 0, \overline{c}_1, c_1, 0, 0, \dots).$$

It follows in particular that assumption A3 is satisfied. Furthermore, since

$$\sum_{j \in \mathbb{Z}^*} \frac{1}{|\operatorname{Re} \lambda_j|} = 2 \sum_{j=1}^{\infty} \frac{1}{\gamma(-\pi/2 + \pi j)^2} < \infty,$$

Theorem 4.1 in [20] implies that, for any $t > 0$, there exists a unique sequence $(p_j)_{j \in \mathbb{Z}^*}$ in $L^2([0,t], \mathbb{C})$ such that

$$\int_0^t e^{\lambda_j s}\, \overline{p}_l(s)\, ds = \begin{cases} 0, & j \neq l, \\ 1, & j = l; \end{cases} \tag{5.8}$$

that is, $(e^{\lambda_j \cdot})_{j \in \mathbb{Z}^*}$ and $(p_j)_{j \in \mathbb{Z}^*}$ are biorthogonal (note that $\overline{p}_j = p_{-j}$ for all $j \in \mathbb{Z}^*$). Consequently,

$$\operatorname{im} R_{t_0} \supset X^+ \quad \forall\, t_0 \in (0, \infty), \tag{5.9}$$

where R_{t_0} is the reachability operator given by (4.8). The inclusion (5.9) shows in particular that A5 holds for every $t_0 > 0$. Since $\sigma(A^+) = \{\lambda_1, \overline{\lambda}_1\}$, condition (4.10) is satisfied, provided that

$$\tau \neq \frac{4k}{\pi \sqrt{1 - \gamma^2}} \quad \forall\, k \in \mathbb{N}. \tag{5.10}$$

Furthermore, since $w(s) \equiv 1$ and $\operatorname{Re} \lambda_1 \neq 0$, (4.9) holds for all $\tau > \delta > 0$, and therefore, we may conclude that assumption A4 is satisfied.

Choose $\tau > 0$ such that (5.10) holds and fix $\delta \in (0, \tau)$. It follows from Theorem 4.4 (combined with Remark 4.5 and (5.9)) that there exists $H \in L^2([0,\delta], \mathbb{R})$ such that the sampled-data feedback control law (3.1) with weighting $w(s) \equiv 1$ achieves closed-loop exponential growth $\omega_{\mathrm{sd}} \leq -9\gamma\pi^2/4$. We now use the construction in the proof of Theorem 4.4 to compute such a hold function H. To this end, note that the operator L_w^+ defined in (4.12) can be represented as

$$L_w^+ = \operatorname{diag}(\dots, 0, 0, \overline{\lambda}, \lambda, 0, 0, \dots), \quad \text{where } \lambda := (e^{\lambda_1(\tau - \delta)} - 1)/\lambda_1.$$

We first find $Q \in \mathcal{B}(\mathbb{C}, X^+)$ such that (4.15) holds. Since Q is of the form

$$Q = (\dots, 0, 0, q_{-1}, q_1, 0, 0, \dots)^T,$$

we do this by computing $q_{-1}, q_1 \in \mathbb{C}$ with the property that the two eigenvalues of the matrix

$$\begin{pmatrix} e^{\overline{\lambda}_1 \tau} & 0 \\ 0 & e^{\lambda_1 \tau} \end{pmatrix} + \begin{pmatrix} q_{-1} \\ q_1 \end{pmatrix} (\overline{c}_1, c_1) \begin{pmatrix} \overline{\lambda} & 0 \\ 0 & \lambda \end{pmatrix} = \begin{pmatrix} e^{\overline{\lambda}_1 \tau} + q_{-1} \overline{c}_1 \overline{\lambda} & q_{-1} c_1 \lambda \\ q_1 \overline{c}_1 \overline{\lambda} & e^{\lambda_1 \tau} + q_1 c_1 \lambda \end{pmatrix}$$

are both equal to 0. A routine calculation leads to

$$q_1 = \frac{-e^{2\lambda_1 \tau}}{(e^{\lambda_1 \tau} - e^{\overline{\lambda}_1 \tau}) c_1 \lambda} = \frac{e^{2\lambda_1 \tau} \lambda_1}{c_1 (e^{\lambda_1 \tau} - e^{\overline{\lambda}_1 \tau})(1 - e^{\lambda_1(\tau - \delta)})}, \quad q_{-1} = \overline{q}_1. \tag{5.11}$$

We now compute $h \in L^2([0,\delta],\mathbb{R})$ such that $R_\delta h = Q$, in which case (4.16) holds for every $\eta > 0$. Using (5.8) to solve $R_\delta h = Q$ for h, we find that

$$h(t) = q_1\overline{p}_1(\delta - t) + \overline{q}_1 p_1(\delta - t) \quad \forall\, t \in [0,\delta].$$

The control law (3.1) with $H = -h$ (and with $w(s) \equiv 1$) achieves closed-loop exponential growth $\omega_{\rm sd} \leq -9\gamma\pi^2/4$. It is shown in [21, section 4] how to construct the functions p_j.

Remark 5.1. If we kept the same form for the boundary control in (5.2) but modified the remaining boundary conditions to other "natural" boundary conditions, identified in [9, 21], we could go through the same process to find a "stabilizing" generalized hold function H. The only difference being that the eigenvalues and eigenvectors would be given by asymptotic formulas—see, e.g., [17] for the formulas for such a beam with one end clamped and the other end free. On the other hand, if the control appears as a bending moment force (e.g., $z_{\xi\xi}(1,t) = u(t)$), then the resulting system will not be well-posed, and our theory does not apply.

REFERENCES

[1] M. Araki, *Recent developments in digital control theory*, in Proceedings of the 12th IFAC World Congress, Sidney, Australia, 1993, pp. 251–260.

[2] A. B. Chammas and C. T. Leondes, *On the design of linear time-invariant systems by periodic output feedback.* II. *Output feedback controllability*, Internat. J. Control, 27 (1978), pp. 895–903.

[3] G. Chen and D. L. Russell, *A mathematical model for linear elastic systems with structural damping*, Quart. Appl. Math., 39 (1982), pp. 433–454.

[4] R. F. Curtain and A. J. Pritchard, *Infinite-Dimensional Linear Systems Theory*, Lecture Notes in Control and Inform. Sci. 8, Springer, Berlin, New York, 1978.

[5] R. F. Curtain and G. Weiss, *Well posedness of triples of operators (in the sense of linear systems theory)*, in *Control and Estimation of Distributed Parameter Systems*, F. Kappel, K. Kunisch, and W. Schappacher, eds., Birkhäuser, Basel, Switzerland, 1989, pp. 41–59.

[6] R. F. Curtain and H. J. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Texts Appl. Math. 21, Springer, New York, 1995.

[7] K.-J. Engel and R. Nagel, *One-Parameter Semigroups for Linear Evolution Equations*, Grad. Texts in Math. 194, Springer, New York, 2000.

[8] L. F. Ho and D. L. Russell, *Admissible input elements for systems in Hilbert space and a Carleson measure criterion*, SIAM J. Control Optim., 21 (1983), pp. 614–640.

[9] W. C. Hurty and M. F. Rubinstein, *Dynamics of Structures*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[10] P. T. Kabamba, *Control of linear systems using generalized sampled-data hold functions*, IEEE Trans. Automat. Control, 32 (1987), pp. 772–783.

[11] T. Kato, *Perturbation Theory for Linear Operators*, 2nd ed., Grundlehren Math. Wiss. 132, Springer, Berlin, New York, 1980.

[12] H. Logemann and A. D. Mawby, *Discrete-time and sampled-data low-gain control of infinite-dimensional linear systems in the presence of input hysteresis*, SIAM J. Control Optim., 41 (2002), pp. 113–140.

[13] H. Logemann, R. Rebarber, and S. Townley, *Stability of infinite-dimensional sampled-data systems,* Trans. Amer. Math. Soc., 355 (2003), pp. 3301–3328.

[14] H. Logemann and E. P. Ryan, *Time-varying and adaptive discrete-time low-gain control of infinite-dimensional linear systems with input nonlinearities*, Math. Control Signals Systems, 13 (2000), pp. 293–317.

[15] H. Logemann and S. Townley, *Discrete-time low-gain control of uncertain infinite-dimensional systems,* IEEE Trans. Automat. Control, 42 (1997), pp. 22–37.

[16] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci. 44, Springer, New York, 1983.

[17] R. Rebarber, *Spectral determination for a cantilever beam*, IEEE Trans. Automat. Control, 34 (1989), pp. 502–510.

[18] R. Rebarber and S. Townley, *Stabilization of distributed parameter systems using piecewise polynomial control*, IEEE Trans. Automat. Control, 42 (1997), pp. 1254–1257.

[19] R. Rebarber and S. Townley, *Generalized sampled-data feedback control of distributed parameter systems*, Systems Control Lett., 34 (1998), pp. 229–240.
[20] R. M. Redheffer, *Completeness of sets of complex exponentials*, Adv. Math., 24 (1977), pp. 1–62.
[21] D. L. Russell, *Mathematical models for the elastic beam and their control-theoretic implications*, in Semigroups, Theory and Applications, Vol. II, H. Brezis, M.G. Crandall, and F. Kappel, eds., Scientific and Technical Longman, Harlow, UK, 1986, pp. 177–216.
[22] D. Salamon, *Realization theory in Hilbert space*, Math. Systems Theory, 21 (1989), pp. 147–164.
[23] D. Salamon, *Infinite-dimensional linear systems with unbounded control and observation: A functional analytic approach*, Trans. Amer. Math. Soc., 300 (1987), pp. 383–431.
[24] E. D. Sontag, *Mathematical Control Theory*, Texts Appl. Math. 6, 2nd ed., Springer, New York, 1998.
[25] O. J. Staffans, *Well-Posed Linear Systems*, Cambridge University Press, Cambridge, UK, 2005.
[26] O. J. Staffans, *Quadratic optimal control of stable well-posed linear systems*, Trans. Amer. Math. Soc., 349 (1997), pp. 3679–3715.
[27] O. J. Staffans and G. Weiss, *Transfer functions of regular linear systems.* II. *The system operator and the Lax-Phillips semigroup*, Trans. Amer. Math. Soc., 354 (2002), pp. 3229–3262.
[28] T. J. Tarn, T. Yang, X. Zeng, and C. Guo, *Periodic output feedback stabilization of neutral systems*, IEEE Trans. Automat. Control, 41 (1996), pp. 511-521.
[29] T. J. Tarn, J. R. Zavgren, and X. Zeng, *Stabilization of infinite-dimensional systems with periodic gains and sampled output*, Automatica J. IFAC, 24 (1988), pp. 95–99.
[30] S. Townley, R. Rebarber, H. J. Zwart, and D. K. Oates, *Stabilization of infinite-dimensional systems by generalized sampled-data control*, in Proceedings of the 3rd International Symposium on Methods and Models in Automation and Robotics, Miedzyzdroje, Poland, 1996, pp. 127–132.
[31] G. Weiss, *Transfer functions of regular linear systems.* I. *Characterizations of regularity*, Trans. Amer. Math. Soc., 342 (1994), pp. 827–854.
[32] G. Weiss, *The representation of regular linear systems on Hilbert spaces*, in Control and Estimation of Distributed Parameter Systems, F. Kappel, K. Kunisch, and W. Schappacher, eds., Birkhäuser, Basel, Switzerland, 1989, pp. 401–416.
[33] G. Weiss, *Admissibility of input elements for diagonal semigroups on* l^2, Systems Control Lett., 10 (1988), pp. 79–82.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 4, pp. 1370–1390

EXISTENCE OF MINIMIZERS FOR POLYCONVEX AND NONPOLYCONVEX PROBLEMS*

GIOVANNI CUPINI† AND ELVIRA MASCOLO†

Abstract. We study the existence of Lipschitz minimizers of integral functionals

$$\mathcal{I}(u) = \int_\Omega \varphi(x, \det Du(x))\, dx,$$

where Ω is an open subset of $\mathbb{R}^N$ with Lipschitz boundary, $\varphi : \Omega \times (0, +\infty) \to [0, +\infty)$ is a continuous function, and $u \in W^{1,N}(\Omega, \mathbb{R}^N)$, $u(x) = x$ on $\partial\Omega$. We consider both the cases of φ convex and nonconvex with respect to the last variable. The attainment results are obtained passing through the minimization of an auxiliary functional and the solution of a prescribed Jacobian equation.

Key words. nonpolyconvex functional, existence of minimizers, Lipschitz regularity, prescribed Jacobian equation

AMS subject classifications. 49J10, 35J60

DOI. 10.1137/040611999

1. Introduction.

In this paper we consider integral functionals

$$\mathcal{I}(u) = \int_\Omega \varphi(x, \det Du(x))\, dx, \tag{1.1}$$

where Ω is a bounded open subset of $\mathbb{R}^N$ with a Lipschitz boundary, $N \geq 2$, $\varphi : \Omega \times (0, +\infty) \to [0, +\infty)$ is a continuous function, and $u \in W^{1,N}(\Omega, \mathbb{R}^N)$.

We aim at proving the existence of Lipschitz solutions to the variational problem

$$\min \left\{ \mathcal{I}(u) \,:\, u \in W^{1,N}(\Omega, \mathbb{R}^N), \quad \det Du > 0 \text{ a.e.}, \quad u(x) = x \text{ on } \partial\Omega \right\}. \tag{1.2}$$

Notice that even if a growth condition from below of the type $t^p \leq \varphi(x, t)$ (which is common in the theory of calculus of variations) is assumed, no coercivity of $\mathcal{I}$ follows in any Sobolev space, preventing us from establishing the existence of minimizers via the direct method. Nevertheless many problems of this type have a solution, and the question of fixing which conditions on φ ensure the existence of solutions is worthy of interest, as is its applications in physics, mainly in elasticity theory and in the problem of the equilibrium of gases (see [17], [5], [6], and [12]). For instance, (1.2) is the variational problem corresponding to a nonhomogeneous elastic material with reference configuration Ω whose stored energy φ is a nonnegative, continuous function depending on the position x in the reference configuration and the size of the deformation of the volume element $\det Du(x) > 0$.

It is well known that an important role is played by the convexity of φ with respect to the last variable: when φ is convex, then $\mathcal{I}$ is said to be a polyconvex functional; if not, then $\mathcal{I}$ is nonpolyconvex. The polyconvex case $\varphi = \varphi(t)$ has been studied by Dacorogna [5] and the nonpolyconvex case by Mascolo and Schianchi [14] and Cellina and Zagatti [4].

*Received by the editors July 20, 2004; accepted for publication (in revised form) April 14, 2005; published electronically October 21, 2005.

http://www.siam.org/journals/sicon/44-4/61199.html

†Dipartimento di Matematica "U. Dini," Università degli Studi di Firenze, Viale Morgagni 67/A, 50134 Firenze, Italy (cupini@math.unifi.it, mascolo@math.unifi.it).

In order to solve (1.2) our strategy is the following: The first step is to look for solutions to the following problem (from now on referred to as the *auxiliary* problem)

(1.3)
$$\min\left\{\mathcal{J}(v)=\int_\Omega \varphi(x,v(x))\,dx\,:\,v\in L^1(\Omega),\quad v>0 \text{ a.e.},\quad \int_\Omega v(x)\,dx=|\Omega|\right\},$$

where $|\Omega|$ stands for the N-dimensional Lebesgue measure of Ω. Then, if v is a solution to (1.3), the second step is to solve in $W^{1,N}(\Omega)$ the boundary value problem

$$\begin{cases}\det Du(x)=v(x) & \text{for a.e. } x \text{ in } \Omega,\\ u(x)=x & \text{on } \partial\Omega.\end{cases} \tag{1.4}$$

A solution u to (1.4) is a solution to (1.2), too. In fact, if $w\in W^{1,N}(\Omega)$, $w(x)=x$ on $\partial\Omega$, then $\det Dw\in L^1(\Omega)$ and $\int_\Omega \det Dw(x)\,dx=|\Omega|$; therefore, if $\det Dw>0$ a.e., then

$$\mathcal{I}(u)=\mathcal{J}(v)\le \mathcal{J}(\det Dw)=\mathcal{I}(w)\,.$$

Following the above scheme, Mascolo in [13] proves the existence of minimizers of (1.2) for smooth domains Ω and $\varphi\in C^2(\overline{\Omega}\times(0,+\infty))$ strictly convex in the last variable.

As far as problem (1.3) is concerned, Ekeland and Temam in [8] prove a relaxation result and Ball and Knowles in [1] obtain an attainment result with the tool of the Young measures; see also Friesecke [10] for related results. The boundary value problem (1.4) may have no solution unless v is sufficiently regular. For instance, the simple continuity of v is not a sufficient condition to get Lipschitz solutions; see the counterexamples independently given by Burago and Kleiner [2] and McMullen [15]. Thus, also the regularity properties of minimizers of the auxiliary problem have to be studied. The pioneering papers on (1.4) are due to Moser [16] and Dacorogna and Moser [7]. In particular, in [7] the authors prove that if v is in $C^{k,\alpha}(\overline{\Omega})$, $k\ge 0$, and $\partial\Omega\in C^{k+3,\alpha}$, then there exists a diffeomorphism of class $C^{k+1,\alpha}(\overline{\Omega})$ solution to (1.4). Later results are due to Rivière and Ye, who prove in [18, Theorem 4] the existence of a bi-Lipschitz homeomorphism u solution to (1.4) under less restrictive assumptions on Ω with v satisfying a Dini-type continuity property. In [19] Ye proves existence results in the framework of the Sobolev spaces.

The plan of the paper is the following. In section 2 we introduce a class of open sets, invariant under bi-Lipschitz homeomorphisms, which is slightly larger than that of open sets with Lipschitz boundaries; see Definition 2.1. In Theorem 2.4 we state the existence of Lipschitz solutions to (1.4) with Ω in this class of open sets and Hölder continuous datum v. It is a variant of the above-cited Theorem 4 in [18], and in the appendix we give the details of the proof. In section 3 we deal with polyconvex functionals. We consider the class of functions φ strictly convex in the last variable satisfying, as a substitute for the growth conditions,

$$\lim_{t\to 0^+} D_t\varphi(x,t)=\lambda_0 \text{ with } \lambda_0\in\mathbb{R}\cup\{-\infty\},\qquad \lim_{t\to+\infty} D_t\varphi(x,t)=+\infty, \tag{1.5}$$

uniformly with respect to x. In Proposition 3.1 we prove that a unique solution v to (1.3) exists and that v is in $L^\infty(\Omega)$. In Proposition 3.5, under more regularity assumptions on φ, we prove that v is Hölder continuous. Therefore, the Lipschitz solution u to (1.4), which exists by Theorem 2.4, is a minimizer of (1.2); see Theorem 3.6. In section 4 we deal with a function φ nonconvex with respect to t, satisfying (1.5).

Denoting φ^{**} the convex envelope of φ with respect to t, we assume that there exist $\alpha, \beta \in L^\infty(\Omega)$, $\beta(x) > \alpha(x)$, $\inf \alpha > 0$ such that for every $x \in \Omega$,

$$t \mapsto \varphi^{**}(x,t) \text{ is affine in } [\alpha(x), \beta(x)]$$

and

$$\varphi(x,\cdot) \equiv \varphi^{**}(x,\cdot) \text{ and } \varphi(x,\cdot) \text{ is strictly convex} \qquad \text{in } (0,\alpha(x)] \text{ and } [\beta(x), +\infty).$$

Under these assumptions in Theorem 4.1 we prove the existence of a bounded solution v to the auxiliary problem (1.3). In section 5 under regularity assumptions on φ we get that v is piecewise Hölder continuous; see Theorem 5.2. In section 6 first we prove that if in (1.4) the datum v is piecewise Hölder continuous, there exists a Lipschitz solution; see Proposition 6.2. Then, solving (1.4) with v the piecewise Hölder continuous solution to the auxiliary problem, in Theorems 6.3 and 6.4 we get a Lipschitz continuous minimizer of functional (1.1). In section 7 we consider special classes of nonpolyconvex functionals. First we consider the class of functionals with a nonconvex φ satisfying $\varphi(x,\alpha(x)) = \varphi(x,\beta(x)) = 0$. This class has been considered by Zagatti [20] (see also Celada and Perrotta [3] for the case $\varphi(x,u,t)$) with the assumption $\int_\Omega \alpha(x)\,dx < |\Omega| < \int_\Omega \beta(x)\,dx$. In [20] and [3] the attainment result is proved using different arguments: the Baire category method and the convex integration method, respectively. Theorems 7.1 and 7.2 are attainment results including the cases $\int_\Omega \alpha\,dx \geq |\Omega|$ and $\int_\Omega \beta(x)\,dx \leq |\Omega|$. Theorem 7.4 deals with a *perturbation* of these functionals; see problem (7.2). We conclude the section considering functionals with φ satisfying the structure condition $\varphi(x,t) = \tilde{\varphi}(|x|,t)$. In this case the existence of bounded radial solutions to (1.3) directly implies the existence of Lipschitz solutions to (1.4).

2. Notation and preliminary results. In the following if Ω is a measurable subset of $\mathbb{R}^N$, then $|\Omega|$ stands for its N-dimensional Lebesgue measure. We write Q in place of $(0,1)^N$ and $B_r(x)$ denotes the ball in $\mathbb{R}^N$ with center at x and radius r. If $\varphi : \Omega \times (0,+\infty) \to [0,+\infty)$, then φ^{**} is the convex envelope of φ with respect to the second variable, i.e., $t \mapsto \varphi^{**}(x,t)$ is the greatest convex function lower than $t \mapsto \varphi(x,t)$. For the sake of simplicity we write $\varphi(x,\cdot)$ instead of $t \mapsto \varphi(x,t)$,

$$D_t^- \varphi(x,s) := \lim_{t \to s^-} \frac{\varphi(x,t) - \varphi(x,s)}{t-s}, \qquad D_t^+ \varphi(x,s) := \lim_{t \to s^+} \frac{\varphi(x,t) - \varphi(x,s)}{t-s},$$

and $\partial\varphi(x,s) := \{d \in \mathbb{R} : \varphi(x,t) \geq \varphi(x,s) + d(t-s) \text{ for every } t \in (0,+\infty)\}$.

We define a class of bounded open subsets of $\mathbb{R}^N$.

DEFINITION 2.1. *We say that a bounded open set Ω of $\mathbb{R}^N$ is of class (L) if $\overline{\Omega}$ has a covering of finitely many open sets Ω_j such that for every j there exists a bi-Lipschitz homeomorphism $\psi_j : \overline{\Omega_j \cap \Omega} \to \overline{Q}$ satisfying*

(a) $\psi_j(\overline{\Omega}_j \cap \partial\Omega) = \{0\} \times [0,1]^{N-1}$, *whenever* $\overline{\Omega}_j \cap \partial\Omega$ *is not empty;*

(b) $\det D\psi_j$ *is Lipschitz continuous and there exists* $A \geq 1$ *such that* $\frac{1}{A} \leq \det D\psi_j \leq A$.

The above definition describes a larger class than that of open sets with Lipschitz boundary, i.e., with the boundary which locally is the graph of a Lipschitz function. This result can be proved in a way similar to that of Proposition A.1 in [7].

LEMMA 2.2. *If a bounded open set Ω of $\mathbb{R}^N$ has a Lipschitz boundary, then it is of class (L).*

An easy consequence of the chain rule for Lipschitz functions is that Definition 2.1 is invariant under bi-Lipschitz homeomorphisms.

LEMMA 2.3. *Let $u_0 : \mathbb{R}^N \to \mathbb{R}^N$ be a bi-Lipschitz homeomorphism, with* $\det Du_0$ *Lipschitz continuous,* $\frac{1}{A} \le \det Du_0 \le A$ *for some* A. *If* Ω *is of class* (L), *then* $u_0(\Omega)$ *is of class* (L), *too.*

On the contrary, there are examples of bounded open sets of $\mathbb{R}^N$ with Lipschitz boundary which are mapped by a bi-Lipschitz homeomorphism $u : \mathbb{R}^N \to \mathbb{R}^N$ onto sets with a not (Lipschitz) continuous boundary; see, e.g., [11, pp. 8–9]. Therefore, the converse of Lemma 2.2 is not true.

Now, we state an existence result of Lipschitz solutions to

$$\begin{cases} \det Du = f & \text{in } \Omega, \\ u(x) = x & \text{on } \partial\Omega \end{cases} \tag{2.1}$$

with f Hölder continuous.

THEOREM 2.4. *Let $\Omega \subset \mathbb{R}^N$ be a bounded connected open set of class (L). Let f be a Hölder continuous function,* $\inf f > 0$, $\int_\Omega f(x)\,dx = |\Omega|$. *Then there exists a bi-Lipschitz homeomorphism $u : \overline{\Omega} \to \overline{\Omega}$ solution to* (2.1).

A similar result is proved in [18, Theorem 4], with a weaker assumption on v, which is assumed to satisfy a Dini-type continuity property, and a regular domain Ω. In [18] the proof is given for cubes only. The proof of Theorem 2.4, based upon the application to open sets of class (L) of the partition method due to Moser [16], is in the appendix.

3. Polyconvex problems: An attainment result. In this section we consider the variational problem

$$\min\left\{\int_\Omega \psi(x, \det Du(x))\,dx \,:\, u \in W^{1,N}(\Omega, \mathbb{R}^N),\quad \det Du > 0 \text{ a.e.},\quad u(x) = x \text{ on } \partial\Omega\right\}, \tag{3.1}$$

where Ω is a bounded open subset of $\mathbb{R}^N$ with a Lipschitz boundary and $\psi : \Omega \times (0, +\infty) \to [0, +\infty)$ is a continuous function.

To get solutions to (3.1), we first consider the following variational problem:

$$\min\left\{\int_\Omega \psi(x, v(x))\,dx \,:\, v \in L^1(\Omega),\quad v > 0 \text{ a.e.},\quad \int_\Omega v(x)\,dx = a\right\},\quad a > 0. \tag{3.2}$$

As far as the problem (3.2) is concerned, the Lipschitz regularity of the boundary of Ω can be dropped.

We prove that there exists a (unique) bounded solution to (3.2) if

(H1) $t \mapsto \psi(x, t)$ is strictly convex for all $x \in \Omega$;

(H2) there exists $\lambda_0 \in \mathbb{R} \cup \{-\infty\}$ such that

$$\lim_{t \to 0^+} D_t^+ \psi(x, t) = \lambda_0, \quad \lim_{t \to +\infty} D_t^- \psi(x, t) = +\infty, \quad \text{uniformly in } x.$$

PROPOSITION 3.1. *Assume that $\psi : \Omega \times (0, +\infty) \to [0, +\infty)$ is a continuous function satisfying* (H1) *and* (H2). *Then for every $\lambda > \lambda_0$ there exists a unique $u_\lambda \in L^\infty(\Omega)$,* $\inf u_\lambda > 0$ *such that*

$$\lambda \in \partial\psi(x, u_\lambda(x)) \quad \forall\, x \in \Omega. \tag{3.3}$$

Moreover, there exists $\lambda_a > \lambda_0$ such that u_{λ_a} is the unique solution to (3.2).

Proof. We proceed as follows: At first we prove that for every $\lambda > \lambda_0$ there exists a function u_λ such that (3.3) holds. Then, we prove that u_λ is in $L^\infty(\Omega)$, $\inf u_\lambda > 0$, and there exists λ_a such that $\int_\Omega u_{\lambda_a}\,dx = a$. Thus, it turns out that u_{λ_a} is a solution to (3.2) and it is unique, because of the strict convexity of the functional.

Step 1. *The definition of u_λ.* Fixing $x \in \Omega$, we define the sets

$$C(x) := \left\{ s \in (0,+\infty) \,:\, D_t^-\psi(x,s) < D_t^+\psi(x,s) \right\}, \quad \Omega_C := \{ x \in \Omega \,:\, C(x) \neq \emptyset \}.$$

Notice that $\partial\psi(x,s) = [D_t^-\psi(x,s), D_t^+\psi(x,s)]$ for all $(x,s) \in \Omega \times (0,+\infty)$.

Suppose that $x \in \Omega \setminus \Omega_C$. From (H1) and the definition of Ω_C, the function $D_t\psi(x,\cdot) : (0,+\infty) \to (\lambda_0,+\infty)$ is well defined, continuous, and strictly increasing. Moreover, it is a surjective function because of (H2). Let $u(x,\cdot)$ be its inverse function, i.e., $u(x,\cdot) : (\lambda_0,+\infty) \to (0,+\infty)$ is such that $u(x,\lambda)$ (from now on denoted by $u_\lambda(x)$) is the unique positive number such that $\lambda = D_t\psi(x,u_\lambda(x))$. $u(x,\cdot)$ is a well defined, strictly increasing, and continuous function.

Now let us consider $x \in \Omega_C$. From (H1), $C(x)$ is (at most) a countable set, so that we denote $C(x) = \{t_n(x)\}_{n\in J(x)}$, where $J(x) \subseteq \mathbb{N}$. As in the above case, if $\lambda \notin \cup_{n\in J(x)}\partial\psi(x,t_n(x))$, we define $u_\lambda(x)$ as the unique positive number such that $D_t\psi(x,u_\lambda(x)) = \lambda$. If instead $\lambda \in \partial\psi(x,t_n(x))$ for some $n \in J(x)$, then we set $u_\lambda(x) = t_n(x)$. Notice that if $u_\lambda(x)$ is chosen greater (less) than $t_n(x)$, then $\lambda < D_t^-\psi(x,u_\lambda(x))$ ($\lambda > D_t^+\psi(x,u_\lambda(x))$). It is easy to prove that for each $x \in \Omega_C$ the function $u(x,\cdot) : (\lambda_0,+\infty) \to (0,+\infty)$ is well defined, increasing, and continuous.

Thus, $u_\lambda : \Omega \to (0,+\infty)$ is the unique function satisfying (3.3) and it is measurable, since

$$\{ x \in \Omega \,:\, u_\lambda(x) < t \} = \left\{ x \in \Omega \,:\, D_t^-\psi(x,t) > \lambda \right\}$$

and $D_t^-\psi(x,t) = \sup_{h<0}(\psi(x,t+h)-\psi(x,t))/h$. By the second limit in (H2) for every $\lambda > \lambda_0$ there exists $R > 0$ such that $D_t^-\psi(x,R) > \lambda$ for every $x \in \Omega$, which implies $u_\lambda(x) < R$ for every $x \in \Omega$. In fact, if $u_\lambda(x) \geq R$ for some x, then by the convexity of ψ with respect to the second variable it would be $D_t^-\psi(x,R) \leq D_t^-\psi(x,u_\lambda(x))$ and by (3.3) we would obtain $D_t^-\psi(x,R) \leq \lambda$, which is a contradiction. Thus, u_λ is in $L^\infty(\Omega)$. The first limit in (H2) implies that for each $\lambda > \lambda_0$ there exists $c(\lambda) > 0$ such that $\sup_{y\in\Omega} D_t^+\psi(y,t) < \lambda$ for every $t < c(\lambda)$. Therefore, it cannot be $u_\lambda(x) < c(\lambda)$, because $\lambda \leq D_t^+\psi(x,u_\lambda(x))$, so that $\inf u_\lambda > 0$.

Step 2. *The definition of λ_a.* Define $\Psi : (\lambda_0,+\infty) \to (0,+\infty)$, $\Psi(\lambda) := \int_\Omega u_\lambda(x)\,dx$, where $u_\lambda(x) = u(x,\lambda)$ is defined as in Step 1. By the monotonicity of u with respect to λ, Ψ is increasing. It holds true that $\lim_{\lambda\to\lambda_0^+} u_\lambda(x) = 0$. In fact, suppose that $\lim_{\lambda\to\lambda_0^+} u_\lambda(x) = \delta(x) > 0$. By (H1), the first limit in (H2), and (3.3), we get

$$\lambda_0 < D_t^-\psi(x,\delta(x)) \leq D_t^-\psi(x,u_\lambda(x)) \leq \lambda\,.$$

Therefore, letting λ go to λ_0^+ we get a contradiction. Analogously it can be proved that $\lim_{\lambda\to+\infty} u_\lambda(x) = +\infty$. Hence,

$$\lim_{\lambda\to\lambda_0^+} \Psi(\lambda) = 0, \quad \lim_{\lambda\to+\infty} \Psi(\lambda) = +\infty. \tag{3.4}$$

From the previous step $\lambda \mapsto u_\lambda(x)$ is continuous and increasing for all x and $u_\lambda \in L^\infty(\Omega)$ for all λ, and therefore Ψ is a continuous function. Thus, there exists $\lambda_a > \lambda_0$

such that $\Psi(\lambda_a) = a$. We claim that u_{λ_a} is a solution to (3.2). In fact, from (H1) and (3.3) for every $w \in L^1(\Omega)$ such that $w > 0$ and $\int_\Omega w(x)\,dx = a$, we have that

$$\psi(x, w(x)) \geq \psi(x, u_{\lambda_a}(x)) + \lambda_a\big(w(x) - u_{\lambda_a}(x)\big) \quad \forall\, x \in \Omega.$$

Thus,

$$\begin{aligned}\int_\Omega \psi(x, w(x))\,dx &\geq \int_\Omega \psi(x, u_{\lambda_a}(x))\,dx + \lambda_a \int_\Omega (w(x) - u_{\lambda_a}(x))\,dx \\ &= \int_\Omega \psi(x, u_{\lambda_a}(x))\,dx\,. \qquad \square\end{aligned}$$

REMARK 3.2. *The growth conditions*

$$\lim_{t\to 0^+} \inf_{y\in\Omega} \psi(y,t) = +\infty\,, \quad \lim_{t\to+\infty} \inf_{y\in\Omega} \frac{\psi(y,t)}{t} = +\infty$$

imply (H2). *If the first limit in* (H2) *is not uniform with respect to* x, *then maybe* $\inf u_\lambda = 0$. *Moreover, the proof of Proposition* 3.1 *works also if we replace* $\lim_{t\to+\infty} D_t^- \psi(x,t) = +\infty$ *with the more general*

$$\lim_{t\to+\infty} D_t^- \psi(x,t) = \lambda_\infty, \qquad \lambda_\infty \in \mathbb{R} \cup \{+\infty\}.$$

It is easy to prove the following refinement of Proposition 3.1.

PROPOSITION 3.3. *Let* $\psi : \Omega \times (0,+\infty) \to [0,+\infty)$ *be a continuous function, differentiable with respect to the last variable,* $D_t\psi \in C(\Omega \times (0,+\infty))$. *If* (H1) *and* (H2) *hold, then the functions* u_λ *in Proposition* 3.1 *are continuous for every* $\lambda > \lambda_0$.

Proof. For every $\lambda > \lambda_0$ let $u_\lambda \in L^\infty(\Omega)$ be as in Proposition 3.1. u_λ is lower semicontinuous. In fact, if

$$\liminf_{x\to x_0} u_\lambda(x) < \alpha < u_\lambda(x_0)\,, \tag{3.5}$$

then (H1) and (3.3) imply $D_t\psi(x_0,\alpha) < \lambda$. By continuity of $D_t\psi$ there exists $\delta > 0$ such that $D_t\psi(x,\alpha) < \lambda$ for every $x \in (x_0 - \delta, x_0 + \delta)$. Then, from (3.3) again we have that $D_t\psi(x,\alpha) < D_t\psi(x, u_\lambda(x))$ for every $x \in (x_0 - \delta, x_0 + \delta)$, which implies $\alpha < u_\lambda(x)$, in contradiction with (3.5). Analogously the upper semicontinuity of u_λ can be proved. $\square$

To get Hölder continuous solutions to (3.2) we require more regularity on ψ:

(H3) there exists $0 < \sigma \leq 1$ such that for every compact $K \subset (0,+\infty)$ and for every $t \in K$ the function $x \mapsto D_t\psi(x,t)$ is of class $C^{0,\sigma}(\Omega)$ with $[D_t\psi(\cdot,t)]_{0,\sigma} \leq k_K$;

(H4) for every $m > 0$ there exists $c_m > 0$ such that

$$\psi(x,t) \geq \psi(x,s) + D_t\psi(x,s)(t-s) + c_m|t-s|^{2+\varepsilon}$$

for every $t > s \geq m$, for every $x \in \Omega$, and for some $\varepsilon \geq 0$.

REMARK 3.4. *Assumption* (H4) *is equivalent to assuming that for every* $m > 0$ *there exists* $\tilde{c}_m > 0$ *such that*

$$D_t\psi(x,t) - D_t\psi(x,s) \geq \tilde{c}_m|t-s|^{1+\varepsilon} \quad \forall\, t > s \geq m \quad \forall\, x \in \Omega. \tag{3.6}$$

Roughly speaking, if $\psi \in C^2$ *satisfies* (H4), *then* $D_{tt}\psi$ *may vanish provided that a suitable growth near the zeros is satisfied; see* (3)(a) *below.*

Notice that if ψ_0 *satisfies* (H4) *and* $\psi_1 = \psi_1(x,t)$ *is such that* $\psi_1(x,\cdot)$ *is convex and* C^1, *then* $\psi = \psi_0 + \psi_1$ *satisfies* (H4), *too. Examples of functions* ψ_0 *satisfying* (H4) *are as follows.*

(1) $\psi_0(t) := (1+t^2)^{p/2}$, $p \geq 2$. *See* [9] *for details.*
(2) $\psi_0(x,t) := |t - a(x)|^p$ *with* $a : \Omega \to \mathbb{R}$ *and* $p \geq 2$.
(3) $\psi_0 : \overline{\Omega} \times (0,+\infty) \to [0,+\infty)$ *of class* C^2*, strictly convex with respect to* t *such that for every* x *there exist at most finitely many positive numbers* $\{s_i(x)\}$ *such that* $D_{tt}\psi_0(x, s_i(x)) = 0$ *and the following hold:*
   (a) *there exist* $\varepsilon, c > 0$ *such that* $D_{tt}\psi_0(x,t) \geq c|t - s_i(x)|^\varepsilon$ *for every* t *in a neighborhood of* $s_i(x)$*;*
   (b) *there exists* $M > 0$ *such that* $\inf\{D_{tt}\psi_0(x,t) : (x,t) \in \Omega \times [M,+\infty)\} > 0$.

PROPOSITION 3.5. *Let* $\psi : \Omega \times (0,+\infty) \to [0,+\infty)$ *be a continuous function, differentiable with respect to the last variable, satisfying* (H1)–(H4)*. Then for each* $\lambda > \lambda_0$*, the function* u_λ *in Proposition* 3.1 *is in* $C^{0,\sigma/(1+\varepsilon)}(\Omega)$*. In particular, for every* $a > 0$ *the unique solution* u_{λ_a} *to* (3.2) *is Hölder continuous.*

Proof. Fix λ and let u_λ, from now on referred to as u, be the correspondent function as described in Proposition 3.1. From the strict convexity of ψ with respect to the last variable and since $\lambda = D_t\psi(x, u(x))$ for every $x \in \Omega$ it is easy to check that u is γ-Hölder continuous with Hölder constant $[u]_\gamma$ if and only if

$$D_t\psi\big(y, u(x) + [u]_{0,\gamma}|x-y|^\gamma\big) - D_t\psi(x, u(x)) \geq 0 \quad \forall\, x, y \in \Omega. \tag{3.7}$$

Fix $x, y \in \Omega$. By (H4) and (3.6) there exist $\varepsilon \geq 0$ and $\tilde{c} > 0$ such that

$$D_t\psi(x,t) - D_t\psi(x,s) \geq \tilde{c}(t-s)^{1+\varepsilon} \quad \forall\, t > s \geq \inf u > 0 \quad \forall\, x \in \Omega. \tag{3.8}$$

Consider the compact interval $K = [\inf u, \|u\|_\infty]$ and let s and t be equal to $u(x)$ and $u(x) + (\frac{k}{\tilde{c}}|x-y|^\sigma)^{1/(1+\varepsilon)}$, respectively, with σ and k_K as in (H3). Using (3.8) and (H3) to estimate $D_t\psi(y,t) - D_t\psi(y,s)$ and $D_t\psi(y,s) - D_t\psi(x,s)$, respectively, we get

$$D_t\psi(y,t) - D_t\psi(x,s) = D_t\psi(y,t) - D_t\psi(y,s) + D_t\psi(y,s) - D_t\psi(x,s) \geq 0.$$

Then u is γ-Hölder continuous with $\gamma = \frac{\sigma}{1+\varepsilon}$.

Thus, for fixed $a > 0$, the solution u_{λ_a} to (3.2), which exists by Proposition 3.1, is Hölder continuous. ∎

Now we are ready to state an existence result of Lipschitz solutions to the polyconvex problem (3.1).

THEOREM 3.6. *Suppose that* Ω *is a bounded open subset of* $\mathbb{R}^N$ *with Lipschitz boundary and let* $\psi : \Omega \times (0,+\infty) \to [0,+\infty)$ *be a continuous function, differentiable with respect to the last variable, satisfying* (H1)–(H4)*. Then there exists a Lipschitz continuous solution to* (3.1)*.*

Proof. Set $a = |\Omega|$ and consider the variational problem (3.2). From Propositions 3.1 and 3.5 such a problem has a (unique) solution $u_{\lambda_a} \in C^{0,\gamma}(\Omega)$, $\gamma > 0$, and $\inf u_{\lambda_a} > 0$. Hence, from Theorem 2.4 there exists a bi-Lipschitz homeomorphism u solving

$$\begin{cases} \det Du = u_{\lambda_a} & \text{in } \Omega, \\ u(x) = x & \text{on } \partial\Omega, \end{cases}$$

and u is a solution to (3.1), too. ∎

4. Nonpolyconvex problems: Attainment result for the auxiliary problem. In this section we consider the variational problem

$$\min\left\{\int_\Omega \varphi(x, v(x))\,dx \,:\, v \in L^1(\Omega),\, v > 0 \text{ a.e.}, \quad \int_\Omega v(x)\,dx = a\right\}, \quad a > 0, \tag{4.1}$$

where Ω is a bounded open subset of $\mathbb{R}^N$, and $\varphi : \Omega \times (0,+\infty) \to [0,+\infty)$ is a continuous function, nonconvex with respect to the last variable t.

Let φ^{**} be the convex envelope of φ with respect to the second variable and define

$$\Omega_A := \{x \in \Omega : t \to \varphi(x,t) \text{ is not strictly convex}\}.$$

We assume that the following assumptions hold:

(K1) Ω_A is a (not empty) measurable set and there exist $\alpha, \beta \in L^\infty(\Omega_A)$, $\beta(x) > \alpha(x)$ for all x, $\inf \alpha > 0$, such that $\varphi(x,\cdot)$ and $\varphi^{**}(x,\cdot)$ both coincide and are strictly convex in $(0,\alpha(x)]$ and $[\beta(x),+\infty)$ for every $x \in \Omega_A$;

(K2) $\varphi^{**}(x,\cdot)$ is affine in $[\alpha(x),\beta(x)]$ for all $x \in \Omega_A$, i.e., for every $\alpha(x) \le t \le \beta(x)$,

$$\varphi^{**}(x,t) = h(x)t + q(x) \text{ with } \ h(x) = \frac{\varphi(x,\beta(x)) - \varphi(x,\alpha(x))}{\beta(x) - \alpha(x)};$$

(K3) there exists $\lambda_0 \in \mathbb{R} \cup \{-\infty\}$ such that

$$\lim_{t\to 0^+} D_t^+ \varphi(x,t) = \lambda_0, \quad \lim_{t\to+\infty} D_t^- \varphi(x,t) = +\infty, \quad \text{uniformly in } x.$$

THEOREM 4.1. *Assume* (K1), (K2), *and* (K3). *Then there exist* $\lambda_a > \lambda_0$ *and* $v_{\lambda_a} \in L^\infty(\Omega)$, $\inf v_{\lambda_a} > 0$ *such that*

(i) $v_{\lambda_a}(x) \notin (\alpha(x),\beta(x))$ *for every* $x \in \Omega_A$;

(ii) $\lambda_a \in \partial\varphi^{**}(x, v_{\lambda_a}(x))$ *for every* $x \in \Omega$;

(iii) $\int_\Omega v_{\lambda_a}(x)\,dx = a$.

In particular, v_{λ_a} *is a solution to* (4.1). *Moreover, if* $\Omega = B_1(0)$ *and* $\varphi(x,t) = \tilde\varphi(|x|,t)$, *then* v_{λ_a} *is a radial function.*

We postpone the proof of Theorem 4.1 to the following lemma.

LEMMA 4.2. *Let* O *be a bounded measurable subset of* $\mathbb{R}^N$. *Let* $\alpha, \beta \in L^1(O)$ *be such that* $\alpha(x) \le \beta(x)$ *for a.e.* x *and suppose*

$$\int_O \alpha(x)\,dx < \kappa < \int_O \beta(x)\,dx\,. \tag{4.2}$$

Then there exists $r > 0$ *such that* $\Theta : O \to \mathbb{R}$, $\Theta(x) := \alpha(x)$ *if* $x \in O \cap B_r(0)$ *and* $\Theta(x) := \beta(x)$ *else, satisfying* $\int_O \Theta(x)\,dx = \kappa$.

Proof. Let R be such that $O \subset B_R(0)$. Consider the functions $\theta_\rho : O \to \mathbb{R}$, $0 \le \rho \le R$, defined as follows: $\theta_0 := \beta$ and if $\rho \ne 0$, then $\theta_\rho(x) := \alpha(x)$, if $x \in O \cap B_r(0)$ and $\theta_\rho(x) := \beta(x)$ else. The continuity of $\rho \to \int_O \theta_\rho(x)\,dx$ and (4.2) imply that there exists $0 < r < R$ such that $\int_O \theta_r(x)\,dx = \kappa$. □

We are now ready to prove Theorem 4.1.

Proof of Theorem 4.1. We divide the proof into three steps. In Step 1 we define a family of functions $v_\lambda^- : \Omega \to (0,+\infty)$, $\lambda > \lambda_0$, such that

$$v_\lambda^-(x) \notin (\alpha(x),\beta(x)) \quad \forall x \in \Omega_A \quad \forall \lambda > \lambda_0 \tag{4.3}$$

and

$$\lambda \in \partial\varphi^{**}\big(x, v_\lambda^-(x)\big) \quad \forall x \in \Omega \quad \forall \lambda > \lambda_0. \tag{4.4}$$

In Step 2 we define a function v_{λ_a} satisfying (i), (ii), and (iii). Finally, in Step 3 we consider the case $\varphi(x,t) = \tilde\varphi(|x|,t)$.

Step 1. *The definition of* v_λ^-. Let us define the function $\psi : \Omega\times(0,+\infty) \to [0,+\infty)$ such that $\psi \equiv \varphi$ in $(\Omega \setminus \Omega_A) \times (0,+\infty)$ and

$$\psi(x,t) := \begin{cases} \varphi(x,t) & \text{if } x \in \Omega_A,\ 0 < t \le \alpha(x), \\ \varphi(x, t+\beta(x)-\alpha(x)), & \\ -\varphi(x,\beta(x)) + \varphi(x,\alpha(x)) & \text{if } x \in \Omega_A,\ t > \alpha(x). \end{cases} \tag{4.5}$$

(K1) and (K2) imply that for every $x \in \Omega_A$

$$D_t^-\varphi(x,\alpha(x)) \le h(x) = \frac{\varphi(x,\beta(x)) - \varphi(x,\alpha(x))}{\beta(x)-\alpha(x)} \le D_t^+\varphi(x,\beta(x)) \tag{4.6}$$

and that ψ satisfies (H1). Moreover, for every $x \notin \Omega_A$ and every $t > 0$ we have $\partial\psi(x,t) = \partial\varphi(x,t) = \partial\varphi^{**}(x,t)$. If instead $x \in \Omega_A$, then

$$\partial\psi(x,t) = \begin{cases} \partial\varphi(x,t) & \text{if } \ 0 < t < \alpha(x), \\ \partial\varphi^{**}(x,\alpha(x)) \cup \partial\varphi^{**}(x,\beta(x)) & \text{if } \ t = \alpha(x), \\ \partial\varphi(x, t+\beta(x)-\alpha(x)) & \text{if } \ t > \alpha(x). \end{cases} \tag{4.7}$$

We claim that (K3) implies that ψ satisfies (H2).

The first limit in (K3) and the assumption $\inf\alpha > 0$ imply $\lim_{t\to 0^+} D_t^+\psi(x,t) = \lambda_0$, uniformly. Let us prove that ψ satisfies the property on the second limit in (H2). Since $\alpha, \beta \in L^\infty(\Omega_A)$, then for every $x \in \Omega$ and $t > \|\alpha\|_{L^\infty(\Omega_A)}$,

$$\begin{aligned} \inf_{y\in\Omega} D_t^-\varphi(y,t) &\le \min\Big\{ \inf_{y\in\Omega_A} D_t^-\varphi(y, t+\beta(y)-\alpha(y)), \inf_{y\in\Omega\setminus\Omega_A} D_t^-\varphi(y,t)\Big\} \\ &= \inf_{y\in\Omega} D_t^-\psi(y,t) \le D_t^-\psi(x,t) \le D_t^-\varphi(x, t+\|\beta-\alpha\|_{L^\infty(\Omega_A)}) \end{aligned}$$

so that by (K3) as t goes to $+\infty$, we get

$$\lim_{t\to+\infty} \inf_{y\in\Omega} D_t^-\psi(y,t) = \lim_{t\to+\infty} D_t^-\psi(x,t) = +\infty \quad \forall x \in \Omega.$$

Since ψ satisfies the assumptions of Proposition 3.1, then for every $\lambda > \lambda_0$ there exists $u_\lambda \in L^\infty(\Omega)$, $\inf u_\lambda > 0$, satisfying (3.3). Moreover, for every $x \in \Omega_A$,

$$\begin{array}{ll} u_\lambda(x) < \alpha(x) & \text{if } \ \lambda < D_t^-\varphi(x,\alpha(x)), \\ u_\lambda(x) = \alpha(x) & \text{if } \ \lambda \in [D_t^-\varphi(x,\alpha(x)), D_t^+\varphi(x,\beta(x))], \\ u_\lambda(x) > \alpha(x) & \text{if } \ \lambda > D_t^+\varphi(x,\beta(x)). \end{array} \tag{4.8}$$

Let us define $v_\lambda^- : \Omega \to (0,+\infty)$,

$$v_\lambda^-(x) := u_\lambda(x) + (\beta(x)-\alpha(x))\chi_{\{y\in\Omega_A\,:\,h(y)<\lambda\}}(x).$$

Since $u_\lambda \in L^\infty(\Omega)$ and $\alpha, \beta \in L^\infty(\Omega_A)$, then $v_\lambda^- \in L^\infty(\Omega)$. From (3.3), (4.6), (4.7), and (4.8) if $x \in \Omega_A$, the following implications hold:

- if $\lambda < D_t^-\varphi(x,\alpha(x))$, then $v_\lambda^-(x) = u_\lambda(x) < \alpha(x)$ and $\lambda \in \partial\psi(x,u_\lambda(x)) = \partial\varphi(x,v_\lambda^-(x))$;
- if $\lambda \in [D_t^-\varphi(x,\alpha(x)), h(x)]$, then $v_\lambda^-(x) = u_\lambda(x) = \alpha(x)$ and $\lambda \in \partial\varphi^{**}(x,\alpha(x))$;
- if $\lambda \in (h(x), D_t^+\varphi(x,\beta(x))]$, then $v_\lambda^-(x) = \beta(x)$ and $\lambda \in \partial\varphi^{**}(x,\beta(x))$;
- if $\lambda > D_t^+\varphi(x,\beta(x))$, then $v_\lambda^-(x) = u_\lambda(x) + \beta(x) - \alpha(x) > \beta(x)$ and $\lambda \in \partial\psi(x,u_\lambda(x)) = \partial\varphi(x,v_\lambda^-(x))$.

Thus (4.3) holds and

$$\lambda \in \partial\varphi^{**}\big(x, v_\lambda^-(x)\big) \tag{4.9}$$

for every $x \in \Omega_A$ and $\lambda > \lambda_0$. When $x \notin \Omega_A$, the equality $v_\lambda^-(x) = u_\lambda(x)$ and (3.3) imply (4.9). Therefore, (4.4) holds true.

Step 2. *The definition of* λ_a *and* v_{λ_a}. Let us define $\Phi : (\lambda_0, +\infty) \to (0, +\infty)$,

$$\Phi(\lambda) := \int_\Omega v_\lambda^-(x)\,dx = \int_\Omega \Big(u_\lambda(x) + (\beta(x) - \alpha(x))\chi_{\{y\in\Omega_A\,:\,h(y)<\lambda\}}(x)\Big)\,dx.$$

As in the proof of (3.4) we have that $\lim_{\lambda\to\lambda_0^+}\Phi(\lambda) = 0$ and $\lim_{\lambda\to+\infty}\Phi(\lambda) = +\infty$.

For each $\lambda > \lambda_0$, define $v_\lambda^+ : \Omega \to (0, +\infty)$,

$$v_\lambda^+(x) := u_\lambda(x) + (\beta(x) - \alpha(x))\chi_{\{y\in\Omega_A\,:\,h(y)\le\lambda\}}(x).$$

For every $\mu > \lambda_0$,

$$\lim_{\lambda\to\mu^-}\Phi(\lambda) = \Phi(\mu), \quad \lim_{\lambda\to\mu^+}\Phi(\lambda) = \int_\Omega v_\mu^+(x)\,dx.$$

Thus, Φ is discontinuous at μ if and only if $|\{y \in \Omega_A\,:\,h(y) = \mu\}| > 0$.

Only one of the following cases is possible:

1. there exists $\lambda_a > \lambda_0$ such that $\Phi(\lambda_a) = a$;
2. there exists $\lambda_a > \lambda_0$ such that $\Phi(\lambda_a) < a = \lim_{\lambda\to\lambda_a^+}\Phi(\lambda)$;
3. there exists $\lambda_a > \lambda_0$ such that $\Phi(\lambda_a) < a < \lim_{\lambda\to\lambda_a^+}\Phi(\lambda)$.

Case 1. As proved in Step 1, $v_{\lambda_a}^-$ satisfies (i), (ii), and $\inf v_{\lambda_a}^- \ge \inf u_{\lambda_a} > 0$. Moreover, by definition of λ_a, (iii) holds. Thus, define $v_{\lambda_a} = v_{\lambda_a}^-$.

Case 2. As above, $v_{\lambda_a}^-$ satisfies (i), (ii), and $\inf v_{\lambda_a}^- \ge \inf u_{\lambda_a} > 0$. It is easy to check that a property analogous to (i) is satisfied by $v_{\lambda_a}^+$ and that $\inf v_{\lambda_a}^+ \ge \inf v_{\lambda_a}^- > 0$. By the very definition of $v_{\lambda_a}^+$ we have also $\int_\Omega v_{\lambda_a}^+\,dx = a$.

Let us prove that $\lambda_a \in \partial\varphi^{**}(x, v_{\lambda_a}^+(x))$ for every x. If $x \notin \Omega_A$ or if $x \in \Omega_A$ and $h(x) \ne \lambda_a$, then $v_{\lambda_a}^-(x) = v_{\lambda_a}^+(x)$ and the above inclusion follows. Suppose that $x \in \Omega_A$ and $h(x) = \lambda_a$. Then $v_{\lambda_a}^-(x) = \alpha(x) < \beta(x) = v_{\lambda_a}^+(x)$ and (K2) implies $\lambda_a \in \partial\varphi^{**}(x, \beta(x)) = \partial\varphi^{**}(x, v_{\lambda_a}^+(x))$.

We have so proved that $\lambda_a \in \partial\varphi^{**}(x, v_{\lambda_a}^+(x))$ for every $x \in \Omega$. Thus, define $v_{\lambda_a} := v_{\lambda_a}^+$.

Case 3. Define $O := \{x \in \Omega_A\,:\,\lambda_a = h(x)\}$ and $\kappa := a - \int_{\Omega\setminus O} v_{\lambda_a}^-(x)\,dx$. The assumption $\Phi(\lambda_a) < a < \lim_{\lambda\to\lambda_a^+}\Phi(\lambda)$ implies

$$\int_O \alpha(x)\,dx = \int_O v_{\lambda_a}^-(x)\,dx < \kappa < \int_\Omega v_{\lambda_a}^+(x)\,dx - \int_{\Omega\setminus O} v_{\lambda_a}^-(x)\,dx = \int_O \beta(x)\,dx.$$

From Lemma 4.2, there exists $\Theta : O \to \mathbb{R}$, $\Theta(x) \in \{\alpha(x), \beta(x)\}$ such that $\int_O \Theta(x)\,dx = \kappa$. Define $v_{\lambda_a} : \Omega \to \mathbb{R}$, $v_{\lambda_a}(x) = v_{\lambda_a}^-(x)$ if $x \notin O$ and $v_{\lambda_a}(x) = \Theta(x)$ else.

It is easy to prove that v_{λ_a} satisfies (i), (ii), (iii), and $\inf v_{\lambda_a} > 0$.

Since $\varphi \ge \varphi^{**}$, then for every $v \in L^1(\Omega)$ such that $v > 0$ a.e. and $\int_\Omega v\,dx = a$, we have that

$$\begin{aligned}\int_\Omega \varphi(x, v(x))\,dx &\ge \int_\Omega \varphi^{**}(x, v(x))\,dx \\ &\ge \int_\Omega \varphi^{**}(x, v_{\lambda_a}(x))\,dx + \lambda_a\int_\Omega (v(x) - v_{\lambda_a}(x))\,dx = \int_\Omega \varphi(x, v_{\lambda_a}(x))\,dx.\end{aligned} \tag{4.10}$$

Thus, v_{λ_a} is a solution to (4.1).

Step 3. *The case* $\varphi(x,t) = \tilde{\varphi}(|x|,t)$. Assume that Ω is the unit ball $B_1(0)$ and that φ has the radial structure $\varphi(x,t) = \tilde{\varphi}(|x|,t)$. It is easy to prove that $\varphi^{**}(x,t) = (\tilde{\varphi})^{**}(|x|,t)$ and that α, β, h are radial functions. Moreover, the sets Ω_A, $\{y \in \Omega_A : h(y) < \lambda\}$ and $\{y \in \Omega_A : h(y) = \lambda\}$ are symmetric sets with respect to the origin. If ψ is defined as in Step 1 above, then it immediately follows that $\psi(x,t) = \tilde{\psi}(|x|,t)$. Looking at the first step of the proof of Proposition 3.1, it turns out that u_λ, satisfying $\partial\psi(x, u_\lambda(x)) = \lambda$, is a radial function for all λ. All these facts allow us to conclude that whenever Cases 1 or 2 in Step 2 hold, i.e., $\Phi(\lambda_a) = a$ or $\Phi(\lambda_a) < a = \lim_{\lambda\to\lambda_a^+} \Phi(\lambda)$, respectively, then v_{λ_a} is a radial function. To prove that v_{λ_a} is radial in the third case it is sufficient to notice that the sets O, $O \cap B_r(0)$, and $O \setminus B_r(0)$ are symmetric with respect to the origin and consequently the function Θ is radial. ∎

5. Nonpolyconvex problems: Regularity result for the auxiliary problem. In this section we prove a regularity result for solutions to the nonconvex variational problem (4.1). Let Ω be a bounded open subset of $\mathbb{R}^N$ and let $\varphi : \Omega\times(0,+\infty) \to [0,+\infty)$ be a continuous function, differentiable with respect to the last variable, $D_t\varphi \in C^{0,\delta}(\Omega \times K)$, $0 < \delta \le 1$, for every compact K in $(0,+\infty)$ such that

(A1) there exist $\alpha, \beta \in C^{0,\delta}(\Omega)$, $\beta(x) > \alpha(x)$ for every x, $\inf \alpha > 0$ such that $\varphi(x,\cdot)$ and $\varphi^{**}(x,\cdot)$ both coincide and are strictly convex in $(0,\alpha(x)]$ and $[\beta(x),+\infty)$ for every $x \in \Omega$;

(A2) $t \to \varphi^{**}(x,t)$ is affine in $[\alpha(x),\beta(x)]$ for every $x \in \Omega$, i.e., for every $\alpha(x) \le t \le \beta(x)$,

$$\varphi^{**}(x,t) = h(x)t + q(x) \text{ with } \ h(x) = \frac{\varphi(x,\beta(x)) - \varphi(x,\alpha(x))}{\beta(x) - \alpha(x)}.$$

Moreover,

$$|\partial\{x : h(x) = \lambda\}| = 0 \quad \forall\, \lambda \in \mathbb{R};$$

(A3) there exists $\lambda_0 \in \mathbb{R} \cup \{-\infty\}$ such that

$$\lim_{t\to 0^+} D_t\varphi(x,t) = \lambda_0, \quad \lim_{t\to+\infty} D_t\varphi(x,t) = +\infty, \quad \text{uniformly in } x;$$

(A4) for every $m > 0$ there exists $c_m > 0$ such that

$$\varphi(x,t) \ge \varphi(x,s) + D_t\varphi(x,s)(t-s) + c_m|t-s|^{2+\varepsilon}$$

for every $s,t \ge m$ such that $s < t \le \alpha(x)$ or $\beta(x) \le s < t$ for every $x \in \Omega$ and some $\varepsilon \ge 0$.

The following result is in the same spirit of Lemma 4.2.

LEMMA 5.1. *Let O be an open set in $\mathbb{R}^N$. Let $\alpha, \beta \in L^1(O)$ be such that $\alpha(x) \le \beta(x)$ for a.e. x and suppose that*

$$\int_O \alpha(x)\,dx < \kappa < \int_O \beta(x)\,dx. \tag{5.1}$$

Then there exists a finite number of balls $B_{\rho_j}(y_j)$, $j = 1,\dots,m$, satisfying

(1) $\overline{B_{\rho_j}(y_j)} \subset\subset O$, $j = 1,\dots,m$;

(2) $\overline{B_{\rho_i}(y_i)} \cap \overline{B_{\rho_j}(y_j)} = \emptyset$ *for every* $i \ne j$;

(3) $\int_O \Theta(x)\,dx = \kappa$,

where $\Theta(x) := \alpha(x)$ *if* $x \in \cup_{1\le j\le m} B_{\rho_j}(y_j)$ *and* $\Theta(x) := \beta(x)$ *else.*

Proof. Since O is open, there exist (at most) countably many pairwise disjoint balls $\{B_{R_j}(y_j)\}_{j\in J}$ in O, and a negligible set $\mathcal{N}$ such that $O = \mathcal{N} \cup (\bigcup_{j\in J} B_{R_j}(y_j))$. Without loss of generality we assume $J = \{1, 2, \dots, m\}$ if $\operatorname{card} J = m \in \mathbb{N}$ and $J = \mathbb{N}$ if J is countable. For every $n \in J$, let us define the function $\theta_n : O \to \mathbb{R}$,

$$\theta_n(x) := \begin{cases} \alpha(x) & \text{if } x \in \displaystyle\bigcup_{1\le j\le n} B_{R_j}(y_j), \\ \beta(x) & \text{else.} \end{cases}$$

If J is finite, then (5.1) implies $\int_O \theta_m(x)\,dx < \kappa$. If $J = \mathbb{N}$, it is easy to check that $\lim_{n\to+\infty} \int_O \theta_n(x)\,dx < \kappa$; thus, there exists $m \in \mathbb{N}$ such that

$$\int_O \theta_m(x)\,dx = \int_{\cup_{1\le j\le m} B_{R_j}(y_j)} \alpha(x)\,dx + \int_{O\setminus\cup_{1\le j\le m} B_{R_j}(y_j)} \beta(x)\,dx < \kappa\,.$$

Aiming at (1) and (2), we slightly reduce the radius of the previously selected balls $\{B_{R_j}(y_j)\}_{1\le j\le m}$. This can easily be done by noticing that

$$\lim_{\varepsilon\to 0^+} \int_{\cup_{j=1}^m B_{R_j}(y_j)\setminus B_{R_j-\varepsilon}(y_j)} (\beta(x) - \alpha(x))\,dx = 0.$$

Thus, there exists $0 < \varepsilon < \min\{R_j \,:\, 1 \le j \le m\}$ such that

$$\int_{\cup_{1\le j\le m} B_{R_j-\varepsilon}(y_j)} \alpha(x)\,dx + \int_{O\setminus\cup_{1\le j\le m} B_{R_j-\varepsilon}(y_j)} \beta(x)\,dx < \kappa\,. \tag{5.2}$$

Set $R := \max\{R_j - \varepsilon \,:\, 1 \le j \le m\}$ and define $\theta : O \times [0, R] \to \mathbb{R}$, $\theta(x, 0) := \beta(x)$ and

$$\theta(x,\rho) := \begin{cases} \alpha(x) & \text{if } x \in \displaystyle\bigcup_{1\le j\le m} (B_{R_j-\varepsilon}(y_j) \cap B_\rho(y_j)), \\ \beta(x) & \text{else} \end{cases}$$

for every $\rho > 0$. From (5.2) we have that

$$\int_O \theta(x, R)\,dx < \kappa < \int_O \theta(x, 0)\,dx = \int_O \beta(x)\,dx.$$

Since $\rho \to \int_O \theta(x,\rho)\,dx$ is a continuous function, there exists $\overline{\rho}$ such that $\int_O \theta(x,\overline{\rho})\,dx = \kappa$. The claim of the theorem follows by defining $\Theta(x) := \theta(x,\overline{\rho})$ and $\rho_j := \min\{R_j - \varepsilon, \overline{\rho}\}$, $1 \le j \le m$. ∎

Let h be as in (A2). For every $\lambda > \lambda_0$ we define

$$\Omega_\lambda^+ := \{x \,:\, h(x) > \lambda\}, \quad \Omega_\lambda^- := \{x \,:\, h(x) < \lambda\}, \quad \Omega_\lambda^= := \{x \,:\, h(x) = \lambda\}. \tag{5.3}$$

Under (A1)–(A4) there exists a piecewise Hölder continuous solution to (4.1).

THEOREM 5.2. *Let* $\varphi : \Omega \times (0, +\infty) \to [0, +\infty)$ *be a continuous function, differentiable with respect to the last variable,* $D_t\varphi(x,t)$ *in* $C^{0,\delta}(\Omega \times K)$ *for every compact* $K \subset (0, +\infty)$. *Suppose that* (A1)–(A4) *hold. Then, with fixed* $a > 0$ *there exist* $\lambda_a > \lambda_0$ *and* $v_{\lambda_a} \in L^\infty(\Omega)$, $\inf v_{\lambda_a} > 0$, *satisfying the following properties:*

(i) $D_t\varphi^{**}(x, v_{\lambda_a}(x)) = \lambda_a$ *for every* $x \in \Omega$;

(ii) $\int_\Omega v_{\lambda_a}(x)\,dx = a$;
(iii) v_{λ_a} *is Hölder continuous in* $\Omega^+_{\lambda_a} \cup \Omega^-_{\lambda_a}$;
(iv) $v_{\lambda_a}(x) < \alpha(x)$ *for all* $x \in \Omega^+_{\lambda_a}$ *and* $v_{\lambda_a}(x) > \beta(x)$ *for all* $x \in \Omega^-_{\lambda_a}$;
(v) *in* $\Omega^=_{\lambda_a}$ *either* $v_{\lambda_a} \equiv \alpha$ *or* $v_{\lambda_a} \equiv \beta$ *or*

$$v_{\lambda_a}(x) = \begin{cases} \alpha(x) & \text{if } x \in \bigcup_{1\le j\le m} B_{\rho_j}(y_j), \\ \beta(x) & \text{if } x \in \Omega^=_{\lambda_a} \setminus \bigcup_{1\le j\le m} B_{\rho_j}(y_j) \end{cases} \tag{5.4}$$

with $B_{\rho_j}(y_j) \subset\subset \operatorname{int}\Omega^=_{\lambda_a}$, $j = 1,\dots,m$ *such that* $\overline{B_{\rho_i}(y_i)} \cap \overline{B_{\rho_j}(y_j)} = \emptyset$ *if* $i \neq j$.

Moreover, v_{λ_a} *is a solution to* (4.1).

Proof. Let $\psi : \Omega \times (0,+\infty) \to [0,+\infty)$ be defined as

$$\psi(x,t) := \begin{cases} \varphi(x,t) & \text{if } \ 0 < t \le \alpha(x)\,,\ x \in \Omega, \\ \varphi(x, t+\beta(x)-\alpha(x)), & \\ -\varphi(x,\beta(x)) + \varphi(x,\alpha(x)) & \text{if } \ t > \alpha(x)\,,\ x \in \Omega. \end{cases} \tag{5.5}$$

It holds true that ψ is a continuous function, differentiable with respect to the last variable, satisfying (H1)–(H4) in section 3, with possibly different constants. By Proposition 3.5 for every $\lambda > \lambda_0$, there exists u_λ such that $u_\lambda \in C^{0,\gamma}(\Omega)$ for some $0 < \gamma \le 1$, $\inf u_\lambda > 0$, and

$$D_t\psi(x, u_\lambda(x)) = \lambda \quad \forall\, x \in \Omega. \tag{5.6}$$

Moreover (see (4.6) and (4.8)),

$$u_\lambda < \alpha \text{ in } \Omega^+_\lambda, \quad u_\lambda = \alpha \text{ in } \Omega^=_\lambda, \quad u_\lambda > \alpha \text{ in } \Omega^-_\lambda. \tag{5.7}$$

Let $\Phi : (\lambda_0, +\infty) \to \mathbb{R}$ be the left-continuous function defined as

$$\Phi(\lambda) := \int_\Omega \Big(u_\lambda(x) + (\beta(x) - \alpha(x))\chi_{\Omega^-_\lambda}(x) \Big)\, dx, \quad \lambda > \lambda_0.$$

We have three different cases:

1. there exists $\lambda_a > \lambda_0$ such that $\Phi(\lambda_a) = a$;
2. there exists $\lambda_a > \lambda_0$ such that $\Phi(\lambda_a) < a = \lim_{\lambda\to\lambda_a^+} \Phi(\lambda)$;
3. there exists $\lambda_a > \lambda_0$ such that $\Phi(\lambda_a) < a < \lim_{\lambda\to\lambda_a^+} \Phi(\lambda)$.

Let us consider the first two cases: since (A1)–(A3) imply (K1)–(K3), then by proceeding as in Theorem 4.1 there exists $v_{\lambda_a} \in L^\infty(\Omega)$, $\inf v_{\lambda_a} > 0$, which satisfies (i) and (ii). Moreover, if case 1 holds, then $v_{\lambda_a} := u_{\lambda_a} + (\beta - \alpha)\chi_{\{h<\lambda_a\}}$, i.e.,

$$v_{\lambda_a} := u_{\lambda_a} \text{ in } \Omega^+_{\lambda_a}, \quad v_{\lambda_a} := \alpha \text{ in } \Omega^=_{\lambda_a}, \quad v_{\lambda_a} := u_{\lambda_a} + \beta - \alpha \text{ in } \Omega^-_{\lambda_a};$$

if instead case 2 holds, then $v_{\lambda_a} := u_{\lambda_a} + (\beta - \alpha)\chi_{\{h\le\lambda\}}$, i.e.,

$$v_{\lambda_a} := u_{\lambda_a} \text{ in } \Omega^+_{\lambda_a}, \quad v_{\lambda_a} := \beta \text{ in } \Omega^=_{\lambda_a}, \quad v_{\lambda_a} := u_{\lambda_a} + \beta - \alpha \text{ in } \Omega^-_{\lambda_a}.$$

Therefore, from the Hölder continuity of α and β, (5.6) and (5.7) it follows that v_{λ_a} satisfies (iii), (iv), and (v). Moreover, reasoning as in (4.10) we get that v_{λ_a} is a solution to (4.1).

Suppose the third case holds. We define v_{λ_a} as in the proof of Theorem 4.1, but using Lemma 5.1 instead of Lemma 4.2. Precisely, since

$$\int_{\Omega^=_{\lambda_a}} \alpha(x)\,dx < \kappa < \int_{\Omega^=_{\lambda_a}} \beta(x)\,dx$$

with

$$\kappa := a - \int_{\Omega\setminus\Omega^{=}_{\lambda_a}} \Big(u_{\lambda_a}(x) + (\beta(x)-\alpha(x))\chi_{\Omega^{-}_{\lambda_a}}(x)\Big)\,dx,$$

then from Lemma 5.1 there exist m balls $B_{\rho_j}(y_j) \subset\subset \operatorname{int}\Omega^{=}_{\lambda_a}$, $j=1,\dots,m$, $\overline{B_{\rho_i}(y_i)} \cap \overline{B_{\rho_j}(y_j)} = \emptyset$ for every $i \neq j$ such that $\Theta : \operatorname{int}\Omega^{=}_{\lambda_a} \to \mathbb{R}$,

$$\Theta := \alpha \quad \text{in } \textstyle\bigcup_{1\le j\le m} B_{\rho_j}(y_j), \qquad \Theta := \beta \quad \text{in } \operatorname{int}\Omega^{=}_{\lambda_a} \setminus \textstyle\bigcup_{1\le j\le m} B_{\rho_j}(y_j)$$

satisfies $\int_{\operatorname{int}\Omega^{=}_{\lambda_a}} \Theta(x)\,dx = \kappa$.

Define v_{λ_a} as follows:

$$v_{\lambda_a}(x) := \begin{cases} u_{\lambda_a}(x) & \text{if } x \in \Omega^{+}_{\lambda_a}, \\ \alpha(x) & \text{if } x \in \bigcup_{1\le j\le m} B_{\rho_j}(y_j), \\ \beta(x) & \text{if } x \in \Omega^{=}_{\lambda_a} \setminus \bigcup_{1\le j\le m} B_{\rho_j}(y_j), \\ u_{\lambda_a}(x) + \beta(x) - \alpha(x) & \text{if } x \in \Omega^{-}_{\lambda_a}. \end{cases}$$

We have that $v_{\lambda_a} \in L^\infty(\Omega)$, $\inf v_{\lambda_a} > 0$, and it satisfies (i)–(v). Moreover, v_{λ_a} is a solution to (4.1). ∎

6. Nonpolyconvex problems: Attainment result in a general setting. In this section we consider the variational problem

$$\min\left\{\int_\Omega \varphi(x, \det Du(x))\,dx \,:\, u \in W^{1,N}(\Omega,\mathbb{R}^N), \quad \det Du > 0 \text{ a.e., } u(x) = x \text{ on } \partial\Omega\right\}, \tag{6.1}$$

where Ω is a bounded open subset of $\mathbb{R}^N$ with Lipschitz boundary and $\varphi : \Omega \times (0,+\infty) \to [0,+\infty)$ is a nonconvex function with respect to the second variable.

Before stating an attainment result for (6.1), we need some preliminary results.

LEMMA 6.1. *Let Ω be a bounded open set with Lipschitz boundary and let $\overline{\Omega} = \cup_{i=1}^m \overline{\Omega}_i$ with $\{\Omega_i\}$ pairwise disjoint open connected sets with Lipschitz boundary.*

Consider $\alpha_i > 0$, $i = 1,\dots,m$, with $\sum_{i=1}^m \alpha_i = |\Omega|$. Then there exists a bi-Lipschitz homeomorphism $u_0 : \overline{\Omega} \to \overline{\Omega}$ such that $\det Du_0 \in C^\infty(\overline{\Omega})$, $\inf \det Du_0 > 0$, and

$$u_0(x) = x \quad \text{on } \partial\Omega, \quad |u_0(\Omega_i)| = \alpha_i, \quad i = 1,\dots,m. \tag{6.2}$$

Moreover, $u_0(\Omega_i)$ is an open set of class (L) for every i.

Proof. Fix $0 < \delta < \min\{\alpha_i/|\Omega_i| \,:\, i = 1,\dots,m\}$. For every $1 \le i \le m$ let $\eta_i \in C^\infty_c(\Omega_i)$ be such that $\int_{\Omega_i} \eta_i(x)\,dx = 1$. Define

$$f(x) = \delta + \sum_{i=1}^m (\alpha_i - \delta|\Omega_i|)\eta_i(x), \quad x \in \overline{\Omega}.$$

Hence, $f \in C^\infty(\overline{\Omega})$, $\inf f > 0$, $\int_{\Omega_i} f(x)\,dx = \alpha_i$ for every i, and $\int_\Omega f(x)\,dx = |\Omega|$. From Theorem 2.4 there exists a bi-Lipschitz homeomorphism $u_0 : \overline{\Omega} \to \overline{\Omega}$ such that

$$\det Du_0 = f \text{ in } \Omega, \qquad u_0(x) = x \text{ on } \partial\Omega.$$

Therefore,

$$|u_0(\Omega_i)| = \int_{\Omega_i} \det Du_0(x)\,dx = \int_{\Omega_i} f(x)\,dx = \alpha_i, \quad i = 1,\dots,m;$$

moreover, Lemma 2.3 implies that $u_0(\Omega_i)$ is an open set of class (L) for each i. ∎

PROPOSITION 6.2. *Let Ω and Ω_i, $i = 1, \dots, m$, be as in Lemma 6.1. Suppose that $g_i : \overline{\Omega_i} \to [c_0, +\infty)$, with $c_0 > 0$, $i = 1, \dots, m$, are Hölder continuous functions satisfying*

$$\sum_{i=1}^{m} \int_{\Omega_i} g_i(x)\, dx = |\Omega|.$$

Then there exists a Lipschitz continuous function $u : \overline{\Omega} \to \overline{\Omega}$ such that

$$u(x) = x \text{ on } \partial\Omega, \quad \det Du(x) = g_i(x) \quad \forall x \in \Omega_i \quad \forall i = 1, \dots, m. \tag{6.3}$$

Proof. By Lemma 6.1 there exists a bi-Lipschitz homeomorphism $u_0 : \overline{\Omega} \to \overline{\Omega}$ such that

$$u_0(x) = x \text{ on } \partial\Omega, \qquad |u_0(\Omega_i)| = \int_{\Omega_i} g_i(x)\, dx$$

and $u_0(\Omega_i)$ is of class (L) for each $i = 1, \dots, m$. Moreover, $f := \det Du_0$ is of class $C^\infty(\overline{\Omega})$ and $\inf f > 0$. Since $\frac{g_i}{f} \circ u_0^{-1}$ is Hölder continuous in $\overline{u_0(\Omega_i)}$ and it satisfies

$$\int_{u_0(\Omega_i)} \frac{g_i}{f} \circ u_0^{-1}(y)\, dy = \int_{\Omega_i} g_i(x)\, dx = |u_0(\Omega_i)|,$$

then from Theorem 2.4 there exists a bi-Lipschitz homeomorphism $z_i : \overline{u_0(\Omega_i)} \to \overline{u_0(\Omega_i)}$ such that

$$\begin{cases} \det Dz_i = \dfrac{g_i}{f} \circ u_0^{-1} & \text{in } u_0(\Omega_i), \\ z_i(y) = y & \text{on } \partial u_0(\Omega_i). \end{cases}$$

Thus, $u_i = z_i \circ u_0$ is a Lipschitz homeomorphism such that

$$\begin{cases} \det Du_i = g_i & \text{in } \Omega_i, \\ u_i = u_0 & \text{on } \partial\Omega_i. \end{cases}$$

Hence, the Lipschitz continuous function $u : \overline{\Omega} \to \overline{\Omega}$ such that $u(x) = u_i(x)$ for every $x \in \overline{\Omega}_i$, $i = 1, \dots, m$, satisfies (6.3). ∎

We are in position to state an existence result for the nonpolyconvex problem (6.1). The sets Ω_λ^+, Ω_λ^-, and $\Omega_\lambda^=$ are defined in (5.3).

THEOREM 6.3. *Let Ω be a bounded open subset of $\mathbb{R}^N$ with Lipschitz boundary and let $\varphi : \Omega \times (0, +\infty) \to [0, +\infty)$ be a continuous function, differentiable with respect to the last variable, $D_t\varphi \in C^{0,\delta}(\Omega \times K)$, $0 < \delta \le 1$, for every compact $K \subset (0, +\infty)$.*

Suppose that (A1)–(A4) *hold and assume that, for every $\lambda > \lambda_0$, Ω_λ^+, Ω_λ^-, and $\operatorname{int} \Omega_\lambda^=$ are either empty or connected open sets with Lipschitz boundary. Then the variational problem* (6.1) *has a Lipschitz continuous solution.*

Proof. From Theorem 5.2, applied with $a = |\Omega|$, there exist $\lambda_a > \lambda_0$ and a solution v_{λ_a} to (4.1) with $\inf v_{\lambda_a} > 0$. Throughout we write v instead of v_{λ_a}.

From Theorem 5.2 v is Hölder continuous in $\Omega_{\lambda_a}^+ \cup \Omega_{\lambda_a}^-$. If $\operatorname{int} \Omega_{\lambda_a}^=$ is empty, we get the thesis applying Proposition 6.2 with $\Omega_1 = \Omega_{\lambda_a}^+$, $\Omega_2 = \Omega_{\lambda_a}^-$, and replacing g_1 and g_2 with the continuous extension of v to $\Omega_{\lambda_a}^+$ and to $\Omega_{\lambda_a}^-$, respectively.

If $\operatorname{int}\Omega^{=}_{\lambda_a}$ is not empty, correspondingly to (v) of Theorem 5.2 we have to consider three cases.

If $v = \alpha$ in $\Omega^{=}_{\lambda_a}$, the thesis follows by applying Proposition 6.2 with $m = 3$, choosing $\Omega_1 = \Omega^{+}_{\lambda_a}$, $\Omega_2 = \Omega^{-}_{\lambda_a}$, $\Omega_3 = \operatorname{int}\Omega^{=}_{\lambda_a}$, and replacing, as above, g_1 and g_2 with the continuous extension of v to $\Omega^{+}_{\lambda_a}$ and $\Omega^{-}_{\lambda_a}$, respectively, and g_3 with α. Analogously, we proceed if $v = \beta$ in $\Omega^{=}_{\lambda_a}$, but defining $g_3 = \beta$.

Now suppose that (5.4) holds. In this case the thesis follows by Proposition 6.2 choosing $\Omega_1 = \Omega^{+}_{\lambda_a}$, $\Omega_2 = \Omega^{-}_{\lambda_a}$, $\Omega_3 = \operatorname{int}\Omega^{=}_{\lambda_a} \setminus \cup_{1\le j\le n} B_{\rho_j}(y_j)$, $\Omega_{3+i} = B_{\rho_i}(y_i)$ for every $i = 1,\dots,n$ and $g_1 = v$, $g_2 = v$, $g_3 = \beta$, $g_{3+i} = \alpha$, for every $i = 1,\dots,n$. ∎

With obvious changes in the proof above, we get the following theorem.

THEOREM 6.4. *Let Ω and φ be as in Theorem* 6.3. *Suppose that* (A1)–(A4) *hold and assume that for every $\lambda > \lambda_0$,*

$$\overline{\Omega^{+}_{\lambda}} = \bigcup_{i=1}^{h} \overline{A_i}, \quad \overline{\Omega^{-}_{\lambda}} = \bigcup_{i=h+1}^{k} \overline{A_i}, \quad \operatorname{int}\Omega^{=}_{\lambda} = \bigcup_{i=k+1}^{l} A_i \tag{6.4}$$

with A_i either empty or pairwise disjoint open connected sets with Lipschitz boundary. Then the variational problem (6.1) *has a Lipschitz continuous solution.*

REMARK 6.5. *The following are examples of sets Ω and functions $h : \Omega \to \mathbb{R}$ such that for every $\lambda \in \mathbb{R}$* (6.4) *holds with either empty or disjoint open sets $\{A_i\}$ with Lipschitz boundary:*

(a) *Ω is a bounded and convex set and h is strictly convex in Ω and constant on $\partial\Omega$;*

(b) *$\Omega = B_1(0)$ and h is a radial function, $h(x) = \tilde{h}(|x|)$, with $\tilde{h}$ piecewise monotone, i.e., there exists $0 = s_0 < s_1 < \cdots < s_m = 1$ such that $\tilde{h}|_{[s_i,s_{i+1}]}$ is monotone for all i.*

7. Nonpolyconvex problems: Some special cases. In this section we consider particular classes of the variational problem (6.1), where Ω is a bounded open subset of $\mathbb{R}^N$ with Lipschitz boundary and $\varphi : \Omega \times (0,+\infty) \to [0,+\infty)$ is a continuous function satisfying (A1) and (A2). We begin considering the case of functions φ such that h in (A2) is a constant. See [20] and [3] for related results.

THEOREM 7.1. *Let $\varphi : \Omega\times(0,+\infty) \to [0,+\infty)$ be a continuous function satisfying* (A1) *and* (A2) *with h constant. If $\int_\Omega \alpha(x)\,dx \le |\Omega| \le \int_\Omega \beta(x)\,dx$, then* (6.1) *has a Lipschitz continuous solution.*

Proof. Consider the auxiliary problem (4.1) with $a = |\Omega|$. If $\int_\Omega \alpha(x)\,dx$ is equal to $|\Omega|$, then α solves (4.1). Then from Theorem 2.4 there exists a Lipschitz homeomorphism u solution to (2.1) with $f = \alpha$. Moreover, u is a solution of (6.1). The same argument works if $\int_\Omega \beta(x)\,dx$ is equal to $|\Omega|$. Of course in this case choose $f = \beta$.

Suppose $\int_\Omega \alpha(x)\,dx < |\Omega| < \int_\Omega \beta(x)\,dx$. Then using Lemma 5.1 with $O = \Omega$, we get that a Lipschitz continuous solution u to (4.1) exists with $u \equiv \alpha$ on pairwise disjoint balls $B_{\rho_j}(y_j) \subset\subset \Omega$, $j = 1,\dots,n$, and with $u \equiv \beta$ outside these balls. The thesis follows by Proposition 6.2 with $m = n+1$, $\Omega_j = B_{\rho_j}(y_j)$, and $g_j = \alpha$ if $j = 1,\dots,m-1$ and with $\Omega_m = \Omega \setminus \bigcup_{j=1}^{n} B_{\rho_j}(y_j)$, $g_m = \beta$. ∎

THEOREM 7.2. *Let $\varphi : \Omega \times (0,+\infty) \to [0,+\infty)$ be a continuous function, differentiable with respect to the last variable, $D_t\varphi \in C^{0,\delta}(\Omega \times K)$, $0 < \delta \le 1$, for every compact $K \subset (0,+\infty)$. Suppose that* (A1), (A2) *with h constant,* (A3), *and* (A4) *hold. If $\int_\Omega \alpha(x)\,dx > |\Omega|$ or $\int_\Omega \beta(x)\,dx < |\Omega|$, then* (6.1) *has a Lipschitz continuous solution.*

Proof. Let $a = |\Omega|$. From Theorem 5.2 there exist $\lambda_a > \lambda_0$ and $v_{\lambda_a} \in L^\infty(\Omega)$ satisfying

$$v_{\lambda_a}(x) \notin (\alpha(x), \beta(x)), \quad D_t\varphi^{**}(x, v_{\lambda_a}(x)) = \lambda_a, \quad \int_\Omega v_{\lambda_a}(x)\,dx = |\Omega|. \tag{7.1}$$

(A1), (A2), and (A3) imply $h = D_t\varphi(x, \alpha(x)) = D_t\varphi(x, \beta(x))$ and the definition of $\{v_\lambda\}$ (see the proofs of Theorems 4.1 and 5.2) gives that $\lambda < h$ if and only if $v_\lambda(x) < \alpha(x)$ for all x, $\lambda > h$ if and only if $v_\lambda(x) > \beta(x)$ for all x. Therefore, if $\int_\Omega \alpha(x)\,dx > |\Omega|$, then $\lambda_a < h$ and $v_{\lambda_a}(x) < \alpha(x)$. Thus, using the notation in (5.3), $\Omega^+_{\lambda_a} = \Omega$. Analogously, if $\int_\Omega \beta(x)\,dx < |\Omega|$, then $\lambda_a > h$ and $v_{\lambda_a}(x) > \beta(x)$, so that $\Omega^-_{\lambda_a} = \Omega$. Therefore, Theorem 5.2 implies that v_{λ_a} is Hölder continuous in Ω. A Lipschitz continuous solution u to

$$\begin{cases} \det Du = v_{\lambda_a} & \text{in } \Omega, \\ u(x) = x & \text{on } \partial\Omega, \end{cases}$$

solution also to (6.1), exists because of Theorem 2.4. ∎

In Propositions 7.3 and 7.4 we deal with a variant of functionals considered above, precisely

$$\min\left\{\int_\Omega \Phi(x, \det Du(x))\,dx : u \in W^{1,N}(\Omega, \mathbb{R}^N),\ \det Du > 0 \text{ a.e.}, \quad u(x) = x \text{ on } \partial\Omega\right\} \tag{7.2}$$

with $\Phi(x,t) = \varphi(x,t) + f(x)t$.

Proposition 7.3. *Let Ω be a bounded open convex set in $\mathbb{R}^N$ and let $\varphi : \Omega \times (0, +\infty) \to [0, +\infty)$ satisfy the assumptions of Theorem 7.2 with $\lambda_0 = -\infty$ in* (A3). *Suppose that $f : \Omega \to (0, +\infty)$ is a strictly convex function, constant on $\partial\Omega$. Then there exists a Lipschitz solution to* (7.2).

Proof. It is easy to see that Φ satisfies the assumptions of Theorem 6.3. Since $\Phi^{**}(x,t) = \varphi^{**}(x,t) + f(x)t$ for every $x \in \Omega$, then in $(0, \alpha(x)]$ and in $[\beta(x), +\infty)$ we have that $\Phi(x,\cdot) = \Phi^{**}(x,\cdot)$. Moreover, for every $t \in [\alpha(x), \beta(x)]$ it holds true that $\Phi^{**}(x,t) = H(x)t + q(x)$ with $H(x) := \mu + f(x)$ and the superlevel, sublevel, and level sets of H satisfy the assumptions in Theorem 6.3. (A3) implies that $D_t\Phi(x,t) = D_t\varphi(x,t) + f(x)$ goes to $-\infty$ as $t \to -\infty$ and goes to $+\infty$ as $t \to +\infty$, uniformly with respect to x. The thesis easily follows from Theorem 6.3. ∎

From now on, Ω is the unit ball B in $\mathbb{R}^N$ centered at the origin.

Proposition 7.4. *Let $\varphi : B \times (0, +\infty) \to [0, +\infty)$ satisfy the assumptions of Theorem 7.2 with $\lambda_0 = -\infty$ in* (A3). *Let $f \in C^{0,\gamma}([0,1])$, $0 < \gamma \le 1$, $f(s) > 0$ for every s, f piecewise monotone. Then there exists a Lipschitz continuous solution to* (7.2) *with $\Phi(x,t) = \varphi(x,t) + f(|x|)t$.*

Proof. Proceeding as in the proof of Proposition 7.3, the thesis easily follows from Remark 6.5(b) and from Theorem 6.4 applied to $\Phi(x,t) = \varphi(x,t) + f(|x|)t$. ∎

Now, we deal with one more class of nonpolyconvex functionals, characterized by an integrand φ with radial structure $\varphi(x,t) = \tilde\varphi(|x|, t)$. Precisely, we deal with the variational problem

$$\min\left\{\int_B \tilde\varphi(|x|, \det Du(x))\,dx \ :\ u \in W^{1,N}(B, \mathbb{R}^N),\ \det Du > 0 \text{ a.e.},\ u(x) = x \text{ on } \partial B\right\} \tag{7.3}$$

and $\tilde{\varphi} : [0,1) \times (0,+\infty) \to [0,+\infty)$ is a continuous function.

THEOREM 7.5. *Let $\tilde{\varphi} : [0,1) \times (0,+\infty) \to [0,+\infty)$ be a continuous function satisfying the following assumptions:*

(i) *there exist $a, b \in L^\infty(0,1)$, $b(s) > a(s) > 0$ for every s, $\inf a > 0$, such that $\tilde{\varphi}(s,\cdot)$ and $\tilde{\varphi}^{**}(s,\cdot)$ both coincide and are strictly convex in $(0,a(s)]$ and $[b(s),+\infty)$ for all $s \in [0,1)$;*

(ii) *$\tilde{\varphi}^{**}(x,\cdot)$ is affine in $[a(s),b(s)]$ for all $s \in [0,1)$;*

(iii) *there exists $\lambda_0 \in \mathbb{R} \cup \{-\infty\}$ such that*

$$\lim_{t\to 0^+} D_t^+ \tilde{\varphi}(s,t) = \lambda_0, \quad \lim_{t\to+\infty} D_t^- \tilde{\varphi}(s,t) = +\infty, \quad \text{uniformly in } s.$$

Then there exists a Lipschitz solution to (7.3).

Proof. Let us define $\varphi(x,t) := \tilde{\varphi}(|x|,t)$ for every $x \in B$. Notice that $\varphi^{**}(x,t) = \tilde{\varphi}^{**}(|x|,t)$ and that assumptions (K1), (K2), and (K3) of Theorem 4.1 holds with $\Omega = \Omega_A = B$, $\alpha(x) = a(|x|)$, and $\beta(x) = b(|x|)$. Let $v \in L^\infty(B)$, $\inf v > 0$, be the radial solution of (4.1). It is a known fact (see, e.g., [15]) that there exists a bi-Lipschitz solution u to (2.1) with $f = v$ and $\Omega = B$. Thus, u is a solution to (7.3), too. ∎

Appendix. Proof of Theorem 2.4. In the following we use the arguments of the proof of Lemma 1 in [16] and the fact, proved in [18], that if $\Omega = (0,1)^N$ and f is Hölder continuous, then there exists a bi-Lipschitz homeomorphism solution to (2.1). We divide the proof into steps.

Step 1. Let Ω be a bounded open connected subset of $\mathbb{R}^N$ of class (L). Thus, there exist m open sets Ω_j such that $\overline{\Omega} \subset \cup_j \Omega_j$ and m bi-Lipschitz homeomorphisms $\psi_j : \overline{\Sigma}_j \to \overline{Q}$, with $\Sigma_j = \Omega \cap \Omega_j$ and $Q = (0,1)^N$ such that $\det D\psi_j \in \mathrm{Lip}(\overline{\Sigma}_j)$ and $\frac{1}{A} < \det D\psi_j < A$ for some $A \geq 1$. Consider a partition of unity $\{\phi_j\}_{j=1}^m$ subordinate to such a covering of $\overline{\Omega}$: $\{\phi_j\}_{j=1}^m$ is a family of smooth and nonnegative functions, $\sum_j \phi_j(x) = 1$ for every $x \in \overline{\Omega}$ and

$$\operatorname{supp} \phi_j \subset\subset \Omega_j \quad \forall\, j = 1,\dots,m. \tag{7.4}$$

Since $\Omega = \cup_{j=1}^m \Sigma_j$ and Ω is connected, we can assume that for every $k = 2,\dots,m$ there exists $\rho(k) < k$ such that $\Sigma_k \cap \Sigma_{\rho(k)}$ is not empty. Define the matrix (α_{hk}), $1 \leq h \leq m$, $2 \leq k \leq m$,

$$\alpha_{hk} = \begin{cases} 1 & \text{if } h = k, \\ -1 & \text{if } h = \rho(k), \\ 0 & \text{else.} \end{cases}$$

Each of the $m-1$ columns contains exactly one pair $+1$, -1 so that $\sum_{k=2}^m \alpha_{hk} = 0$ for every h.

Define $\eta_k \in C_c^\infty(\Sigma_k \cap \Sigma_{\rho(k)})$ such that $\int_\Omega \eta_k(x)\,dx = 1$. Let $g \in C^{0,\alpha}(\overline{\Omega})$ be such that $\int_\Omega g(x)\,dx = 0$. Define the Hölder continuous functions $g_h : \overline{\Omega} \to \mathbb{R}$, $1 \leq h \leq m$,

$$g_h := g\phi_h|_{\overline{\Omega}} - \sum_{k=2}^m \lambda_k \alpha_{hk} \eta_k,$$

where $\lambda_2,\dots,\lambda_m$ are real numbers solutions of the following system of m equations

$$\sum_{k=2}^m \lambda_k \alpha_{hk} = \int_\Omega g\phi_h\,dx, \qquad h = 1,\dots,m. \tag{7.5}$$

Since the rank of (α_{hk}) is $m-1$ and both $\sum_{h=1}^m \sum_{k=2}^m \lambda_k \alpha_{hk}$ and $\sum_{h=1}^m \int_\Omega g\phi_h\,dx$ are equal to 0, then system (7.5) is uniquely solvable.

We claim that $\operatorname{supp} g_h \subseteq \overline{\Sigma}_h$. In fact $\operatorname{supp}\ \phi_h|_{\overline{\Omega}} \subseteq \overline{\Sigma}_h$ and, since $\alpha_{hk} \neq 0$ if and only if $h = k$ or $h = \rho(k)$,

$$\operatorname{supp} \lambda_k \alpha_{hk} \eta_k \subset \Sigma_k \cap \Sigma_{\rho(k)} \subseteq \Sigma_h$$

for every $k = 2, \dots, m$. Moreover, from (7.5) there exists $M > 0$ depending on Ω, $\{\phi_j\}_j$, and $\{\eta_j\}_j$ only such that $\sup |g_h| \leq M \sup |g|$.

Step 2. Let Ω, $\{\Sigma_j\}_j$, $\{\psi_j\}_j$, $\{\phi_j\}_j$, $\{\eta_j\}_j$, m, and M be as above. Let f in (2.1) be such that $\sup |f-1| < m^{-1}M^{-1}$. Define m Hölder continuous functions g_h reasoning as in the previous step with g replaced by $f-1$. For every $j = 1, \dots, m+1$ define $f_j : \overline{\Omega} \to (0, +\infty)$,

$$f_j(x) := \begin{cases} 1 & \text{if } j = 1, \\ 1 + \sum_{h=1}^{j-1} g_h(x) & \text{if } j > 1. \end{cases}$$

In particular $f_{m+1} = f$. Notice that each f_j is a Hölder continuous function, and since $\sup |f-1| < m^{-1}M^{-1}$, then $\inf f_j > 0$. Fixed $j = 1, \dots, m$, we have that

$$f_{j+1} - f_j = 0 \quad \text{in } \overline{\Omega} \setminus \overline{\Sigma}_j, \qquad \int_\Omega f_j(x)\,dx = |\Omega|, \qquad \int_{\Sigma_j} f_{j+1}(x)\,dx = \int_{\Sigma_j} f_j(x)\,dx. \tag{7.6}$$

Define $f_j^*, f_{j+1}^* : \overline{Q} \to (0, +\infty)$,

$$f_j^* := f_j\big(\psi_j^{-1}\big) \det D\psi_j^{-1}, \quad f_{j+1}^* := f_{j+1}\big(\psi_j^{-1}\big) \det D\psi_j^{-1},$$

so that $f_j^*, f_{j+1}^* \in C^{0,\alpha}(\overline{Q})$ and $\int_Q f_j^*\,dx = \int_Q f_{j+1}^*\,dx$.

As proved in [18] there exist two bi-Lipschitz homeomorphisms $v_j, w_j : \overline{Q} \to \overline{Q}$ solutions to

$$\begin{cases} \det Dv_j = \frac{f_j^*}{\int_Q f_j^*\,dx} & \text{in } Q, \\ v_j(y) = y & \text{on } \partial Q, \end{cases} \quad \text{and} \quad \begin{cases} \det Dw_j = \frac{f_{j+1}^*}{\int_Q f_j^*\,dx} & \text{in } Q, \\ w_j(y) = y & \text{on } \partial Q, \end{cases}$$

respectively. Let us consider $\varphi_j : \overline{Q} \to \overline{Q}$, $\varphi_j(y) := (v_j^{-1} \circ w_j)(y)$. Then

$$\det D\varphi_j(y) = \det Dv_j^{-1}(w_j(y)) \det Dw_j(y) = \frac{f_{j+1}^*(y)}{f_j^*(\varphi_j(y))} \quad \forall\, y \in \overline{Q}$$

so that

$$f_j\big((\psi_j^{-1} \circ \varphi_j)(y)\big) \det D\psi_j^{-1}(\varphi_j(y)) \det D\varphi_j(y) = f_{j+1}\big(\psi_j^{-1}(y)\big) \det D\psi_j^{-1}(y) \quad \forall y \in \overline{Q}.$$

Using the invertibility of ψ_j the equality above implies that

$$f_j(u_j(x)) \det Du_j(x) = f_{j+1}(x) \quad \forall\, x \in \overline{\Sigma}_j, \tag{7.7}$$

where $u_j : \overline{\Sigma}_j \to \overline{\Sigma}_j$ is the Lipschitz continuous function defined as $u_j(x) := (\psi_j^{-1} \circ \varphi_j \circ \psi_j)(x)$.

Since $\varphi_j(\psi_j(x)) = \psi_j(x)$ for all $x \in \partial\Sigma_j$, we have that $u_j(x) = x$ for every $x \in \partial\Sigma_j$. Then $\tilde{u}_j : \overline{\Omega} \to \mathbb{R}$, $j = 1, \dots, m$,

$$\tilde{u}_j(x) := \begin{cases} u_j(x) & \text{if } x \in \overline{\Sigma}_j, \\ x & \text{else} \end{cases}$$

is Lipschitz continuous and from (7.6) and (7.7)

$$f_j(\tilde{u}_j(x)) \det D\tilde{u}_j(x) = f_{j+1}(x) \quad \forall\, x \in \overline{\Omega}.$$

Iterating this argument on j and recalling that $f_1 = 1$ and $f_{m+1} = f$, we get that $\tilde{u}_1 \circ \cdots \circ \tilde{u}_m$ is a Lipschitz solution to (2.1).

Step 3. Now we suppose that f in (2.1) satisfies $\sup |f - 1| \geq m^{-1}M^{-1}$. There exists $c_1 > 0$ and $0 < t_1 < 1$ such that $\int_\Omega c_1 f^{t_1}(x)\, dx = |\Omega|$ and $\sup |c_1 f^{t_1} - 1| < m^{-1}M^{-1}$. Applying the same arguments described in Step 2 to $g := c_1 f^{t_1} - 1$, we obtain a Lipschitz function u_1 satisfying (2.1) with f replaced by $c_1 f^{t_1}$. Applying again this procedure to $g := c_2 f^{t_2} - c_1 f^{t_1}$, with a suitable choice of c_2 and t_2 in such a way that $t_1 < t_2 \leq 1$, $\int_\Omega c_2 f^{t_2}\, dx = |\Omega|$ and $\sup |c_2 f^{t_2} - c_1 f^{t_1}| < m^{-1}M^{-1}$, we get u_2 Lipschitz solution to

$$\begin{cases} c_1 f^{t_1}(u_2) \det Du_2 = c_2 f^{t_2} & \text{in } \Omega, \\ u_2(x) = x & \text{on } \partial\Omega. \end{cases}$$

Hence, $u_1 \circ u_2$ solves (2.1) with f replaced by $c_2 f^{t_2}$. It can be proved that the exponents $\{t_i\}$ can be chosen such that in finitely many steps, say n, we get $t_n = 1$. The existence of a Lipschitz continuous solution to (2.1) follows.

REFERENCES

[1] J.M. Ball and G. Knowles, *Young measures and minimization problems of mechanics,* in Elasticity: Mathematical Methods and Applications, The Ian N. Sneddon 70th Birthday Volume, G. Eason and R.W. Ogden, eds., Ellis Horwood, Chichester, UK, 1990, pp. 1–20.

[2] D. Burago and B. Kleiner, *Separated nets in Euclidean space and Jacobians of biLipschitz maps,* Geom. Funct. Anal., 8 (1998), pp. 304–314.

[3] P. Celada and S. Perrotta, *Vectorial Hamilton–Jacobi equations with rank-one affine dependence on the gradient,* Nonlinear Anal., 41 (2000), pp. 383–404.

[4] A. Cellina and S. Zagatti, *An existence result in a problem of the vectorial case of the calculus of variations,* SIAM J. Control Optim., 33 (1995), pp. 960–970.

[5] B. Dacorogna, *A relaxation theorem and its applications to the equilibrium of gases,* Arch. Ration. Mech. Anal., 77 (1981), pp. 359–385.

[6] B. Dacorogna, *Direct Methods in the Calculus of Variations,* Appl. Math. Sci. 78, Springer, Berlin, 1989.

[7] B. Dacorogna and J. Moser, *On a partial differential equation involving the Jacobian determinant,* Ann. Inst. H. Poincaré Anal. Non Linéaire, 7 (1990), pp. 1–26.

[8] I. Ekeland and R. Temam, *Analyse convexe et problèmes variationnels,* Dunod, Gauthier-Villars, Paris, 1974.

[9] I. Fonseca, N. Fusco, and P. Marcellini, *An existence result for a nonconvex variational problem via regularity,* ESAIM Control Optim. Calc. Var., 7 (2002), pp. 69–95.

[10] G. Friesecke, *A necessary and sufficient condition for nonattainment and formation of microstructure almost everywhere in scalar variational problems,* Proc. Roy. Soc. Edinburgh, 124 (1994), pp. 437–471.

[11] P. Grisvard, *Elliptic Problems in Nonsmooth Domains,* Monogr. Stud. Math. 24, Pitman, Boston, 1985.

[12] P. Marcellini, *The stored-energy for some discontinuous deformations in nonlinear elasticity,* in Partial Differential Equations and the Calculus of Variations, Progr. Nonlinear Differential Equations Appl., Birkhäuser, Boston, 1989, pp. 767–786.

[13] E. Mascolo, *Existence results for a class of noncoercive polyconvex integrals,* Boll. Un. Mat. Ital. A (7), 5 (1991), pp. 97–107.

[14] E. Mascolo and R. Schianchi, *Existence theorems for nonconvex problems,* J. Math. Pures Appl., 62 (1983), pp. 349–359.

[15] C.T. McMullen, *Lipschitz maps and nets in Euclidean space,* Geom. Funct. Anal., 8 (1998), pp. 304–314.

[16] J. Moser, *On the volume elements on a manifold,* Trans. Amer. Math. Soc., 120 (1965), pp. 286–294.

[17] R.W. Odgen, *Large deformation isotropic elasticity: On the correlation of theory and experiment for compressible rubberlike solids,* Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 328 (1972), pp. 567–583.

[18] T. Rivière and D. Ye, *Resolution of the prescribed volume form equation,* NoDEA Nonlinear Differential Equations Appl., 3 (1996), pp. 323–369.

[19] D. Ye, *Prescribing the Jacobian determinant in Sobolev spaces,* Ann. Inst. H. Poincaré Anal. Non Linéaire, 11 (1994), pp. 275–296.

[20] S. Zagatti, *On the Dirichlet problem for vectorial Hamilton–Jacobi equations,* SIAM J. Math. Anal., 29 (1998), pp. 1481–1491.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 4, pp. 1391–1418

ε-OPTIMAL BIDDING IN AN ELECTRICITY MARKET WITH DISCONTINUOUS MARKET DISTRIBUTION FUNCTION*

EDWARD J. ANDERSON† AND HUIFU XU‡

Abstract. This paper investigates the optimal bidding strategy (supply function) for a generator offering power into a wholesale electricity market. The model has three characteristics: the uncertainties facing the generator are described by a single probability function, namely the market distribution function; the supply function to be chosen is nondecreasing but need not be smooth; the objective function is the expected profit which can be formulated as a Stieltjes integral along the generator's supply curve. In previous work the market distribution function has been assumed smooth, but in practice this assumption may not be satisfied. This paper focuses on the case that the market distribution function is not continuous, and hence an optimal supply function may not exist. We consider a modified optimization problem and show the existence of an optimal solution for this problem. Then we show constructively how such an optimum can be approximated with an ε-optimal supply function by undercutting when the generator does not hold a hedging contract (and possibly overcutting when the generator has a hedging contract). Our results substantially extend previous work on the market distribution model.

Key words. electricity markets, discontinuous market distribution function, R-semicontinuity, ε-optimal supply function

AMS subject classifications. 90C46, 65K10, 49K30

DOI. 10.1137/S0363012903425556

1. Introduction.

In recent years many countries have carried out substantial restructuring of their electricity industries. Though each country has adopted its own solution, the trend has been towards increased market mechanisms, particularly at the wholesale level. It is important to understand the operation of these electricity markets, and yet the special features that are characteristic of wholesale electricity markets make this a challenging task.

We begin by sketching the fundamentals of the way that a wholesale market for electricity works. Generators compete to supply electricity to users (primarily retailers providing electricity to consumers). The price paid fluctuates as demand (and supply) varies. The price is determined through a process that is a type of sealed bid auction. In each time period each generator submits a bid, which we refer to as a supply function $S(p)$, which gives the quantity of electricity that the generator is willing to supply for any price p (strictly, this is a price per megawatt hour and the quantity is measured in megawatts). The supply function is increasing (not necessarily strictly) and often has to satisfy other restrictions imposed by the market operator. The spot price is determined from the combined supply functions of all the generators, and is such that supply at the spot price is just sufficient to meet demand. In practice this has to take account of the location of both the generators and the demand within the network, but we will ignore location effects in this paper. A generator needs to decide on the supply function to offer into the market in order to maximize profits. The

*Received by the editors April 8, 2003; accepted for publication (in revised form) March 24, 2005; published electronically November 4, 2005.
http://www.siam.org/journals/sicon/44-4/42555.html
†Australian Graduate School of Management, University of New South Wales, Sydney NSW 2052, Australia (eddiea@agsm.edu.au).
‡School of Mathematics, University of Southampton, Southampton SO17 1BJ, United Kingdom (h.xu@maths.soton.ac.uk).

demand at any time is uncertain and the offers of other generators are also unknown.

A number of authors have used equilibrium concepts to look at the operation of an electricity market. An important set of papers is that by Green and Newbery [8], Green [9], Newbery [12], and Green [7], in which they analyze the experience in the pool market of England and Wales using the concept of an equilibrium in supply functions (see Klemperer and Meyer [10]). The concept of supply function equilibria has also been applied in this context by Rudkevich [13, 14], Anderson and Philpott [1], and Baldick, Grant, and Kahn [5]. The equilibrium models usually assume the smoothness or piecewise linearity of generators' supply functions and consequently it becomes relatively easy to find the optimal choice of strategy by a single generator. However, this may not be the case when we allow general supply functions.

Some recent papers have looked in detail at the optimal strategy for a generator offering power in an electricity spot market. The conclusions depend largely on the models that are used to describe both the generator's objective and constraints, and the market mechanisms. Anderson and Philpott [2] study strategies for generators making offers into an electricity market when either or both of the demand and the offers of competing generators are stochastic. They introduce the *market distribution function* and use it to describe the residual demand for a generator. The market distribution function ψ is a function of price p and quantity q and the value $\psi(q,p)$ represents the probability that a generator is not fully dispatched if it offers a quantity q of electricity at price p. The advantage of this approach is that in many circumstances a single function $\psi(q,p)$ is enough to determine a generator's expected profit given any particular offer curve.

Anderson and Philpott [2] explore the problem of finding an offer curve that maximizes the expected value of the profit made by an individual generator. The offer curve is simply a monotonic continuous curve in the two-dimensional (quantity, price) space. This curve need not be smooth; indeed, in practice it will often take the form of a series of steps. Anderson and Philpott show that the problem of maximizing expected profit is, in some circumstances, equivalent to maximizing a line integral along the offer curve of the market distribution function and they derive necessary conditions for a supply offer curve to be optimal. Anderson and Xu [4] study the same model and extend the analysis to include necessary conditions of a higher order in the presence of horizontal and/or vertical sections in an offer curve. They also derive sufficient conditions for an offer curve to be locally optimal. Neame, Philpott, and Pritchard [11] use Anderson and Philpott's model to study a generator's optimal choice of supply offer curve under the assumption that the generator is a price taker.

All of this work has been carried out under the assumption that the market distribution function $\psi(q,p)$ is continuously differentiable in both price p and quantity q. In this paper we address the problem of finding optimal offer strategies when ψ may be discontinuous in price.

In many electricity wholesale markets, a generator's offer curve consists of a finite number of steps. For instance, in Australia generator bids are restricted to have no more than ten prices. We can model a generator's offer strategy as a step function of price. In this case the market clearing price will not have a continuous distribution; instead, the distribution of clearing price will be concentrated at certain prices. The consequence of this is that the market distribution function will be discontinuous at these prices. The probability of not being fully dispatched if an offer is made of a quantity q at a price p (i.e., the value of ψ) will increase discontinuously as the price moves from just below the offer of another generator to the same price as the other

generator, when the dispatch will be shared between the two generators. In fact this discontinuity happens in both directions of price movement, since as the price moves from being the same as the other generator to just above this level, there is another discontinuous jump in the probability of not being fully dispatched. Previous work in this area has generally made the assumption that the market has a large number of participants, with offer prices well diversified, and so the market distribution function for a generator is nearly continuous and can be approximated with a continuous function.

In this paper we look in more detail at what happens with a discontinuous market distribution function. In the most natural models for this, the existence of an optimal solution for a generator will not be guaranteed (and indeed will occur only rarely). The lack of an optimal solution just reflects the actual difficulty associated with undercutting that can occur in practice. Suppose we know that another generator is offering power at \$30 per MWh, and we have to choose our best offer curve. For example, we might suppose that the other generator is nonstrategic and always submits the same offer. Power we offer at any price up to \$30 will be used in preference to the other generator, power at \$30 will involve a sharing out of the demand between us and the other generator, and power offered at any price higher than \$30 will be dispatched only when the power from the other generator is insufficient to meet demand. This leads to a profit function that is discontinuous in price (in both directions) if we choose to offer power at a single price. A typical good solution to this problem involves offering some power at a price just below the \$30 mark. The closer to \$30 the better for us, but the price must remain below \$30 in order to avoid having to share dispatch. In other words, as we indicated, there is no optimal solution unless we take explicit account of the discretization that may be forced on us by market rules such as, for example, the restriction that we use whole numbers of cents as prices.

In theory at least, this type of undercutting behavior, when translated into the framework of a Nash equilibrium, with different generators all engaged in the same process, will lead to very competitive outcomes as generators repeatedly lower their offer prices towards their true marginal costs. This is essentially the same kind of argument that gives rise to low consumer prices in the Bertrand equilibrium of classical microeconomics. But it can be argued that this is misleading, since markets usually operate in the form of a sealed bid auction, with participants unaware of the bids of other generators. This leads to the possibility of less competitive outcomes through the use of strategies which randomize over the prices offered (see von der Fehr and Harbord [15]).

In this paper, however, we will not discuss equilibrium solutions. Instead we seek to characterize solutions which approach optimality in the undercutting case. In practice it is not unusual for a generator to know the prices at which one or more of the other generators will offer power. For example, there may be nonstrategic generators who offer some quantity of energy at fixed prices which do not vary from day to day. It may be surprising that this occurs, since it is clear that this policy will not in general be optimal for the nonstrategic generator. One explanation is that more complex randomized policies may offer only a limited improvement in profit. As a concrete example of this behavior, consider the Australian market, in which a single generation unit offers power at 10 different price points, set for a 24-hour period (with quantities offered at each price point set separately for each half-hour). These price points are not usually varied from one day to the next; moreover, complete information on all bids is freely available one day after the event (see the web site www.nemmco.com).

For example, power from the Bayswater (coal-fired) power station in New South Wales is offered at a number of price points, but these have included the price \$22.89 for many months on end.

In order to deal with the undercutting behavior, our approach is to alter the model to ensure that the limit of undercutting solutions is a solution with the limiting value. The fundamental idea is to suppose that the generator we are interested in has automatic priority of dispatch when there are other generators offering at the same price. This will ensure that the market distribution function has sufficient continuity properties to guarantee that there will be an optimal solution. Though the exact optimal value is unachievable, the generator can operate in a way that gets as close as it likes to this value. From a practical point of view, establishing the supremum value, and the limiting solution which achieves this, is useful since it enables the generator to find a good solution near the limit, and also bounds the opportunity cost of accepting a suboptimal outcome. In practice market rules imply further restrictions on bids offered, but knowledge of the best possible limit solution will help to guide the selection of a suitable bidding policy.

There is another complication we need to consider. In most cases a generator will have hedging contracts for a significant part of its output. As we shall see, this can have the effect of reversing some of the incentives for the generator. If the generator has contracted for a larger quantity than will actually be dispatched in a certain period, then the generator will benefit from lower prices. The consequence is that, in this case, it will usually be optimal to "overcut" another generator's offer. Hence, to use the same example as above, if another generator has offered some quantity at \$30 per MWh, then we could well decide to make an offer of some amount of power a little above this level (and the closer to \$30 the better). In the case that offers have to be made in whole numbers of cents this would lead to an offer at \$30.01.

We can summarize this paper as follows. We first demonstrate the existence of an optimal solution for a modified problem (section 2). The modified problem differs from the original problem through the method used to determine the sharing of dispatch when two generators offer power at the same price: essentially, the optimizing generator is given the ability to choose its best sharing rule. Care is needed in determining the precise form of the objective function in these circumstances (Theorem 2.9) and in establishing the appropriate form of continuity in order to show the existence result (Lemma 2.7). Then we explore the necessary conditions for an optimal solution to this modified problem (section 3). Next we show how to use an optimal solution for the modified problem to generate an ε-optimal solution for the original problem (section 4). Finally, we illustrate all this with an example (section 5).

2. Problem formulation and fundamentals. In this section we will introduce some notation and formulate the problem that we shall consider.

We consider the behavior of a single generator, which we call A, and we let $R(q, p)$ be the profit for generator A if it is dispatched q at a clearing price p. Usually $R(q, p)$ has three components. First there is the cost, $C(q)$, of generating a quantity q of electricity, which is often taken to be an increasing convex function. Second there is the money, pq, paid to the generator through the market clearing mechanism. Finally, we must also consider the hedging contracts entered into by the generator. These are financial instruments which do not involve the actual generation of electricity; the money paid under the contract is tied to the pool price. If the generator enters into a contract at a strike price f for a quantity Q, and the actual spot price is p, then the generator will pay an amount $Q(p - f)$ to the other party in the contract. The

contracts we consider are two-way contracts for differences, so if the spot price is lower than the contract strike price, then the generator will receive an amount $Q(f-p)$. Contracts of this sort are a common feature of electricity markets operating with a "pool" structure in which prices for all traded electricity are determined through a combined pricing and dispatch mechanism (such as the markets operating in Australia and New Zealand and the old pool arrangements in England and Wales). Note that this is a different environment than that of markets which are based on bilateral contracts, such as in the new trading arrangements in England and Wales.

Thus we arrive at the following expression for the profit to generator A as a function of spot price p and dispatched quantity q:

$$R(q,p) = pq - C(q) + Q(f-p). \tag{2.1}$$

We will not assume any particular functional form for the function R. However, throughout this paper we will assume that R has continuous bounded partial derivatives, R_q and R_p, and is strictly concave in q for fixed p. Thus we have $R_q(\cdot,p)$ strictly decreasing for each fixed p. In the case that (2.1) holds, this assumption will be satisfied provided that the marginal cost of generation is strictly increasing, since $R_q(q,p) = p - C'(q)$.

Next we consider the market dispatch mechanism. We will restrict attention to the case where there is a single node. We consider this from the point of view of generator A. We model the sequence of events in this way. First generator A submits a supply function $S_A(p)$, which gives the total amount of power that generator A is prepared to supply as a function of the price p. Then all the other generators submit their supply functions, which we collectively write as $S_B(p)$—this is the total amount of power that all the other generators are prepared to supply as a function of price. We take both S_A and S_B as right-continuous increasing functions (not necessarily strictly increasing). Where there is a discontinuity in S_A, a jump up occurs at a certain price p, and this corresponds to a certain quantity of power being offered at p and all available at that price. Hence right-continuity is a natural assumption here.

Finally, a demand occurs, where demand at this node is given by a function $D(p)$ of price. We suppose that from the point of view of generator A, both $S_B(p)$ and $D(p)$ are uncertain and must be modeled as stochastic. The market clears at the lowest price p for which $S_A(p) + S_B(p) \geq D(p)$.

In the model we are considering, which corresponds to the most common type of pool market, all generators are paid this clearing price for all the electricity that they are dispatched. This is a type of uniform price auction mechanism. There are other (discriminatory) auction price mechanisms that have been proposed.

Throughout this paper, we assume that a generator's supply function (or equivalently supply curve) is nondecreasing. It can be step-like or strictly increasing, or both. Rather than dealing with a supply function $S_A(\cdot)$ directly, it is convenient to model the offer using a continuous curve $\mathbf{s} = \{(\hat{q}(\tau), \hat{p}(\tau)), 0 \leq \tau \leq T\}$, in which the components $\hat{q}(\tau)$ and $\hat{p}(\tau)$ are continuous monotonic increasing function of τ, and $\hat{q}(\tau)$ and $\hat{p}(\tau)$ trace, respectively, the quantity and price components. Without loss of generality we may take $\hat{q}(0) = \hat{p}(0) = 0$ and $\hat{p}(T) \leq p_M$, where p_M is a bound on the price of any offer. It is quite common for electricity markets to have a cap on prices; for example, this is \$10,000 per MWh in Australia. We also assume that q_M is a bound on the generation capacity of generator A, and thus $\hat{q}(T) \leq q_M$.

In all markets there are restrictions on the form of offers made into the market; we have already mentioned the need for offers to consist of step functions in many

cases. But in this paper we will not include any constraints on the form of offers. Our perspective is that a generator which has a specific optimal offer curve will usually be able to approximate this within the rules of the market. Owners of generators will often be offering power from more than one generation set in a coordinated way, and this can also increase their flexibility.

We use a single market distribution function $\psi(q,p)$ to describe the uncertainty in the market. Following Anderson and Philpott [2], $\psi(q,p)$ is defined as the probability of generator A not being fully dispatched if it offers an amount of generation q at a price p. Different generators will have different market distribution functions, but we just write ψ rather than ψ_A for this function. It turns out that when $\psi(q,p)$ is continuous, knowledge of the single function ψ is enough to determine the expected profit for a generator. When ψ is continuous, Anderson and Philpott [2] have demonstrated that the expected profit if a generator offers in a supply curve $\mathbf{s}$ can be expressed as a Stieltjes integral along the line $\mathbf{s}$:

$$v(\mathbf{s}) = \int_{\mathbf{s}} R(q,p)\ d\psi(q,p). \tag{2.2}$$

The generator's aim is to choose an optimal supply curve $\mathbf{s}$ so that $v(\mathbf{s})$ is maximized. Note that the market distribution function ψ is assumed to be known. Anderson and Philpott [3] have proposed a Bayesian inference method to estimate ψ given data on the market behavior in previous days. Note that although the setting is a stochastic one, this formulation of the problem of maximizing expected profit has converted the objective function into a deterministic optimization problem.

When the function ψ is discontinuous we need a different form of the fundamental relationship (2.2), and this will be derived in Theorem 2.9 below.

2.1. Discontinuous ψ function. Previous work in this area has assumed that the market distribution function is continuous. In this paper, rather than requiring ψ to be continuous, we assume that ψ may be discontinuous at a finite number of prices. Since ψ is a type of probability distribution function, a discontinuity in its value corresponds to a single price at which there is a jump in the probability of being fully dispatched. For this to occur two things have to happen. First some other generator has to make an offer which contains a "step," a distinct tranche of energy at a given price, and second this price has to be determined in advance (in other words, it cannot be drawn from a continuous distribution). The first condition may be met because of market rules which only allow step function offers, but for a discontinuity in ψ it is also necessary to be able to predict the prices at which other generators make offers.

We illustrate this with an example.

Example 2.1. Suppose that just two generators A and B are offering power into the spot market. Generator B is nonstrategic: its offer does not vary and is known in advance from previous market data. Thus the only uncertainty is in relation to the level of demand. Suppose that the generator B offers 200 MW at a price of \$10 per MWh, 300 MW at a price of \$14, and 300 MW at a price of \$18. Thus generator B's supply function is

$$S_B(p) = \begin{cases} 0 & \text{for } 0 \le p < 10, \\ 200 & \text{for } 10 \le p < 14, \\ 500 & \text{for } 14 \le p < 18, \\ 800 & \text{for } p \ge 18. \end{cases}$$

Consider generator A offering 100 MW at price \$10 per MWh. We suppose that demand, which can be a function of price, is uncertain. If the market clears at price \$10 with a total demand of 300 MW, then all the power offered at this price is dispatched. However, if the demand is below 300 MW at price \$10, then market rules will impose some sharing of dispatch should the market clear at this price. Suppose that the market rules share dispatch proportionately to the quantity offered at that price, so that one third of demand is met from generator A and two thirds from generator B. Thus neither of the generators gets fully dispatched at price \$10. On the other hand, if generator A offers 100 MW at price $\$10 - \varepsilon$, where $\varepsilon > 0$ is small, then it is fully dispatched provided that the demand at price \$10 is greater than or equal to 100 MW. Therefore the probability of not being fully dispatched if generator A offers at price \$10 with a quantity of 100 MW is strictly greater than the probability of not being fully dispatched if A offers at price $\$10 - \varepsilon$ with a quantity of 100 MW; i.e.,

$$\lim_{\epsilon \downarrow 0} \psi(100, 10 - \varepsilon) < \psi(100, 10).$$

This example motivates us to consider discontinuities in the functions $S_B(p)$. We write $\mathcal{P}$ for the entire set of prices at which the other generators may make significant offers, and hence at which there may be discontinuities in $S_B(p)$. Let $\mathcal{P} \equiv \{p^1, \ldots, p^n\}$, where $0 < p^1 < \cdots p^n \leq p_M$. We assume that the prices in $\mathcal{P}$ are known in advance.

For clarity we write $\omega_1 \in \Omega_1$ for the realizations of the demand, and $\omega_2 \in \Omega_2$ for realizations of the other generator offers. More formally, we assume a probability space $(\Omega_1 \times \Omega_2, \mathcal{F}, \Pr)$. The demand need not be independent of other generator offers. We shall assume that for every realization ω_1, the demand, $D(p, \omega_1)$, is a continuously differentiable decreasing function of p. Moreover we assume that for every realization ω_2, the total of the other generator offers, $S_B(p, \omega_2)$, is a continuously differentiable increasing function of p except at points in $\mathcal{P}$. We will normally omit the explicit dependence on ω_1 and ω_2, and write $D(p)$ and $S_B(p)$.

Observe that in any realization (of demand and other generator offers) for which $D(p) < q + \lim_{\varepsilon \downarrow 0} S_B(p - \varepsilon)$, an offer of q at price p cannot be fully dispatched (since if it were fully dispatched, then the price is at least p and so the other generators would be dispatched at least $\lim_{\varepsilon \to 0} S_B(p - \varepsilon)$, giving a contradiction). Hence

$$\psi(q, p) \geq \Pr(D(p) < q + \lim_{\varepsilon \downarrow 0} S_B(p - \varepsilon)). \tag{2.3}$$

If $p \notin \mathcal{P}$, then for every realization of other generator offers $\lim_{\varepsilon \downarrow 0} S_B(p - \varepsilon) = S_B(p)$ and thus $\psi(q, p) \geq \Pr(D(p) < q + S_B(p))$.

On the other hand, in any realization for which an offer of q at price p is not fully dispatched we can show that $D(p) < q + \lim_{\varepsilon \downarrow 0} S_B(p + \varepsilon)$ (since in this case the clearing price is p or less, and so the maximum quantity dispatched from the other generators is $\lim_{\varepsilon \downarrow 0} S_B(p + \varepsilon)$). Thus

$$\psi(q, p) \leq \Pr(D(p) < q + \lim_{\varepsilon \downarrow 0} S_B(p + \varepsilon)) \tag{2.4}$$

and $\psi(q, p) \leq \Pr(D(p) < q + S_B(p))$ when $p \notin \mathcal{P}$. Hence, except at points of discontinuity in S_B,

$$\psi(q, p) = \Pr(D(p) < q + S_B(p)). \tag{2.5}$$

We may use this as a definition of $\psi(q,p)$ for $p \notin \mathcal{P}$, but for $p \in \mathcal{P}$ the value of ψ depends on the sharing rule.

Since in any realization of demand and other generator offers in which $D(p) < q + S_B(p)$ the same inequality holds for any higher value of p or q, we can deduce that $\psi(q,p)$ is increasing in both its arguments at prices $p \notin \mathcal{P}$. Moreover we can use (2.3) and (2.4) to show that $\psi(q,p)$ is also increasing in p at prices $p \in \mathcal{P}$.

Note that since ψ is monotonic increasing in p and bounded, the two limits $\lim_{\delta \downarrow 0} \psi(q, p^j + \delta)$ and $\lim_{\delta \downarrow 0} \psi(q, p^j - \delta)$ will both exist. For convenience, we will use the following notation: for $j = 1, \ldots, n$,

$$\psi_+(q,p^j) = \lim_{\delta \downarrow 0} \psi(q, p^j + \delta),$$

$$\psi_-(q,p^j) = \lim_{\delta \downarrow 0} \psi(q, p^j - \delta),$$

$$\Phi(q,p^j) \equiv \psi_+(q,p^j) - \psi_-(q,p^j). \tag{2.6}$$

Thus $\Phi(q,p^j)$ is the jump in the probability of dispatch that takes place if the generator offers an amount q at a price just below p^j in comparison with what happens if the price is increased to be just above p^j.

It is important to consider the expected return for a generator offering a curve $\mathbf{s}$ when the market distribution function is discontinuous: in the continuous case we have the expression (2.2). In general we would expect to have, in addition to an integral, a sum of discrete values $R(q,p)$ at points (q,p) on $\mathbf{s}$ at which there is a jump in the value of ψ. This is indeed what happens when the curve $\mathbf{s}$ is strictly increasing. We will show later that if we define $q^j(\mathbf{s})$ as the point at which the curve $\mathbf{s}$ crosses the discontinuity p^j, then

$$v(\mathbf{s}) = \int_{\mathbf{s}^C} R(q,p)\, d\psi(q,p) + \sum_{j=1}^{n} R(q^j(\mathbf{s}), p^j) \Phi(q^j(\mathbf{s}), p^j), \tag{2.7}$$

where $\mathbf{s}^C$ is the part of curve $\mathbf{s}$ excluding the points $(q^j(\mathbf{s}), p^j)$. However, when the curve $\mathbf{s}$ has a horizontal section at one of the prices p^j, things are more complex.

2.2. Sharing rules. If we suppose that the generator is offering power at the same price p^j as another generator, then we cannot calculate the expected profit without knowledge of the market rules concerning the sharing of dispatch between two generators offering at the same price. Moreover, the value of ψ at p^j gives just the probability of complete dispatch, whereas the sharing rules imply more information than this. Specifically the values of ψ might not be enough to determine a generator's expected profit. It may be that two different sharing rules give the same ψ values but different expected profit. To illustrate this we return to Example 2.1.

Example 2.2. Suppose as before that generator B offers 200 MW at price \$10 and 300 MW at \$14, while generator A offers 100 MW at \$10. Suppose that sharing of dispatch between two generators offering at the same price is proportional to the offers made at that price. Suppose now that generator A has costs of \$8 per MWh and demand is uniformly distributed between 0 MW and 500 MW. Thus with probability 0.4 demand is greater than 300 MW, the market clears at \$14, and the profit to generator A is \$600 per hour. On the other hand, with probability 0.6 the market will clear at \$10 and generator A will be only partially dispatched. It is not hard to see that the total expected profit per hour is given by

$$v = 0.6 \int_0^{100} 2\frac{x}{100} dx + 0.4 \times 600 = 300.$$

Consider now a sharing rule which gives priority to generator B. In this case with probability 0.4 the demand is less than 200 MW and generator A is not dispatched at all (while generator B is partially dispatched). We have the following expression for expected profit:

$$v = 0.2 \int_0^{100} 2\frac{x}{100} dx + 0.4 \times 600 = 260.$$

Notice that in this second case, too, generator A is not fully dispatched unless demand is greater than 300 MW. So both these sharing rules have the same value for $\psi(100, 10)$, which is the probability of generator A not being fully dispatched with this offer. Indeed the two rules will give the same value of $\psi(q, 10)$ for any value of q.

In order to make further progress we need to consider specific sharing rules. We will write $v(\mathbf{s},\mathcal{L})$ for the expected profit when an offer curve of $\mathbf{s}$ is used together with market sharing rules defined by $\mathcal{L}$. We will investigate the particular choice of sharing rule which is best for generator A.

Suppose that generator A uses the supply function $S_A(p)$ and the other generators use the supply function $S_B(p)$. We write $S_B(p_-)$ for the limit $\lim_{\varepsilon \downarrow 0} S_B(p - \varepsilon)$ and $S_A(p_-)$ for the limit $\lim_{\varepsilon \downarrow 0} S_A(p - \varepsilon)$.

We are interested in the sharing rule to be applied when the market clears at price p^j. The market clears at this price if and only if $D(p^j)$ satisfies

$$S_A(p^j_-) + S_B(p^j_-) \le D(p^j) \le S_A(p^j) + S_B(p^j). \tag{2.8}$$

A sharing rule $\mathcal{L}$ is any method for determining the dispatch quantity $\gamma_A(\mathcal{L})$ for generator A in this case. Though this is not made explicit in the notation, the sharing rule is applied at a particular price p^j; and in general we need to define a sharing rule for each price $p \in \mathcal{P}$. Notice that $\gamma_A(\mathcal{L})$ is a function of the demand, but we suppress this dependence in the notation.

A feasible sharing rule has to satisfy the following inequalities:

$$S_A(p^j_-) \le \gamma_A(\mathcal{L}) \le S_A(p^j), \tag{2.9}$$

$$S_B(p^j_-) \le D(p^j) - \gamma_A(\mathcal{L}) \le S_B(p^j). \tag{2.10}$$

The right-hand inequalities correspond to the restriction that no generator can be dispatched more than it offers at price p^j. The left-hand inequalities correspond to the restriction that any power offered at prices less than p^j must be completely dispatched.

More generally, we can make the following definition.

DEFINITION 2.3. *Let the market clear at price $p \in (0, p^M)$, and thus*

$$S_A(p_-) + S_B(p_-) \le D(p) \le S_A(p) + S_B(p).$$

Then $\mathcal{L}$ is a feasible sharing rule if it determines uniquely the respective dispatch quantities $\gamma_A \in [S_A(p_-), S_A(p)]$ for generator A, and $\gamma_B \in [S_B(p_-), S_B(p)]$ for the other generators, such that

$$\gamma_A + \gamma_B = D(p).$$

Notice that unless two generators both offer power at the price p there will only be one possible choice for γ_A and γ_B. Thus the only prices at which the sharing rule needs to be defined are p^j, $j = 1, 2, \ldots, n$.

We let $q^*(p^j)$ be the value of q at which $R(q,p^j)$ achieves its maximum over $[0,q_M]$. Our assumptions on R imply that this is unique. We have $q^*(p^j)=0$ if $R_q(0,p^j)<0$, $q^*(p^j)=q_M$ if $R_q(q_M,p^j)>0$, and $R_q(q^*(p^j),p^j)=0$ otherwise. Notice that $q^*(p^j)$ is not affected by any hedging contracts.

Now we define the sharing rule $\mathcal{L}^*$ as follows. Let

$$q^j=\begin{cases} S_A(p^j) & \text{if } S_A(p^j)<q^*(p^j),\\ q^*(p^j) & \text{if } S_A(p^j_-)\le q^*(p^j)\le S_A(p^j),\\ S_A(p^j_-) & \text{if } q^*(p^j)<S_A(p^j_-).\end{cases}$$

It is easy to see that q^j maximizes $R(q,p^j)$ subject to $S_A(p^j_-)\le q\le S_A(p^j)$. Since we also require that (2.10) be satisfied, we define the dispatch amount γ_A from generator A under $\mathcal{L}^*$ (when the price is p^j) as

$$\gamma_A(\mathcal{L}^*)=\begin{cases} (a) & D(p^j)-S_B(p^j_-) & \text{if } D(p^j)-S_B(p^j_-)<q^j,\\ (b) & D(p^j)-S_B(p^j) & \text{if } D(p^j)-S_B(p^j)>q^j,\\ (c) & q^j & \text{otherwise.}\end{cases}$$

The lemma below demonstrates that $\mathcal{L}^*$ is the *best* choice of sharing rule for generator A, in the sense that no other sharing rule will produce such a large profit for A.

LEMMA 2.4. *$\mathcal{L}^*$ is a feasible sharing rule, and $v(\mathbf{s},\mathcal{L}^*)\ge v(\mathbf{s},\mathcal{L})$ for every feasible sharing rule $\mathcal{L}$.*

Proof. We consider the profit when the clearing price is $p^j\in\mathcal{P}$, since if the clearing price is not in $\mathcal{P}$ no sharing rule will be needed. We write I_A for the interval $[S_A(p^j_-),S_A(p^j)]$ and I_B for the interval $[D(p^j)-S_B(p^j),D(p^j)-S_B(p^j_-)]$. The length of interval I_A is the offer from A at price p^j, while I_B is the range of possible residual demand for A at this price. Therefore a feasible sharing rule has γ_A in both I_A and I_B.

We wish to establish that γ_A is the unique optimal solution to

$$\max_{q\in I_A\cap I_B} R(q,p^j). \tag{2.11}$$

From (2.8) we observe that I_A and I_B will overlap, so the feasible set for the maximization problem is nonempty. Observe also that q^j is the unique optimal solution to the problem

$$\max_{q\in I_A} R(q,p^j),$$

which implies that, within interval I_A, $R(\cdot,p^j)$ is strictly increasing for $q\le q^j$ and strictly decreasing for $q>q^j$.

We consider the three cases in the definition of $\gamma_A(\mathcal{L}^*)$. In case (a) q^j falls to the right of the interval I_B, hence the right end point of I_B, $D(p^j)-S_B(p^j_-)$, is the optimal solution of (2.11). Similarly in case (b) q^j falls to the left of the interval I_B, and hence the left-hand end point of interval I_B, $D(p^j)-S_B(p^j)$, is the optimal solution of (2.11). In case (c) q^j is in I_B and is therefore the optimal solution of (2.11). This shows $\gamma_A(\mathcal{L}^*)$ is the optimal solution of (2.11). ∎

2.3. *R*-semicontinuous ψ function. We need to define a specific type of discontinuity behavior for the function ψ. In fact, at some points in the (q,p) plane we

need ψ to be continuous from above, and at other points to be continuous from below, depending on the characteristics of the function R.

DEFINITION 2.5. *Suppose that the market distribution function $\psi(q,p)$ is continuous at all prices $p \notin \mathcal{P}$. ψ is called R-semicontinuous if $\psi_-(q,p^j) = \psi(q,p^j)$ when $R_q(q,p^j) \geq 0$ and $\psi_+(q,p^j) = \psi(q,p^j)$ when $R_q(q,p^j) < 0$, $j = 1,\ldots,n$.*

With the form of profit function given in (2.1) we can see that an R-semicontinuous market distribution function will have the property of being continuous from below (in the (q,p) plane) when $p > C'(q)$, and will be continuous from above when the reverse inequality holds. We will show that ψ will be R-semicontinuous when the sharing rule $\mathcal{L}^*$ is applied.

Though we have assumed that both demand and other generator offers are well behaved in any given realization, we also need to have the realizations of demand and offers in some sense continuously distributed through the appropriate spaces.

Assumption 2.6 (continuity). The function $q \longmapsto \Pr(D(p) < q + S_B(p))$ is continuous on $[0, q^M]$, and the function $q \longmapsto \Pr(D(p) < q + S_B(p))$ is continuous on $[0, p^M] \backslash \mathcal{P}$.

This assumption implies, from (2.5), that $\psi(q,p)$ is continuous at all prices $p \notin \mathcal{P}$.

LEMMA 2.7. *Under Assumption* 2.6*, if the sharing rule $\mathcal{L}^*$ is used, then the market distribution function ψ is R-semicontinuous.*

Proof. We consider an offer of an amount q by generator A at a price p^j, where $\psi(q,\cdot)$ is discontinuous. We suppose that there is no other offer by generator A. Thus $S_A(p^j) = q$ and $S_A(p^j_-) = 0$.

Suppose first that $R_q(q,p^j) \geq 0$, so $q \leq q^*(p^j)$. From the definition of q^j, we have $q^j = q$, thus $\mathcal{L}^*$ will choose to dispatch an amount q, if this is possible when the constraints due to the demand realization are considered.

Recall that $\psi(q,p^j)$ is defined as the probability of not being fully dispatched when A makes an offer of q at p^j. Under $\mathcal{L}^*$, the probability of not being fully dispatched is the probability of either the market clearing at a price below p^j or clearing at price p^j but $D(p^j) - S_B(p^j_-) < q$, which means generator B's offer at p^j is not dispatched at all, and the residual demand for generator A falls below q. In this case A gets dispatched $D(p^j) - S_B(p^j_-)$. This is exactly the case (a) in the definition of $\gamma_A(\mathcal{L}^*)$.

Define the event

$$H = \{(\omega_1,\omega_2) : D(p^j,\omega_1) < q + S_B(p^j_-,\omega_2)\}.$$

Then

$$\psi(q,p^j) = \Pr(H).$$

In what follows, we show $\psi(q,\cdot)$ is continuous at $p = p^j$ from below. We write G_ε for the event that an offer of q at price $p^j - \varepsilon$ is not fully dispatched; i.e.,

$$G_\varepsilon = \{(\omega_1,\omega_2) : D(p^j - \varepsilon,\omega_1) < q + S_B(p^j - \varepsilon,\omega_2)\}.$$

Then the G_ε are monotonically increasing sets as ε decreases to zero, with, say, a limit G. D is a continuous function of p in each realization, and so if $(\omega_1,\omega_2) \in H$, then for some choice of $\varepsilon_0 > 0$, $D(p^j - \varepsilon,\omega_1) < q + S_B(p^j - \varepsilon,\omega_2)$ for $0 < \varepsilon < \varepsilon_0$. Therefore $H \subset G$. From the axioms of probability, $\Pr(G) = \lim_{\varepsilon\to 0} \Pr(G_\varepsilon)$. Thus $\psi(q,p^j) \leq \lim_{\varepsilon\to 0} \psi(q,p^j - \varepsilon)$. But since ψ is increasing in p, there must be equality here, i.e., $\psi_-(q,p^j) = \psi(q,p^j)$, as required.

Now consider the case that $R_q(q,p^j) < 0$ and we show $\psi(q,\cdot)$ is continuous at $p = p^j$ from above.

In this case $q > q^*(p^j)$. The probability of not being fully dispatched under $\mathcal{L}^*$ is the probability of either the market clearing at a price below p^j or of clearing at p^j with $D(p^j) - S_B(p^j) < q$. Thus $\psi(q, p^j) = \Pr(J)$ where

$$J = \{(\omega_1, \omega_2) : D(p^j, \omega_1) < q + S_B(p^j, \omega_2)\}.$$

We write F_ε for the event that an offer of q at price $p^j + \varepsilon$ is not fully dispatched. Then F_ε is monotonically decreasing as ε decreases to zero, with a limit F, say. So every realization in F is in every F_ε for $\varepsilon < \varepsilon_0$, where ε_0 depends on the realization; i.e., for every $(\omega_1, \omega_2) \in F$, $D(p^j + \varepsilon, \omega_1) < q + S_B(p^j + \varepsilon, \omega_2)$ for $\varepsilon > 0$. Thus from the continuity of D and S_B, $F \subset \{(\omega_1, \omega_2) : D(p^j, \omega_1) \le q + S_B(p^j, \omega_2)\}$. Thus from Assumption 2.6

$$\begin{aligned}\Pr(F) &\le \Pr((\omega_1, \omega_2) : D(p^j, \omega_1) \le q + S_B(p^j, \omega_2)) \\ &= \Pr((\omega_1, \omega_2) : D(p^j, \omega_1) < q + S_B(p^j, \omega_2)) = \psi(q, p^j).\end{aligned}$$

Now, since $\lim_{\varepsilon \to 0} \Pr(F_\varepsilon) = \Pr(F)$, we have established that $\psi_+(q, p^j) \le \psi(q, p^j)$, and the monotonicity of ψ shows that these are equal. □

The reverse implication does not hold: we can have an R-semicontinuous ψ without using the sharing rule $\mathcal{L}^*$.

2.4. Expected profit. We let $\Psi = \{(q, p) : 0 < \psi(q, p) < 1\}$. In line with Definition 2.5, we can divide Ψ into two regions Ψ_+ and Ψ_-, where

$$\Psi_+ = \{(q, p) \in \Psi : R_q(q, p) \ge 0\}, \qquad \Psi_- = \{(q, p) \in \Psi : R_q(q, p) < 0\}.$$

In the case that ψ is R-semicontinuous, to calculate the expected profit from a supply curve $\mathbf{s}$ when it has a segment on the horizontal line $\{(q, p^j) : q \in [0, q^M]\}$, we need to think of it as part of the region below that line in the set Ψ_+ and part of the region above that line in the set Ψ_-. This motivates the following definitions:

$$\begin{aligned}\Psi^j &= \{(q, p) \in \Psi : 0 \le q \le q_M,\ p^{j-1} < p < p^j\} \cup \{(q, p^j) \in \Psi_+\} \\ &\quad \cup \{(q, p^{j-1}) \in \Psi_-\}, \qquad j = 2, \dots, n, \\ \Psi^1 &= \{(q, p) \in \Psi : 0 \le q \le q_M,\ 0 \le p < p^1\} \cup \{(q, p^1) \in \Psi_+\}, \\ \Psi^{n+1} &= \{(q, p) \in \Psi : 0 \le q \le q_M,\ p^n < p \le p^M\} \cup \{(q, p^n) \in \Psi_-\}, \\ \mathbf{s}^j &= \mathbf{s} \cap \Psi^j.\end{aligned}$$

It is not hard to see that the values q^j which we introduced in relation to the sharing rule $\mathcal{L}^*$ also define the points at which an offer curve $\mathbf{s}$ moves from Ψ^j to Ψ^{j+1}. Thus, writing q^j as a function of $\mathbf{s}$,

$$q^j(\mathbf{s}) = \sup\{q : (q, p^j) \in \mathbf{s}^j\}$$

for $j = 1, \dots, n$. From the monotonicity of the offer curve $\mathbf{s}$, and because R_q is decreasing, we can also write

$$q^j(\mathbf{s}) = \inf\{q : (q, p^j) \in \mathbf{s}^{j+1}\}.$$

Under the assumptions of Lemma 2.7, we can take ψ as R-semicontinuous and made up from a number of different pieces ψ^j, where ψ^j is defined on the interval between p^{j-1} and p^j and is well behaved on that interval. Thus we let

$$\psi^j(q, p) = \begin{cases} \psi_+(q, p^{j-1}) & \text{for } p = p^{j-1}, \\ \psi(q, p) & \text{for } p^{j-1} < p < p^j, \\ \psi_-(q, p^j) & \text{for } p = p^j \end{cases}$$

for $j = 2, \ldots, n$, and

$$\psi^1(q,p) = \begin{cases} \psi(q,p) & \text{for } 0 \le p < p^1, \\ \psi_-(q,p^1) & \text{for } p = p^1, \end{cases}$$

$$\psi^{n+1}(q,p) = \begin{cases} \psi_+(q,p^n) & \text{for } p = p^n, \\ \psi(q,p) & \text{for } p^n < p \le p^M. \end{cases}$$

We need to make an assumption on the behavior of the function ψ.

Assumption 2.8 (continuous differentiability). $\psi(q,p)$ is continuously differentiable for $p \notin \mathcal{P}$ and each ψ^j can be extended to a continuously differentiable function on an open set W^j which contains the closure of the set Ψ^j.

THEOREM 2.9. *Suppose that a generator offers a curve* $\mathbf{s}$ *and Assumptions* 2.6 *and* 2.8 *are satisfied. If sharing rule* $\mathcal{L}^*$ *is used, then the expected profit for the generator is*

$$v(\mathbf{s}) = \sum_{j=1}^{n+1} \int_{\mathbf{s}^j} R(q,p)\ d\psi^j(q,p) + \sum_{j=1}^{n} R(q^j(\mathbf{s}),p^j)\Phi(q^j(\mathbf{s}),p^j), \tag{2.12}$$

where Φ *is defined in* (2.6).

Proof. To simplify our presentation we prove the theorem for $n = 1$ with just one price discontinuity at p^1. The case with $n > 1$ can be dealt with similarly. We suppose that generator A uses an offer curve $\mathbf{s}$ which we take as $\mathbf{s} = \{(x(\tau), y(\tau))\}$ in parameter form.

We write γ_A for the dispatch quantity from generator A given the offer curve $\mathbf{s}$. We start by showing that $\psi(x(\tau), y(\tau))$ is the probability that γ_A is less than $x(\tau)$. In the case that ψ is continuous in a neighborhood of $(x(\tau), y(\tau))$, this is straightforward and is implicitly established in [2]. But when $y(\tau) = p^1$ we need to be more careful. Observe that from the definition of $\mathcal{L}^*$, if $x(\tau) < q^1$, then

$$\Pr(\gamma_A < x(\tau)) = \Pr(D(p^1) - S_B(p^1_-) < x(\tau)).$$

But the probability of an offer of $x(\tau)$ at price p^1 is not fully dispatched under $\mathcal{L}^*$ with the same probability. Hence $\Pr(\gamma_A < x(\tau)) = \psi(x(\tau), y(\tau))$ as required. The case when $x(\tau) \ge q^1$ can be dealt with similarly.

We let τ^1 be such that $y(\tau^1) = p^1$ and $x(\tau^1) = q^1$. We consider the expected profit on a segment, $\mathbf{s}_\delta \equiv \{(x(\tau), y(\tau)) : \tau^1 - \delta < \tau \le \tau^1 + \delta\}$, of curve $\mathbf{s}$. From our observation on $\psi(x(\tau), y(\tau))$ we know that the probability that the market clears at a price p and quantity q on the offer curve in the segment $\mathbf{s}_\delta$ is given by $\psi(x(\tau^1 + \delta), y(\tau^1 + \delta)) - \psi(x(\tau^1 - \delta), y(\tau^1 - \delta))$.

The expected profit from the line segment $\mathbf{s}_\delta$ is bounded above (below) by this probability multiplied by the supremum (infimum) of R over the set $\mathbf{s}_\delta$. Since R is continuously differentiable, for δ small enough, the expected profit from segment $\mathbf{s}_\delta$ is

$$v(\mathbf{s}_\delta) = R(x(\tau^1), y(\tau^1))(\psi(x(\tau^1 + \delta), y(\tau^1 + \delta)) - \psi(x(\tau^1 - \delta), y(\tau^1 - \delta))) + o(\delta).$$

The total expected profit from the offer curve $\mathbf{s}$ can be written as $v(\mathbf{s}) = v(\mathbf{s}^1_\delta) + v(\mathbf{s}_\delta) + v(\mathbf{s}^2_\delta)$, where $\mathbf{s}^1_\delta$, $\mathbf{s}^2_\delta$ are the other components of $\mathbf{s}$ created when $\mathbf{s}_\delta$ is removed. So $\mathbf{s}^1_\delta = \{(x(\tau), y(\tau)) : \tau \le \tau^1 - \delta\}$ and $\mathbf{s}^2_\delta = \{(x(\tau), y(\tau)) : \tau \ge \tau^1 + \delta\}$. These components lie entirely within the regions where ψ is continuously differentiable (and

given by ψ^i), using Assumption 2.8. Using the result of Anderson and Philpott [2] we know that

$$v(\mathbf{s}^i_\delta) = \int_{\mathbf{s}^i_\delta} R(q,p)\ d\psi^i(q,p), \qquad i = 1,2.$$

By driving δ to zero, we have

$$\begin{aligned}\lim_{\delta\to 0} v(\mathbf{s}_\delta) &= R(x(\tau^1),y(\tau^1))(\psi^2(x(\tau^1),y(\tau^1)) - \psi^1(x(\tau^1),y(\tau^1)))\\ &= R(q^1,p^1)\Phi(q^1,p^1),\end{aligned}$$

and

$$\lim_{\delta\to 0}\int_{\mathbf{s}^i_\delta} R(q,p)\ d\psi^i(q,p) = \int_{\mathbf{s}^i} R(q,p)\ d\psi^i(q,p), \qquad i = 1,2.$$

This completes the proof. □

2.5. Existence of an optimal solution. Having established the objective function formula (2.12), our approach to showing that an optimal solution exists is to concentrate on the formal problem of maximizing (2.12) given an R-semicontinuous market distribution function ψ.

In order to discuss the optimality of a continuous offer curve, we need to compare the line integrals on two distinct curves. When ψ is continuous, Anderson and Philpott [2] use Green's theorem and observe that

$$\int\int_{\mathcal{S}} Z(q,p)\ dpdq = \int_{\mathcal{C}} R(q,p)\ d\psi(q,p),$$

where $\mathcal{S}$ is a region enclosed by a curve $\mathcal{C}$ and

$$Z(q,p) = \begin{cases} R_q\psi_p - R_p\psi_q, & (q,p)\in\Psi,\\ 0 & \text{otherwise.}\end{cases} \tag{2.13}$$

Clearly this result will not hold when the curve $\mathcal{C}$ crosses one of the lines of discontinuity at $p\in\mathcal{P}$.

Our approach will be to calculate the change in v that arises from a change in offer curve $\mathbf{s}$ by applying Green's theorem separately to each region Ψ^j together with a calculation of the change that arises across the lines of discontinuity. We need to start with a lemma that can be established using an integration by parts argument.

LEMMA 2.10. *Suppose $p^j\in\mathcal{P}$ and $0\le q_1<q_2\le q_M$. Then, under Assumption 2.8,*

$$\begin{aligned}&\int_{q_1}^{q_2} R(q,p^j)\ d\psi^j(q,p^j) - \int_{q_1}^{q_2} R(q,p^j)\ d\psi^{j+1}(q,p^j) + R(q_2,p^j)\Phi(q_2,p^j)\\ &\qquad -R(q_1,p^j)\Phi(q_1,p^j)\\ &= \int_{q_1}^{q_2}\Phi(q,p^j)R_q(q,p^j)\ dq.\end{aligned}$$

Proof. Let

$$v_1 = \int_{q_1}^{q_2} R(q,p^j)\ d\psi^j(q,p^j) + R(q_2,p^j)\Phi(q_2,p^j)$$

and

$$v_2 = \int_{q_1}^{q_2} R(q, p^j)\ d\psi^{j+1}(q, p^j) + R(q_1, p^j)\Phi(q_1, p^j).$$

From Assumption 2.8, both $\psi^j(\cdot, p^j)$ and $\psi^{j+1}(\cdot, p^j)$ are continuously differentiable, and we have already assumed that $R(\cdot, p^j)$ is continuously differentiable. Integrating both v_1 and v_2 by parts, we obtain

$$\begin{aligned} v_1 - v_2 &= R(q_2, p^j)\psi^j(q_2, p^j) - R(q_1, p^j)\psi^j(q_1, p^j) - \int_{q_1}^{q_2} \psi^j(q, p^j)R_q(q, p^j)\ dq \\ &\quad - R(q_2, p^j)\psi^{j+1}(q_2, p^j) + R(q_1, p^j)\psi^{j+1}(q_1, p^j) + \int_{q_1}^{q_2} \psi^{j+1}(q, p^j)R_q(q, p^j)\ dq \\ &\quad + R(q_2, p^j)\Phi(q_2, p^j) - R(q_1, p^j)\Phi(q_1, p^j) \\ &= \int_{q_1}^{q_2} \Phi(q, p^j)R_q(q, p^j)\ dq, \end{aligned}$$

as required. □

Note that v_1 is the expected return of the generator for offering $q_2 - q_1$ at price just under p^j, and v_2 is the expected return of the generator for offering $q_2 - q_1$ at price just above p^j. The lemma states that the difference between these two values can be expressed as the integral of $\Phi(q, p^j)R_q(q, p^j)$ with respect to q from q_1 to q_2. Since R_q is the marginal profit, $\Phi(q, p^j)R_q(q, p^j)$ represents the difference of the marginal profits between the offer of q at just above p^j and the offer of q at just below p^j.

Anderson and Philpott [2] and Anderson and Xu [4] treat $v(\mathbf{s})$ in (2.2) as an objective function and investigate the necessary and sufficient conditions for an offer curve $\mathbf{s}$ to be a local maximum. When ψ is continuously differentiable on Ψ, Anderson and Xu prove that there exists a maximum over the set of curves that are considered. However, the existence result is not straightforward when ψ is not continuous, and our first result is to confirm that a maximum does exist provided that ψ satisfies the conditions we have given.

A generator need not offer all its generation capacity into the market; the offer curve will start at some point $(0, \hat{p}(0))$ and finish at $(\hat{q}(T), \hat{p}(T))$. However, the clearing price is determined as though the offer curve began with a vertical segment from the origin to $(0, \hat{p}(0))$ and finished with a vertical segment from $(\hat{q}(T), \hat{p}(T))$ to $(\hat{q}(T), p_M)$. Hence we assume that Λ, the set of possible offer curves, has these characteristics. The following result has been established by Anderson and Xu [4].

LEMMA 2.11. *Let Λ be the set of monotonic continuous curves starting at the origin and ending on the closed line segment from $(0, p_M)$ to (q_M, p_M). Then Λ is compact under the Hausdorff metric:*

$$|\mathbf{s}_1 - \mathbf{s}_2|_H = \max_{(q_1,p_1)\in\mathbf{s}_1} \min_{(q_2,p_2)\in\mathbf{s}_2} \sqrt{(q_1 - q_2)^2 + (p_1 - p_2)^2}.$$

We need some sort of compactness result such as this to ensure the existence of an optimal solution; once compactness is established in some topology, then the existence result follows provided that we have a suitable continuity property in that topology. Our next result uses compactness to establish that the problem of finding a curve $\mathbf{s}$ which maximizes the expected profit $v(\mathbf{s})$ in (2.12) has an optimal solution. But before we prove this theorem we need to establish a preliminary lemma (which is required because $q^j(\mathbf{s})$ is not a continuous function of $\mathbf{s}$).

LEMMA 2.12. *If* $\mathbf{s}_k \to \mathbf{s}$ *in the Hausdorff metric and for some* j: $\lim_{k\to\infty} q^j(\mathbf{s}_k) = q_0$, *then*

$$\int_{q^j(\mathbf{s})}^{q_0} \Phi(q,p^j)R_q(q,p^j)\ dq \le 0. \tag{2.14}$$

Proof. Observe that if $q^j(\mathbf{s}) = q_0$ there is nothing to prove, so we suppose these two are unequal. Since $\mathbf{s}_k \to \mathbf{s}$ in the Hausdorff metric we can deduce that $\min_{(q,p)\in\mathbf{s}} \left((q_0-q)^2 + (p^j-p)^2\right)^{1/2} = 0$ and hence that $(q_0,p^j) \in \mathbf{s}$. Thus, from monotonicity, all of the line interval $(q^j(\mathbf{s}),p^j)$ to (q_0,p^j) is in $\mathbf{s}$. First we suppose that $q^j(\mathbf{s}) < q_0$. Then, from the definition of q^j, this line interval lies in Ψ^{j+1} and hence is part of Ψ_-, where $R_q < 0$. Since $\Phi(q,p^j) \ge 0$, the inequality (2.14) follows. On the other hand, if $q^j(\mathbf{s}) > q_0$, then the line interval (q_0,p^j) to $(q^j(\mathbf{s}),p^j)$ lies in Ψ^j and hence is part of Ψ_+, where $R_q \ge 0$. Again, we have shown the desired inequality (2.14) after noting that the limits of the integral are reversed. ∎

THEOREM 2.13 (existence). *Let* Λ *be defined as above and let* v *be the expected return function given in* (2.12). *Under Assumptions* 2.6 *and* 2.8, *if the market distribution function is* R*-semicontinuous, then* v *achieves its maximum on* Λ.

Proof. Let $v^* = \sup_{\mathbf{s}\in\Lambda} v(\mathbf{s})$, which exists since R is bounded and ψ lies between 0 and 1. For every $k > 0$, there exists a supply curve $\mathbf{s}_k \in \Lambda$ such that $v^* - v(\mathbf{s}_k) \le \frac{1}{k}$. Since Λ is a compact set, there exists $\mathbf{s}^* \in \Lambda$ such that $|\mathbf{s}_k - \mathbf{s}^*|_H \to 0$ (we can take a subsequence if necessary). In addition we shall arrange that for each j, $\lim_{k\to\infty} q^j(\mathbf{s}_k)$ exists. We want to prove that $v(\mathbf{s}^*) = v^*$. We will do this by showing that $v(\mathbf{s}_k) \to v(\mathbf{s}^*)$, using Green's theorem on each of the Ψ^j regions together with Lemma 2.10 for the crossovers from one Ψ^j to the next.

We define $\mathbf{s}^{*j} = \mathbf{s}^* \cap \Psi^j$ and $\mathbf{s}_k^j = \mathbf{s}_k \cap \Psi^j$. Thus

$$v(\mathbf{s}^*) = \sum_{j=1}^{n+1} \int_{\mathbf{s}^{*j}} R(q,p)\ d\psi^j(q,p) + \sum_{j=1}^{n} R(q^j(\mathbf{s}^*),p^j)\Phi(q^j(\mathbf{s}^*),p^j),$$

$$v(\mathbf{s}_k) = \sum_{j=1}^{n+1} \int_{\mathbf{s}_k^j} R(q,p)\ d\psi^j(q,p) + \sum_{j=1}^{n} R(q^j(\mathbf{s}_k),p^j)\Phi(q^j(\mathbf{s}_k),p^j).$$

Let $\operatorname{sign}(q,p)$ be a function such that $\operatorname{sign}(q,p) = 1$ if (q,p) is located below the curve $\mathbf{s}^*$; $\operatorname{sign}(q,p) = -1$ if (q,p) is located above the curve $\mathbf{s}^*$; and $\operatorname{sign}(q,p) = 0$ if (q,p) is located on the curve $\mathbf{s}^*$. Now, using Green's theorem

$$\begin{aligned}
&\int_{\mathbf{s}_k^j} R(q,p)\ d\psi^j(q,p) - \int_{\mathbf{s}^{*j}} R(q,p)\ d\psi^j(q,p) \\
&= \int\int_{A_k^j} \operatorname{sign}(q,p)Z(q,p)\ dq\ dp + \int_{q^j(\mathbf{s}^*)}^{q^j(\mathbf{s}_k)} R(q,p^j)\ d\psi^j(q,p^j) \\
&\quad - \int_{q^{j-1}(\mathbf{s}^*)}^{q^{j-1}(\mathbf{s}_k)} R(q,p^{j-1})\ d\psi^j(q,p^j),
\end{aligned}$$

where A_k^j is the area between $\mathbf{s}^{*j}$ and $\mathbf{s}_k^j$, and Z is given by (2.13). Let A_k be the entire area between $\mathbf{s}^*$ and $\mathbf{s}_k$. Then

$$v(\mathbf{s}_k) - v(\mathbf{s}^*) = \int\int_{A_k} \operatorname{sign}(q,p)Z(q,p)\ dq\ dp$$

$$+\sum_{j=1}^{n}\int_{q^j(\mathbf{s}^*)}^{q^j(\mathbf{s}_k)} R(q,p^j)\ d\psi^j(q,p^j) - \sum_{j=1}^{n}\int_{q^j(\mathbf{s}^*)}^{q^j(\mathbf{s}_k)} R(q,p^j)\ d\psi^{j+1}(q,p^j)$$
$$+\sum_{j=1}^{n} R(q^j(\mathbf{s}_k),p^j)\Phi(q^j(\mathbf{s}_k),p^j) - \sum_{j=1}^{n} R(q^j(\mathbf{s}^*),p^j)\Phi(q^j(\mathbf{s}^*),p^j)$$
$$= \int\int_{A_k} \operatorname{sign}(q,p)Z(q,p)\ dq\ dp + \sum_{j=1}^{n}\int_{q^j(\mathbf{s}^*)}^{q^j(\mathbf{s}_k)} \Phi(q,p^j)R_q(q,p^j)\ dq$$

using Lemma 2.10.

Since $|\mathbf{s}_k - \mathbf{s}^*|_H \to 0$ and Z is bounded from Assumption 2.6, the area integral approaches zero as $k \to \infty$. Also from Lemma 2.12, we know that

$$\lim_{k\to\infty}\int_{q^j(\mathbf{s}^*)}^{q^j(\mathbf{s}_k)} \Phi(q,p^j)R_q(q,p^j)\ dq \le 0.$$

Thus $v(\mathbf{s}^*) \ge \lim_{k\to\infty} v(\mathbf{s}_k) = v^*$, but from the definition of v^*, $v(\mathbf{s}^*) \le v^*$, and thus we have established the desired equality. □

Using this theorem and the results of Theorem 2.9 and Lemma 2.7, we have the immediate corollary.

COROLLARY 2.14. *Under Assumptions* 2.6 *and* 2.8*, if the sharing rule* $\mathcal{L}^*$ *is used, then there is an optimal supply curve.*

3. Necessary conditions for optimality. From Theorem 2.13, we know that the problem of maximizing profit using the sharing rule $\mathcal{L}^*$ is well defined. This is equivalent to the following maximization problem:

$$\max_{\mathbf{s}\in\Lambda}\ v(\mathbf{s}) \equiv \sum_{j=1}^{n+1}\int_{\mathbf{s}^j} R(q,p)\ d\psi^j(q,p) + \sum_{j=1}^{n} R(q^j(\mathbf{s}),p^j)\Phi(q^j(\mathbf{s}),p^j). \tag{3.1}$$

In this section, we discuss necessary conditions for an offer curve to be optimal for this problem.

When ψ is continuously differentiable, optimality conditions were derived by Anderson and Philpott [2] and extended by Anderson and Xu [4]. Let $\mathbf{s} = \{(\hat{q}(\tau),\hat{p}(\tau)) : 0 \le \tau \le T\}$ be the offer curve. The main tool that is used in investigating the optimality conditions of $\mathbf{s}$ is the line integral of Z along $\mathbf{s}$, which is defined by

$$w(\tau) = \int_0^\tau Z(\hat{q}(t),\hat{p}(t))(\hat{q}'(t) + \hat{p}'(t))\ dt.$$

When ψ is not continuously differentiable, we need to use a different approach.

We take ψ as R-semicontinuous and we define, for $(q,p) \in \Psi^j$, the function $Z^j(q,p) = R_q\psi_p^j - R_p\psi_q^j$. This will make Z^j match Z in the interior of Ψ^j and be defined by continuity for points in Ψ^j that lie on its boundary.

Given a monotonic continuous, piecewise smooth offer curve $\mathbf{s} = \{(\hat{q}(\tau),\hat{p}(\tau)), 0 \le \tau \le T\}$, for each τ we let $J(\tau)$ be the index of the region Ψ^j in which $(\hat{q}(\tau),\hat{p}(\tau))$ lies and we let τ^j be the parameter value at which the curve moves from Ψ^j to Ψ^{j+1}, and thus $\hat{q}(\tau^j) = q^j$. Then we define

$$w(\tau) = \int_0^\tau Z^{J(\tau)}(\hat{q}(t),\hat{p}(t))(\hat{q}'(t) + \hat{p}'(t))\ dt + \sum_{j=1}^{J(\tau)-1} \Phi(q^j,p^j)R_q(q^j,p^j).$$

THEOREM 3.1 (first order necessary conditions). *Suppose that* $\mathbf{s} = \{\hat{q}(\tau), \hat{p}(\tau), 0 \leq \tau \leq T\}$ *is an offer curve and Assumptions* 2.6 *and* 2.8 *are satisfied. Suppose that there exist* m *numbers* $0 \leq \tau_1 < \tau_2 < \cdots < \tau_m \leq T$ *with* $0 < \hat{q}(\tau) < q_M$ *and* $0 < \hat{p}(\tau) < p_M$ *for* $\tau_1 < \tau < \tau_m$. *Suppose further that on each section* (τ_{i-1}, τ_i), $i = 2, \ldots, m$, $\mathbf{s}$ *is either strictly increasing in both components, or horizontal, or vertical, with different characteristics in successive segments and with* $\tau_1(\tau_m)$ *the smallest (largest) parameter value such that* $(\hat{q}(\tau), \hat{p}(\tau)) \in \Psi$. *If* $\mathbf{s}$ *is optimal for* (3.1), *then* $w(\tau_1) = 0$ *and* $w(\tau_m) = w(T)$. *Moreover, for each interval* I *being one of* (τ_{i-1}, τ_i), $i = 2, \ldots, m$, *one of the following holds:*

(i) $\mathbf{s}$ *is strictly increasing in both components and* $w(\tau) = w(\tau_{i-1})$ *for* $\tau \in I$.

(ii) $\mathbf{s}$ *is horizontal on* I. *For* $\tau \in I$ *with* $(\hat{q}(\tau), \hat{p}(\tau)) \in \Psi_+$, *then* $w(\tau) \leq w(\tau_{i-1})$; *for* $\tau \in I$ *with* $(\hat{q}(\tau), \hat{p}(\tau)) \in \Psi_-$, *then* $w(\tau) \leq w(\tau_i)$. *Moreover, if* $\hat{p}(\tau_i) \notin \mathcal{P}$, *then* $w(\tau_{i-1}) = w(\tau_i)$.

(iii) $\mathbf{s}$ *is vertical on* I, $w(\tau_{i-1}) = w(\tau_i)$, *and* $w(\tau) \geq w(\tau_i)$ *for* $\tau \in I$.

Proof. We begin by looking at the w values at τ_1 and τ_m. First, we prove $w(\tau_1) = 0$ (the proof that $w(\tau_m) = w(T)$ is similar). By assumption, for any $\tau < \tau_1$, $(\hat{q}(\tau), \hat{p}(\tau))$ is located outside the Ψ region where Z and $w(\tau)$ are zero. Note that if $\hat{p}(\tau_1) \notin \mathcal{P}$, then w is continuous at τ_1 and $w(\tau_1) = 0$. Thus we only need to consider the case that $\hat{p}(\tau_1) = p^j \in \mathcal{P}$. This means that the lower boundary of Ψ contains a horizontal section $p = p^j$ and the point $(\hat{q}(\tau_1), \hat{p}(\tau_1))$ is located on the horizontal section. We consider three cases according to whether the point $(\hat{q}(\tau_1), \hat{p}(\tau_1))$ is located in Ψ_+, in Ψ_-, or on the line separating these regions. In the latter case $R_q(\hat{q}(\tau_1), \hat{p}(\tau_1)) = 0$, and hence $w(\tau_1) = 0$. If $(\hat{q}(\tau_1), \hat{p}(\tau_1))$ is in Ψ_+, then $J(\tau_1) = j - 1$ and $w(\tau_1) = 0$ by the definition of the w function. Thus we are left with the case when $(\hat{q}(\tau_1), \hat{p}(\tau_1))$ is in Ψ_-, when $J(\tau_1) = j$. By definition, since Z is zero outside Ψ,

$$w(\tau_1) = R_q(\hat{q}(\tau_1), p^j)\Phi(\hat{q}(\tau_1), p^j) \leq 0.$$

Suppose for a contradiction that $w(\tau_1) < 0$. Since $\Phi(q, p^j) = \psi(q, p^j_+)$ for all q with (q, p_j) at the boundary, this implies that $\psi(\hat{q}(\tau_1), p^j_+) > 0$. By Assumption 2.8, $\psi(q, p^j_+)$ is continuous in q, and so there exists $\delta > 0$ such that $\psi(\hat{q}(\tau_1) - \delta, p^j_+) > 0$. Consider another supply curve $\mathbf{r}$ which enters Ψ at a point $(\hat{q}(\tau_1) - \delta, p^j_+)$ and then goes horizontally until it reaches the point $(\hat{q}(\tau_1), p^j_+)$ and then joins $\mathbf{s}$ to the end. Using Lemma 2.10, it is easy to verify the difference between the expected profits of the two supply curves,

$$E(\mathbf{s}) - E(\mathbf{r}) = \int_{\hat{q}(\tau_1)-\delta}^{\hat{q}(\tau_1)} R_q(x, p^j)\Phi(x, p^j)dx = \delta w(\tau_1) + o(\delta) < 0,$$

for δ sufficiently small. This contradicts the optimality of $\mathbf{s}$ and establishes $w(\tau_1) = 0$.

Part (i). This part of the theorem amounts to the statement that if $\hat{q}(\tau)$ and $\hat{p}(\tau)$ are both increasing in an interval $\tau \in (\tau_A, \tau_B)$ and we choose a point $(\hat{q}, \hat{p}) = (\hat{q}(\tau^*), \hat{p}(\tau^*))$ in this interval, then $Z(\hat{q}, \hat{p}) = 0$ if $(\hat{q}, \hat{p})$ is in the interior of a Ψ^j, and $\Phi(\hat{q}, \hat{p})R_q(\hat{q}, \hat{p}) = 0$ if $\hat{p} \in \mathcal{P}$. The first statement is proved in Anderson and Philpott [2], but for convenience we will repeat their argument here. We begin by defining a small perturbation of $\mathbf{s}$ around the point $(\hat{q}(\tau^*), \hat{p}(\tau^*))$. Reparameterizing $\mathbf{s}$ if necessary, we can assume that $\hat{q}'(\tau^*) > 0$. Let

$$\mathbf{r}_\delta(\tau) = \begin{cases} (\hat{q}(2\tau - (\tau^* - \delta)), \hat{p}(\tau^* - \delta)), & \tau^* - \delta \leq \tau \leq \tau^*, \\ (\hat{q}(\tau^* + \delta), \hat{p}(2\tau - (\tau^* + \delta)), & \tau^* \leq \tau \leq \tau^* + \delta, \\ (\hat{q}(\tau), \hat{p}(\tau)), & \text{otherwise.} \end{cases}$$

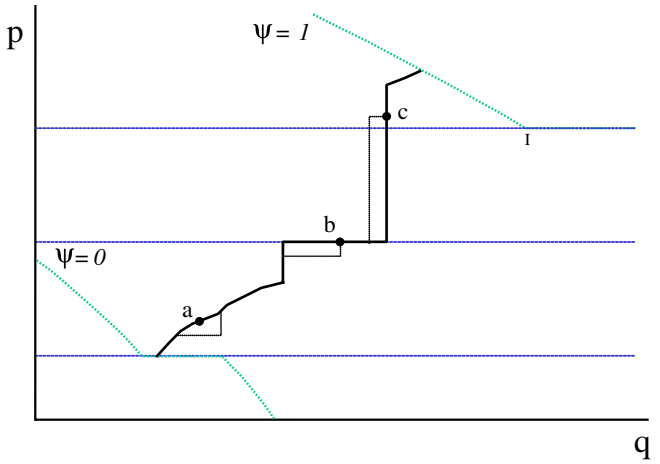


FIG. 1. *Perturbations for first order optimality.*

This perturbation is illustrated in Figure 1 at the point marked a.

In the case that $\hat{p}(\tau^*) \notin \mathcal{P}$ we can avoid any discontinuities within the perturbation by taking δ small enough. In this case

$$v(\mathbf{r}_\delta) - v(\mathbf{s}) = \int \int_{A(\delta)} Z(q,p)\ dqdp,$$

where $A(\delta)$ is the region between the two curves. Since $\mathbf{s}$ is optimal and Z is continuous in this region we obtain the conclusion that $Z(\hat{q}(\tau^*), \hat{p}(\tau^*)) \leq 0$, since otherwise we have $v(\mathbf{r}_\delta) > v(\mathbf{s})$ for δ small enough. If we reverse the direction of the perturbation (going above $(\hat{q}(\tau^*), \hat{p}(\tau^*))$ rather than below it) we can show that $Z(\hat{q}(\tau^*), \hat{p}(\tau^*)) \geq 0$. The two inequalities show that $Z = 0$ as required.

Now suppose that $\hat{p}(\tau^*) = p^j \in \mathcal{P}$. Using Lemma 2.10 and the usual Green's theorem argument we have

$$\begin{aligned} v(\mathbf{r}_\delta) - v(\mathbf{s}) = &\int \int_{A^j(\delta)} Z(q,p)\ dqdp + \int \int_{A^{j+1}(\delta)} Z(q,p)\ dqdp \\ &+ \int_{\hat{q}(\tau^*)}^{\hat{q}(\tau^*+\delta)} \Phi(q,p^j) R_q(q,p^j)\ dq, \end{aligned}$$

where $A^i(\delta) = A(\delta) \cap \Psi^i$. But we have just shown that $Z(q,p) = 0$ along the $\mathbf{s}$ curve (except where $\mathbf{s}$ crosses the p^j line). The continuity of Z implies that both the first two integrals are $o(\delta)$. Thus the continuity of Φ and R_q will imply that

$$v(\mathbf{r}_\delta) - v(\mathbf{s}) = (\hat{q}(\tau^* + \delta) - \hat{q}(\tau^*))\Phi(\hat{q}(\tau^*), p^j) R_q(\hat{q}(\tau^*), p^j) + o(\delta).$$

Since $\hat{q}'(\tau^*) > 0$, $\hat{q}(\tau^*+\delta) - \hat{q}(\tau^*)$ is $O(\delta)$, and thus $\Phi(\hat{q}(\tau^*), p^j) R_q(\hat{q}(\tau^*), p^j) \leq 0$ from the optimality of $\mathbf{s}$. Again reversing the perturbation shows $\Phi(\hat{q}(\tau^*), p^j) R_q(\hat{q}(\tau^*), p^j) \geq 0$, and thus

$$\Phi(\hat{q}(\tau^*), p^j) R_q(\hat{q}(\tau^*), p^j) = 0,$$

as required.

Part (ii). As before we consider a point with parameter $\tau^* \in (\tau_{i-1}, \tau_i)$, a horizontal section. We will need to use two different types of perturbation. Suppose first that $(\hat{q}(\tau_{i-1}), \hat{p}(\tau_{i-1})) \in \Psi_+$. There are two possibilities: $\mathbf{s}$ is vertical immediately before τ_{i-1} and $\mathbf{s}$ is strictly increasing immediately before τ_{i-1}. We only consider the first case in detail.

Let $\delta > 0$ be small and define $\tau_{i-1}(-\delta) = \hat{p}^{-1}(\hat{p}(\tau_{i-1}) - \delta)$ so that this is the parameter value at which $\mathbf{s}$ reaches a price level δ below $\hat{p}(\tau_{i-1})$. Let $\mathbf{r}_\delta$ be the perturbation of $\mathbf{s}$ which moves a horizontal section from $\hat{q}(\tau_{i-1})$ to $\hat{q}(\tau^*)$ down by an amount δ. Thus

$$\mathbf{r}_\delta(\tau) = \begin{cases} (\hat{q}(\tau), \hat{p}(\tau)), & 0 \le \tau < \tau_{i-1}(-\delta), \\ (\hat{q}(\tau_{i-1} - \tau_{i-1}(-\delta) + \tau), \hat{p}(\tau_{i-1}) - \delta), & \tau_{i-1}(-\delta) \le \tau < \tau^* - \tau_{i-1} + \tau_{i-1}(-\delta), \\ (\hat{q}(\tau^*), \hat{p}(\tau + \tau_{i-1} - \tau^*)), & \tau^* - \tau_{i-1} + \tau_{i-1}(-\delta) \le \tau < \tau^*, \\ (\hat{q}(\tau), \hat{p}(\tau)), & \tau^* \le \tau \le T. \end{cases}$$

This perturbation is illustrated in Figure 1, with $(\hat{q}(\tau^*), \hat{p}(\tau^*))$ shown as the point marked b. If $(\hat{q}(\tau^*), \hat{p}(\tau^*)) \in \Psi_+$, then this perturbation does not involve a discontinuity in Z, and thus

$$\begin{aligned} v(\mathbf{r}_\delta) - v(\mathbf{s}) &= \int_{\hat{p}(\tau_{i-1})-\delta}^{\hat{p}(\tau_{i-1})} \int_{\hat{q}(\tau_{i-1})}^{\hat{q}(\tau^*)} Z(q,p)\ dqdp \\ &= \delta \int_{\hat{q}(\tau_{i-1})}^{\hat{q}(\tau^*)} Z(q, \hat{p}(\tau_{i-1}))\ dq + o(\delta). \end{aligned}$$

Since $\mathbf{s}$ is optimal, we obtain the conclusion that $\int_{\hat{q}(\tau_{i-1})}^{\hat{q}(\tau^*)} Z(q, \hat{p}(\tau_{i-1}))dq \le 0$, since otherwise we have $v(\mathbf{r}_\delta) > v(\mathbf{s})$ for δ small enough. Thus $w(\tau^*) \le w(\tau_{i-1})$. In the case that $\mathbf{s}$ is strictly increasing immediately before τ_{i-1}, we need to make a slightly more complex definition for $\mathbf{r}_\delta(\tau)$ and the area over which Z is integrated is no longer rectangular, but the basic argument is the same.

Now if $(\hat{q}(\tau^*), \hat{p}(\tau^*)) \in \Psi_-$, then $(\hat{q}(\tau_i), \hat{p}(\tau_i)) \in \Psi_-$ and we choose a perturbation that moves a horizontal section of $\mathbf{s}$ from $\hat{q}(\tau^*)$ to $\hat{q}(\tau_i)$ upward by an amount δ. If $(\hat{q}(\tau^*), \hat{p}(\tau^*)) \in \Psi_-$, then the continuity of Z for this perturbation implies that $\int_{\hat{q}(\tau^*)}^{\hat{q}(\tau_i)} Z(q, \hat{p}(\tau_i))dq \ge 0$ and hence $w(\tau^*) \le w(\tau_i)$.

When $\hat{p}(\tau_i) \notin \mathcal{P}$ we have continuity for Z without having to restrict ourselves to $(\hat{q}(\tau^*), \hat{p}(\tau^*)) \in \Psi_+$ for a perturbation downwards at the beginning of the horizontal section, or $(\hat{q}(\tau^*), \hat{p}(\tau^*)) \in \Psi_-$ for a perturbation upwards at the end of the horizontal section. Hence we can take $\tau^* = \tau_i$ for the first argument and $\tau^* = \tau_{i-1}$ for the second argument to show that both the inequalities $w(\tau_i) \le w(\tau_{i-1})$ and $w(\tau_{i-1}) \le w(\tau_i)$ hold, and hence that there is equality.

Part (iii). Suppose $\mathbf{s}$ is vertical on the interval between τ_{i-1} and τ_i. We establish the result using perturbations of either end of the interval. We begin with a perturbation that moves the lower part of the interval to the left. There are two possibilities: $\mathbf{s}$ is horizontal immediately before τ_{i-1} or $\mathbf{s}$ is strictly increasing immediately before τ_{i-1}. The first case makes it slightly simpler to give an explicit perturbation, and we restrict ourselves to this.

Let $\delta > 0$ be small and define $\tau_{i-1}(-\delta) = \hat{q}^{-1}(\hat{q}(\tau_{i-1}) - \delta)$ so that this is the parameter value at which $\mathbf{s}$ reaches a quantity $\hat{q}(\tau_{i-1}) - \delta$. Let $\mathbf{r}_\delta$ be the perturbation of $\mathbf{s}$ which moves a vertical section from $\hat{p}(\tau_{i-1})$ to $\hat{p}(\tau^*)$ to the left by an amount δ;

thus:

$$\mathbf{r}_\delta(\tau) = \begin{cases} (\hat{q}(\tau), \hat{p}(\tau)), & 0 \le t < \tau_{i-1}(-\delta), \\ (\hat{q}(\tau_{i-1}) - \delta), \hat{p}(\tau_{i-1} - \tau_{i-1}(-\delta) + t), & \tau_{i-1}(-\delta) \le t < \tau^* - \tau_{i-1} + \tau_{i-1}(-\delta), \\ (\hat{q}(t + \tau_{i-1} - \tau^*), \hat{p}(\tau^*)), & \tau^* - \tau_{i-1} + \tau_{i-1}(-\delta) \le t < \tau^*, \\ (\hat{q}(\tau), \hat{p}(\tau)), & \tau^* \le t \le T. \end{cases}$$

This perturbation is illustrated in Figure 1, with $(\hat{q}(\tau^*), \hat{p}(\tau^*))$ shown as the point marked c. In general this perturbation may involve a number of different regions Ψ^j. Suppose that $(\hat{q}(\tau_{i-1}), \hat{p}(\tau_{i-1})) \in \Psi^f$ and $(\hat{q}(\tau^*), \hat{p}(\tau^*)) \in \Psi^g$. Write $A(\delta)$ for the region between the two curves, $\mathbf{s}$ and $\mathbf{r}_\delta$, and $A^j(\delta) = A(\delta) \cap \Psi^j$. Then

$$\begin{aligned} v(\mathbf{s}) - v(\mathbf{r}_\delta) &= \sum_{j=f}^{g} \int\int_{A^j(\delta)} Z(q,p)\ dqdp + \sum_{j=f}^{g-1} \int_{\hat{q}(\tau_{i-1})-\delta}^{\hat{q}(\tau_{i-1})} \Phi(q, p^j) R_q(q, p^j)\ dq \\ &= \delta \int_{\hat{p}(\tau_{i-1})}^{p^f} Z(\hat{q}(\tau_{i-1}), p)\ dp + \delta \sum_{j=f+1}^{g-1} \int_{p^{j-1}}^{p^j} Z(\hat{q}(\tau_{i-1}), p)\ dp \\ &\quad + \delta \int_{p^{g-1}}^{\hat{p}(\tau^*)} Z(\hat{q}(\tau_{i-1}), p)\ dp + \delta \sum_{j=f}^{g-1} \Phi(\hat{q}(\tau_{i-1}), p^j) R_q(\hat{q}(\tau_{i-1}), p^j) + o(\delta) \\ &= \delta(w(\tau^*) - w(\tau_{i-1})) + o(\delta). \end{aligned}$$

As $\mathbf{s}$ is optimal we obtain the conclusion that $w(\tau^*) \ge w(\tau_{i-1})$, since otherwise we have $v(\mathbf{r}_\delta) > v(\mathbf{s})$ for δ small enough.

The other perturbation to be considered involves a section of the vertical segment from τ^* to τ_i, which moves to the right. The argument in this case is exactly the same and we can show that $w(\tau^*) \ge w(\tau_i)$. Moreover, since these results also apply with $\tau^* = \tau_{i-1}$ and with $\tau^* = \tau_i$, we see that $w(\tau_i) = w(\tau_{i-1})$. □

In many cases there will only be a single solution which satisfies the necessary conditions, and hence the optimal solution can be identified without further computation. Later we will illustrate the application of these conditions on an example, but first it is helpful to give some more general discussion.

In practice, the nature of the optimal solution will be quite dependent on the form of the $Z = 0$ curve. If, as is usually the case, this is a monotonic increasing curve, then the optimal solution will typically follow it for much of its length, with some small variations introduced by the discontinuities. We see this behavior in the example we consider in the next section.

On a vertical section of the offer curve we must have w values greater than at the end points of the section. In the case when neither of the end points is on a horizontal price discontinuity, then this will imply that the beginning (bottom) of the vertical section is in a region where $Z > 0$ and the end of the section is in a region where $Z < 0$. If the bottom end point, say $(\hat{q}(\tau_{i-1}), \hat{p}(\tau_{i-1}))$, lies on a horizontal price discontinuity, say p^j, then we need to consider two cases. First suppose that $(\hat{q}(\tau_{i-1}), \hat{p}(\tau_{i-1})) \in \Psi_-$, which in turn implies $(\hat{q}(\tau_{i-1}), \hat{p}(\tau_{i-1})) \in \Psi^{j+1}$. Then $J(\tau_{i-1}) = j + 1$ and $w(\tau_{i-1})$ already incorporates the jump at this discontinuity; thus the vertical section must begin in a region where $Z > 0$ to avoid contradicting the necessary conditions. The other case occurs when $(\hat{q}(\tau_{i-1}), \hat{p}(\tau_{i-1})) \in \Psi_+$, in which case $R_q \ge 0$ and the jump in value is positive. In this case, we can draw no immediate conclusion on the sign of Z at the start of the vertical section. The same kind of argument shows that if the top end of a vertical section is at a price discontinuity, then when this point is in Ψ_+

we can conclude that the vertical section finishes in a region where $Z < 0$, and no conclusion can be drawn when the point is in Ψ_-.

Now consider a horizontal section that does not coincide with a price discontinuity. In this case the condition of the theorem simply says that $w(\tau)$ is less than the w values at both end points. So the left-hand end of the horizontal section must be in a region where $Z < 0$ and the right-hand end in a region where $Z > 0$.

When the horizontal section runs along a price discontinuity the situation is a little more complex. Suppose first that the left-hand end of the horizontal section is in Ψ_+. Then the necessary conditions for optimality imply that w is decreasing and hence that the beginning of the horizontal section is in a region where $Z \leq 0$. In the same way, if the right-hand end of the horizontal section is in Ψ_-, then we can deduce that this end point is in a region where $Z \geq 0$.

4. Construction of an approximate optimal supply function through undercutting and overcutting. We have found a form of sharing rule for which there will be an optimal solution. However, this form of sharing rule will not occur in practice. Our eventual aim is to have a way of generating ε-optimal solutions for problems with arbitrary sharing rules.

We suppose that we have found an optimal solution $\mathbf{s}^*$ for the modified problem with sharing rule $\mathcal{L}^*$. The next step is to create an ε-optimal solution for the problem using undercutting and overcutting. We define the solution $\mathbf{s}^*(\delta)$ for any $\delta > 0$ by following $\mathbf{s}^*$ except at the prices in $\mathcal{P}$. In essence, where $\mathbf{s}^* = p^j$ and lies in Ψ_+, we undercut the solution and set $\mathbf{s}^*(\delta) = \mathbf{s}^* - \delta$. Where $\mathbf{s}^* = p^j$ and lies in Ψ_-, we overcut the solution and set $\mathbf{s}^*(\delta) = \mathbf{s}^* + \delta$. To make this definition more precise involves some messy technical details, which we will give below. We shall assume that the equation $R_q(q,p) = 0$ defines a single monotonically increasing line which divides Ψ_+ and Ψ_-, and we write this in the form $p = \Gamma(q)$. In the case in which R is given by (2.1), Γ is just C'. If $\mathbf{s}^*$ moves from the region Ψ_+ to the region Ψ_- at a single value p^j, then $\mathbf{s}^*(\delta)$ follows the line $p = \Gamma(q)$ to join the undercutting section to the overcutting one.

We need to go back to the individual component functions $\hat{q}(\tau)$ and $\hat{p}(\tau)$, say, that define $\mathbf{s}^*$. We let $q_\delta(\tau) = \hat{q}(\tau)$ for every τ. We have the following definitions for $p_\delta(\tau)$:

$$p_\delta(\tau) = \begin{cases} p^j - \delta & \text{for } \{\tau : p^j - \delta \leq \hat{p}(\tau) \leq p^j, (\hat{q}(\tau), p^j - \delta) \in \Psi_+\}, \\ p^j + \delta & \text{for } \{\tau : p^j \leq \hat{p}(\tau) \leq p^j + \delta, (\hat{q}(\tau), p^j + \delta) \in \Psi_-\}, \\ \Gamma(\hat{q}(\tau)) & \text{for } \{\tau : p^j - \delta \leq \hat{p}(\tau) \leq p^j, (\hat{q}(\tau), p^j - \delta) \notin \Psi_+, (\hat{q}(\tau), p^j) \in \Psi_+\}, \\ \Gamma(\hat{q}(\tau)) & \text{for } \{\tau : p^j \leq \hat{p}(\tau) \leq p^j + \delta, (\hat{q}(\tau), p^j + \delta) \notin \Psi_-, (\hat{q}(\tau), p^j) \in \Psi_-\}, \\ \hat{p}(\tau) & \text{otherwise.} \end{cases}$$

As it stands this defines $p_\delta(\tau)$ in such a way that it may not be continuous. We need to make the definition of p_δ continuous by filling in these (vertical) gaps. Suppose that

$$\hat{p}(\tau_{0-}) = \lim_{\tau \uparrow \tau_0} \hat{p}(\tau) = \lim_{\tau \downarrow \tau_0} \hat{p}(\tau) - \eta$$

for some $\eta > 0$. Then we define

$$(\widetilde{p}_\delta(\tau), \widetilde{q}_\delta(\tau)) = \begin{cases} (p_\delta(\tau), q_\delta(\tau)) & \text{for } \tau < \tau_0, \\ (p_\delta(\tau_{0-}) + \tau - \tau_0, q_\delta(\tau_0)) & \text{for } \tau_0 \leq \tau \leq \tau_0 + \eta, \\ (p_\delta(\tau - \gamma), q_\delta(\tau - \gamma)) & \text{for } \tau > \tau_0 + \eta. \end{cases}$$

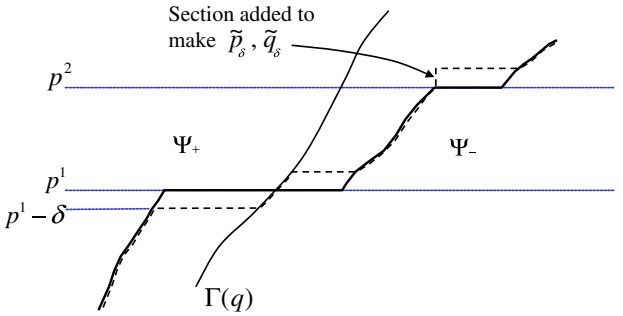


FIG. 2. *Construction of* $\mathbf{s}^*(\delta)$.

This removes one of the discontinuities, and we can continue in the same way to remove each of the other discontinuities (at most one of which is introduced at each p^j). Then $\mathbf{s}^*(\delta)$ is defined by $(\tilde{p}_\delta(\tau), \tilde{q}_\delta(\tau))$. Figure 2 illustrates this construction.

Our key result is that $\mathbf{s}^*(\delta)$ is ε-optimal for small enough δ. We are now in a position to prove this.

THEOREM 4.1. *Suppose that* $\mathbf{s}^*$ *maximizes the expected profits of the generator when the sharing rule* $\mathcal{L}^*$ *is used. Then*

$$\lim_{\delta \downarrow 0} u(\mathbf{s}^*(\delta)) = \sup_{\mathbf{s} \in \Lambda} u(\mathbf{s}),$$

where $u(\mathbf{r})$ *denotes the expected profits of the generator given a supply curve* $\mathbf{r}$ *with some other sharing rule* $\mathcal{L}$.

Proof. We write $v(\mathbf{r})$ for the expected profit of the generator given a supply curve $\mathbf{r}$ with the ideal sharing rule $\mathcal{L}^*$. From Lemma 2.4 we know that $v(\mathbf{s}) \geq u(\mathbf{s})$. Thus

$$v(\mathbf{s}^*) \geq \sup_{\mathbf{s} \in \Lambda} u(\mathbf{s}).$$

Moreover, as each $\mathbf{s}^*(\delta) \in \Lambda$, $\lim_{\delta \downarrow 0} u(\mathbf{s}^*(\delta)) \leq \sup_{\mathbf{s} \in \Lambda} u(\mathbf{s})$. So it is enough to show $\lim_{\delta \downarrow 0} u(\mathbf{s}^*(\delta)) = v(\mathbf{s}^*)$.

Now it is not hard to see that $q^j(\mathbf{s}^*(\delta)) = q^j(\mathbf{s}^*)$ for each j. This is a result of the construction we have used for $\mathbf{s}^*(\delta)$. Thus

$$v(\mathbf{s}^*) - v(\mathbf{s}^*(\delta)) = \sum_{j=1}^{n+1} \left(\int_{\mathbf{s}^{*j}} R(q,p)\ d\psi^j(q,p) - \int_{\mathbf{s}^*(\delta)^j} R(q,p)\ d\psi^j(q,p) \right).$$

Since the end points of the segments $\mathbf{s}^{*j}$ and $\mathbf{s}^*(\delta)^j$ coincide within each Ψ^j, we can use Green's theorem within this region to show that the difference between the integrals tends to zero as $\delta \to 0$. So $v(\mathbf{s}^*) = \lim_{\delta \to 0} v(\mathbf{s}^*(\delta))$. But as $\mathbf{s}^*(\delta)$ does not contain a tranche offered at any p^j the sharing rule used will not affect the expected profit, and hence $v(\mathbf{s}^*(\delta)) = u(\mathbf{s}^*(\delta))$ for each δ. □

5. An example. In order to illustrate the ideas we have discussed above we return to the example we considered before. We now suppose that the market demand is given by $D(p) + \epsilon$ where $D(p) = 800 - \frac{10}{3}p^2$, and the random shock ϵ ranges uniformly over $[0, 2300]$. We wish to find the optimal offer curve for generator A. For $p \neq 10, 14, 18$, we can derive the market distribution function for generator A:

$$\psi(q,p) = \frac{1}{2300}(q - D(p) + S_B(p))$$

and

$$\Psi = \left\{ (q,p) : 0 < \frac{1}{2300}(q - D(p) + S_B(p)) < 1 \right\}.$$

The values of ψ at $p = 10, 14, 18$ will depend on the market sharing rules, which we do not need to specify.

We need to specify a contract position and the cost of generation. We suppose that generator A has contracts for a total quantity of 800 MW at a strike price of \$15 per MWh (that is, $Q = 800$ MW), and we take the total capacity for the generators to offer into the market as 1100 MW. We also take the costs generator A incurs for generating an amount q MWh as nonlinear and given by $C(q) = 10q + 0.004q^2$. Thus the profit function (in \$ per hour) is

$$R(q,p) = qp - 10q - 0.004q^2 + 800(15 - p).$$

We wish to find an optimal supply curve for generator A so that its expected profit is maximized. However, since ψ here is discontinuous, an optimal supply curve may not exist.

Note that

$$R_q(q,p) = p - 10 - 0.008q.$$

Thus $q^*(p^j) = 125(p^j - 10)$, and

$$\Psi_+ = \{(q,p) \in \Psi : q \le 125(p - 10)\},$$
$$\Psi_- = \{(q,p) \in \Psi : q > 125(p - 10)\}.$$

The optimal sharing rule $\mathcal{L}^*$ is defined according to the rules set out earlier. Essentially, the aim of the sharing rule is to obtain a dispatch of $125(p^j - 10)$ for generator A at each of the prices $p^j = 10, 14, 18$ (or as close to this figure as possible).

In what follows, we derive the optimal supply curve, assuming the sharing rule $\mathcal{L}^*$, using Theorem 3.1. Note that for $p \neq 10, 14$, or 18,

$$Z(q,p) = R_q\psi_p - R_p\psi_q = \frac{1}{2300}\left[(p - 10 - 0.008q)\left(\frac{20}{3}p\right) - q + 800\right].$$

In Figure 3 we show the upper and lower boundaries of the region Ψ (i.e., where $\psi = 0$ and $\psi = 1$) together with the curve (in fact a parabola) where $Z = 0$ and the straight line $p = C'(q) = 10 + 0.008q$.

We will try to identify an optimal offer curve which satisfies the first order necessary conditions that are derived in Theorem 3.1. To do this we consider tracing a curve starting at the lower boundary $\psi = 0$ and finishing at $\psi = 1$. According to part (i) of Theorem 3.1, the optimal offer curve must follow the $Z = 0$ line at any point where it is neither horizontal nor vertical, and it is natural to start by considering a solution which follows this line. The argument given below shows that, for this example, there will be only one solution that satisfies the necessary optimality conditions. This is often the case for this type of problem (but not always). If there is more than one solution satisfying the optimality conditions, then the expected profits for the different offer curves need to be compared directly in order to find a global optimum.

Observe, though, that a solution which follows the line $Z = 0$ from the point $B = (507.2, 9.372)$ where it crosses $\psi = 0$ through the point D cannot be optimal.

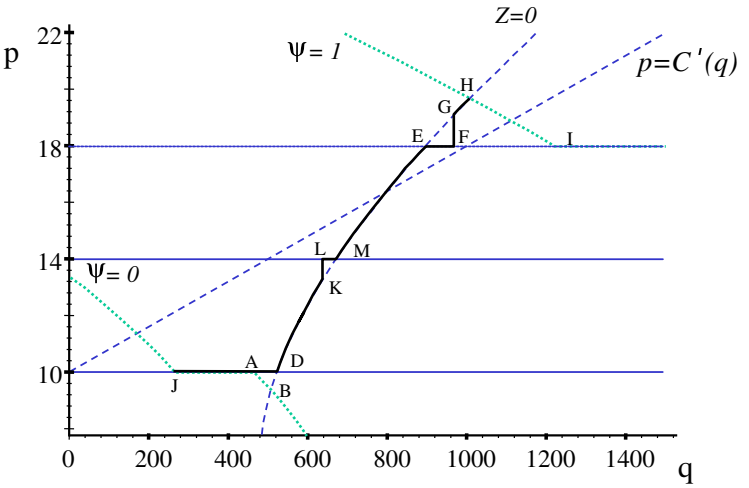


FIG. 3. *An optimal solution for the example.*

The line $Z = 0$ crosses $p = 10$ at $D = (521.7, 10)$, and at this point there is a discontinuity in w which would contradict the necessary condition (i) of the theorem. In fact, $\Phi(521.7, 10)R_q(521.7, 10) = -0.363$.

Thus we need to consider the possibility of a vertical segment finishing at the $p = 10$ line. In order to satisfy the conditions of the theorem, the integral of Z along the vertical section would have to exactly match the jump down that occurs at $p = 10$. For convenience we define

$$W_1(q, a, b) = \int_a^b Z(q, p)dp + \Phi(q, b)R_q(q, b)$$

to be the w integral over a vertical segment at q starting at $p = a$ and finishing at $p = b$ where there is supposed to be a discontinuity. Thus in this case we are interested in finding a starting point (q, a) for which $W_1(q, a, 10) = 0$. The possibilities here are to begin by following the line $Z = 0$ from the point B, but to start a vertical segment before reaching D, or to start from some point between A and B with a vertical section. However, it is not hard to check that in all cases $W_1(q, a, 10)$ will be negative.

Therefore we next consider a horizontal segment starting from some point in JA. This is in the region Ψ_-, and thus the necessary conditions will just imply that $w(\tau)$ is less than the w value at the next corner point. This condition will be satisfied since Z is positive in this region. However, this same condition will ensure that this horizontal section of the optimal offer curve does not go beyond D where Z changes sign. In fact, the necessary conditions imply that the horizontal segment starts at $J = (266.6, 10)$.

Now we consider the possibility of a vertical section which finishes on the horizontal line $p = 14$. Again, using the necessary conditions will imply that we should start this vertical segment at a point (q, a) where $W_1(q, a, 14) = 0$. This equation defines a curve which crosses the $Z = 0$ line at the point $K = (643.7, 13.4)$. Again we can establish that the horizontal section must run from $L = (643.7, 14)$ to the point $M = (671.8, 14)$ on the $Z = 0$ line and no further.

The solution then has to follow the $Z = 0$ line until the point $E = (898.0, 18)$. At this point it starts a horizontal segment. Since the solution is now in the Ψ_+ region,

the start of this horizontal section remains in the region "below the line" including $14 < p < 18$. For this reason there is no jump in the w value until the line leaves the horizontal, and there cannot be a solution with a vertical segment ending at the $p = 18$ line which satisfies the necessary conditions.

To make our discussion here easier we define

$$W_2(q, a, b) = \Phi(q, a)R_q(q, a) + \int_a^b Z(q, p)dp,$$

which is the w integral over a vertical segment starting at $(q, a) \in \Psi_+$ when a is a price discontinuity. In this case, we need to continue with a vertical segment starting at a point $(q, 18)$ and ending at a point (q, b), either on the $Z = 0$ line or on the upper boundary $\psi = 1$, with $W_2(q, 18, b) = 0$. Now the curve defined by $W_2(q, 18, b) = 0$ intersects the $Z = 0$ line at $G = (921.2, 18.36)$. The optimal solution then continues along the $Z = 0$ line until crossing the upper boundary at $H = (1009.6, 19.675)$. The complete optimal solution is shown in Figure 1.

Having found the optimal solution to the problem when $\mathcal{L}^*$ is used, it is straightforward to generate ε-optimal solutions for the case where we do not have the ideal sharing rule when prices coincide. We should follow the solution described above, but overcutting slightly for horizontal sections of the offer curve in Ψ_- and undercutting in Ψ_+. It will not matter what the offer looks like outside the region Ψ. If we choose to undercut and overcut by an amount of \$0.01, we end up with the following offer schedule:

(a) an amount of 266.6 MW at price \$0 (or any price below \$10),
(b) an amount of 521.7–266.6 = 255. 1 MW at price \$10.01,
(c) an amount of 643.7–521.7 = 122.0 MW in a smooth curve rising to a price of \$13.40,
(d) an amount of 671.8–643.7 = 28. 1 MW at a price of \$14.01,
(e) an amount of 898.0–671.8 = 226. 2 MW in a smooth curve rising to a price of \$17.99,
(f) an amount of 921.2–898.0 = 23.2 MW at a price of \$17.99,
(g) an amount of 1009.6–921.2 = 88.4 MW in a curve from a price of \$18.36 to \$19.67,
(h) an amount of 1100–1009.6 = 90.4 MW at \$50 (or any price above \$19.67).

The offer schedule above now needs to be altered in line with market rules. In the case that only step functions are allowed as offers, then the smooth curves of (c), (e), and (g) will need to be approximated with step functions. In the case that piecewise linear offers are required, then these curves would be approximated by one or more linear segments.

6. Discussion. Work on optimal offer policies and on Nash equilibria in an electricity market setting has often confronted the issue of undercutting (though our discussion of overcutting solutions is new). The essential problem is that the possibility of undercutting on price will in many models lead to highly competitive (Bertrand-type) equilibrium solutions, with no possibility of supporting an equilibrium in which generators offer at prices above their marginal costs. However, this idealized behavior is very far from that which is observed in actual markets around the world. Different authors have suggested a variety of methods to address the issue.

Using supply functions as a model for the offer procedure is one approach which avoids the difficulty of undercutting. In this framework we usually assume that there

is no single price at which a generator offers a significant quantity of power, and this allows us to formulate models in which Nash equilibria exist for supply functions.

An alternative approach which has been suggested by von der Fehr and Harbord [15], and which has been used by Wolfram [16] and Brunekreeft [6], is to assume that offers are made at one or a small number of prices but that these prices are not revealed in advance to the other players. This allows a type of mixed strategy to be played which chooses prices according to a continuous distribution. This clearly rules out the possibility of undercutting, since even though we know the strategy of the other player, we do not know a price which we can then undercut.

In this paper we do not try to establish equilibrium conditions; instead we concentrate on the question of evaluating the optimal (or ε-optimal) offer strategy. From a generator's viewpoint this is valuable if the generator wishes to achieve the maximum one period profit. The analysis we give can then point to the best possible policy which is likely to involve some part of the offer either just below or just above other players' prices. But the analysis is also valuable if the generator decides to adopt a less aggressive policy, since it indicates the degree of suboptimality involved in adopting any other (nonundercutting) solution. In practice generators will also need to build an offer according to specific market rules: again we can think of the optimal supply function strategy as setting a benchmark against which other policies can be compared. Sometimes, as in Australia and New Zealand, these market rules imply that a step function is used, in which case a step function approximation to the type of policy shown in Figure 3 should be constructed. Other markets allow an offer to be piecewise linear, which will enable a much closer approximation to be achieved.

It is natural to ask whether the type of analysis we give here could be extended to an equilibrium analysis. In fact it is not possible to construct an exact Nash equilibrium in offers with the type of undercutting behavior we have analyzed. However, an interesting area for further research is the existence of an ε-equilibrium, in which player i submits an offer S_i (being a step function satisfying the market rules) in such a way that the expected profit for player i, $v_i(S_i)$, is within ε of the best possible expected profit for player i given the offers of the other generators. Such a step function ε-equilibrium will not be unique, and so there will be the usual conceptual problems of coordination on nonunique equilibria, coupled here with additional difficulties in coordinating on an appropriate value of ε. Nevertheless, such ε-equilibrium might be arrived at in practice through repeated adjustment of generator offers in response to the other generators, but where generators prefer not to change their offer strategy unless this will lead to an increase in expected profit of at least ε.

REFERENCES

[1] E. J. Anderson and A. B. Philpott, *Using supply functions for offering generation into an electricity market*, Oper. Res., 50 (2002), pp. 477–489.

[2] E. J. Anderson and A. B. Philpott, *Optimal offer construction in electricity markets*, Math. Oper. Res., 27 (2002), pp. 82–100.

[3] E. J. Anderson and A. B. Philpott, *Estimation of electricity market distribution functions*, Ann. Oper. Res., 121 (2003), pp. 21–32.

[4] E. J. Anderson and H. Xu, *Necessary and sufficient conditions for optimal offers in electricity markets*, SIAM J. Control Optim., 41 (2002), pp. 1212–1228.

[5] R. Baldick, R. Grant, and E. Kahn, *Theory and application of linear supply function equilibrium in electricity markets*, Journal of Regulatory Economics, 25 (2004), pp. 143–167.

[6] G. Brunekreeft, *A multiple-unit, multiple-period auction in the British electricity spot market*, Energy Economics, 23 (2001), pp. 99–118.

[7] R. J. Green, *The electricity contract market in England and Wales*, J. Industrial Economics, 47 (1999), pp. 107–124.
[8] R. J. Green and D. M. Newbery, *Competition in the British electricity spot market*, J. Political Economy, 100 (1992), pp. 929–953.
[9] R. J. Green, *Increasing competition in the British electricity spot market*, J. Industrial Economics, 44 (1996), pp. 205–216.
[10] P. D. Klemperer and M. A. Meyer, *Supply function equilibria in oligopoly under uncertainty*, Econometrica, 57 (1989), pp. 1243–1277.
[11] P. J. Neame, A. B. Philpott, and G. Pritchard, *Offer stack optimization in electricity pool markets*, Oper. Res., 51 (2003), pp. 397–408.
[12] D. M. Newbery, *Competition, contracts and entry in the electricity spot market*, RAND J. Economics, 29 (1998), pp. 726–749.
[13] A. Rudkevich, *Supply Function Equilibrium in Power Markets: Learning All the Way*, TCA Technical Paper 1299-1702, Tabors Caramanis and Associates, Cambridge, MA, 1999.
[14] A. Rudkevich, *Supply function equilibrium: Theory and applications*, in Proceedings of the Hawaii International Conference on Systems Science (HICSS-36), Waikoloa, HI, 2003.
[15] N. H. M. von der Fehr and D. Harbord, *Spot market competition in the UK electricity industry*, Economic Journal, 103 (1993), pp. 531–546.
[16] C. D. Wolfram, *Strategic bidding in a multiunit auction: An empirical analysis of bids to supply electricity in England and Wales*, RAND J. Economics, 29 (1998), pp. 703–725.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 4, pp. 1419–1435

APPROXIMATION OF SOLUTIONS OF RICCATI EQUATIONS*

PAVEL BUBÁK†, CORNELIS V. M. VAN DER MEE‡, AND ANDRÉ C. M. RAN§

Abstract. This paper deals with two interrelated issues. One is an invariant subspace approach to finding solutions for the algebraic Riccati equation for a class of infinite dimensional systems. The second is approximation of the solution of the algebraic Riccati equation by finite dimensional approximants. The theory of exponentially dichotomous operators and bisemigroups is instrumental in our approach.

Key words. Riccati equation, matrix approximation, exponential dichotomy

AMS subject classifications. 93C25, 49N10, 47D06

DOI. 10.1137/S0363012903436843

1. Introduction.

The goal of this paper is twofold. The first goal is to use the theory of exponentially dichotomous operators and bisemigroups to derive a result from the existence of solutions to an algebraic Riccati equation of the type occurring in LQ-optimal control. This approach allows one to mimic the finite dimensional approach to algebraic Riccati equations; that is, it allows one to use an invariant subspace argument to obtain the extremal solutions to the algebraic Riccati equation. This topic is dealt with in section 2. It is a continuation of earlier work in this direction presented in [18, 19].

The second goal is to use the results obtained in section 2 to discuss finite dimensional approximations of the solutions of the algebraic Riccati equation and of the corresponding closed loop semigroup. Our results in this direction are presented in section 3.

The work on which this paper reports is loosely based on the work done by the first author for his masters thesis, in combination with work on the perturbation of bisemigroup generators of the last two authors [19].

Finite dimensional approximations of solutions of algebraic Riccati equations and of the corresponding closed loop semigroups are the topic of several earlier contributions; see [2, 8, 14, 12, 13, 20]. In comparison with [14] we do not discuss the algebraic Riccati equation coming from H^∞-control theory, but rather confine ourselves to the one stemming from LQ-optimal control. The result we obtain is, in this special case, the same, under slightly different assumptions, but with a completely different, and in our view, more transparent proof. In [20] attention was also focused on the algebraic Riccati equation from LQ-optimal control. The assumptions there are seriously weaker than the ones imposed in previous works. In particular, instead of exponential stability (or exponential stabilizability) in [20] strong stabilizability is assumed. Instead, we consider exponentially dichotomous operators, which allows us to deal with Hamiltonian operators of linear systems that have no spectrum within a strip

*Received by the editors October 28, 2003; accepted for publication (in revised form) February 2, 2005; published electronically November 4, 2005.
http://www.siam.org/journals/sicon/44-4/43684.html
†Doktorand MUUK, Sokolovska 83, 180 00 Praha, Czech Republic (bubu@karlin.mff.cuni.cz).
‡Dipartimento di Matematica e Informatica, Università di Cagliari, Viale Merello 92, 09123 Cagliari, Italy (cornelis@bugs.unica.it). This author's research was supported by MIUR under COFIN grant 2004015437, and by INdAM.
§Afdeling Wiskunde en Informatica, FEW, Vrije Universiteit, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands (ran@cs.vu.nl).

about the imaginary axis and for which one of the off-diagonal operators is compact. However, we obtain stronger results on the closed loop approximants (compare our Theorem 3.4 with Theorem 4.2 in [20]) in return for our stronger assumptions. Again, our methods of proof are quite different from the ones in [20].

Our approach is part of a long tradition of studying stability results for solutions of Riccati equations by performing a stability analysis of certain invariant subspaces of the Hamiltonian operator, while also linking these to stable factorizations of a transfer function [3, 17]. Structural similarities between these interlocking problems and state space approaches to solve convolution equations [5] and stationary transport equations [10] have naturally led to the formal study of exponentially dichotomous operators [4], results on their perturbation [19], and its present stability analysis of Hamiltonian operators of autonomous linear systems.

In [19] we have linked the left and right canonical Wiener–Hopf factorizability of a transfer function built from the Hamiltonian operator

$$\begin{pmatrix} A_0 & -D \\ -Q & -A_1 \end{pmatrix} \tag{1.1}$$

of a linear system to the existence of the stable and anti-stable solution of a Riccati equation, under hardly more than the assumption that $-A_0$ and $-A_1$ generate exponentially decaying C_0-semigroups on a general Banach space. Even though not stated explicitly, stability results for these solutions of the Riccati equations are expected (and thus now conjectured) to hold if the transfer function has a left or right canonical Wiener–Hopf factorization. Using the well-known fact that this is true for positive selfadjoint transfer functions on a Hilbert space, we naturally arrive at the basic outline of the present paper. The stability analysis itself appears to be straightforward.

Our present approach may be viewed as a tool to derive stability results for Riccati equations starting from the Hamiltonian operator, where the derivation of the latter is standard system theory [7, 15]. Many of the existing results (but not all; see [20]) can thus be derived in a transparent way, but the present approach potentially leads to useful applications to delay systems where the underlying spaces are L^1 [11].

When dealing with Hamiltonian operators of the type (1.1) with $D = BR^{-1}B^*$, $Q = C^*C$, and $A_0 = A_1^* = A$, the infinitesimal generator of a C_0-semigroup on a separable Hilbert space $\mathcal{H}$, it is sufficient to require the exponential stabilizability of (A, B) or the exponential detectability of (C, A) to arrive at an exponentially dichotomous operator on $\mathcal{H}\dot{+}\mathcal{H}$ after a similarity implementing state feedback or output injection (e.g., see [7]). Thus for the purpose of this article it is sufficient to deal with Hamiltonian operators that are exponentially dichotomous.

Let us conclude the introduction with some notations and definitions. By $\mathcal{D}(A)$, $\operatorname{Ker} A$, and $\operatorname{Im} A$ we denote the domain, kernel, and range of a linear operator A, respectively, and by $I_{\mathcal{H}}$ the identity operator on a Hilbert space $\mathcal{H}$. By $\mathcal{H} \stackrel{\text{def}}{=} \mathcal{H}_1\dot{+}\mathcal{H}_2$ we denote the orthogonal direct sum of the Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ and by $A \stackrel{\text{def}}{=} A_1\dot{+}A_2$ the linear operator on $\mathcal{H}$ with domain $\{(x_1, x_2) : x_j \in \mathcal{D}(\mathcal{H}_j), j = 1, 2\}$ defined by $A(x_1, x_2) = (A_1x_1, A_2x_2)$.

2. Preliminaries. A closed and densely defined linear operator $-S$ on a Hilbert space $\mathcal{H}$ is called *exponentially dichotomous* [4] if for some bounded projection P commuting with S, the restrictions of S to $\operatorname{Im} P$ and of $-S$ to $\operatorname{Ker} P$ are the infinitesimal generators of exponentially decaying C_0-semigroups. We then define the *bisemigroup*

generated by $-S$ as

$$E(t;-S) = \begin{cases} e^{-tS}(I-P), & t > 0 \\ -e^{-tS}P, & t < 0. \end{cases}$$

Its *separating projection* P is given by $P = -E(0^-;-S) = I_{\mathcal{H}} - E(0^+;-S)$. One easily verifies [4] the existence of $\varepsilon > 0$ such that $\{\lambda \in \mathbb{C} : |\operatorname{Re}\lambda| \le \varepsilon\}$ is contained in the resolvent set $\rho(S)$ of S and for every $x \in \mathcal{H}$

$$(\lambda - S)^{-1}x = -\int_{-\infty}^{\infty} e^{\lambda t}E(t;-S)x\,dt, \qquad |\operatorname{Re}\lambda| \le \varepsilon. \tag{2.1}$$

As a result, $\|(\lambda - S)^{-1}x\| \to 0$ as $\lambda \to \infty$ in $\{\lambda \in \mathbb{C} : |\operatorname{Re}\lambda| \le \varepsilon'\}$ for some $\varepsilon' \in (0, \varepsilon]$.

We have the following perturbation result given also in [19]. We shall give its proof for selfcontainedness.

THEOREM 2.1. *Let $-S_0$ be exponentially dichotomous, Γ be a compact operator, and $-S = -S_0 + \Gamma$, where $\mathcal{D}(S) = \mathcal{D}(S_0)$. Suppose the imaginary axis is contained in the resolvent set of S. Then $-S$ is exponentially dichotomous. Moreover, $E(t;-S) - E(t;-S_0)$ is a compact operator, also in the limits as $t \to 0^{\pm}$.*

Proof. There exists $\varepsilon > 0$ such that

$$\int_{-\infty}^{\infty} e^{\varepsilon|t|}\|E(t;-S_0)\|\,dt < \infty. \tag{2.2}$$

Using the resolvent identity

$$(\lambda - S)^{-1} - (\lambda - S_0)^{-1} = -(\lambda - S_0)^{-1}\Gamma(\lambda - S)^{-1}, \qquad |\operatorname{Re}\lambda| \le \varepsilon,$$

for some $\varepsilon > 0$, we obtain the convolution integral equation

$$E(t;-S)x - \int_{-\infty}^{\infty} E(t-\tau;-S_0)\Gamma E(\tau;-S)x\,d\tau = E(t;-S_0)x, \tag{2.3}$$

where $x \in \mathcal{H}$ and $0 \ne t \in \mathbb{R}$. In (2.3), the convolution kernel $E(\cdot;-S_0)\Gamma$ is continuous in the norm except for a jump discontinuity in $t = 0$, as a result of the strong continuity (except for the jump) of $E(\cdot;-S_0)$ and the compactness of Γ. Further, (2.2) implies that $e^{\varepsilon|\cdot|}E(\cdot;-S_0)\Gamma$ is Bochner integrable.

The symbol of the convolution integral equation (2.3), which equals $I_{\mathcal{H}} + (\lambda - S_0)^{-1}\Gamma = (\lambda - S_0)^{-1}(\lambda - S)$, tends to $I_{\mathcal{H}}$ in the norm as $\lambda \to \infty$ in the strip $|\operatorname{Re}\lambda| \le \varepsilon$, since Γ is compact and $(\lambda - S_0)^{-1}$ tends to zero strongly. Moreover, it is a compact perturbation of the identity which, by definition, only takes invertible values on the imaginary axis. Thus there exists $\varepsilon_0 \in (0, \varepsilon]$ such that the symbol only takes invertible values on the strip $|\operatorname{Re}\lambda| \le \varepsilon_0$.

Before proceeding with the proof we now state the Bochner–Phillips theorem [6, 9]:

- Let $\mathcal{A}_0$ be a Banach algebra, $\mathcal{A}$ its natural extension to a Banach algebra with unit element, and $W_{\mathcal{A}_0}$ the Banach algebra of all ordered pairs (A_∞, A), where $A_\infty \in \mathcal{A}$ and A is a Bochner integrable function from $\mathbb{R}$ into $\mathcal{A}_0$, endowed with the norm

$$\|(A_\infty, A)\|_{W_{\mathcal{A}_0}} \stackrel{def}{=} \|A_\infty\|_{\mathcal{A}} + \int_{-\infty}^{\infty} \|A(t)\|_{\mathcal{A}_0}\,dt.$$

Then (A_∞, A) is invertible in $W_{\mathcal{A}_0}$ if and only if A_∞ and all of the Fourier transform values

$$A_\infty + \int_{-\infty}^{\infty} e^{i\lambda t} A(t)\, dt, \qquad \lambda \in \mathbb{R},$$

are invertible elements of $\mathcal{A}$. In that case the inverse (B_∞, B) is given by $B_\infty = (A_\infty)^{-1}$ and

$$B_\infty + \int_{-\infty}^{\infty} e^{i\lambda t} B(t)\, dt = \left[A_\infty + \int_{-\infty}^{\infty} e^{i\lambda t} A(t)\, dt \right]^{-1}, \qquad \lambda \in \mathbb{R}.$$

We now apply this result in two different situations: (i) $\mathcal{A} = \mathcal{A}_0 = L(\mathcal{H})$ is the Banach algebra of bounded linear operators on $\mathcal{H}$, and (ii) $\mathcal{A}_0 = K(\mathcal{H})$ is the Banach algebra of compact operators on $\mathcal{H}$ and $\mathcal{A} = \{\lambda I_{\mathcal{H}} + K : \lambda \in \mathbb{C}, K \in K(\mathcal{H})\}$. We then also use that an element $(A_\infty, A) \in W_{L(\mathcal{H})}$ induces a bounded linear operator on $BC(\mathbb{R}^-; \mathcal{H}) \oplus BC(\mathbb{R}^*; \mathcal{H})$, the bounded continuous functions from $\mathbb{R}$ into $\mathcal{H}$ with a jump discontinuity at $t = 0$, by convolution.

By the Bochner–Phillips theorem, the convolution equation (2.3) has a unique solution $u(\cdot; x) = E(\cdot; -S)x$ with the following properties:

1) $E(\cdot; -S)$ is strongly continuous, except for a jump discontinuity at $t = 0$,
2) $\int_{-\infty}^{\infty} e^{\varepsilon_0 |t|} \|E(t; -S)\|\, dt < \infty$; hence $E(\cdot; -S)$ is exponentially decaying,
3) $E(t; -S) - E(t; -S_0)$ is a compact operator, also in the limits as $t \to 0^{\pm}$, and
4) the identity (2.1) holds.

As result [4], $-S$ is exponentially dichotomous. □

The set $\theta = (A_0, Q, D; \mathcal{H})$ is called a *triple* if $\mathcal{H}$ is a complex Hilbert space, A_0 generates a strongly continuous semigroup on $\mathcal{H}$ of negative exponential type, and Q and D are bounded selfadjoint operators on $\mathcal{H}$. Then obviously $-S_0 = (-A_0) \dot{+} A_0^*$ is exponentially dichotomous on $\mathcal{H} \dot{+} \mathcal{H}$ and $P_0 = I_{\mathcal{H}} \dot{+} 0$ is the separating projection of the corresponding bisemigroup $E(\cdot; -S_0)$. The triple θ is called *semicompact* if D is a compact operator on $\mathcal{H}$, and *compact* if both D and Q are compact operators on $\mathcal{H}$. The triple θ is called *positive semidefinite* if Q and D are positive semidefinite selfadjoint, and *antipodal* if one of Q and D is positive semidefinite selfadjoint and the other is negative semidefinite selfadjoint.

Theorem 2.1 can be used to prove the following more specific result.

THEOREM 2.2. *Let $\theta = (A_0, Q, D; \mathcal{H})$ be a positive semidefinite and semicompact triple. Then the block matrix operator $-S$ defined on $\mathcal{H} \dot{+} \mathcal{H}$ by*

$$S = \begin{bmatrix} A_0 & -D \\ -Q & -A_0^* \end{bmatrix},$$

is exponentially dichotomous.

Proof. Suppose (2.2) is satisfied. Let us define the operator

$$S_Q = \begin{bmatrix} A_0 & 0 \\ -Q & -A_0^* \end{bmatrix}, \qquad \mathcal{D}(S_Q) = \mathcal{D}(S_0).$$

Consider the unique and positive semidefinite solution X of the Lyapunov equation (e.g., [7, (1.12)–(1.13)])

$$A_0^* X + X A_0 = -Q,$$

given by

$$Xx = \int_0^\infty e^{\tau A_0^*} Q e^{-\tau A_0} x \, d\tau, \qquad x \in \mathcal{H}.$$

Note that

$$\begin{bmatrix} I & 0 \\ -X & I \end{bmatrix} S_Q \begin{bmatrix} I & 0 \\ X & I \end{bmatrix} = S_0.$$

So S_Q and S_0 are similar. Hence $-S_Q$ is exponentially dichotomous, and we obtain

$$E(\cdot; -S_Q) = \begin{bmatrix} I & 0 \\ X & I \end{bmatrix} E(\cdot; -S_0) \begin{bmatrix} I & 0 \\ -X & I \end{bmatrix}.$$

We also see that the separating projection P_Q of $E(\cdot; -S_Q)$ is given by

$$P_Q = \begin{bmatrix} I & 0 \\ X & 0 \end{bmatrix}.$$

Next, we remark that $S - S_Q$ is a compact operator. Hence by Theorem 2.1, to prove that $-S$ is exponentially dichotomous, it suffices to prove that $-S$ does not have imaginary eigenvalues. Indeed, let λ be an imaginary eigenvalue of $-S$. Then there exist $x \in \mathcal{D}(A_0)$ and $y \in \mathcal{D}(A_0^*)$ such that

$$\begin{aligned} (\lambda + A_0)x - Dy &= 0, \\ -Qx + (\lambda - A_0^*)y &= 0. \end{aligned}$$

Then, since λ is purely imaginary, we have

$$\langle Qx, x\rangle + \langle Dy, y\rangle = \langle (\lambda - A_0^*)y, x\rangle + \langle (\lambda + A_0)x, y\rangle = 0,$$

which implies $Qx = Dy = 0$. But then $(\lambda - A_0)x = (\lambda + A_0^*)y = 0$ for some imaginary λ, and hence $x = y = 0$, as claimed. □

Let

$$P = -E(0^-; -S) = I_{\mathcal{H}\dot{+}\mathcal{H}} - E(0^+; -S)$$

denote the separating projection of $E(\cdot; -S)$. Consider the indefinite scalar product generated by

$$\mathcal{J}_1 = \begin{bmatrix} 0 & -I_{\mathcal{H}} \\ -I_{\mathcal{H}} & 0 \end{bmatrix}$$

on $\mathcal{H}\dot{+}\mathcal{H}$. Since $\mathcal{J}_1 S + S^* \mathcal{J}_1 = 2(Q\dot{+}D)$, the real part $\frac{1}{2}(S + \mathcal{J}_1^{-1} S^* \mathcal{J}_1)$ of S with respect to the indefinite scalar product generated by $\mathcal{J}_1$ is positive semidefinite selfadjoint whenever $\theta = (A_0, Q, D; \mathcal{H})$ is a positive semidefinite triple. Hence, in this case it is clear that $\operatorname{Im} P$ is a $\mathcal{J}_1$-nonpositive and $\operatorname{Ker} P$ is a $\mathcal{J}_1$-nonnegative S-invariant subspace of $\mathcal{H}\dot{+}\mathcal{H}$ (cf. [1, section 3.2]). Also [1], since iS is selfadjoint with respect to the indefinite scalar product generated by

$$\mathcal{J}_2 = \begin{bmatrix} 0 & iI_{\mathcal{H}} \\ -iI_{\mathcal{H}} & 0 \end{bmatrix},$$

it is clear that $\operatorname{Im} P$ and $\operatorname{Ker} P$ are $\mathcal{J}_2$-neutral S-invariant subspaces of $\mathcal{H}\dot{+}\mathcal{H}$ (i.e., on these subspaces the sesquilinear form $(x,y)\mapsto(\mathcal{J}_2x,y)$ is trivial).

Further, with X as above and $\theta=(A_0,Q,D;\mathcal{H})$ a positive semidefinite triple, we have

$$-S_X \stackrel{def}{=} \begin{bmatrix} I_{\mathcal{H}} & 0 \\ -X & I_{\mathcal{H}} \end{bmatrix} (-S) \begin{bmatrix} I_{\mathcal{H}} & 0 \\ X & I_{\mathcal{H}} \end{bmatrix} = -S_0 + \begin{bmatrix} I_{\mathcal{H}} \\ -X \end{bmatrix} D \begin{bmatrix} X & I_{\mathcal{H}} \end{bmatrix},$$

which implies that $(A_0-DX, XDX, -D;\mathcal{H})$ is an antipodal compact triple.

We need the following definitions. Suppose W is a continuous function from the extended imaginary axis $i(\mathbb{R}\cup\{\infty\})$ into $\mathcal{L}(\mathcal{H})$. Then by a *left canonical* (*Wiener–Hopf*) *factorization* of W we mean a representation of W of the form

$$W(\lambda)=W_+(\lambda)W_-(\lambda), \qquad \operatorname{Re}\lambda=0,$$

in which $W_\pm(\pm\lambda)$ is continuous on the closed right half-plane (the point at ∞ included), is analytic on the open right half-plane, and takes only invertible values for λ in the closed right half-plane (the point at infinity included). Obviously, such an operator function only takes invertible values on the extended imaginary axis. By a *right canonical* (*Wiener–Hopf*) *factorization* we mean a representation of W of the form

$$W(\lambda)=W_-(\lambda)W_+(\lambda), \qquad \operatorname{Re}\lambda=0,$$

where $W_\pm(\lambda)$ is as above.

THEOREM 2.3. *Let $\theta=(A_0,Q,D;\mathcal{H})$ be a positive semidefinite and semicompact triple. Then we have the following decompositions:*

$$Im\,P\dot{+}Ker\,P_0=\mathcal{H}\dot{+}\mathcal{H}, \tag{2.4}$$

$$Ker\,P\dot{+}Im\,P_0=\mathcal{H}\dot{+}\mathcal{H}. \tag{2.5}$$

Proof. Let us introduce the operators

$$V=P_0P+(I-P_0)(I-P), \tag{2.6}$$

$$V_Q=P_0P_Q+(I-P_0)(I-P_Q). \tag{2.7}$$

Then

$$V_Q=\begin{bmatrix} I_{\mathcal{H}} & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} I_{\mathcal{H}} & 0 \\ X & 0 \end{bmatrix}+\begin{bmatrix} 0 & 0 \\ 0 & I_{\mathcal{H}} \end{bmatrix}\begin{bmatrix} 0 & 0 \\ -X & I_{\mathcal{H}} \end{bmatrix}=\begin{bmatrix} I_{\mathcal{H}} & 0 \\ -X & I_{\mathcal{H}} \end{bmatrix},$$

so that V_Q is invertible. On the other hand, the identity

$$\begin{aligned} &E(t;-S)-\int_{-\infty}^{\infty}E(t-\tau;-S_Q)\begin{bmatrix} 0 & D \\ 0 & 0 \end{bmatrix}E(\tau;-S)\,d\tau \\ &=E(t;-S)-\int_{-\infty}^{\infty}E(t-\tau;-S_Q)\begin{bmatrix} 0 & D^{1/2} \\ 0 & 0 \end{bmatrix}\left(E(\tau;-S^*)\begin{bmatrix} 0 & 0 \\ D^{1/2} & 0 \end{bmatrix}\right)^* d\tau \\ &=E(t;-S_Q), \end{aligned}$$

which is analogous to (2.3) and where the integrand is norm continuous in τ, implies that

$$P-P_Q=-\int_{-\infty}^{\infty}E(-\tau;-S_Q)\begin{bmatrix} 0 & D \\ 0 & 0 \end{bmatrix}E(\tau;-S)\,d\tau,$$

is compact. Further,

$$V - V_Q = [P_0 - (I - P_0)]\,(P - P_Q)$$

implies that $V - V_Q$ is compact. As a result, V is a Fredholm operator of index zero.

Now the operator V satisfies the identities

$$\begin{aligned}\operatorname{Ker} V &= [\operatorname{Im} P \cap \operatorname{Ker} P_0] \dot{+} [\operatorname{Ker} P \cap \operatorname{Im} P_0],\\ \operatorname{Im} V &= [\operatorname{Im} P + \operatorname{Ker} P_0] \cap [\operatorname{Ker} P + \operatorname{Im} P_0].\end{aligned}$$

So, in order to establish (2.4) and (2.5) it suffices to prove that

$$\operatorname{Im} P \cap \operatorname{Ker} P_0 = \operatorname{Ker} P \cap \operatorname{Im} P_0 = \{0\}.$$

Indeed, the operator function

$$\begin{aligned}W(\lambda) &= I + \begin{bmatrix} Q^{1/2} & 0 \\ 0 & D^{1/2}\end{bmatrix} (\lambda - S_0)^{-1} \begin{bmatrix} 0 & D^{1/2} \\ Q^{1/2} & 0\end{bmatrix}\\ &= I + \begin{bmatrix} Q^{1/2} & 0 \\ 0 & D^{1/2}\end{bmatrix} \begin{bmatrix} 0 & (\lambda + A_0^*)^{-1} \\ (\lambda - A_0)^{-1} & 0\end{bmatrix} \begin{bmatrix} Q^{1/2} & 0 \\ 0 & D^{1/2}\end{bmatrix}\end{aligned} \tag{2.8}$$

has the identity operator as its real part for imaginary λ and hence

$$\sup_{\operatorname{Re}\lambda = 0} \|I_{\mathcal{H}} - cW(\lambda)\| < 1$$

for some $c > 0$.

Also, W belongs to the Wiener algebra in the sense that there exists a norm measurable operator function $L(\cdot)$ for which W is equal to I plus the Fourier transform of L, and with L having only compact operators as its values such that

$$\int_{-\infty}^{\infty} e^{\varepsilon|t|} \|L(t)\|\, dt < \infty,$$

because of the norm continuity of $L(t)$ for $t \in \mathbb{R} \setminus \{0\}$ and the exponential decay of $\|L(t)\|$ as $t \to \pm\infty$. As a result [9], cW and hence W has left and right canonical factorizations

$$W(\lambda) = W_-^{(l)}(\lambda) W_+^{(l)}(\lambda) = W_+^{(r)}(\lambda) W_-^{(r)}(\lambda), \qquad |\operatorname{Re}\lambda| \le \varepsilon, \tag{2.9}$$

for some $\varepsilon > 0$, where $W_-^{(l)}(\lambda)$, $W_-^{(r)}(\lambda)$ and their inverses are analytic in the half-plane $\operatorname{Re}\lambda < \varepsilon$ and tend to the identity in the norm as $\lambda \to \infty$ in this half-plane and $W_+^{(l)}(\lambda)$, $W_+^{(r)}(\lambda)$ and their inverses are analytic in the half-plane $\operatorname{Re}\lambda > -\varepsilon$ and tend to the identity in the norm as $\lambda \to \infty$ in this half-plane. Using

$$S = S_0 - \begin{bmatrix} 0 & D^{1/2} \\ Q^{1/2} & 0\end{bmatrix} \begin{bmatrix} Q^{1/2} & 0 \\ 0 & D^{1/2}\end{bmatrix}$$

and

$$W(\lambda)^{-1} = I - \begin{bmatrix} Q^{1/2} & 0 \\ 0 & D^{1/2}\end{bmatrix} (\lambda - S)^{-1} \begin{bmatrix} 0 & D^{1/2} \\ Q^{1/2} & 0\end{bmatrix},$$

we obtain

$$W(\lambda)^{-1}\begin{bmatrix} Q^{1/2} & 0 \\ 0 & D^{1/2} \end{bmatrix}(\lambda - S_0)^{-1} = \begin{bmatrix} Q^{1/2} & 0 \\ 0 & D^{1/2} \end{bmatrix}(\lambda - S)^{-1}. \tag{2.10}$$

Letting $x \in \operatorname{Ker} P_0 \cap \operatorname{Im} P$, we substitute the first of the factorizations (2.9) into (2.10), observe that the left- and right-hand sides of the resulting identity

$$W_-^{(l)}(\lambda)^{-1}\begin{bmatrix} Q^{1/2} & 0 \\ 0 & D^{1/2} \end{bmatrix}(\lambda - S_0)^{-1}x = W_+^{(l)}(\lambda)\begin{bmatrix} Q^{1/2} & 0 \\ 0 & D^{1/2} \end{bmatrix}(\lambda - S)^{-1}x$$

are analytic in λ for $\operatorname{Re}\lambda < \varepsilon$ and $\operatorname{Re}\lambda > -\varepsilon$, respectively, apply Liouville's theorem, and obtain

$$\begin{bmatrix} Q^{1/2} & 0 \\ 0 & D^{1/2} \end{bmatrix}(\lambda - S_0)^{-1}x = \begin{bmatrix} Q^{1/2} & 0 \\ 0 & D^{1/2} \end{bmatrix}(\lambda - S)^{-1}x = 0.$$

Next, we employ the equality

$$\begin{aligned}(\lambda - S)^{-1}x - (\lambda - S_0)^{-1}x &= -(\lambda - S)^{-1}\begin{bmatrix} 0 & D \\ Q & 0 \end{bmatrix}(\lambda - S_0)^{-1}x \\ &= -(\lambda - S)^{-1}\begin{bmatrix} 0 & D^{1/2} \\ Q^{1/2} & 0 \end{bmatrix}\cdot\begin{bmatrix} Q^{1/2} & 0 \\ 0 & D^{1/2} \end{bmatrix}(\lambda - S_0)^{-1}x = 0\end{aligned}$$

to enable the application of Liouville's theorem to the analytic continuation of $(\lambda - S)^{-1}x = (\lambda - S_0)^{-1}x$ and conclude that $x = 0$. As a result, $\operatorname{Ker} P_0 \cap \operatorname{Im} P = \{0\}$, as claimed. In a similar way we prove that $\operatorname{Im} P_0 \cap \operatorname{Ker} P = \{0\}$. □

Theorem 2.3 implies that $(\lambda - S_0)^{-1}(\lambda - S)$ has left and right canonical factorizations (as in (2.9)). Letting

$$\Gamma = \begin{bmatrix} 0 & D \\ Q & 0 \end{bmatrix},$$

these factorizations have the following form ([3, Chapter 1])

$$(\lambda - S_0)^{-1}(\lambda - S) = \big[I + (\lambda - S_0)^{-1}(I - \mathcal{P})\Gamma\big]\big[I + \mathcal{P}(\lambda - S_0)^{-1}\Gamma\big],$$

where

$$\begin{aligned}(I + (\lambda - S_0)^{-1}(I - \mathcal{P})\Gamma)^{-1} &= I - (I - \mathcal{P})(\lambda - S)^{-1}\Gamma, \\ (I + \mathcal{P}(\lambda - S_0)^{-1}\Gamma)^{-1} &= I - (\lambda - S)^{-1}\mathcal{P}\Gamma.\end{aligned}$$

Here $\mathcal{P}$ is either the projection of $\mathcal{H}\dot{+}\mathcal{H}$ onto $\operatorname{Im} P$ along $\operatorname{Ker} P_0$ (for the right canonical factorization) or the projection of $\mathcal{H}\dot{+}\mathcal{H}$ onto $\operatorname{Ker} P$ along $\operatorname{Im} P_0$ (for the left canonical factorization).

The following result has been established in [18] for a positive semidefinite triple, without assuming the compactness of D. As a result, in [18] one does not get the compactness of Π_+, only its boundedness.

THEOREM 2.4. *Let $(A_0, Q, D; \mathcal{H})$ be a positive semidefinite and semicompact triple. Then there exist unique positive semidefinite selfadjoint operators $-\Pi_+$ and Π_- on $\mathcal{H}$, where Π_+ is compact and Π_- is bounded, such that*

(1) *the image and kernel of the separating projection P of $E(\cdot;-S)$ are graph subspaces in the sense that*

$$\mathrm{Im}\, P = \mathrm{Im} \begin{bmatrix} I_{\mathcal{H}} \\ \Pi_- \end{bmatrix}, \qquad \mathrm{Ker}\, P = \mathrm{Im} \begin{bmatrix} \Pi_+ \\ I_{\mathcal{H}} \end{bmatrix}, \tag{2.11}$$

(2) Π_- *maps $\mathcal{D}(A_0)$ into $\mathcal{D}(A_0^*)$ and Π_+ maps $\mathcal{D}(A_0^*)$ into $\mathcal{D}(A_0)$,*

(3) Π_- *is a solution of the operator Riccati equation*

$$\Pi A_0 x + A_0^* \Pi x + Qx - \Pi D \Pi x = 0, \qquad x \in \mathcal{D}(A_0), \tag{2.12}$$

and Π_+ is a solution of the operator Riccati equation

$$A_0 \Pi x + \Pi A_0^* x + \Pi Q \Pi x - Dx = 0, \qquad x \in \mathcal{D}(A_0^*), \tag{2.13}$$

(4) *and $A_0 - D\Pi_-$ and $A_0 + \Pi_+ Q$ are the infinitesimal generators of exponentially decaying C_0-semigroups on $\mathcal{H}$.*

Proof. According to Theorem 2.3, there exist bounded projections $\mathcal{P}^{(l)}$ and $\mathcal{P}^{(r)}$ on $\mathcal{H}\dot{+}\mathcal{H}$ such that $\mathcal{P}^{(l)}$ projects $\mathcal{H}\dot{+}\mathcal{H}$ onto $\mathrm{Ker}\, P$ along $\mathrm{Im}\, P_0$ and $\mathcal{P}^{(r)}$ projects $\mathcal{H}\dot{+}\mathcal{H}$ onto $\mathrm{Im}\, P$ along $\mathrm{Ker}\, P_0$. Hence there exist bounded linear operators Π_- and Π_+ on $\mathcal{H}$, so-called angular operators (cf. [3, Chapter 5]), such that

$$\mathcal{P}^{(l)} = \begin{bmatrix} I_{\mathcal{H}} & 0 \\ \Pi_- & 0 \end{bmatrix}, \qquad \mathcal{P}^{(r)} = \begin{bmatrix} 0 & \Pi_+ \\ 0 & I_{\mathcal{H}} \end{bmatrix}. \tag{2.14}$$

As a result, there exist bounded linear operators Π_- and Π_+ on $\mathcal{H}$ such that (2.11) is true.

One easily proves that

$$\mathcal{P}^{(l)} = V^{-1}(I_{\mathcal{H}} - P_0), \qquad \mathcal{P}^{(r)} = V^{-1} P_0,$$

where V is given by (2.6). Since the projections P_0 and $I - P_0$ commute with $(\lambda - S_0)^{-1}$ and P and $I - P$ commute with $(\lambda - S)^{-1}$ whenever $|\mathrm{Re}\, \lambda| \leq \varepsilon$ for some $\varepsilon > 0$, the invertible operator V maps $\mathcal{D}(S_0) = \mathcal{D}(S) = \mathcal{D}(A_0)\dot{+}\mathcal{D}(A_0^*)$ onto itself. Consequently, $\mathcal{P}^{(l)}$ and $\mathcal{P}^{(r)}$ map this domain into itself and hence Π_- maps $\mathcal{D}(A_0)$ into $\mathcal{D}(A_0^*)$ and Π_+ maps $\mathcal{D}(A_0^*)$ into $\mathcal{D}(A_0)$.

The Riccati equations (2.12) and (2.13) follow from the identities

$$S \begin{bmatrix} I_{\mathcal{H}} \\ \Pi_- \end{bmatrix} x = \begin{bmatrix} I_{\mathcal{H}} \\ \Pi_- \end{bmatrix} (A_0 - D\Pi_-)x, \qquad S \begin{bmatrix} \Pi_+ \\ I_{\mathcal{H}} \end{bmatrix} y = \begin{bmatrix} \Pi_+ \\ I_{\mathcal{H}} \end{bmatrix} (-A_0^* - Q\Pi_+)y, \tag{2.15}$$

where $x \in \mathcal{D}(A_0)$ and $y \in \mathcal{D}(A_0^*)$, in the standard way. Furthermore, since S is exponentially dichotomous with separating projection P and

$$\mathrm{Ker}\, P = \mathrm{Im} \begin{bmatrix} I_{\mathcal{H}} \\ \Pi_- \end{bmatrix}, \qquad \mathrm{Im}\, P = \mathrm{Im} \begin{bmatrix} \Pi_+ \\ I_{\mathcal{H}} \end{bmatrix},$$

we immediately have part (4) of Theorem 2.4.

Now remark that Π_- and Π_+ are selfadjoint (because of the $\mathcal{J}_2$-neutrality of $\mathrm{Im}\, P$ and $\mathrm{Ker}\, P$), while $-\Pi_+$ and Π_- are positive semidefinite (because of the $\mathcal{J}_1$-nonpositivity of $\mathrm{Im}\, P$ and the $\mathcal{J}_1$-nonnegativity of $\mathrm{Ker}\, P$).

Finally, from the compactness of $V - V_Q$ and hence from the compactness of

$$V^{-1} - V_Q^{-1} = \begin{bmatrix} I_{\mathcal{H}} & \Pi_+ \\ \Pi_- & I_{\mathcal{H}} \end{bmatrix} - \begin{bmatrix} I_{\mathcal{H}} & 0 \\ X & I_{\mathcal{H}} \end{bmatrix} = \begin{bmatrix} 0 & \Pi_+ \\ \Pi_- - X & 0 \end{bmatrix},$$

where V and V_Q are given by (2.6) and (2.7), it follows directly that Π_+ and $\Pi_- - X$ are compact operators. ∎

In [16] a closely related existence result was obtained under the assumption that the spectrum of the block matrix operator S only consists of algebraically and geometrically simple eigenvalues and does not have finite accumulation points.

3. Approximation. Letting $\mathcal{H}_n$ be a sequence of closed linear subspaces of $\mathcal{H}$, there exist unique operators $\pi_n : \mathcal{H} \to \mathcal{H}_n$ and $\imath_n : \mathcal{H}_n \to \mathcal{H}$ such that $\imath_n\pi_n$ is the orthogonal projection of $\mathcal{H}$ onto $\mathcal{H}_n$ and $\pi_n\imath_n$ is the identity operator on $\mathcal{H}_n$. We assume that $\imath_n\pi_n$ tends to $I_{\mathcal{H}}$ in the strong sense.

Starting from a given triple $\theta = (A_0, Q, D; \mathcal{H})$, we define $Q_n = \pi_n Q \imath_n$, $D_n = \pi_n D \imath_n$, which are selfadjoint on $\mathcal{H}_n$ and positive semidefinite whenever Q and D are positive semidefinite. Let A_{0n} be a generator of a strongly continuous semigroup on $\mathcal{H}_n$ of negative exponential type. Then a sequence of triples $\theta_n = (A_{0n}, Q_n, D_n; \mathcal{H}_n)$ is called an *approximant* to the triple θ if the following condition holds: for some $\varepsilon > 0$ we have the approximation

$$\lim_{n\to\infty} e^{\varepsilon|t|} \, \|\hat{\imath}_n E(t; -S_{0n})\hat{\pi}_n x - E(t; -S_0)x\|_{\mathcal{H}} = 0 \tag{3.1}$$

for every $x \in \mathcal{H}$, uniformly in $t \in \mathbb{R} \setminus \{0\}$. Here $\hat{\pi}_n = \pi_n \dot{+} \pi_n$, $\hat{\imath}_n = \imath_n \dot{+} \imath_n$ and $S_{0n} = A_{0n} \dot{+} (-A_{0n}^*)$ on $\mathcal{H}_n \dot{+} \mathcal{H}_n$. The sequence of triples θ_n is called a *finite dimensional approximant* to θ if it is an approximant to θ and the spaces $\mathcal{H}_n = \pi_n[\mathcal{H}]$ are finite dimensional.

We remark that it is easily seen that $\imath_n Q_n \pi_n$ converges to Q strongly, while $\imath_n D_n \pi_n$ converges to D in norm because of the compactness of D.

THEOREM 3.1. *Let $\theta_n = (A_{0n}, Q_n, D_n; \mathcal{H}_n)$ be a sequence of triples approximant to the positive semidefinite semicompact triple $\theta = (A_0, Q, D; \mathcal{H})$. Put*

$$S_n = \begin{bmatrix} A_{0n} & -D_n \\ -Q_n & -A_{0n}^* \end{bmatrix}.$$

Then

$$\lim_{n\to\infty} \|\hat{\imath}_n E(t; -S_n)\hat{\pi}_n x - E(t; -S)x\|_{\mathcal{H}} = 0 \tag{3.2}$$

for every $x \in \mathcal{H} \dot{+} \mathcal{H}$, uniformly in $t \in \mathbb{R} \setminus \{0\}$.

Proof. Consider the sequence of triples $\theta_n^Q = (A_{0n}, Q_n, 0; \mathcal{H}_n)$ approximant to the positive semidefinite triple $\theta = (A_0, Q, 0; \mathcal{H})$. Put

$$S_n^Q = \begin{bmatrix} A_{0n} & 0 \\ -Q_n & -A_{0n}^* \end{bmatrix}.$$

In analogy with (2.3) we obtain

$$E(t; -S_n^Q)x = E(t; -S_{0n})x + \int_{-\infty}^{\infty} E(t - \tau; -S_{0n})\Gamma_Q E(\tau; -S_n^Q)x \, d\tau.$$

Because of (3.1), we see that $\|E(t;-S_{0n})\|$ has a finite upper bound which is independent of $t \in \mathbb{R} \setminus \{0\}$ and $n \in \mathbb{N}$. Using dominated convergence, we take the limit under the integral sign and find that for some $\varepsilon > 0$

$$\lim_{n\to\infty} e^{\varepsilon|t|} \left\| \hat{\imath}_n E(t;-S_n^Q)\hat{\pi}_n x - E(t;-S^Q)x \right\|_{\mathcal{H}} = 0 \tag{3.3}$$

for every $x \in \mathcal{H}$, uniformly in $t \in \mathbb{R} \setminus \{0\}$.

Next, in analogy with (2.3) we have

$$E(t;-S_n) - \int_{-\infty}^{\infty} E(t-\tau;-S_n^Q)\Gamma_n^D E(\tau;-S_n)\,d\tau = E(t;-S_n^Q),$$

where $\Gamma_n^D = \begin{pmatrix} 0 & D_n \\ 0 & 0 \end{pmatrix}$. This integral equation implies that

$$\begin{aligned} \hat{\imath}_n E(t;-S_n)\hat{\pi}_n x - \int_{-\infty}^{\infty} \hat{\imath}_n E(t-\tau;-S_n^Q)\Gamma_n^D E(\tau;-S_n)\hat{\pi}_n x\,d\tau \\ = \hat{\imath}_n E(t;-S_n^Q)\hat{\pi}_n x, \end{aligned} \tag{3.4}$$

where $x \in \mathcal{H}\dot{+}\mathcal{H}$. Note that $\Gamma_n^D = \hat{\pi}_n \Gamma_D \hat{\imath}_n$, so that (3.4) can be written in the form

$$\begin{aligned} \hat{\imath}_n E(t;-S_n)\hat{\pi}_n x - \int_{-\infty}^{\infty} \hat{\imath}_n E(t-\tau;-S_n^Q)\hat{\pi}_n \Gamma_D \cdot \hat{\imath}_n E(\tau;-S_n)\hat{\pi}_n x\,d\tau \\ = \hat{\imath}_n E(t;-S_n^Q)\hat{\pi}_n x, \end{aligned}$$

where $x \in \mathcal{H}\dot{+}\mathcal{H}$. Equation (3.3) and the compactness of Γ_D imply that for some $\varepsilon > 0$

$$\lim_{n\to\infty} e^{\varepsilon|t|} \left\| \hat{\imath}_n E(t;-S_n^Q)\hat{\pi}_n \Gamma_D - E(t;-S^Q)\Gamma_D \right\|_{\mathcal{H}} = 0,$$

uniformly in $t \in \mathbb{R} \setminus \{0\}$. Because of the unique solvability of (3.4) on the complex Banach space of bounded continuous $\mathcal{H}_n$-valued functions on the real line with a possible jump discontinuity in $t = 0$, in combination with (3.3), we obtain (3.2) as claimed. □

Let

$$X_n = \int_0^{\infty} e^{\tau A_{0n}^*} Q_n e^{\tau A_{0n}}\,d\tau$$

be the unique solution of the Lyapunov equation

$$A_{0n}^* X_n + X_n A_{0n} = -Q_n.$$

Using dominated convergence one easily proves that, under the hypotheses of Theorem 3.1,

$$\lim_{n\to\infty} \|\imath_n X_n \pi_n x - Xx\| = 0, \qquad x \in \mathcal{H}. \tag{3.5}$$

Similarly, the unique solution

$$Y_n = \int_0^{\infty} e^{\tau A_{0n}} D_n e^{\tau A_{0n}^*}\,d\tau$$

of the Lyapunov equation

$$A_{0n}Y_n + Y_nA_{0n}^* = -D_n$$

has the property that

$$\lim_{n\to\infty} \|\imath_n Y_n \pi_n x - Yx\| = 0, \qquad x \in \mathcal{H},$$

where

$$Y = \int_0^\infty e^{\tau A_0} D\, e^{\tau A_0^*}\, d\tau$$

is the unique solution of the Lyapunov equation

$$A_0Y + YA_0^* = -D.$$

Let us now prove the strong stability of Π_- and the stability of Π_+ in the norm if a positive semidefinite and semicompact triple is approximated by a sequence of triples in the sense of the above definition. The obvious way to do so is to study the operator Wiener–Hopf equation

$$u(t;x) - \int_0^\infty E(t-\tau;-S_0)\Gamma u(\tau;x)\, d\tau = E(t;-S_0)x, \tag{3.6}$$

where $x \in \operatorname{Ker} P_0$ and $t > 0$, or the operator Wiener–Hopf equation

$$v(t;x) - \int_{-\infty}^0 E(t-\tau;-S_0)\Gamma v(\tau;x)\, d\tau = E(t;-S_0)x, \tag{3.7}$$

where $x \in \operatorname{Im} P$ and $t < 0$. Unfortunately, their integral kernel $E(\cdot;-S_0)\Gamma$ is, in general, not Bochner integrable. If it were, one would trivially obtain the solutions of (3.6) and (3.7) as follows:

$$u(t;x) = E(t;-S)\mathcal{P}^{(l)}x, \qquad v(t;x) = E(t;-S)\mathcal{P}^{(r)}x.$$

Let us therefore introduce the modified operator convolution kernel

$$\begin{aligned} K(t;-S_0) &= \begin{bmatrix} Q^{1/2} & 0 \\ 0 & D^{1/2} \end{bmatrix} E(t;-S_0) \begin{bmatrix} 0 & D^{1/2} \\ Q^{1/2} & 0 \end{bmatrix} \\ &= \begin{cases} \begin{bmatrix} 0 & -Q^{1/2}e^{-tA_0}D^{1/2} \\ 0 & 0 \end{bmatrix}, & t < 0, \\ \\ \begin{bmatrix} 0 & 0 \\ D^{1/2}e^{tA_0^*}Q^{1/2} & 0 \end{bmatrix}, & t > 0. \end{cases} \end{aligned}$$

Note that $K(t;-S_0)$ is compact and norm continuous in $t \neq 0$. This integral kernel satisfies

$$\begin{bmatrix} Q^{1/2} & 0 \\ 0 & D^{1/2} \end{bmatrix} E(t;-S_0)\Gamma = K(t;-S_0) \begin{bmatrix} Q^{1/2} & 0 \\ 0 & D^{1/2} \end{bmatrix}$$

and leads to operator Wiener–Hopf equations with Bochner integrable convolution kernel and symbol $W(\lambda)$ defined by (2.8). Indeed, these equations are given by

$$w(t;x) - \int_0^\infty K(t-\tau;-S_0)w(\tau;x)\,d\tau = \begin{bmatrix} Q^{1/2} & 0 \\ 0 & D^{1/2} \end{bmatrix} E(t;-S_0)x, \tag{3.8}$$

where $x \in \operatorname{Ker} P_0$ and $t > 0$, and by

$$z(t;x) - \int_{-\infty}^0 K(t-\tau;-S_0)z(\tau;x)\,d\tau = \begin{bmatrix} Q^{1/2} & 0 \\ 0 & D^{1/2} \end{bmatrix} E(t;-S_0)x, \tag{3.9}$$

where $x \in \operatorname{Im} P$ and $t < 0$. Equations (3.8) and (3.9) are uniquely solvable, because their symbol $W(\lambda)$ has left and right canonical Wiener–Hopf factorizations. Once (3.8) and (3.9) have been solved, we have

$$u(t;x) = E(t;-S_0)x + \int_0^\infty E(t-\tau;-S_0)\begin{bmatrix} 0 & D^{1/2} \\ Q^{1/2} & 0 \end{bmatrix} w(\tau;x)d\tau \tag{3.10}$$

for $x \in \operatorname{Ker} P_0$ and $t > 0$, and

$$v(t;x) = E(t;-S_0)x + \int_{-\infty}^0 E(t-\tau;-S_0)\begin{bmatrix} 0 & D^{1/2} \\ Q^{1/2} & 0 \end{bmatrix} z(\tau;x)d\tau \tag{3.11}$$

for $x \in \operatorname{Im} P_0$ and $t < 0$. We then finally obtain

$$\mathcal{P}^{(l)}x = u(0^+;x), \qquad \mathcal{P}^{(r)}x = -v(0^-;x),$$

and hence [cf. (2.14)]

$$\Pi_- x = \begin{bmatrix} 0 & I_{\mathcal{H}} \end{bmatrix} u(0^+;x), \quad \Pi_+ x = -\begin{bmatrix} I_{\mathcal{H}} & 0 \end{bmatrix} v(0^-;x). \tag{3.12}$$

THEOREM 3.2. *Let $\theta_n = (A_{0n}, Q_n, D_n; \mathcal{H}_n)$ be a sequence of triples approximant to the positive semidefinite semicompact triple $\theta = (A_0, Q, D; \mathcal{H})$. Then*

$$\lim_{n\to\infty} \|\imath_n \Pi_{-,n} \pi_n x - \Pi_- x\| = 0, \tag{3.13}$$

and

$$\lim_{n\to\infty} \|\imath_n \Pi_{+,n} \pi_n x - \Pi_+ x\| = 0 \tag{3.14}$$

for every $x \in \mathcal{H}$.

Proof. From (3.1), the strong convergence $\imath_n Q_n^{1/2} \pi_n \to Q^{1/2}$ and the compactness of $D^{1/2}$, we obtain for some $\varepsilon > 0$

$$\lim_{n\to\infty} e^{\varepsilon|t|}\, \|\hat{\imath}_n K(t;-S_{0n})\hat{\pi}_n - K(t)\| = 0,$$

uniformly in $t \in \mathbb{R} \setminus \{0\}$, and hence for some $\varepsilon > 0$

$$\lim_{n\to\infty} \int_{-\infty}^\infty e^{\varepsilon|t|}\, \|\hat{\imath}_n K(t;-S_{0n})\hat{\pi}_n - K(t)\|\, dt = 0.$$

Thus, using (3.1) and the unique solvability of (3.8), we get for some $\varepsilon > 0$

$$\lim_{n\to\infty} e^{\varepsilon|t|}\, \|\hat{\imath}_n w_n(t;\hat{\pi}_n x) - w(t;x)\| = 0$$

for every $x \in \mathcal{H}\dot{+}\mathcal{H}$, uniformly in $t \in \mathbb{R}^+$. Similarly, for some $\varepsilon > 0$ we have

$$\lim_{n\to\infty} e^{\varepsilon|t|} \|\hat{\imath}_n z_n(t;\hat{\pi}_n x) - z(t;x)\| = 0$$

for every $x \in \mathcal{H}\dot{+}\mathcal{H}$, uniformly in $t \in \mathbb{R}^-$. With the help of (3.1), (3.10), and (3.11), we find for some $\varepsilon > 0$

$$\lim_{n\to\infty} e^{\varepsilon|t|} \|\hat{\imath}_n u_n(t;\hat{\pi}_n x) - u(t;x)\| = 0$$

for every $x \in \mathcal{H}\dot{+}\mathcal{H}$, uniformly in $t \in \mathbb{R}^+$, as well as

$$\lim_{n\to\infty} e^{\varepsilon|t|} \|\hat{\imath}_n v_n(t;\hat{\pi}_n x) - v(t;x)\| = 0$$

for every $x \in \mathcal{H}\dot{+}\mathcal{H}$, uniformly in $t \in \mathbb{R}^-$. Using (3.12) we then easily obtain (3.13) and (3.14). □

The following result strengthens the convergence properties stated in Theorem 3.2.

THEOREM 3.3. *Let $\theta_n = (A_{0n}, Q_n, D_n; \mathcal{H}_n)$ be a sequence of triples approximant to the positive semidefinite semicompact triple $\theta = (A_0, Q, D; \mathcal{H})$. Then*

$$\lim_{n\to\infty} \|\imath_n(\Pi_{-,n} - X_n)\pi_n - (\Pi_- - X)\| = 0, \tag{3.15}$$

$$\lim_{n\to\infty} \|\imath_n \Pi_{+,n}\pi_n - \Pi_+\| = 0. \tag{3.16}$$

Proof. From (3.8) and $E(t;-S_0) = 0\dot{+}e^{tA_0^*}$ for $t > 0$ it is clear that for every $t > 0$ the right-hand side of (3.8) can be viewed as the result of applying a compact operator to a vector $x \in \mathcal{H}$. Since (3.1) and the compactness of $D^{1/2}$ imply that for some $\varepsilon > 0$

$$\lim_{n\to\infty} \left\| \imath_n D_n^{1/2} e^{tA_{0n}^*}(I - P_{0n})\pi_n - D^{1/2}e^{tA_0^*}(I - P_0) \right\| = 0, \qquad t > 0,$$

we have

$$\lim_{n\to\infty} \left\| \hat{\imath}_n \begin{bmatrix} Q_n^{1/2} & 0 \\ 0 & D_n^{1/2} \end{bmatrix} E(t;-S_{0n})\hat{\pi}_n - \begin{bmatrix} Q^{1/2} & 0 \\ 0 & D^{1/2} \end{bmatrix} E(t;-S_0) \right\| = 0,$$

and this allows one to sharpen the derivation of (3.14) and to obtain (3.16) instead.

To prove (3.15), we replace (3.7), (3.9), and (3.11) by

$$v_Q(t;x) - \int_{-\infty}^{0} E(t-\tau;-S_Q)\Gamma_D v_Q(\tau;x)\, d\tau = E(t;-S_Q)x, \tag{3.17}$$

$$z_Q(t;x) - \int_{-\infty}^{0} K(t-\tau;-S_Q) z_Q(\tau;x)\, d\tau = \begin{bmatrix} 0 & 0 \\ 0 & D^{1/2} \end{bmatrix} E(t;-S_Q)x, \tag{3.18}$$

$$v_Q(t;x) = E(t;-S_Q)x + \int_{-\infty}^{0} E(t-\tau;-S_Q) \begin{bmatrix} 0 & D^{1/2} \\ 0 & 0 \end{bmatrix} z_Q(\tau;x) d\tau\,, \tag{3.19}$$

respectively, where $x \in \mathcal{H}$, $\Gamma_D = \begin{pmatrix} 0 & D \\ 0 & 0 \end{pmatrix}$, and the convolution kernel $K(t;-S_Q)$ satisfies

$$K(t;-S_Q) \begin{bmatrix} 0 & 0 \\ 0 & D^{1/2} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & D^{1/2} \end{bmatrix} E(t;-S_Q).$$

Repeating the proof of (3.16) with the help of (3.17), (3.18), and (3.19) we obtain (3.15). □

It remains to consider the approximation of the C_0-semigroups generated by $A_0 - D\Pi_-$ and $A_0^* + Q\Pi_+$. Indeed, from (2.15) we easily derive the identity

$$S\begin{bmatrix} I_{\mathcal{H}} & \Pi_+ \\ \Pi_- & I_{\mathcal{H}} \end{bmatrix} = \begin{bmatrix} A_0 & -D \\ -Q & -A_0^* \end{bmatrix}\begin{bmatrix} I_{\mathcal{H}} & \Pi_+ \\ \Pi_- & I_{\mathcal{H}} \end{bmatrix} = \begin{bmatrix} I_{\mathcal{H}} & \Pi_+ \\ \Pi_- & I_{\mathcal{H}} \end{bmatrix}\begin{bmatrix} A_0 - D\Pi_- & 0 \\ 0 & -A_0^* - Q\Pi_+ \end{bmatrix},$$

where $A_0 - D\Pi_-$ and $A_0^* + Q\Pi_+$ both generate exponentially decaying C_0-semigroups on $\mathcal{H}$. Writing down the analogous identity for resolvent operators and applying the inverse Laplace transform, we get

$$\begin{bmatrix} I_{\mathcal{H}} & \Pi_+ \\ \Pi_- & I_{\mathcal{H}} \end{bmatrix}^{-1} E(t;-S)\begin{bmatrix} I_{\mathcal{H}} & \Pi_+ \\ \Pi_- & I_{\mathcal{H}} \end{bmatrix} = \begin{cases} e^{t(A_0 - D\Pi_-)} \dot{+} 0_{\mathcal{H}}, & t > 0, \\ 0_{\mathcal{H}} \dot{+} \left(-e^{-t(A_0^* + Q\Pi_+)}\right), & t < 0, \end{cases} \tag{3.20}$$

where $0_{\mathcal{H}}$ denotes the zero operator on $\mathcal{H}$.

THEOREM 3.4. *Let $\theta_n = (A_{0n}, Q_n, D_n; \mathcal{H}_n)$ be a sequence of triples approximant to the positive semidefinite semicompact triple $\theta = (A_0, Q, D; \mathcal{H})$. Then for $t > 0$ we have*

$$\lim_{n\to\infty} \left\| \imath_n e^{t(A_{0n} - D_n\Pi_{-,n})}\pi_n - e^{t(A_0 - D\Pi_-)} \right\| = 0, \tag{3.21}$$

$$\lim_{n\to\infty} \left\| \imath_n e^{t(A_{0n}^* + Q_n\Pi_{+,n})}\pi_n - e^{t(A_0^* - Q\Pi_+)} \right\| = 0, \tag{3.22}$$

uniformly in t on compact intervals of either $[0,\infty)$ or $(-\infty, 0]$.

Proof. Because of (3.2) and (3.20), it suffices to prove that for each $x \in \mathcal{H}$ we have

$$\lim_{n\to\infty} \|[\hat{\imath}_n M_n \hat{\pi}_n - M]\, x\| = 0, \tag{3.23}$$

$$\lim_{n\to\infty} \left\| \left[\hat{\imath}_n M_n^{-1} \hat{\pi}_n - M^{-1}\right] x \right\| = 0, \tag{3.24}$$

where

$$M = \begin{bmatrix} I_{\mathcal{H}} & \Pi_+ \\ \Pi_- & I_{\mathcal{H}} \end{bmatrix}, \qquad M_n = \begin{bmatrix} I_{\mathcal{H}_n} & \Pi_{+,n} \\ \Pi_{-,n} & I_{\mathcal{H}_n} \end{bmatrix}.$$

Indeed, (3.13), (3.16), and the compactness of the operator Π_+ imply that

$$\lim_{n\to\infty} \|\imath_n(I_{\mathcal{H}_n} - \Pi_{-,n}\Pi_{+,n})\pi_n - (I_{\mathcal{H}} - \Pi_-\Pi_+)\| = 0.$$

Now note that $I_{\mathcal{H}} - \Pi_-\Pi_+$ is invertible, as a result of the existence of the projection P [cf. (2.11)]. Thus

$$\lim_{n\to\infty} \left\|\imath_n(I_{\mathcal{H}_n} - \Pi_{-,n}\Pi_{+,n})^{-1}\pi_n - (I_{\mathcal{H}} - \Pi_-\Pi_+)^{-1}\right\| = 0,$$

and by taking the adjoint

$$\lim_{n\to\infty} \left\|\imath_n(I_{\mathcal{H}_n} - \Pi_{+,n}\Pi_{-,n})^{-1}\pi_n - (I_{\mathcal{H}} - \Pi_+\Pi_-)^{-1}\right\| = 0.$$

We now easily show that

$$M^{-1} = \begin{bmatrix} (I_{\mathcal{H}} - \Pi_+\Pi_-)^{-1} & -(I_{\mathcal{H}} - \Pi_+\Pi_-)^{-1}\Pi_+ \\ -(I_{\mathcal{H}} - \Pi_-\Pi_+)^{-1}\Pi_- & (I_{\mathcal{H}} - \Pi_-\Pi_+)^{-1} \end{bmatrix}. \tag{3.25}$$

Hence from (3.25) and the analogous expression for M_n^{-1}, we now easily derive (3.23) and (3.24). □

4. Conclusions and remarks. In this paper, exponentially dichotomous block matrix operators on $\mathcal{H}\dot{+}\mathcal{H}$ have been studied as additive perturbations of exponentially dichotomous operators of the type $A_0\dot{+}(-A_0^*)$. This allows a considerable range of LQ-optimal control theory applications; for instance, the example in [16] concerning the heat equation could be dealt with in this way. (We did not do so explicitly, because we expect no better results than the ones one can expect for a Hamiltonian that is a Riesz spectral operator, and taking as approximations for $\mathcal{H}_n$ the spaces spanned by the first n vectors in a properly constructed Riesz basis of eigenvectors of the Hamiltonian. The results would be no better than the ones already existing in the literature.)

Also we considered (possibly finite dimensional) approximations. In connection with the latter topic there are still many open questions. Questions that are natural from a numerical analysis point of view come to mind; for instance, how is the speed of convergence in the results described in section 3 tied to the speed of convergence in (3.1) and to the speed of convergence of $\imath_n D\pi_n$ to D? What about Lipschitz estimates and relative error bounds? All these points are open problems, although an analysis of our proofs may provide some answers.

Finally, delay systems defy application of the existing results. In order to be able to deal with applications to delay systems we would need criteria for exponential dichotomy, where the linear operator is not a perturbation of a naturally given exponentially dichotomous operator, and where a Banach space setting is adopted.

Acknowledgments. The authors would like to express their appreciation for the strenuous efforts of the referees to improve the contents and presentation of this paper. Cornelis V. M. van der Mee is greatly indebted to the Department of Mathematics of Vrije Universiteit, Amsterdam for its hospitality during visits in which the major part of his research was conducted.

REFERENCES

[1] T. Ya. Azizov and I. S. Iokhvidov, *Linear Operators in Spaces with an Indefinite Metric*, John Wiley, Chichester, 1989.

[2] H. T. Banks and K. Kunisch, *The linear regulator problem for parabolic systems*, SIAM J. Control Optim., 22 (1984), pp. 684–698.

[3] H. Bart, I. Gohberg, and M. A. Kaashoek, *Minimal Factorization of Matrix and Operator Functions*, Birkhäuser, Basel, 1979.

[4] H. Bart, I. Gohberg, and M. A. Kaashoek, *Wiener–Hopf factorization, inverse Fourier transforms and exponentially dichotomous operators*, J. Funct. Anal., 68 (1986), pp. 1–42.

[5] H. Bart, I. Gohberg, and M. A. Kaashoek, *Wiener–Hopf equations with symbols analytic in a strip*, Constructive Methods of Wiener–Hopf Factorization Oper. Theory Adv. Appl., I. Gohberg and M. A. Kaashoek, eds., Birkhäuser, Basel, Switzerland, 1986, pp. 39–74.

[6] S. Bochner and R. S. Phillips, *Absolutely convergent Fourier expansions for non-commutative normed rings*, Ann. of Math., 43 (1942), pp. 409–418.

[7] R. F. Curtain and H. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Springer, New York, 1995.

[8] J. S. Gibson, *Linear-quadratic optimal control of hereditary differential systems: Infinite dimensional Riccati equations and numerical approximations*, SIAM J. Control Optim., 21 (1983), pp. 95–139.

[9] I. C. Gohberg and J. Leiterer, *The factorization of operator functions with respect to a contour.* II. *The canonical factorization of operator functions close to the identity*, Math. Nach., 54 (1972), pp. 41–74 (in Russian). Appendix.

[10] W. Greenberg, V. Protopopescu, and C. V. M. van der Mee, *Boundary Value Problems in Abstract Kinetic Theory*, Birkhäuser, Verlag, Basel, Switzerland, 1987.

[11] J. K. Hale and S. M. Verduyn-Lunel, *Introduction to Functional Differential Equations*, Applied Mathematical Sciences 99, Springer, New York, 1993.

[12] K. Ito, *Strong convergence and convergence rates of approximating solutions for algebraic Riccati equations in Hilbert spaces*, Distributed Parameter Systems, F. Kappel, K. Kunisch, and W. Schappacher, eds., Springer, New York, 1987, pp. 151–166.

[13] K. Ito, *Finite-dimensional compensators for infinite-dimensional systems via Galerkin-type approximation*, SIAM J. Control Optim., 28 (1990), pp. 1251–1269.

[14] K. Ito and K. A. Morris, *An approximation theory of solutions to operator Riccati equations for H^∞ control*, SIAM J. Control Optim., 36 (1998), pp. 82–99.

[15] B. van Keulen, *H^∞-Control for Distributed Parameter Systems: A State-space Approach*, Birkhäuser, Boston, 1993.

[16] C. R. Kuiper and H. J. Zwart, *Connections between the algebraic Riccati equation and the Hamiltonian for Riesz-spectral systems*, J. Math. Systems, Estim. Control, 6 (1996), pp. 1–48.

[17] P. Lancaster and L. Rodman, *Algebraic Riccati Equations*, Oxford University Press, New York, 1995.

[18] H. Langer, A. C. M. Ran, and B. A. van de Rotten, *Invariant subspaces of infinite dimensional Hamiltonians and solutions of the corresponding Riccati equations*, Linear Operators and Matrices, Oper. Theory Adv. Appl., I. Gohberg and H. Langer, eds., Birkhäuser, Basel and Boston, 2002, pp. 235–254.

[19] C. V. M. van der Mee and A. C. M. Ran, *Perturbation results for exponentially dichotomous operators on general Banach spaces,* J. Funct. Anal., 210 (2004), pp. 193–213.

[20] J. C. Oostveen, R. F. Curtain, and K. Ito, *An approximation theory for strongly stabilizing solutions to the operator LQ Riccati equation*, SIAM J. Control Optim., 38 (2000), pp. 1909–1937.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 4, pp. 1436–1473

ON EXISTENCE OF LIMIT OCCUPATIONAL MEASURES SET OF A CONTROLLED STOCHASTIC DIFFERENTIAL EQUATION*

VIVEK BORKAR† AND VLADIMIR GAITSGORY‡

Abstract. We establish that, under certain conditions, the set of occupational measures as well as the set of mathematical expectations of occupational measures generated by the admissible controls and the corresponding solutions of a controlled stochastic differential equation (CSDE) converge (with the time horizon tending to infinity) to a set called limit occupational measures set (LOMS) and we show that this limit set coincides with the set of stationary marginal distributions of the CSDE. We also demonstrate the applicability of our results for averaging of singularly perturbed CSDE.

Key words. singularly perturbed controlled stochastic differential equations, occupational measures, averaging method, limit occupational measures sets, approximation of slow motions

AMS subject classifications. 34E15, 34C29, 34A60, 93C70

DOI. 10.1137/S0363012904443476

1. Introduction.

In this paper we establish that, under certain conditions, the set of occupational measures as well as the set of mathematical expectations of occupational measures generated by the admissible controls and the corresponding solutions of a controlled stochastic differential equation (CSDE) converge (with the time horizon tending to infinity) to a set called limit occupational measures set (LOMS) and we show that this limit set coincides with the set of stationary marginal distributions of the CSDE.

The motivation for our study is the applicability of results to averaging of singularly perturbed CSDE. We show that, given a singularly perturbed CSDE, the slow components of its state variables are approximated by the solutions of the averaged system in which the controls take values in the LOMS of the system describing the fast dynamics. In the deterministic control setting, a similar approach was used in [4], [5], [6], [7], [8], [26], [27], [28] (see also [18], [19], [24], [25], [30], [46], [51] for related results). The current paper is based on a combination of ideas developed in the deterministic setting and also on results of [11], [13], and [49] which describe the set of stationary marginal distributions of the CSDE.

Note that singularly perturbed problems of control and optimization have been considered in both deterministic and stochastic literature (see [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [12], [17], [18], [19], [20], [21], [22], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [35], [36], [37], [38], [39], [41], [42], [43], [44], [45], [46], [48], [50], [51], [53] and references therein). Singularly perturbed CSDE, in particular, have been studied in [2], [3], [12], [32], [33], and [38], where earlier references can also be found. In [2], [3], and [12] the Hamilton–Jacobi–Bellman (HJB) equations corresponding to

*Received by the editors May 5, 2004; accepted for publication (in revised form) February 8, 2005; published electronically November 4, 2005.

http://www.siam.org/journals/sicon/44-4/44347.html

†School of Technology and Computer Science, Tata Institute of Fundamental Research, Homi Bhabha Rd., Mumbai 400005, India (borkar@tifr.res.in). This project was partly supported by grant III.5(157)/99-ET from the Department of Science and Technology, Government of India.

‡School of Mathematics, University of South Australia, Mawson Lakes Campus, Mawson Lakes SA, 5095, Australia (v.gaitsgory@unisa.edu.au). This project was supported by the Australian Research Council Grant DP0346099.

singularly perturbed CSDE were analyzed. In [2], in particular, it was shown that the optimal value function of the problem of optimal control of singularly perturbed CSDE with a fairly general structure (the only structural constraint was the periodicity in fast variables) converges to a viscosity solution of the HJB equation in which the fast variables are averaged out. In [12], results concerning asymptotic behavior of singularly perturbed CSDE with nongenerate diffusion were obtained. In [32] and [33], singularly perturbed CSDE linear in fast variables were studied and the limit behavior of the attainability sets was described. In [38], weak convergence methods were used to establish a number of important results concerning mainly the case when the fast dynamics are not controlled.

The results obtained in this paper can be used for an approximation of the slow dynamics of singularly perturbed CSDE having a general structure (that is, in particular, nonlinear and nonperiodic in fast variables, and having a controlled fast dynamics). It allows one to treat stochastic nondegenerate and degenerate diffusion cases (as well as a purely deterministic case) in a similar manner and also to deal with the situation when the classical approach, based on equating the singular perturbation parameter to zero, may not lead to a correct approximation of the slow dynamics.

The paper is organized as follows: In section 2 we introduce some notations and define the LOMS of the CSDE as the limit towards which converges the set of mathematical expectations of occupational measures generated by the controls and solutions of the CSDE. In section 3 we identify necessary and sufficient conditions for the LOMS to exist (Theorems 3.2 and 3.3) and also sufficient conditions for every element of the LOMS to be asymptotically approximated (in mean) by an occupational measure obtained with some admissible control (Theorem 3.4). In section 4 we establish that if the LOMS exists, it coincides with the set of marginal stationary distributions of the CSDE (Theorem 4.1) and show that every occupational measure converges to this set in mean (Theorem 4.2). The proofs for sections 3 and 4 are contained in sections 6 and 7.

In section 5 we demonstrate the applicability of above mentioned results to averaging of singularly perturbed CSDE (Theorem 5.1). The proofs for section 5 are contained in section 8.

2. Preliminaries. For a compact set U and m dimensional Euclidean space R^m, $\mathcal{P}(U\times R^m)$ and $\mathcal{P}(U\times\bar{R}^m)$ will stand for the spaces of probability measures defined on the σ- algebras of Borel subsets of $U\times R^m$ and $U\times\bar{R}^m$, respectively, with $\bar{R}^m$ being the one point compactification of R^m (see, e.g., [23, p. 126]). Note that any probability measure μ on $U\times R^m$ may be identified with the unique probability measure on $U\times\bar{R}^m$ that restricts to μ on $U\times R^m$ and perforce assigns zero probability to its complement. Conversely, any probability measure μ on $U\times\bar{R}^m$, assigning probability one to $U\times R^m$, defines a unique probability measure on $U\times R^m$. Thus, $\mathcal{P}(U\times R^m)$ can be considered as a subset of $\mathcal{P}(U\times\bar{R}^m)$ consisting of the probability measures μ on $U\times\bar{R}^m$ with $\mu(U\times R^m)=1$.

The set $\mathcal{P}(U\times\bar{R}^m)$ will be treated as a compact metric space with a metric $\rho(\cdot,\cdot)$ consistent with its weak convergence topology which is metrizable and compact. There are many ways of how $\rho(\cdot,\cdot)$ can be introduced. In this paper the following definition will be used (in most of the cases): for any $\mu',\mu''\in\mathcal{P}(U\times\bar{R}^m)$,

$$\rho(\mu',\mu'')\stackrel{def}{=}\sum_{i=1}^{\infty}2^{-i}\left|\int f_i(u,y)\mu'(du,dy)-\int f_i(u,y)\mu''(du,dy)\right|, \tag{2.1}$$

where $f_i(u,y), i=1,2,\dots$, is the sequence of Lipschitz continuous functions which is

dense in the unit ball of $C(U \times \bar{R}^m)$ (the space of continuous functions defined on $U \times \bar{R}^m$). Using the metric ρ, one can define the Hausdorff metric ρ_H on the set of subsets of $\mathcal{P}(\bar{R}^m \times U)$ as follows: $\forall \mathcal{M}_i \subset \mathcal{P}(U \times \bar{R}^m), i = 1, 2,$

$$\rho_H(\mathcal{M}_1, \mathcal{M}_2) \stackrel{def}{=} \max \left\{ \sup_{\mu \in \mathcal{M}_1} \rho(\mu, \mathcal{M}_2), \sup_{\mu \in \mathcal{M}_2} \rho(\mu, \mathcal{M}_1) \right\}, \tag{2.2}$$

where (here and in what follows)

$$\rho(\mu, \mathcal{M}_i) \stackrel{\text{def}}{=} \inf_{\mu' \in \mathcal{M}_i} \rho(\mu, \mu'). \tag{2.3}$$

Remark 1. Note that, if $\mathcal{M}_1$ and/or $\mathcal{M}_2$ are not closed, then from the fact that $\rho_H(\mathcal{M}_1, \mathcal{M}_2) = 0$ it does not follow that $\mathcal{M}_1 = \mathcal{M}_2$. That is, $\rho_H(\cdot, \cdot)$ is, in fact, a semimetric. By some abuse of terminology we still will refer to it as to a metric keeping in mind that its equality to zero is equivalent to the equality of the closures of the corresponding sets.

We will be dealing with a CSDE

$$dy(\tau) = a(u(\tau), y(\tau))d\tau + b(y(\tau))dW(\tau) \tag{2.4}$$

with the initial conditions

$$y(0) = y_0, \tag{2.5}$$

where:

- the functions $a(u, y) : U \times R^m \to R^m$ and $b(y) : R^m \to R^{m \times m}$ are continuous and satisfy Lipschitz conditions in y, with $a(u, y)$ satisfying it uniformly with respect to $u \in U$;
- U is a compact metric space;
- $W(\cdot)$ is an R^m-valued standard Brownian motion;
- y_0 is an R^m-valued random variable independent of $W(\cdot)$;
- *admissible controls* $u(\cdot)$ are $U-$valued random processes progressively measurable with respect to a right continuous and complete filtration $\{\mathcal{F}_\tau\} \subset \mathcal{F}$ of σ-fields (with $(\Omega, \mathcal{F}, \mathcal{P})$ being a given probability space) such that:
  - $\{y_0$ and $W(\theta); \theta \le \tau\}$ is measurable with respect to $\mathcal{F}_\tau$ for $\tau \ge 0$,
  - For $\tau' \ge \tau \ge 0$, $W(\tau') - W(\tau)$ is independent of $\mathcal{F}_\tau$.

Let $S > 0$, $u(\cdot)$ be an admissible control and $y(\cdot)$ be the corresponding solution of the CSDE (2.4) on the interval $[0, S]$. Define the occupational measure $\mu_S^{u(\cdot), y(\cdot)}$ generated by the pair $(u(\cdot), y(\cdot))$ on this interval by taking

$$\mu_S^{u(\cdot), y(\cdot)}(Q) \stackrel{def}{=} \frac{1}{S} meas\{\tau \ : \ (u(\tau), y(\tau)) \in Q\} \tag{2.6}$$

for any Borel subset Q of $U \times \bar{R}^m$, with *meas* standing for the Lebesgue measure on $[0, S]$. Note that $\mu_S^{u(\cdot), y(\cdot)}$ is uniquely defined by

$$\int f_i(u, y) \mu_S^{u(\cdot), y(\cdot)}(du, dy) = \frac{1}{S} \int_0^S f_i(u(\tau), y(\tau)) d\tau, \quad i = 1, 2, \ldots, \tag{2.7}$$

where $f_i(\cdot)$ are as in (2.1). From (2.7) it follows that, for any fixed $\mu \in \mathcal{P}(U \times \bar{R}^m)$, the value of the metric $\rho(\mu, \mu_S^{u(\cdot), y(\cdot)})$ is a random variable, which allows one to easily verify

that $\mu_S^{u(\cdot),y(\cdot)}$ is a $\mathcal{P}(U \times \bar{R}^m)$- valued random variable. Define also the mathematical expectation $E[\mu_S^{u(\cdot),y(\cdot)}]$ of $\mu_S^{u(\cdot),y(\cdot)}$ as the probability measure on $\bar{R}^m \times U$ such that

$$E[\mu_S^{u(\cdot),y(\cdot)}](Q) \stackrel{def}{=} \frac{1}{S} E[meas\{\tau \ : \ (y(\tau), u(\tau)) \in Q\}] \tag{2.8}$$

for any Borel subset Q of $U \times \bar{R}^m$. By (2.7) and (2.8),

$$\int f_i(u,y) E[\mu_S^{u(\cdot),y(\cdot)}](du,dy) = E\left[\int_0^S f_i(u,y)\mu_S^{u(\cdot),y(\cdot)}(du,dy)\right]$$

$$= E\left[\frac{1}{S}\int_0^S f_i(u(\tau),y(\tau))d\tau\right], \quad i = 1,2,\ldots, \tag{2.9}$$

with $E[\mu_S^{u(\cdot),y(\cdot)}]$ being uniquely defined by these equations.

Denote by $\mathcal{M}(S,y_0)$ and $E[\mathcal{M}(S,y_0)]$ the collections of the occupational measures and their mathematical expectations:

$$\mathcal{M}(S,y_0) \stackrel{\text{def}}{=} \bigcup_{(u(\cdot),\ y(\cdot))} \{\mu_S^{u(\cdot),y(\cdot)}\}, \quad E[\mathcal{M}(S,y_0)] \stackrel{\text{def}}{=} \bigcup_{(u(\cdot),\ y(\cdot))} \{E[\mu_S^{u(\cdot),y(\cdot)}]\}, \tag{2.10}$$

where the unions are over all admissible controls and corresponding solutions of (2.4) with the initial condition (2.5).

By analogy with the deterministic setting (see [26], [27], and [28]), we introduce the following definition.

DEFINITION. *A convex and compact set $\mathcal{M} \subset \mathcal{P}(U \times R^m)$ will be called LOMS of the CSDE* (2.4) *with respect to the initial conditions having probability distributions from a given class $\mathcal{C}$ if, for any initial conditions with the distribution from this class,*

$$\rho_H(E[\mathcal{M}(S,y_0)],\mathcal{M}) \leq \ \nu^{\mathcal{C}}(S), \qquad \lim_{S\to\infty} \nu^{\mathcal{C}}(S) = 0. \tag{2.11}$$

The following assumption about the solutions of (2.4) will be used throughout the paper.

Assumption 1. There exists $\alpha > 0$ such that any solution of (2.4) obtained with an admissible control satisfies the inequality

$$\sup_{\tau,u(\cdot)} E[||y(\tau)||^\alpha] \leq c_1(E[||y_0||^\alpha] + 1), \ c_1 \ = \text{const}. \tag{2.12}$$

As an example let us consider the case when the CSDE (2.4) is linear. That is,

$$a(u,y) \stackrel{\text{def}}{=} A_1 y + A_2 u, \quad b(y) \stackrel{\text{def}}{=} A_3, \tag{2.13}$$

where U is a compact subset of R^s (for some natural s) and $A_i, i = 1,2,3$, are matrices of the corresponding dimensions. In this case the solution of (2.4) can be presented in the form

$$y(\tau) = e^{A_1\tau}y(0) + \int_0^\tau e^{A_1(\tau-\tau')}A_2u(\tau')d\tau' + \int_0^\tau e^{A_1(\tau-\tau')}A_3dW(\tau') \tag{2.14}$$

and it is easy to verify that Assumption 1 will be valid with $\alpha = 2$ if the eigenvalues of A_1 have negative real parts. Note that, for general nonlinear systems, sufficient conditions for Assumption 1 to be valid can be derived from the existence of the corresponding Liapunov functions (see, e.g., [11], [16], and [34] for classical results on the uncontrolled case).

3. Strong and weak h-approximation conditions. Let $h(u,y) : U \times \bar{R}^m \to R^j$ be defined by

$$h(u,y) \stackrel{\text{def}}{=} (f_1(u,y), f_2(u,y), \dots, f_j(u,y)), \tag{3.1}$$

where, as above, $f_i(\cdot)$ are as in the definition of the metric $\rho(\cdot,\cdot)$ (see (2.1)). In some instances (e.g., in the definitions below or in Lemma 3.1 and Theorems 3.2, 3.4(i)) we will consider j, and hence $h(\cdot)$, as being fixed. In other cases (e.g., in Theorems 3.3 and 3.4(ii)), the reference "*for every* $h(u,y)$ *as in* (3.1)" will be used in order to indicate that j can be any positive integer: $j = 1, 2, \dots$.

DEFINITION. *We shall say that the CSDE* (2.4) *satisfies* strong h-approximation condition (S-h-AC) *if, for any initial condition* y_0' *and admissible control* $u'(\cdot)$*, corresponding to any other initial condition* y_0'' *there exists an admissible control* $u''(\cdot)$ *such that the solutions* $y'(\cdot)$ *and* $y''(\cdot)$ *of the CSDE* (2.4) *(obtained with* y_0'*,* $u'(\cdot)$ *and* y_0''*,* $u''(\cdot)$*, respectively) satisfy the inequality*

$$E\left[\left\|\frac{1}{S}\int_0^S h(u'(\tau),y'(\tau))d\tau - \frac{1}{S}\int_0^S h(u''(\tau),y''(\tau))d\tau\right\|\right]$$

$$\le \nu_h(S)(1 + E[||y_0'||^\alpha] + E[||y_0''||^\alpha]) \tag{3.2}$$

for some monotone decreasing $\nu_h(\cdot) : [0,\infty) \to [0,\infty)$ *such that* $\lim_{S\to\infty} \nu_h(S) = 0$ *(*α *is the same as in Assumption* 1*).*

DEFINITION. *We shall say that the CSDE* (2.4) *satisfies* weak h-approximation condition (W-h-AC) *if*

$$\left\|E\left[\frac{1}{S}\int_0^S h(u'(\tau),y'(\tau))d\tau\right] - E\left[\frac{1}{S}\int_0^S h(u''(\tau),y''(\tau))d\tau\right]\right\|$$

$$\le \nu_h(S)(1 + E[||y_0'||^\alpha] + E[||y_0''||^\alpha]), \tag{3.3}$$

where y_0', y_0''*,* $u'(\cdot), u''(\cdot)$*,* $y'(\cdot), y''(\cdot)$*,* $\nu_h(\cdot)$ *and* α *are as above.*

Note that, in the linear case (2.13), one can take $u''(\cdot) = u'(\cdot)$ and obtain (see (2.14)) that

$$E[||y'(\tau) - y''(\tau)||] = E[||e^{A_1\tau}(y_0' - y_0'')||] \le ||e^{A_1\tau}||(E[||y_0'||] + E[||y_0''||]). \tag{3.4}$$

Since $h(\cdot)$ satisfies Lipschitz conditions, the validity of S-h-AC will follow from (3.4) (with $\nu_h(S) = O(\frac{1}{S})$ and $\alpha \ge 1$) if the eigenvalues of A_1 have negative real parts, in which case $||e^{A_1\tau}|| \le \beta_1 e^{-\beta_2\tau}$, with β_1, β_2 being positive constants. Note that A_3 in (2.13) can be degenerate or, in fact, it can be zero (the deterministic case). Note also that a Liapunov-type stability condition which leads to the validity of a similar estimate (and, thus, leads to the fulfillment of S-h-AC with $u''(\cdot) = u'(\cdot)$) for a nonlinear CSDE can be found in [11].

Remark 2. Note that W-h-AC is an auxiliary condition which is introduced in order to simplify our consideration. It is obvious that it is implied by S-h-AC, but we were unable to construct an example in which W-h-AC is satisfied while S-h-AC is not. We leave the question of whether it is possible to construct such an example (or whether W-h-AC and S-h-AC are equivalent) open. Note that, in case

of the uncontrolled dynamics (U consists only of one point; say, $U = \{\bar{u}\}$), S-h-AC is implied by W-h-AC and, hence, W-h-AC and S-h-AC are equivalent. In fact, as is noticed later (see Remark 4 on page 10), if W-h-AC is satisfied, then there exists a nonrandom vector $\tilde{h}$ such that

$$E\left[\left\|\frac{1}{S}\int_0^S h(\bar{u}, y(\tau))d\tau - \tilde{h}\right\|\right] \le \bar{\nu}^{(C,\alpha)}(S), \qquad \lim_{S\to\infty} \bar{\nu}^{(C,\alpha)}(S) = 0 \tag{3.5}$$

for any solution $y(\cdot)$ of (2.4) which has the initial condition satisfying the inequality $E[||y(0)||^\alpha] \le C = \text{const}$. It follows that, for any two solutions $y'(\cdot)$ and $y''(\cdot)$ of (2.4) with the initial conditions satisfying a similar inequality,

$$E\left[\left\|\frac{1}{S}\int_0^S h(\bar{u}, y'(\tau))d\tau - \frac{1}{S}\int_0^S h(\bar{u}, y''(\tau))d\tau\right\|\right] \le 2\bar{\nu}^{(C,\alpha)}(S).$$

Therefore, S-h-AC is satisfied.

For $h(u, y)$ as in (3.1), let $V_h(S, y_0)$ stand for the collection of random variables

$$V_h(S, y_0) \stackrel{\text{def}}{=} \bigcup_{(u(\cdot),\ y(\cdot))} \left\{\frac{1}{S}\int_0^S h(u(\tau), y(\tau))d\tau\right\} = \bigcup_{\mu\in\mathcal{M}(S,y_0)} \left\{\int h(u, y)\mu(du, dy)\right\} \tag{3.6}$$

and $E[V_h(S, y_0)]$ stand for the set of the corresponding mathematical expectations

$$\begin{aligned} E[V_h(S, y_0)] &\stackrel{\text{def}}{=} \bigcup_{(u(\cdot),\ y(\cdot))} \left\{E\left[\frac{1}{S}\int_0^S h(u(\tau), y(\tau))d\tau\right]\right\} \\ &= \bigcup_{\mu\in E[\mathcal{M}(S,y_0)]} \left\{\int h(u, y)\mu(du, dy)\right\}, \end{aligned} \tag{3.7}$$

where, in both (3.6) and (3.7), the first unions are over the admissible controls and corresponding solutions of (2.4) with the initial conditions (2.5).

Next, we introduce the Hausdorff metric $d_H^E(\cdot, \cdot)$ on collections of random variables as follows.

Definition. *Let V_1 and V_2 be two collections of integrable random variables defined on the same probability space and taking values in R^j. Then*

$$d_H^E(V_1, V_2) \stackrel{def}{=} \max\left\{\sup_{\zeta\in V_1} d^E(\zeta, V_2),\ \sup_{\zeta\in V_2} d^E(\zeta, V_1)\right\}, \tag{3.8}$$

with

$$d^E(\zeta, V_2) \stackrel{\text{def}}{=} \inf_{\zeta'\in V_2} E[||\zeta-\zeta'||]\ \ \forall\zeta\in V_1, \quad d^E(\zeta, V_1) \stackrel{\text{def}}{=} \inf_{\zeta'\in V_1} E[||\zeta-\zeta'||]\ \ \forall\zeta\in V_2, \tag{3.9}$$

where (here and in what follows) $||\cdot||$ is the Euclidean norm in R^j.

It is easy to see that d_H^E is nonnegative, symmetric, and satisfies the triangle inequality. For the constant valued collections of random variables, which can be viewed as just subsets of R^j, the definition above is reduced to the "standard" definition of the Hausdorff metric (semimetric) in R^j:

$$d_H(V_1, V_2) = \max\left\{\sup_{\zeta\in V_1} d(\zeta, V_2), \sup_{\zeta\in V_2} d(\zeta, V_1)\right\}, \qquad d(\zeta, V_i) = \inf_{\zeta'\in V_i} ||\zeta-\zeta'||,\ \ i = 1, 2. \tag{3.10}$$

Note that, as in the case with $\rho_H(\cdot,\cdot)$ (see Remark 1 on page 3), the equality $d_H(\cdot,\cdot) = 0$ is equivalent to the fact that the closures of the corresponding subsets of R^j are equal.

LEMMA 3.1. *S-h-AC is equivalent to the fulfillment of the inequality*

$$d_H^E(V_h(S,y_0'),V_h(S,y_0'')) \le \nu_h(S)(1+E[||y_0'||^\alpha]+E[||y_0''||^\alpha]), \tag{3.11}$$

and W-h-AC is equivalent to the fulfillment of the inequality

$$d_H(E[V_h(S,y_0')],E[V_h(S,y_0'')]) \le \nu_h(S)(1+E[||y_0'||^\alpha]+E[||y_0''||^\alpha]) \tag{3.12}$$

for any initial conditions y_0' and y_0''.

Proof. The proof is obvious. □

DEFINITION. *We shall say that the initial condition* (2.5) *has a probability distribution belonging to the class (C,α) if*

$$E[||y_0||^\alpha] \le C = \text{const}. \tag{3.13}$$

THEOREM 3.2. *Let Assumption* 1 *be valid. If the CSDE* (2.4) *satisfies W-h-AC, then there exists a convex and compact set $V_h \subset R^j$ such that, for any initial condition y_0 with the probability distribution from the class (C,α),*

$$d_H(E[V_h(S,y_0)],V_h) \le \ \nu_h^{C,\alpha}(S), \qquad \lim_{S\to\infty}\nu_h^{C,\alpha}(S)=0. \tag{3.14}$$

Conversely, if there exists V_h such that (3.14) *is valid for any initial condition y_0 with the probability distribution from the class (C,α), then W-h-AC is satisfied for any initial conditions y_0', y_0'' with the probability distributions from this class.*

Proof. The fact that the validity of (3.14) implies W-h-AC is obvious since from (3.14) it follows that

$$d_H(E[V_h(S,y_0')],E[V_h(S,y_0'')]) \le d_H(E[V_h(S,y_0')],V_h)+d_H(V_h,E[V_h(S,y_0'')])$$

$$\le 2\nu_h^{C,\alpha}(S),$$

which, by (3.12), leads to the fulfillment of W-h-AC. The proof of the fact that W-h-AC implies the existence of a convex and compact set V_h which satisfies (3.14) for any initial condition y_0 with the probability distribution from the class (C,α) is given in section 6. □

THEOREM 3.3. *Let Assumption* 1 *be valid. If the CSDE* (2.4) *satisfies W-h-AC for any vector function $h(u,y)$ as in* (3.1). *Then the LOMS $\mathcal{M}$ of the CSDE* (2.4) *with respect to the initial conditions having the probability distribution from the class (C,α) exists. That is, for any initial condition y_0 with the probability distribution from this class, the estimate is valid:*

$$\rho_H(E[\mathcal{M}(S,y_0)],\mathcal{M}) \le \ \nu^{C,\alpha}(S), \qquad \lim_{S\to\infty}\nu^{C,\alpha}(S)=0. \tag{3.15}$$

Also, the LOMS $\mathcal{M}$ allows the representation

$$\mathcal{M} \stackrel{\text{def}}{=} \{\mu\in\mathcal{P}(U\times R^m) \mid \int h(u,y)\mu(du,dy)\in V_h \ \ \forall h(u,y) \ \ \textit{as in} \ (3.1)\}, \tag{3.16}$$

where V_h are convex and compact sets the existence of which (for every $h(u,y)$ as in (3.1)*) is established by Theorem* 3.2.

Conversely, if there exists a convex and compact set $\mathcal{M} \subset \mathcal{P}(U \times R^m)$ *which satisfies* (3.15) *with any initial condition* y_0 *having the probability distribution from the class* (C, α)*, then W-h-AC is satisfied for any vector function* $h(\cdot)$ *as in* (3.1) *and any initial conditions* y_0', y_0'' *with the probability distributions from this class. Also, for any* $h(\cdot)$ *as in* (3.1)*, the estimate* (3.14) *is valid with*

$$V_h = \bigcup_{\mu \in \mathcal{M}} \left\{ \int h(u, y) \mu(du, dy) \right\}. \tag{3.17}$$

Proof of Theorem 3.3 is in the end of section 6.

THEOREM 3.4. *Let Assumption* 1 *be valid.* (i) *If the CSDE* (2.4) *satisfies S-h-AC, then, for any initial condition* y_0 *with the probability distribution from the class* (C, α)*,*

$$\sup_{\zeta \in V_h} d^E(\zeta, V_h(S, y_0)) \le \tilde{\nu}_h^{(C,\alpha)}(S), \qquad \lim_{S \to \infty} \tilde{\nu}_h^{(C,\alpha)}(S) = 0, \tag{3.18}$$

where V_h *is as in* (3.14) *and* $d^E(\cdot, \cdot)$ *is defined by* (3.9).

(ii) *If the CSDE* (2.4) *satisfies S-h-AC for any vector function* $h(\cdot)$ *as in* (3.1)*, then, for any initial condition* y_0 *with the probability distribution from the class* (C, α)*,*

$$\sup_{\mu \in \mathcal{M}} \rho^E(\mu, \mathcal{M}(S, y_0)) \le \tilde{\nu}^{(C,\alpha)}(S), \qquad \lim_{S \to \infty} \tilde{\nu}^{(C,\alpha)}(S) = 0, \tag{3.19}$$

where

$$\rho^E(\mu, \mathcal{M}(S, y_0)) \stackrel{\text{def}}{=} \inf_{\mu' \in \mathcal{M}(S, y_0)} E[\rho(\mu, \mu')]. \tag{3.20}$$

Proof of Theorem 3.4 is in section 6.

In conclusion of this section let us consider the following simple result which is used in the proof of Theorem 3.4.

PROPOSITION 3.5. *Let* V_i*,* $i = 1, \dots, k$ *be collections of random variables defined on the same probability space such that any element* $\zeta_i \in V_i$ *is independent from any element* $\zeta_j \in V_j$ *for* $i \ne j$*. Assume also that*

$$E[||\zeta_i||^2] \le \bar{c} = \text{const} \qquad \forall \zeta_i \in V_i, \qquad i = 1, 2 \dots k.$$

Then

$$d_H^E \left(\frac{1}{k} \sum_1^k V_i, \ \frac{1}{k} \sum_1^k E[V_i] \right) \le \sqrt{\frac{\bar{c}}{k}}, \tag{3.21}$$

where $E[V_i]$ *stands for the set of mathematical expectations of the elements of* V_i *and* $d_H^E(\cdot, \cdot)$ *is defined in* (3.8).

Proof. Take an arbitrary element $\zeta \in \frac{1}{k} \sum_1^k V_i$. By definition it is presented in the form $\zeta = \frac{1}{k} \sum_1^k \zeta_i$, where $\zeta_i \in V_i$. Consider

$$\bar{\zeta} \stackrel{\text{def}}{=} E[\zeta] = \frac{1}{k} \sum_1^k E[\zeta_i] \in \frac{1}{k} \sum_1^k E[V_i].$$

Due to the independence of ζ_i, $i = 1, \dots, k$,

$$E[||\zeta - \bar{\zeta}||] \le \sqrt{E[||\zeta - \bar{\zeta}||^2]} = \sqrt{\frac{1}{k^2}\sum_1^k E[||\zeta_i - E[\zeta_i]||^2]} \le \sqrt{\frac{\bar{c}}{k}}. \tag{3.22}$$

Now take an arbitrary $\bar{\zeta} \in \frac{1}{k}\sum_1^k E[V_i]$. By definition, there exist $\zeta_i \in V_i$ such that $\bar{\zeta} = \frac{1}{k}\sum_1^k E[\zeta_i]$. Define $\zeta \stackrel{\text{def}}{=} \frac{1}{k}\sum_1^k \zeta_i \in \frac{1}{k}\sum_1^k V_i$. Similarly to (3.22), one can establish that $E[||\zeta - \bar{\zeta}||] \le \sqrt{\frac{\bar{c}}{k}}$. This completes the proof of the proposition. ∎

4. Representation of the limit occupational measures set. Let $C_0^2(R^m)$ be the space of twice continuously differentiable functions $f(y) : R^m \to R^1$ which vanish at infinity along with their first and second derivatives and let $\mathcal{D}$ be a countable dense set in $C_0^2(R^m)$. Let $L : C_0^2(R^m) \to C_b(U \times R^m)$ be the operator defined as follows:

$$(Lf)(y,u) \;=\; \frac{1}{2} tr(b(y)b^T(y)\nabla^2 f(y)) + \langle \nabla f(y), a(u,y)\rangle \quad \forall f \in C_0^2(R^m). \tag{4.1}$$

Define the set of probability measures $D \subset \mathcal{P}(R^s \times A)$ by

$$D \;=\; \{\mu \in \mathcal{P}(U \times R^m) : \int (Lf)(u,y)\mu(du,dy) = 0 \;\; \forall f \in \mathcal{D}\}. \tag{4.2}$$

and introduce the following assumption.

Assumption 2. For some $\alpha > 0$,

$$\int ||y||^\alpha \mu(du,dy) \le c_2 = \text{const} \qquad \forall \mu \in D. \tag{4.3}$$

Note that the set D is convex and it is easy to verify that it is compact if Assumption 2 is satisfied. In fact, from this assumption it follows that D is tight and, hence, by Prohorov's theorem (see, e.g., Theorem 2.3.1, p. 25 in [15]), it is relatively compact in $\mathcal{P}(U \times R^m)$. Also, D is closed. This implied the compactness.

In [13] and [49] it was shown that, under some mild conditions, the set D represents the set of marginal distributions of stationary relaxed solutions of (2.4). In the following theorem it is established that, if W-h-AC is satisfied for any $h(u,y)$ as in (3.1)), then the LOMS of the CSDE (2.4) exists (the existence being implied by Theorem 3.3) and coincides with D.

THEOREM 4.1. *Let Assumptions* 1 *and* 2 *be satisfied with* $\alpha \ge 2$. *Then,*

(i) *The estimate*

$$\rho_H\left(\bigcup_{\{y_0\}\in(C,\alpha)} \{E[\mathcal{M}(S,y_0)]\},\; D\right) \le \bar{\nu}^{(C,\alpha)}(S), \qquad \lim_{S\to\infty} \bar{\nu}^{(C,\alpha)}(S) = 0, \tag{4.4}$$

is valid, where the union is over all initial conditions with the probability distribution from the class (C,α).

(ii) *If W-h-AC is satisfied for any* $h(u,y)$ *as in* (3.1), *then the LOMS* $\mathcal{M}$ *of the CSDE* (2.4) *with respect to the initial conditions having the probability distribution from the class* (C,α) *exists and is equal to* D*:*

$$\mathcal{M} = D. \tag{4.5}$$

Proof. The statement (i) of the theorem is proved in section 7 on the basis of Theorem 4.2 stated below. The validity of (ii) is proved on the basis of (i) as follows. By (3.15),

$$\rho_H\left(\bigcup_{\{y_0\}\in(C,\alpha)}\{E[\mathcal{M}(S,y_0)]\},\ \mathcal{M}\right)\le\nu^{C,\alpha}(S).$$

If now one assumes that (4.4) is valid, it will follow that

$$\rho_H(\mathcal{M},D)\le\nu^{C,\alpha}(S)+\bar\nu^{(C,\alpha)}(S)\quad\Rightarrow\quad\rho_H(\mathcal{M},D)=0.$$

The latter implies (4.5) since both $\mathcal{M}$ and D are compact. □

THEOREM 4.2. *Let Assumption* 1 *be satisfied with* $\alpha\ge 2$. *Then, for any initial condition* y_0 *with the probability distribution from the class* (C,α),

$$\sup_{\mu\in\mathcal{M}(S,y_0)}E[\rho(\mu,D)]\le\bar\nu^{(C,\alpha)}(S),\qquad\lim_{S\to\infty}\bar\nu^{(C,\alpha)}(S)=0,\tag{4.6}$$

where $\rho(\mu,D)$ *is defined as in* (2.3).

Proof of Theorem 4.2 is in section 7.

COROLLARY 4.3. *Let Assumptions* 1 *and* 2 *be valid with* $\alpha\ge 2$. *If the CSDE* (2.4) *satisfies W-h-AC for any vector function* $h(\cdot)$ *as in* (3.1), *then, for any initial condition* y_0 *with the probability distribution from the class* (C,α),

$$\rho_H(E[\mathcal{M}(S,y_0)],D)\le\ \nu^{C,\alpha}(S),\qquad\lim_{S\to\infty}\nu^{(C,\alpha)}(S)=0.\tag{4.7}$$

If the CSDE (2.4) *satisfies S-h-AC for any vector function* $h(\cdot)$ *as in* (3.1), *then, for any initial condition* y_0 *with the probability distribution from the class* (C,α),

$$\sup_{\mu\in D}\rho^E(\mu,\mathcal{M}(S,y_0))\le\tilde\nu^{(C,\alpha)}(S),\qquad\lim_{S\to\infty}\tilde\nu^{(C,\alpha)}(S)=0.\tag{4.8}$$

Proof. The proof follows immediately from Theorems 3.3, 3.4, and 4.1(ii). □

Remark 3. The estimate (4.8) is, in fact, equivalent to the validity of S-h-AC for any vector function $h(\cdot)$. Let us show that if (4.8) is satisfied, then, for any $y_0',u'(\cdot)$ and any y_0'', there exists $u''(\cdot)$ such that (3.2) is valid (with y_0',y_0'' being assumed to be from the class (C,α)). Let $\mu_S^{u'(\cdot)y'(\cdot)}$ be the occupational measure generated by the pair $(u'(\cdot),y'(\cdot))$ on the interval $[0,S]$. By (4.6), there exists $\mu_S'\in D$ such that $E[\rho(\mu_S^{u'(\cdot)y'(\cdot)},\mu_S')]\le\bar\nu^{(C,\alpha)}(S)$. From (4.8) it follows, in turn, that the estimate $E[\rho(\mu_S',\mu_S'')]\le\tilde\nu^{(C,\alpha)}(S)$ is valid for some $\mu_S''\in\mathcal{M}(S,y_0'')$. By definition, μ_S'' is an occupational measure generated on the interval $[0,S]$ by some admissible control $u''(\cdot)$ and the corresponding solution $y''(\cdot)$ of the CSDE (2.4) which satisfies the initial condition $y''(0)=y_0''$. That is, $\mu_S''=\mu_S^{u''(\cdot)y''(\cdot)}$ and $E[\rho(\mu_S^{u'(\cdot)y'(\cdot)},\mu_S^{u''(\cdot)y''(\cdot)})]\le\bar\nu^{(C,\alpha)}(S)+\tilde\nu^{(C,\alpha)}(S)$. The latter estimate implies the validity of S-h-AC for any $h(\cdot)$ as in (3.1).

Remark 4. Under Assumptions 1 and 2, one can show that, corresponding to any extreme point μ of D, there exists an admissible control $u_\mu(\cdot)$ and the corresponding solution $y_\mu(\cdot)$ of the CSDE (2.4) such that, for any $h(\cdot)$ as in (3.1), there almost surely exists the limit

$$\lim_{S\to\infty}\frac{1}{S}\int_0^S h(u_\mu(\tau),y_\mu(\tau))d\tau=\int h(u,y)\mu(du,dy).\tag{4.9}$$

We do not give the proof of this statement in the paper (it is based on results of [13], [49], and the ergodic theory). Let us note only that, for the uncontrolled case mentioned in Remark 2 on page 5, it follows that, if W-h-AC is satisfied, then the value of the integral on the right-hand side of (4.9) is the same for any extreme points μ of D and, hence, it is the same for all elements of D. Denoting this value as $\tilde{h}$ and using (4.6), one can easily verify the validity of (3.5).

Let $g(u,y) : U \times R^m \to R^n$ be continuous and satisfy Lipschitz conditions in y. Define the collection of R^n-valued random variables $V_g(S,y_0)$ similarly to (3.6) with the replacement of $h(\cdot)$ by $g(\cdot)$. That is,

$$V_g(S,y_0) \stackrel{\text{def}}{=} \bigcup_{(u(\cdot),y(\cdot))} \left\{ \frac{1}{S}\int_0^S g(u(\tau),y(\tau))d\tau \right\} = \bigcup_{\mu\in\mathcal{M}(S,y_0)} \left\{ \int g(u,y)\mu(du,dy) \right\}, \tag{4.10}$$

where, as in (3.6), the first union is over all admissible controls and corresponding solutions of the CSDE (2.4). Define also the set $V_g \subset R^n$ by

$$V_g \stackrel{\text{def}}{=} \bigcup_{\mu\in D} \left\{ \int g(u,y)\mu(du,dy) \right\}, \tag{4.11}$$

where the union is over elements of D. Note that from the convexity of D it follows that V_g is convex. Also, the following two corollaries used in averaging of singularly perturbed CSDE (see section 5 below) are valid.

Corollary 4.4. *Let Assumptions* 1 and 2 *be satisfied with* $\alpha \geq 2$. *Then the set* V_g *is compact and there exists a function* $\bar{\nu}_g^{(C,\alpha)}(S)$, *tending to zero as* S *tends to infinity, such that for any initial condition* y_0 *with the distribution from the class* (C,α),

$$\sup_{v\in V_g(S,y_0)} E[d(v,V_g)] \leq \bar{\nu}_g^{(C,\alpha)}(S). \tag{4.12}$$

Corollary 4.5. *Let Assumptions* 1 and 2 *be valid with* $\alpha \geq 2$ *and let the CSDE* (2.4) *satisfies S-h-AC for any vector function* $h(u,y)$ *as in* (3.1). *Then there exists a function* $\tilde{\nu}_g^{(C,\alpha)}(S)$, *tending to zero as* S *tends to infinity, such that for any initial condition* y_0 *with the probability distribution from the class* (C,α),

$$\sup_{v\in V_g} d^E(v, V_g(S,y_0)) \leq \tilde{\nu}_g^{(C,\alpha)}(S). \tag{4.13}$$

Proofs of Corollaries 4.4 *and* 4.5. The proofs follow from Theorems 4.1(ii), (4.2), and Theorems 4.1(ii), (3.4), respectively, if $g(\cdot) = h(\cdot)$ (with $h(\cdot)$ being as in (3.1)). In the general case, the corollaries are proved in section 7. □

5. Application in averaging of singularly perturbed controlled stochastic differential equations. Consider the following singularly perturbed CSDE:

$$dy^\epsilon(t) = \frac{1}{\epsilon}a(u(t),y^\epsilon(t))dt + \frac{1}{\sqrt{\epsilon}}b(y^\epsilon(t))dB_1(t). \tag{5.1}$$

$$dz^\epsilon(t) = g(u(t),y^\epsilon(t),z^\epsilon(t))dt + \sigma(z^\epsilon(t))dB_2(t), \tag{5.2}$$

with the initial conditions

$$y^\epsilon(0) = y_0, \qquad z^\epsilon(0) = z_0, \tag{5.3}$$

where:

- ϵ is a small positive parameter;
- the functions $b(\cdot) : R^m \to R^{m\times m}$ and $\sigma(\cdot) : R^n \to R^{n\times n}$ satisfy Lipschitz conditions; the functions $a(u,y) : U \times R^m \to R^m$ and $g(u,y,z) : U \times R^m \times R^n \to R^n$ are continuous and satisfy Lipschitz conditions, respectively, in y and (y,z) uniformly with respect to $u \in U$;
- U is a compact metric space;
- $B_1(\cdot)$ and $B_2(\cdot)$ are R^m- and R^n-valued independent standard Brownian motions;
- y_0 and z_0 are R^m- and R^n-valued random variables which have bounded fourth moments and are independent of $B_1(\cdot), B_2(\cdot)$;
- *admissible controls* $u(\cdot)$ are U-valued random processes progressively measurable with respect to a right continuous and complete filtration $\{\hat{\mathcal{F}}_t\}$ of σ-fields such that
  - $\{y_0, z_0 \ \& \ B_1(\theta), B_2(\theta); \theta \le t\}$ is measurable with respect to $\hat{\mathcal{F}}_t$ for $t \ge 0$,
  - For $t' \ge t \ge 0$, $B_1(t') - B_1(t)$ and $B_2(t') - B_2(t)$ are independent of $\hat{\mathcal{F}}_t$.

Let us set $\tau = \frac{t}{\epsilon}, y(\tau) = y^\epsilon(\epsilon\tau), u'(\tau) = u^\epsilon(\epsilon\tau), W(\tau) = \frac{1}{\sqrt{\epsilon}} B_1(\epsilon\tau)$ and $\{\mathcal{F}_\tau\} = \{\hat{\mathcal{F}}_{\epsilon\tau}\}$. The subsystem (5.1) takes then the form of the CSDE (2.4) and is called *the associated system*. Assuming that the associated system satisfies Assumptions 1 and 2, let us define the averaged CSDE by

$$dz(t) = \tilde{g}(\mu(t), z(t))dt + \sigma(z(t))dB_2(t), \tag{5.4}$$

where:

- $\tilde{g}(\mu, z) \stackrel{\text{def}}{=} \int_{R^m \times U} g(u,y,z)\mu(du,dy) : \mathcal{P}(R^m \times U) \times R^n \to R^n$;
- the Brownian motion $B_2(\cdot)$ and the initial condition $z(0) = z_0$ are the same as in (5.2);
- *admissible controls* $\mu(\cdot)$ are $\{\hat{\mathcal{F}}_t\}$-progressive D-valued random processes (D being defined in (4.2)).

Let $G(\cdot) : R^n \to R$ be Lipschitz continuous and $T > 0$. Consider the problem of optimal control

$$\inf_{u(\cdot), y^\epsilon(\cdot), z^\epsilon(\cdot)} E[G(z^\epsilon(T))] \stackrel{\text{def}}{=} G^*_\epsilon, \tag{5.5}$$

where inf is over the admissible controls and the corresponding solutions of the singularly perturbed CSDE (5.1) and (5.2). Consider also the problem

$$\inf_{\mu(\cdot), z(\cdot)} E[G(z(T))] \stackrel{\text{def}}{=} G^*_{av}, \tag{5.6}$$

where inf is over the admissible controls and the corresponding solutions of the averaged CSDE (5.4). Note that, if $\sigma(\cdot) \equiv 0$, then the averaged system (5.4) is purely deterministic and the minimization in (5.6) can be restricted to open loop controls (a similar phenomenon was dealt with in [1], where the fast dynamics were defined by a Markov decision process).

THEOREM 5.1. (i) *Let the associated system satisfy Assumptions* 1 *and* 2 *with* $\alpha = 4$. *Then, corresponding to any admissible solution* $(y^\epsilon(\cdot), z^\epsilon(\cdot))$ *of the singularly perturbed CSDE* (5.1) and (5.2), *there exists an admissible solution* $z(\cdot)$ *of the averaged CSDE* (5.4) *such that*

$$\max_{t\in[0,T]} E[||z^\epsilon(t) - z(t)||^2] \le \tilde{\nu}(\epsilon), \qquad \lim_{\epsilon\to 0} \tilde{\nu}(\epsilon) = 0 \tag{5.7}$$

and the following inequality is valid:

$$\liminf_{\epsilon\to 0} G^*_\epsilon \geq G^*_{av}. \tag{5.8}$$

(ii) *If, in addition, the associated system satisfies S-h-AC for any vector function* $h(\cdot)$ *as in* (3.1)*, then, corresponding to an arbitrary admissible solution* $z(\cdot)$ *of the averaged CSDE* (5.4)*, there exists an admissible solution* $(y^\epsilon(\cdot), z^\epsilon(\cdot))$ *of the singularly perturbed CSDE* (5.1) and (5.2) *such that* (5.7) *is valid and*

$$\lim_{\epsilon\to 0} G^*_\epsilon \;=\; G^*_{av}. \tag{5.9}$$

Proof. The proof's details are outlined in section 8. □

Remark 5. The approximation of the z-components of the state variables of the CSDE (5.1) and (5.2) by the solutions of the averaged system (5.4) stated in Theorem 5.1 has many similarities with the classical relaxed control setting (see [52]). In contrast to the latter, however, the approximation established in Theorem 5.1 is asymptotic (that is valid when the small parameter tends to zero) and also the controls used in (5.4) take values in the LOMS (and not in the space of all probability measures defined on the control set).

Remark 6. Note that the conditions of Theorem 5.1 can be relaxed. Namely, the theorem remains valid if Assumptions 1 and 2 are satisfied with $\alpha > 2$ and also if they are satisfied with $\alpha = 2$ (to prove the result in the latter case, one needs to impose some additional conditions; in particular, one needs to assume that there exists an integrable random variable η such that the solution of the associated system satisfy the inequality $||y(\tau)||^2 \leq \eta \ \forall \tau \geq 0$). Note also that a statement similar to Theorem 5.1 is valid for singularly perturbed CSDE in which the fast subsystem may depend on the slow state variables. The proof of such a statement is in many ways similar to one outlined in section 8 but it is more technically involved and we do not include it in the paper.

Let us consider a special case when $b(y) = 0$. That is, the associated system is deterministic and it can be written in the form

$$\frac{dy(\tau)}{d\tau} = a(u(\tau), y(\tau)). \tag{5.10}$$

Assume that there exist positive definite matrices F_1 and F_2 such that, for any y', y'' and any $u \in U$,

$$(a(u, y') - a(u, y''))^T F_1 (y' - y'') \leq -(y' - y'')^T F_2 (y' - y''), \tag{5.11}$$

Note that (5.11) is satisfied if $a(u, y) = A_1 y + A_2 u$ (as in (2.13)), with the eigenvalues of A_1 having negative real parts. Taking $y^T F_1 y$ as a Liapunov function, one can easily verify (see, e.g., [27]) that solutions $y'(\tau)$ and $y''(\tau)$ of (5.10) obtained with the same control and with initial conditions $y'(0) = y'_0,\ y''(0) = y''_0$, satisfy the inequality

$$||y'(\tau) - y''(\tau)|| \leq \beta_1 e^{-\beta_2 \tau} ||y'_0 - y''_0|| \quad \forall \tau \geq 0, \tag{5.12}$$

where β_1 and β_2 are some positive constants. This implies the validity of S-h-AC. From (5.12) it follows (see Theorem 3.1(ii) in [25]) that there exists a compact set $Y \subset R^m$ such that any solution $y(\cdot)$ of (5.10) satisfies the inequality

$$\min_{y\in Y} ||y(\tau) - y|| \leq \beta_1 e^{-\beta_2 \tau} \min_{y\in Y} ||y(0) - y|| \qquad \forall \tau \geq 0, \tag{5.13}$$

where β_1, β_2 are as in (5.12) (that is, Y is forward invariant with respect to the solutions of (5.10) and is a global attractor for these solutions). Using the inequality (5.13), it is straightforward to verify that Assumption 1 is satisfied with an arbitrary positive α. Also, using this inequality, one can establish that $\mu(U \times Y) = 1 \ \ \forall \mu \in \mathcal{M}$, where, as above, $\mathcal{M}$ is the limit occupational measures set. Hence, this set allows the representation (see (4.2) and Theorem 4.1(ii))

$$\mathcal{M} = D = \{\mu \in \mathcal{P}(U \times Y) : \int_{U \times Y} \langle \nabla f(y), a(u,y) \rangle \mu(du, dy) = 0 \ \forall \ f \in \mathcal{D}\}, \tag{5.14}$$

where $\mathcal{P}(U \times Y)$ is the space of probability measures defined on Borel subsets of $U \times Y$. Note that the representation of the LOMS in the form (5.14) is consistent with one obtained for the deterministic case in [26] and that Assumption 2 is satisfied automatically in this case.

As an example, let us consider the singularly perturbed CSDE (5.1) and (5.2), in which: $y = (y_1, y_2)$ (that is, $y \in R^2$); U is a square in R^2: $U = \{(u_1, u_2) : \ |u_i| \leq 1, \ i = 1, 2\}$;

$$b(y) = 0, \qquad a(u,y) = (-y_1 + u_1, \ -y_2 + u_2); \tag{5.15}$$

and the slow dynamics are one dimensional ($z \in R^1$) with $z_0 = 0$ (zero initial condition) and with

$$\sigma(z) = \sigma = \text{const}, \qquad g(u,y,z) = g(u,y) \stackrel{\text{def}}{=} y_2 u_1 - y_1 u_2. \tag{5.16}$$

Consider the optimal control problem (5.5) with $G(z) = z$. Using (5.15) and (5.16), it is easy to verify that, if the fast subsystem (5.1) is multiplied by ϵ and, then ϵ is formally equated to zero, the resulting slow dynamics become uncontrolled and the value of the objective function is equal to zero. The limit of the optimal value of (5.5) is, however, strictly less than zero: $\lim_{\epsilon \to 0} G^*_\epsilon < 0$. This is evidenced by the fact that, if the rapidly oscillating controls $u_1(t) = cos(\frac{t}{\epsilon})$, $u_2(t) = sin(\frac{t}{\epsilon})$ are used, then the value of the objective function can be verified to be equal to $-0.5T + O(\epsilon) < 0$. Thus, the classical approach based on the equating of the singular perturbation parameter to zero is not applicable in the given example. The averaged problem is equivalent in this case to the infinite dimensional linear program

$$\min_{\mu \in D} \int_{U \times Y} g(u,y) \mu(du, dy) \stackrel{\text{def}}{=} g^*, \tag{5.17}$$

with $G^*_{av} = g^* T$, where D is as in (5.14) and $g(u,y)$ is defined in (5.16). Note that the solution of the problem (5.17) has been found numerically by approximating the problem with finite dimensional linear programs (using the approach proposed in [29]). The optimal value g^* of (5.17), in particular, was found to be approximately equal to -0.7679. One may conclude, therefore (by (5.9)), that $\lim_{\epsilon \to 0} G^*_\epsilon \approx -0.7679T$.

In some cases the averaged system can be equivalent to the system obtained via equating of the singular parameter to zero. To illustrate that, let us assume that the associated system is linear (that is, (2.13) is true, with eigenvalues of A_1 having negative real parts). Let us assume also that the slow subsystem (5.2) is linear in y and u. That is,

$$g(u,y,z) \stackrel{\text{def}}{=} A_4(z) y + A_5(z) u, \tag{5.18}$$

with $A_4(z), A_5(z)$ being matrices functions of the corresponding dimensions, then the averaged system becomes equivalent to

$$dz(t) = A_4(z(t))\bar{y}(t) + A_5(z(t))\bar{u}(t) + \sigma(z(t))dB_2(t), \tag{5.19}$$

where

$$(\bar{u}(t), \bar{y}(t)) \in \Omega \stackrel{\text{def}}{=} \left\{ (\bar{u}, \bar{y}) \mid (\bar{u}, \bar{y}) = \int (u, y)\mu(du, dy),\ \mu \in D \right\}. \tag{5.20}$$

Under the assumption that U is convex, it can be shown (although, we do not do it in this paper) that the set Ω defined above can be represented in the form

$$\Omega = \{(\bar{u}, \bar{y}) \mid \bar{y} = -A_1^{-1}A_2\bar{u},\ \bar{u} \in U\}. \tag{5.21}$$

and that (5.19) is equivalent to the system

$$dz(t) = (-A_4(z(t))A_1^{-1}A_2 + A_5(z(t)))\bar{u}(t) + \sigma(z(t))dB_2(t), \quad \bar{u}(t) \in U. \tag{5.22}$$

Note that the system (5.22) can be obtained via multiplying (5.1) by ϵ, then formally equating ϵ to zero and expressing y as a function of u, and then substituting the result into (5.2).

6. Proofs for section 3. Proofs of Theorems 3.2 and 3.4 are based on a number of lemmas stated below.

LEMMA 6.1. *Let a function $\psi(S) : (0, \infty) \to R^1$ be such that, for some monotone decreasing function $\nu(S)$, $lim_{S\to\infty}\nu(S) = 0$, the following inequalities are valid:*

$$|\psi(S) - \psi(kS)| \le \nu(S), \quad k = 1, 2, \ldots. \tag{6.1}$$

Let also

$$|\psi(S') - \psi(S'')| \le \frac{\alpha|S'' - S'|}{\max\{S', S''\}} \quad \forall S', S'' > 0, \qquad \alpha = \text{const}. \tag{6.2}$$

Then there exists a limit

$$lim_{S\to\infty}\psi(S) \stackrel{\text{def}}{=} A \tag{6.3}$$

and the estimate

$$|\psi(S) - A| \le \nu(S) \quad \forall S > 0 \tag{6.4}$$

is valid.

Proof. To establish the existence of the limit it is sufficient to show that, corresponding to any $\delta > 0$, there exists $S_\delta > 0$ such that, for any $S'' \ge S' \ge S_\delta$,

$$|\psi(S'') - \psi(S')| \le \delta. \tag{6.5}$$

Note that from (6.1) it follows that, for any $k_2 \ge k_1 \ge 1$,

$$|\psi(S) - \psi\left(\frac{k_2}{k_1}S\right)| \le |\psi(S) - \psi(k_2S)| + |\psi(k_2S) - \psi\left(\frac{k_2}{k_1}S\right)| \le \nu(S) + \nu\left(\frac{k_2}{k_1}S\right) \le 2\nu(S).$$

Choose integer $k_2 \geq k_1 \geq 1$ in such a way that

$$0 \leq \frac{S''}{S'} - \frac{k_2}{k_1} \leq \frac{\delta}{2\alpha} \Rightarrow 0 \leq S'' - \frac{k_2}{k_1} S' \leq \frac{\delta}{2\alpha} S'.$$

Then, by (6.2),

$$|\psi(S'') - \psi(S')| \leq |\psi(S'') - \psi\left(\frac{k_2}{k_1} S'\right)| + |\psi\left(\frac{k_2}{k_1} S'\right) - \psi(S')| \leq \frac{\alpha |S'' - \frac{k_2}{k_1} S'|}{\frac{k_2}{k_1} S'} + 2\nu(S')$$

$$\leq \frac{\delta}{2} + 2\nu(S).$$

Choosing S_δ to be such that $\nu(S_\delta) = \frac{\delta}{2}$, one establishes (6.5). Thus the limit (6.3) exists. Passing to the limit as $k \to \infty$ in (6.1), one obtains the estimate (6.4). □

In the following lemmas, it is always supposed that Assumption 1 is satisfied.

LEMMA 6.2. *Let $h(\cdot)$ be as in* (3.1) *and the constant c_h be defined by*

$$c_h \stackrel{\text{def}}{=} \max_{(u,y) \in U \times \bar{R}^m} ||h(u,y)||. \tag{6.6}$$

Then the following estimates are valid:

$$\sup_{\zeta \in V_h(S,y_0)} E[||\zeta||] \leq c_h \qquad \Rightarrow \qquad \sup_{\zeta \in E[V_h(S,y_0)]} ||\zeta|| \leq c_h; \tag{6.7}$$

$$d_H^E(V_h(S',y_0), V_h(S'',y_0)) \leq \frac{2c_h |S'' - S'|}{\max\{S', S''\}}; \tag{6.8}$$

$$d_H(E[V_h(S',y_0)], E[V_h(S'',y_0)]) \leq \frac{2c_h |S'' - S'|}{\max\{S', S''\}}. \tag{6.9}$$

Proof. Note that (6.7) is obvious and that (6.9) follows from (6.8) since

$$d_H(E[V_1], E[V_2]) \leq d_H^E(V_1, V_2)$$

for any collections of random variables V_1 and V_2 such that $E[||\zeta||] < \infty \ \ \forall \zeta \in V_i, i = 1, 2$. Let us prove (6.8). Assume that $S'' \geq S'$. Then, by (6.7), for any admissible control $u(\cdot)$ and corresponding solution $y(\cdot)$ of the CSDE (2.4),

$$E\left[\left\| \frac{1}{S'} \int_0^{S'} h(u(\tau), y(\tau)) d\tau - \frac{1}{S''} \int_0^{S''} h(u(\tau), y(\tau)) d\tau \right\|\right]$$

$$\leq E\left[\left\| \left(\frac{1}{S'} - \frac{1}{S''}\right) \int_0^{S'} h(u(\tau), y(\tau)) d\tau \right\|\right] + E\left[\left\| \frac{1}{S''} \int_{S'}^{S''} h(u(\tau), y(\tau)) d\tau \right\|\right]$$

$$\leq \frac{2c_h (S'' - S')}{S''}.$$

This implies (6.8). □

LEMMA 6.3. *Let y_0 be fixed (nonrandom) and $\Psi(p, S, y_0)$ be the support function of the set $E[V_h(S, y_0)]$:*

$$\Psi_h(p, S, y_0) \stackrel{\text{def}}{=} \sup_{v \in E[V_h(S,y_0)]} \{p^T v\}.$$

If the CSDE (2.4) *satisfies W-h-AC, then there exists a convex, positively homogeneous and Lipschitz continuous function* $\Psi_h(p)$ *such that*

$$|\Psi_h(p,S,y_0)-\Psi_h(p)|\le c\nu_h(S)||p||(1+||y_0||^\alpha), \tag{6.10}$$

where $\nu_h(S)$ *is the function introduced in* (3.2) *and* $c=1+c_1$ (c_1 *is the constants from* (2.12)).

Proof. Note, first, that $\Psi_h(p,S,y_0)$ allows also the representation

$$\Psi_h(p,S,y_0)=\frac{1}{S}\sup_{(u(\cdot),y(\cdot))}\left\{E\left[\int_0^S p^T h(u(\tau),y(\tau))d\tau|y(0)=y_0\right]\right\},$$

where the sup is over all admissible controls and corresponding solutions of (2.4).

From (6.7) it follows that

$$|\Psi_h(p,S,y_0)|\le c_h||p||,\quad |\Psi_h(p',S,y_0)-\Psi_h(p'',S,y_0)|\le c_h||p'-p''|| \tag{6.11}$$

and from (6.9) it follows that

$$|\Psi_h(p,S',y_0)-\Psi_h(p,S'',y_0)|\le\frac{2c_h||p|||S''-S'|}{\max\{S',S''\}}. \tag{6.12}$$

Also, by (3.12),

$$|\Psi_h(p,S,y_0')-\Psi_h(p,S,y_0'')|\le\nu_h(S)||p||(1+||y_0'||^\alpha+||y_0''||^\alpha). \tag{6.13}$$

Note that if (6.10) is established, then the fact that $\Psi_h(p)$ is convex, positively homogeneous and Lipschitz continuous will follow from the fact that $\Psi_h(p,S,y_0)$ is convex, positively homogeneous and Lipschitz continuous in p (see (6.11)).

By Lemma 6.1, to establish (6.10), it is sufficient to verify the validity of the following estimates:

$$|\Psi(p,kS,y_0)-\Psi(p,S,y_0)|\le c\nu_h(S)||p||(1+||y_0||^\alpha), k=1,2,\dots. \tag{6.14}$$

For $k=1$ it is obvious. Assume that

$$|\Psi(p,(k-1)S,y_0)-\Psi(p,S,y_0)|\le c\nu_h(S)||p||(1+||y_0||^\alpha) \tag{6.15}$$

and show the validity of (6.14) using the induction. Define the collection of random variables $W_h(S,y_0)$ as follows:

$$W_h(S,y_0)\stackrel{\text{def}}{=}\bigcup_{(u(\cdot),y(\cdot))}\left\{\left(\frac{1}{S}\int_0^S h(u(\tau),y(\tau))d\tau,\ y(S)\right)\right\}, \tag{6.16}$$

where the union is over all admissible controls and corresponding solutions of the CSDE (2.4). Using dynamic programming, one can obtain

$$\Psi_h(p,kS,y_0)=\sup_{(\zeta,\eta)\in W_h((k-1)S,y_0)}\left\{\frac{k-1}{k}E[p^T\zeta]+\frac{1}{k}E[\Psi_h(p,S,\eta)]\right\}. \tag{6.17}$$

By (6.13) and (2.12), for any η such that $(\zeta,\eta)\in W_h((k-1)S,y_0)$,

$$|E[\Psi_h(p,S,\eta)]-\Psi_h(p,S,y_0)\ |\le E[|\Psi_h(p,S,\eta)-\Psi_h(p,S,y_0)|]$$

$$\leq \nu_h(S)||p||(1+E[||\eta||^\alpha]+||y_0||^\alpha) \leq \nu_h(S)||p||(1+c_1(1+||y_0||^\alpha)+||y_0||^\alpha)$$

$$= c\nu_h(S)||p||(1+||y_0||^\alpha),$$

with the constant c being as defined in the statement of the lemma. Hence, using (6.17), one can obtain

$$\left|\Psi_h(p,kS,y_0) - \left(\frac{k-1}{k}\Psi_h(p,(k-1)S,y_0)+\frac{1}{k}\Psi_h(p,S,y_0)\right)\right|$$

$$= \left|\Psi_h(p,kS,y_0) - \sup_{(\zeta,\eta)\in W_h((k-1)S,y_0)}\left\{\frac{k-1}{k}E[p^T\zeta]+\frac{1}{k}\Psi_h(p,S,y_0)\right\}\right|$$

$$\leq \frac{1}{k}\sup_{(\zeta,\eta)\in W_h((k-1)S,y_0)}\{|E[\Psi_h(p,S,\eta)]-\Psi_h(p,S,y_0)|\} \leq \left(\frac{1}{k}\right)c\nu_h(S)||p||(1+||y_0||^\alpha).$$

From (6.15) it follows, on the other hand, that

$$\left|\frac{k-1}{k}\Psi_h(p,(k-1)S,y_0) - \frac{k-1}{k}\Psi_h(p,S,y_0))\right| \leq \left(\frac{k-1}{k}\right)c\nu_h(S)||p||(1+||y_0||^\alpha)$$

$$\Rightarrow \qquad |\Psi_h(p,kS,y_0)-\Psi_h(p,S,y_0)|$$

$$\leq \left|\Psi_h(p,kS,y_0) - \left(\frac{k-1}{k}\Psi_h(p,(k-1)S,y_0\right)+\frac{1}{k}\Psi_h(p,S,y_0))\right|$$

$$+\left|\frac{k-1}{k}\Psi_h(p,(k-1)S,y_0) - \frac{k-1}{k}\Psi_h(p,S,y_0)\right| \leq \left(\frac{1}{k}+\frac{k-1}{k}\right)c\nu_h(S)||p||(1+||y_0||^\alpha).$$

The latter implies (6.14). □

LEMMA 6.4. *Let the CSDE* (2.4) *satisfy W-h-AC and let V_h be a convex and compact subset of R^j defined by*

$$V_h \stackrel{\text{def}}{=} \{v \mid p^T v \leq \Psi_h(p)\ \forall p \in R^j\}. \tag{6.18}$$

Then, for any y_0 with the probability distribution from the class (C,α),

$$d_H(coE[V_h(S,y_0)],V_h) \leq c(1+C)\nu_h(S), \tag{6.19}$$

where co stands for the convex hull of the corresponding set.

Comment. The notation (6.18) anticipates the fact that this set will coincide with the set V_h, the existence of which is claimed in Theorem 3.2.

Proof. Note that the fact that the set V_h is convex and compact follows from its definition in the form (6.18) and from the continuity of $\Psi_h(p)$.

Let y_0 be random. The support functions of both $E[V_h(S,y_0)]$ and $coE[V_h(S,y_0)]$ are equal to $E[\Psi_h(p,S,y_0)]$:

$$\sup_{v\in coE[V_h(S,y_0)]}\{p^Tv\} = \sup_{v\in E[V_h(S,y_0)]}\{p^Tv\} = E[\Psi_h(p,S,y_0)].$$

The support function for V_h is $\Psi_h(p)$ (see Corollary 13.2.1 in [47]). Hence (see, e.g., Lemma П2.9, p. 207 in [24]),

$$d_H(coE[V_h(S,y_0)],V_h) \le \sup_{||p||\le 1} |E[\Psi_h(p,S,y_0)] - \Psi_h(p)|. \tag{6.20}$$

By (6.10),

$$| E[\Psi_h(p,S,y_0)] - \Psi_h(p) | \le c\nu_h(S)||p||(1+E[||y_0||^\alpha]) \le c\nu_h(S)||p||(1+C) \tag{6.21}$$

for any y_0 with the probability distribution from the class (C,α). This and (6.20) imply (6.19). □

LEMMA 6.5. *For any $S>0$ and $k=1,2,\ldots$, there exists a collection of random variables $V'_h(kS,y_0)$ such that*

$$V'_h(kS,y_0) \subset V_h(kS,y_0) \qquad \Rightarrow \qquad E[V'_h(kS,y_0)] \subset E[V_h(kS,y_0)] \tag{6.22}$$

and such that: (i) *The estimate*

$$d_H(E[V'_h(kS,y_0)], coE[V_h(S,y_0)]) \le \frac{\bar c_h}{k} + c\nu_h(S)(1+E[||y_0||^\alpha]), \tag{6.23}$$

is valid if W-h-AC is satisfied; and (ii) *The estimate*

$$d^E_H(V'_h(kS,y_0), coE[V_h(S,y_0)] \le \frac{c_h}{\sqrt{k}} + \frac{\bar c_h}{k} + c\nu_h(S)(1+E[||y_0||^\alpha]), \tag{6.24}$$

is valid if S-h-AC is satisfied ($c_h, \bar c_h$ and c being constants).

Proof. The following three parts detail the proof.

Part I: *Construction of $V'_h(kS,y_0)$.* Consider the CSDE (2.4) on the interval $[0,kS]$ $(k=1,2,\ldots)$. Denote by $\{u(\cdot)\}^{0,kS}$ the family of admissible controls on the interval $[0,kS]$ such that the restriction of any control from this family to the interval $((i-1)S,iS]$ $(k\ge i\ge 1)$ is conditionally independent on $\mathcal{F}_{(i-1)S}$ conditioned on $y((i-1)S)$.

Define $V'(kS,y_0)$ as the collection of random variables

$$V'_h(kS,y_0) \stackrel{\text{def}}{=} \bigcup_{\{u(\cdot),y(\cdot)\}^{0,kS}} \left\{ \frac{1}{kS}\int_0^{kS} h(u(\tau),y(\tau))d\tau \right\} \tag{6.25}$$

and $E[V'(kS,y_0)]$ as the set of corresponding mathematical expectations

$$E[V'_h(kS,y_0)] \stackrel{\text{def}}{=} \bigcup_{\{u(\cdot),y(\cdot)\}^{0,kS}} \left\{ E\left[\frac{1}{kS}\int_0^{kS} h(u(\tau),y(\tau))d\tau \right] \right\}, \tag{6.26}$$

where, in both cases, the union is over the controls from $\{u(\cdot)\}^{0,kS}$ and the corresponding solutions of the CSDE (2.4) on the interval $[0,kS]$. Note that, by definition,

$$V'_h(S,y_0) = V_h(S,y_0), \qquad E[V'_h(S,y_0)] = E[V_h(S,y_0)] \tag{6.27}$$

and that the inclusions (6.22) are valid for $k=2,3,\ldots$.

Let $\{u(\cdot),y(\cdot)\}_\eta^{(i-1)S,iS}$ be the family of restrictions to the interval $((i-1)S,iS]$ of the controls $\{u(\cdot)\}^{0,kS}$ and the solutions of the CSDE (2.4) which are obtained with

these controls and satisfy the initial conditions $y((i-1)S) \stackrel{\text{def}}{=} \eta$. Define the collection of random variables $V'_h((i-1)S, iS, \eta)$ $(i = 1, \dots, k)$:

$$V_h((i-1)S,\ iS,\ \eta) \stackrel{\text{def}}{=} \bigcup_{\{u(\cdot),y(\cdot)\}_\eta^{(i-1)S,iS}} \left\{ \frac{1}{S} \int_{(i-1)S}^{iS} h(u(\tau), y(\tau)) d\tau \right\} \tag{6.28}$$

and the set of corresponding mathematical expectations $E[V_h((i-1)S, iS, \eta)]$:

$$E[V_h((i-1)S, iS, \eta)] \stackrel{\text{def}}{=} \bigcup_{\{u(\cdot),y(\cdot)\}_\eta^{(i-1)S,iS}} \left\{ E\left[\frac{1}{S} \int_{(i-1)S}^{iS} h(u(\tau), y(\tau)) d\tau \right] \right\}. \tag{6.29}$$

It is easy to verify that

$$V'_h(kS, y_0) = \bigcup_{(\zeta,\eta)\in W'_h((k-1)S,y_0)} \left\{ \frac{k-1}{k}\zeta + \frac{1}{k} V_h((k-1)S,\ kS,\ \eta) \right\} \tag{6.30}$$

$$\Rightarrow\ E[V'_h(kS, y_0)] = \bigcup_{(\zeta,\eta)\in W'_h((k-1)S,y_0)} \left\{ \frac{k-1}{k} E[\zeta] + \frac{1}{k} E[V_h((k-1)S,\ kS,\ \eta)] \right\}, \tag{6.31}$$

where

$$W'_h((k-1)S, y_0) \tag{6.32}$$

$$\stackrel{\text{def}}{=} \bigcup_{\{u(\cdot),y(\cdot)\}^{0,(k-1)S}} \left\{ \left(\frac{1}{(k-1)S} \int_0^{(k-1)S} h(u(\tau), y(\tau)) d\tau,\ y((k-1)S) \right) \right\},$$

the union being over the controls from the family of restrictions of the elements of $\{u(\cdot)\}^{0,kS}$ to the interval $[0, (k-1)S]$ and corresponding solutions of the CSDE (2.4).

Part II: *Proof of Lemma* 6.5(i). Using induction, let us show that

$$d_H(E[V'_h(kS, y_0)],\ \frac{1}{k}\sum_1^k E[V_h((i-1)S,\ iS,\ y_0)] \le c\nu_h(S)(1+E[||y_0||^\alpha]),\quad k = 1, 2, \dots. \tag{6.33}$$

For $k = 1$ it is immediate since, by definition, $E[V_h(S, y_0)] = E[V_h(0,\ S,\ y_0)]$. Assume that the estimate

$$d_H(E[V'_h((k-1)S,\ y_0)], \frac{1}{k-1}\sum_1^{k-1} E[V_h((i-1)S,\ iS,\ y_0)]) \le c\nu_h(S)(1+E[||y_0||^\alpha]) \tag{6.34}$$

is valid. From Assumption 1 and W-h-AC (see (3.12)) it follows that, for any η such that $(\zeta, \eta) \in W'_h((k-1)S, y_0)$,

$$d_H(E[V_h((k-1)S,\ kS,\ \eta)],\ E[V_h((k-1)S,\ kS,\ y_0)]) \le \nu_h(S)(1+E[||\eta||^\alpha]+E[||y_0||^\alpha])$$

$$\le \nu_h(S)(1 + c_1(1 + E[||y_0||^\alpha]) + E[||y_0||^\alpha]) = c\nu_h(S)(1 + E[||y_0||^\alpha]).$$

This and (6.31) lead to the estimate

$$d_H(E[V'(kS, y_0)],\ \frac{k-1}{k}E[V'((k-1)S,\ y_0)] + \frac{1}{k}E[V_h((k-1)S,\ kS,\ y_0)])$$

$$= d_H\left(E[V'(kS,y_0)],\ \bigcup_{(\zeta,\eta)\in W'_h((k-1)S,y_0)}\left\{\frac{k-1}{k}E[\zeta]+\frac{1}{k}E[V_h((k-1)S,\ kS,\ y_0)]\right\}\right)$$
$$\leq \left(\frac{1}{k}\right)c\nu_h(S)(1+E[||y_0||^\alpha]).$$

Using the estimate above and (6.34), one can further obtain that

$$d_H(E[V'(kS,y_0)],\ \frac{1}{k}\sum_1^k E[V_h((i-1)S,\ iS,\ y_0)])$$
$$\leq d_H\left(\frac{k-1}{k}E[V'((k-1)S,\ y_0)]+\frac{1}{k}E[V_h((k-1)S,\ kS,\ y_0)]\right),$$
$$\frac{1}{k}\sum_1^k E[V_h((i-1)S,\ iS,\ y_0)]\ +\ \left(\frac{1}{k}\right)c\nu_h(S)(1+E[||y_0||^\alpha])$$
$$\leq\ d_H\left(\frac{k-1}{k}E[V'((k-1)S,\ y_0)],\frac{k-1}{k}\frac{1}{k-1}\sum_1^{k-1}E[V_h((i-1)S,\ iS,\ y_0)]\right)$$
$$+\left(\frac{1}{k}\right)c\nu_h(S)(1+E[||y_0||^\alpha])\leq\left(\frac{k-1}{k}\right)c\nu_h(S)(1+E[||y_0||^\alpha])+\left(\frac{1}{k}\right)c\nu_h(S)(1+E[||y_0||^\alpha])$$
$$= c\nu_h(S)(1+E[||y_0||^\alpha]).$$

Thus, (6.33) is established.

Since

$$E[V_h((i-1)S,\ iS,\ y_0)]=E[V_h(S,y_0)]\quad\forall i=1,\dots,k,$$

(6.33) is equivalent to

$$d_H(E[V'_h(kS,y_0)],\ \frac{1}{k}\sum_1^k E[V_h((S,y_0)])\leq c\nu_h(S)(1+E[||y_0||^\alpha]).$$

By Shapley–Folkman's theorem (see, e.g., [24, p. 204])

$$d_H\left(\frac{1}{k}\sum_1^k E[V_h(S,y_0)],\ coE[V_h(S,y_0)]\right)\leq\frac{2(j+1)c_h}{k},\tag{6.35}$$

where c_h is as in (6.6) and j is the dimension of the Euclidean space containing the subsets above. These imply (6.23) with $\bar{c}_h\stackrel{\text{def}}{=}2(j+1)c_h$.

Part III: *Proof of Lemma* 6.5(ii). Let $y_0^0\stackrel{\text{def}}{=}y_0$ and let y_0^i $i=1,\dots,k-1$ have the same probability distribution as y_0 and be independent of y_0 and among themselves (and also independent of $W(\cdot)$). Using induction, let us show that

$$d_H^E\left(V'_h(kS,y_0),\ \frac{1}{k}\sum_1^k V_h((i-1)S,\ iS,\ y_0^{i-1})\right)\leq c\nu_h(S)(1+E[||y_0||^\alpha]),\quad k=1,2,\dots.\tag{6.36}$$

For $k = 1$, the above expression is obviously true. Assume that

$$d_H^E\left(V_h'((k-1)S, y_0),\ \frac{1}{k-1}\sum_1^{k-1} V_h((i-1)S,\ iS,\ y_0^{i-1})\right) \le c\nu_h(S)(1+E[||y_0||^\alpha]). \tag{6.37}$$

From Assumption 1 and S-h-AC (see (3.11)) it follows that, for any η such that $(\zeta, \eta) \in W_h'((k-1)S, y_0)$,

$$d_H^E\left(V_h((k-1)S,\ kS,\ \eta),\ V_h((k-1)S,\ kS,\ y_0^{k-1})\right) \le \nu_h(S)(1+E[||\eta||^\alpha]+E[||y_0||^\alpha])$$

$$\le \nu_h(S)(1+c_1(1+E[||y_0||^\alpha])+E[||y_0||^\alpha]) = c\nu_h(S)(1+E[||y_0||^\alpha]).$$

Hence, by (6.30),

$$d_H^E\left(V'(kS, y_0),\ \frac{k-1}{k}V'((k-1)S,\ y_0)+\frac{1}{k}V_h((k-1)S,\ kS,\ y_0^{k-1})\right)$$

$$= d_H^E\left(V'(kS, y_0), \bigcup_{(\zeta,\eta)\in W_h'((k-1)S,y_0)}\left\{\frac{k-1}{k}\zeta+\frac{1}{k}V_h((k-1)S,\ kS,\ y_0^{k-1})\right\}\right)$$

$$\le \left(\frac{1}{k}\right)\sup_{(\zeta,\eta)\in W_h'((k-1)S,y_0)} d_H^E(V_h((k-1)S,\ kS,\ \eta),\ V_h((k-1)S,\ kS,\ y_0^{k-1}))$$

$$\le \left(\frac{1}{k}\right)c\nu_h(S)(1+E[||y_0||^\alpha]).$$

Using this estimate and (6.37), one obtains

$$d_H^E\left(V'(kS, y_0),\ \frac{1}{k}\sum_1^k V_h((i-1)S,\ iS,\ y_0^{i-1})\right)$$

$$\le d_H^E\left(\frac{k-1}{k}V'((k-1)S,\ y_0)+\frac{1}{k}V_h((k-1)S,\ kS,\ y_0^{k-1}),\ \frac{1}{k}\sum_1^k V_h((i-1)S,\ iS,\ y_0^{i-1})\right)$$

$$+\left(\frac{1}{k}\right)c\nu_h(S)(1+E[||y_0||^\alpha])$$

$$\le d_H^E\left(\frac{k-1}{k}V'((k-1)S,\ y_0),\ \frac{k-1}{k}\frac{1}{k-1}\sum_1^{k-1} V_h((i-1)S,\ iS,\ y_0^{i-1})\right)$$

$$+\left(\frac{1}{k}\right)c\nu_h(S)(1+E[||y_0||^\alpha]) \le \left(\frac{k-1}{k}\right)c\nu_h(S)(1+E[||y_0||^\alpha])+\left(\frac{1}{k}\right)c\nu_h(S)(1+E[||y_0||^\alpha])$$

$$= c\nu_h(S)(1+E[||y_0||^\alpha]).$$

This proves the validity of the estimate (6.36).

The elements of $V_h((i-1)S,\ iS,\ y_0^{i-1}), i = 1, 2, \ldots, k$, are mutually independent and, by (6.6),

$$E[||\zeta||^2] \le c_h^2 \qquad \forall\zeta \in V_h((i-1)S,\ iS,\ y_0^{i-1}), \quad i = 1, 2, \ldots, k.$$

Hence, one can use Proposition 3.5 and the fact that $E[V_h((i-1)S,\ iS,\ y_0^{i-1})] = E[V_h(S, y_0)]$ to obtain

$$d_H^E\left(\frac{1}{k}\sum_1^k V_h((i-1)S,\ iS,\ y_0^{i-1}),\ \frac{1}{k}\sum_1^k E[V_h((i-1)S,\ iS,\ y_0^{i-1})]\right)$$
$$= d_H^E\left(\frac{1}{k}\sum_1^k V_h((i-1)S,\ iS,\ y_0^{i-1}),\ \frac{1}{k}\sum_1^k E[V_h(S, y_0)]\right) \le \frac{c_h}{\sqrt{k}}.$$

The last estimate along with (6.35) and (6.33) imply (6.24). □

COROLLARY 6.6. *For any y_0 with the probability distribution from the class (C, α),*

$$d_H(E[V_h'(kS, y_0)], V_h) \le \frac{\bar{c}_h}{k} + 2c(1+C)\nu_h(S) \tag{6.38}$$

if W-h-AC is satisfied and

$$d_H^E(V_h'(kS, y_0), V_h) \le \frac{c_h}{\sqrt{k}} + \frac{\bar{c}_h}{k} + 2c(1+C)\nu_h(S) \tag{6.39}$$

if S-h-AC is satisfied.

Proof. The estimates follow from (6.19), (6.23) and (6.19), (6.24), respectively. □

LEMMA 6.7. *For any y_0 with the probability distribution from the class (C, α),*

$$\sup_{\zeta\in V_h} d(\zeta,\ E[V_h(S, y_0)]) \le \nu_h^1(S), \qquad \lim_{S\to\infty} \nu_h^1(S) = 0 \tag{6.40}$$

if W-h-AC is satisfied, and

$$\sup_{\zeta\in V_h} d^E(\zeta,\ V_h(S, y_0)) \le \nu_h^2(S), \qquad \lim_{S\to\infty} \nu_h^2(S) = 0 \tag{6.41}$$

if S-h-AC is satisfied.

Proof. Using (6.9) with $S'' = S$ and $S' = \lfloor S^{\frac{1}{2}}\rfloor S^{\frac{1}{2}}$, one can obtain

$$d_H(E[V_h(\lfloor S^{\frac{1}{2}}\rfloor S^{\frac{1}{2}}, y_0)],\ E[V_h(S, y_0)]) \le \frac{2c_h}{S^{\frac{1}{2}}}, \tag{6.42}$$

where $\lfloor S^{\frac{1}{2}}\rfloor$ stands for the integer part of $S^{\frac{1}{2}}$. Hence,

$$\sup_{\zeta\in V_h} d(\zeta,\ E[V_h(S, y_0)]) \le \sup_{\zeta\in V_h} d(\zeta,\ E[V_h(\lfloor S^{\frac{1}{2}}\rfloor S^{\frac{1}{2}}, y_0)]) + \frac{2c_h}{S^{\frac{1}{2}}}. \tag{6.43}$$

From (6.22) and (6.38) (with the replacement of S by $S^{\frac{1}{2}}$ and the replacement of k by $\lfloor S^{\frac{1}{2}}\rfloor$) it follows, on the other hand, that

$$\sup_{\zeta\in V_h} d(\zeta,\ E[V_h(\lfloor S^{\frac{1}{2}}\rfloor S^{\frac{1}{2}}, y_0)]) \le \sup_{\zeta\in V_h} d(\zeta,\ E[V_h'(\lfloor S^{\frac{1}{2}}\rfloor S^{\frac{1}{2}}, y_0)])$$
$$\le d_H(E[V_h'(\lfloor S^{\frac{1}{2}}\rfloor S^{\frac{1}{2}}, y_0)],\ V_h) \le \frac{\bar{c}_h}{\lfloor S^{\frac{1}{2}}\rfloor} + 2c(1+C)\nu_h(S^{\frac{1}{2}})$$

This and (6.43) imply (6.40) with $\nu_h^1(S) \stackrel{\text{def}}{=} \frac{2c_h}{S^{\frac{1}{2}}} + \frac{\bar{c}_h}{\lfloor S^{\frac{1}{2}}\rfloor} + 2c(1+C)\nu_h(S^{\frac{1}{2}})$.

To establish (6.41), one can use (6.8) and obtain, similarly to (6.42), that

$$d_H^E(V_h(\lfloor S^{\frac12}\rfloor S^{\frac12}, y_0),\ V_h(S, y_0)) \le \frac{2c_h}{S^{\frac12}},$$

$$\Rightarrow \quad \sup_{\zeta\in V_h} d^E(\zeta,\ V_h(S, y_0)) \le \sup_{\zeta\in V_h} d^E(\zeta,\ V_h(\lfloor S^{\frac12}\rfloor S^{\frac12}, y_0)]) + \frac{2c_h}{S^{\frac12}}. \tag{6.44}$$

By (6.22) and (6.39) (with S being replaced by $S^{\frac12}$ and k by $\lfloor S^{\frac12}\rfloor$ as above),

$$\sup_{\zeta\in V_h} d^E(\zeta,\ V_h(\lfloor S^{\frac12}\rfloor S^{\frac12}, y_0)) \le \sup_{\zeta\in V_h} d^E(\zeta,\ V_h'(\lfloor S^{\frac12}\rfloor S^{\frac12}, y_0))$$

$$\le d_H^E(V_h'(\lfloor S^{\frac12}\rfloor S^{\frac12}, y_0),\ V_h) \le \frac{c_h}{\sqrt{\lfloor S^{\frac12}\rfloor}} + \frac{\bar c_h}{\lfloor S^{\frac12}\rfloor} + 2c(1+C)\nu_h(S^{\frac12}).$$

This and (6.44) imply (6.41) with $\nu_h^2(S)$ being equal to the sum of $\frac{2c_h}{S^{\frac12}}$ and the right-hand side of the last inequality. □

Proof of Theorem 3.2. By (6.19),

$$sup_{\zeta\in E[V_h(S,y_0)]} d(\zeta, V_h) \le sup_{\zeta\in coE[V_h(S,y_0)]} d(\zeta, V_h) \le d_H(coE[V_h(S, y_0)], V_h)$$

$$\le c(1+C)\nu_h(S).$$

Comparing this estimate and (6.40), one obtains (3.14) with $\nu_h^{C,\alpha}(S) \stackrel{\text{def}}{=} \max\{c(1+C)\nu_h(S), \nu_h^1(S)\}$. □

Proof of Theorem 3.4. The statement (i) of the theorem is established by (6.41), with $\tilde\nu_h^{C,\alpha}(S) \stackrel{\text{def}}{=} \nu_h^2(S)$. Let us prove the statement (ii). Assume it is not true. Then there exists a number $\delta > 0$, and sequences $\mu_l \in \mathcal{M}$ and $S_l, l = 1, 2, \dots$, ($S_l \to \infty$ as $l \to \infty$) such that, for any $\mu' \in \mathcal{M}(S_l, y_0)$,

$$E[\rho(\mu_l, \mu')] = \sum_{i=1}^{\infty} 2^{-i} E\left[\left|\int f_i(u,y)\mu_l(du,dy) - \int f_i(u,y)\mu'(du,dy)\right|\right] \ge \delta \tag{6.45}$$

$$\Rightarrow \sum_{i=1}^{N} 2^{-i} E[|\int f_i(u,y)\mu_l(du,dy) - \int f_i(u,y)\mu'(du,dy)|] \ge \frac{\delta}{2} \quad \forall \mu' \in \mathcal{M}(S_l, y_0)$$

for N large enough. Hence, for any $\mu' \in \mathcal{M}(S_l, y_0)$,

$$E\left[\sqrt{\sum_{i=1}^{N}\left|\int f_i(u,y)\mu_l(du,dy) - \int f_i(u,y)\mu'(du,dy)\right|^2}\right] \ge \frac{c_N\delta}{2}, \quad c_N = \text{const.} \tag{6.46}$$

Let $h(u,y)$ be defined by (3.1) with $j = N$ and let $\zeta_l \stackrel{\text{def}}{=} \int h(u,y)\mu_l(du,dy)$; by (3.17), $\zeta_l \in V_h$. Also, by (3.6), $V_h(S_l, y_0)$ is the union of all $\zeta' \stackrel{\text{def}}{=} \int h(u,y)\mu'(du,dy)$ over $\mu' \in \mathcal{M}(S_l, y_0)$. The estimate (6.46) is equivalent, thus, to

$$E[||\zeta_l - \zeta'||] \ge \frac{c_N\delta}{2} \quad \forall \zeta' \in V_h(S_l, y_0) \qquad \Leftrightarrow \qquad d^E(\zeta_l, V_h(S_l, y_0)) \ge \frac{c_N\delta}{2}.$$

The latter contradicts (6.41). □

Proof of Theorem 3.3. Let Assumption 1 and W-h-AC be satisfied for any $h(\cdot)$ as in (3.1). Then, by Theorem 3.2, for any such $h(\cdot)$ and any initial condition y_0 with the probability distribution from the class (C,α), there exists a convex and compact set V_h such that (3.14) is satisfied. From Corollary 3.7 in [28] (see also Theorem 3.1(i) in [27] and more general results in [7]) it follows that

$$\rho_H(E[\mathcal{M}(S,y_0)],\bar{\mathcal{M}}) \le \ \nu^{C,\alpha}(S) \tag{6.47}$$

for some $\nu^{C,\alpha}(S)$ tending to zero as S tends to infinity, where

$$\bar{\mathcal{M}} \stackrel{\text{def}}{=} \{\mu \in \mathcal{P}(U\times \bar{R}^m) \mid \int h(u,y)\mu(du,dy) \in V_h \ \ \forall h(u,y) \ \ as\ in\ (3.1)\}.$$

It is easy to verify that the set $\bar{\mathcal{M}}$ is convex and compact. Hence, the estimate (3.15) will be established if one shows that $\mathcal{M}=\bar{\mathcal{M}}$. Since, by definition, $\mathcal{M}\subset\bar{\mathcal{M}}$, it is enough to show that $\bar{\mathcal{M}}\subset\mathcal{M}$.

For $N=1,2,\dots$, let $Y_N \stackrel{\text{def}}{=} \{y\in R^m \mid ||y||\le N\}$ and $Y_N^c \stackrel{\text{def}}{=} \{y\in R^m \mid ||y||>N\}$. By (2.8) and (2.10), for any $\mu\in E[\mathcal{M}(S,y_0)]$, there exists an admissible control $u(\cdot)$ and the corresponding solution $y(\cdot)$ of the CSDE (2.4) such that

$$\mu(U\times Y_N^c) = \frac{1}{S}\int_0^S E[\chi_{Y_N^c}(y(\tau))]d\tau \ \le \ \frac{1}{N^\alpha}\frac{1}{S}\int_0^S E[\chi_{Y_N^c}(y(\tau))\ ||y(\tau)||^\alpha]d\tau,$$

where $\chi_{Y_N^c}(y(\tau))$ is the indicator function of Y_N^c. Hence, using Assumption I and the fact that y_0 has the probability distribution from the class (C,α), one obtains that, for any $\mu\in E[\mathcal{M}(S,y_0)]$,

$$\mu(U\times Y_N^c) \le \frac{1}{N^\alpha}\frac{1}{S}\int_0^S E[||y(\tau)||^\alpha]d\tau \le \frac{1}{N^\alpha}c_1(C+1)$$
$$\Rightarrow \qquad \mu(U\times Y_N) \ge 1-\frac{1}{N^\alpha}c_1(C+1).$$

Take an arbitrary $\mu\in\bar{\mathcal{M}}$. By (6.47), there exists a sequence $\mu_i\in E[\mathcal{M}(S_i,y_0)]$ (S_i tends to infinity as i tends to infinity) such that

$$\lim_{i\to\infty}\rho(\mu_i,\mu)=0 \quad \Rightarrow \quad \mu(U\times Y_N)\ge \limsup_{i\to\infty}\mu_i(U\times Y_N)\ge 1-\frac{1}{N^\alpha}c_1(C+1)$$
$$\Rightarrow \qquad \mu(U\times R^m)=1.$$

The latter implies that $\mu\in\mathcal{M}$ and, hence, $\bar{\mathcal{M}}\subset\mathcal{M}$.

Using the second representation for $E[V_h(S,y_0)]$ in (3.7), it is straightforward to verify that the validity of (3.15) implies the validity of (3.14) with V_h as in (3.17) for any $h(\cdot)$ as in (3.1). The fact that W-h-AC is satisfied for any such $h(\cdot)$ follows now from Theorem 3.2. □

7. Proofs for sections 4. *Proof of Theorem* 4.2. Assume that (4.6) is not valid. Then there exist a number $\delta>0$ and initial conditions y_0 with the probability distribution from the class (C,α) such that, for some S_i, $\lim_{i\to\infty}S_i=\infty$, and some $\mu_i\in\mathcal{M}(S_i,y_0)$,

$$E[\inf_{\mu'\in D}\rho(\mu_i,\mu')]\ge\delta \quad i=1,2,\dots, \qquad \Rightarrow \qquad E[\inf_{\mu'\in D}\ \rho(\mu^*,\mu')]\ge\delta, \tag{7.1}$$

where it is assumed (without loss of generality) that $\mu_i \to \mu^*$ in law as $i \to \infty$. From the fact that $\mu_i \in \mathcal{M}(S_i, y_0)$ it follows that there exist an admissible control $u^i(\cdot)$ and the corresponding solution $y^i(\cdot)$ of the CSDE (2.4) (with the initial conditions $y^i(0) = y_0$) such that μ_i is the occupational measure of the pair $(u^i(\cdot), y^i(\cdot))$ on the interval $[0, S_i]$ (see (2.6)). Hence, for any $f \in \mathcal{D}$

$$\frac{1}{S_i}(f(y^i(S) - f(y(0))) = \int_{U\times\bar{R}^m} (Lf)(u,y)\mu_i(du,dy)$$

$$+\frac{1}{S_i}\int_0^{S_i} \langle \nabla f(y^i(\tau)), b(y^i(\tau))dW(\tau)\rangle,$$

where, in order for the integration in the first term of the right-hand side to be legitimate, the definition of $(Lf) : U\times R^m \to R^1$ (see (4.1) above) is formally extended to $(Lf) : U \times \bar{R}^m \to R^1$ by setting $(Lf)(\cdot,\infty) \stackrel{\text{def}}{=} 0$.

The left-hand side and the variance of the second term on the right-hand side of the above expression tend to zero as $S_i \to \infty$ (this can be easily derived from the fact that the probability distribution of y_0 belongs to the class (C, α) and from that Assumption 1 is satisfied with $\alpha \geq 2$). Hence, the first term on the right-hand side tends to zero in law.

By Skorohod's theorem (see, e.g., [15, p. 23]), there exist $\mathcal{P}(U \times \bar{R}^m)$-valued random variables $\tilde{\mu}_i$ and $\tilde{\mu}^*$ defined on a common probability space such that they agree in law with μ_i and μ^*, respectively, and such that

$$\lim_{i\to\infty} \rho(\tilde{\mu}_i, \tilde{\mu}^*) = 0 \quad a.s. \tag{7.2}$$

$$\Rightarrow \quad \lim_{i\to\infty} \int_{U\times\bar{R}^m} (Lf)(u,y)\tilde{\mu}_i(dxdu) = \int_{U\times\bar{R}^m} (Lf)(u,y)\tilde{\mu}^*(du,dy) \quad a.s.$$

$$\Rightarrow \quad \int_{U\times\bar{R}^m} (Lf)(u,y)\tilde{\mu}^*(du,dy) = 0 \ \ a.s. \quad \Rightarrow \quad \int_{U\times\bar{R}^m} (Lf)(u,y)\mu^*(du,dy) = 0 \ \ a.s.$$

Since $\mathcal{D}$ is countable, the last expression is valid for all $f \in \mathcal{D}$ outside a common zero probability set. Hence, if one establishes that $\mu^* \in \mathcal{P}(U \times R^m)$ a.s., it will follow that $\mu^* \in D$ a.s. and, thus, it will contradict (7.1).

To complete the proof of the theorem, one needs to show now that $\mu^* \in \mathcal{P}(U\times R^m)$ a.s. That is, one needs to show that

$$\mu^*(U \times R^m) = 1 \quad a.s. \tag{7.3}$$

From Assumption 1 it follows that, for any $\delta > 0$, there exists a compact set $Y_\delta \subset R^m$ such that

$$E[\tilde{\mu}_i(U \times Y_\delta)] = E[\mu_i(U \times Y_\delta)] \geq 1 - \delta,$$

where it is also taken into account that $\tilde{\mu}_i$ and μ_i agree in law. By (7.2),

$$\tilde{\mu}^*(U \times Y_\delta) \geq \limsup_{i\to\infty} \tilde{\mu}_i(U \times Y_\delta) \quad a.s.$$

$$\Rightarrow \quad E\left[\tilde{\mu}^*(U \times Y_\delta)\right] \geq E[\limsup_{i\to\infty} \tilde{\mu}_i(U \times Y_\delta)] \geq \limsup_{i\to\infty} E[\tilde{\mu}_i(U \times Y_\delta)] \geq 1 - \delta.$$

Since μ^* and $\tilde{\mu}^*$ agree in law,

$$E[\tilde{\mu}^*(U \times Y_\delta)] = E[\mu^*(U \times Y_\delta)] \quad \Rightarrow \quad E[\mu^*(U \times Y_\delta)] \geq 1 - \delta.$$

Since δ can be arbitrary small, the latter implies that $E[\mu^*(U \times R^m)] = 1$ which, in turn, implies the validity of (7.3). This completes the proof of the theorem. □

Proof of Theorem 4.1(i). It can be easily verified that $\rho(\mu, D)$ is a convex function of μ. Hence, by (4.6),

$$\rho(E[\mu], D) \leq E[\rho(\mu, D)] \leq \bar{\nu}^{(C,\alpha)}(S) \quad \forall \mu \in \mathcal{M}(S, y_0)$$

$$\Rightarrow \sup_{\mu \in E[\mathcal{M}(S,y_0)]} \rho(\mu, D) \leq \bar{\nu}^{(C,\alpha)}(S).$$

Since the above estimate is uniform with respect to the initial conditions y_0 which have the probability distribution from the class (C, α), it follows that

$$\sup_{\{y_0\} \in (C,\alpha)} \left\{ \sup_{\mu \in E[\mathcal{M}(S,y_0)]} \rho(\mu, D) \right\} \leq \bar{\nu}^{(C,\alpha)}(S).$$

In Lemma 7.1 below it is shown that

$$\sup_{\mu \in D} \rho \left(\mu, \bigcup_{\{y_0\} \in (C,\alpha)} E[\mathcal{M}(S, y_0)] \right) = 0. \tag{7.4}$$

These imply (4.4). □

LEMMA 7.1. *The equation* (7.4) *is valid if the conditions of Theorem* 4.1 *are satisfied.*

Proof. For any $h(\cdot)$ as in (3.1), define the set D_h by

$$D_h \stackrel{\text{def}}{=} \bigcup_{\mu \in D} \left\{ \int h(u, y) \mu(du, dy) \right\}. \tag{7.5}$$

As follows from Lemma 3.5 in [28], the validity of (7.4) will be established if one shows that

$$\sup_{v \in D_h} d \left(v, \bigcup_{\{y_0\} \in (C,\alpha)} \{E[V_h(S, y_0)]\} \right) = 0. \tag{7.6}$$

Take an arbitrary element $v \in D_h$. By definition, there exists $\mu \in D$ such that $v = \int h(u, y) \mu(du, dy)$. From results in [13] and [49] it follows that there exists m-dimensional standard Brownian motion $W'(\cdot)$ and a stationary $\mathcal{P}(U) \times R^m$ - valued random process $(\lambda'(\tau), y'(\tau))$ such that

$$dy'(\tau) = \tilde{a}(\lambda'(\tau), y'(\tau))dt + b(y'(\tau))dW'(\tau), \qquad \tilde{a}(\lambda, y) \stackrel{\text{def}}{=} \int a(u, y)\lambda(du), \tag{7.7}$$

and

$$E \left[\int h(u, y'(\tau))\lambda'(\tau)(du) \right] = \int h(u, y)\mu(du, dy) = v \quad \forall \tau \geq 0, \tag{7.8}$$

$$E[||y'(\tau)||^{\alpha}] = \int ||y||^{\alpha}\mu(du,dy) \le c_2 \quad \forall \tau \ge 0, \tag{7.9}$$

with $W'(\cdot)$ being independent of $y'(0)$ and $\lambda'(\cdot)$ being nonanticipative (i.e., for $\bar{\tau} > \tau, W'(\bar{\tau}) - W'(\tau)$ is independent of $\{y'(0)$ and $W'(\theta), \lambda'(\theta), \theta \le \tau\}$); ($c_2$ is the constant from Assumption 2). Using Filippov type chattering lemma for CSDE (see, e.g., [14, p. 15]), one can establish that there exists a sequence of admissible controls $u^i(\cdot)$ and the corresponding sequence of solutions $y^i(\cdot)$ of the CSDE (2.4) (considered with $W'(\cdot)$ instead of $W(\cdot)$) such that $y^i(0) = y'(0)$ and such that

$$\lim_{i\to\infty} E\left[\left\| \frac{1}{S}\int_0^S h(u^i(\tau), y^i(\tau))d\tau - \frac{1}{S}\int_0^S \int h(u, y'(\tau))\lambda'(\tau)(du)d\tau \right\|\right] = 0. \tag{7.10}$$

From (7.8) and (7.10) it follows that

$$\lim_{i\to\infty} \left\| E\left[\frac{1}{S}\int_0^S h(u^i(\tau), y^i(\tau))d\tau\right] - v \right\| = 0. \tag{7.11}$$

Since

$$E\left[\frac{1}{S}\int_0^S h(u^i(\tau), y^i(\tau))d\tau\right] \in E[V_h(S, y'(0))] \subset \bigcup_{\{y_0\}\in(C,\alpha)} \{E[V_h(S, y_0)]\}$$

(the last inclusion being due to the fact that, by (7.9), $y'(0)$ has the probability distribution from the class (C, α) with $C \ge c_2$), one can use (7.11) to obtain that

$$dist\left(v, \bigcup_{\{y_0\}\in(C,\alpha)} \{E[V_h(S, y_0)]\}\right) = 0.$$

As v is an arbitrary element of D_h, this implies (7.6). □

The proofs of Corollaries 4.4 and 4.5 are based on the following result.

LEMMA 7.2. *A sequence $\mu^k \in \mathcal{P}(U \times R^m), k = 1, 2, \ldots,$ converges to $\mu \in \mathcal{P}(U \times R^m)$ in the metric ρ defined in* (2.1) *(that is, $\lim_{k\to\infty} \rho(\mu^k, \mu) = 0$) if and only if*

$$\lim_{k\to\infty} \int f(u,y)\mu^k(du,dy) = \int f(u,y)\mu(du,dy)$$

for any bounded continuous function $f(u, y) : U \times R^m \to R^1$.

Proof. follows from Theorem 2.1.1 in [15]. □

Proof of Corollary 4.4. By Assumption 2 (see (4.3)), for any $\mu \in D$ and $N \ge 1$,

$$N^{\alpha-1}\int_{||y||\ge N} ||y||\mu(du,dy) \le \int_{||y||\ge N} ||y||^{\alpha}\mu(du,dy) \le \int ||y||^{\alpha}\mu(du,dy) \le c_2$$

$$\int_{||y||\ge N} ||y||\mu(du,dy) \le \frac{c_2}{N^{\alpha-1}} \qquad \forall \mu \in D. \tag{7.12}$$

Let $\xi_N(\theta) : [0,\infty) \to [0,1]$ be a continuous function such that $\xi_N(\theta) = 1$ for $\theta \in [0, N]$ and such that $\xi_N(\theta) = 0$ for $\theta \in [N+1, \infty)$. Let $g_N(u,y) \stackrel{\text{def}}{=} g(u,y)\xi_N(||y||)$. According to these definitions, $g_N(u,y) = g(u,y)$ for $||y|| \le N$ and also

$$||g_N(u,y)|| \le ||g(u,y)||, \quad ||g_N(u,y)|| \le \max_{u\in U, ||y||\le N} ||g(u,y)|| \qquad \forall (u,y) \in U \times Y.$$

Due to the Lipschitz continuity of $g(u,y)$,

$$||g(u,y)|| \le a_1 + a_2||y|| \quad \Rightarrow \quad ||g_N(u,y)|| \le a_1 + a_2||y|| \qquad \forall y \in R^m, \tag{7.13}$$

where $a_i = \text{const}, i = 1,2$. By (7.12), it is implied that

$$\left\| \int g(u,y)\mu(du,dy) - \int g_N(u,y)\mu(du,dy) \right\| \le \frac{a_3}{N^{\alpha-1}} \ \ \forall \mu \in D, \quad a_3 = \text{const}. \tag{7.14}$$

From (7.12) and (7.13) it follows that the set V_g is bounded. Let us prove that it is closed by showing that, if $\mu_k \in D$ and the limit $\lim_{k\to\infty} \int g(u,y)\mu_k(du,dy)$ exists, then this limit belongs to V_g. Assume the above limit does exist. Due to the fact that D is compact, one may also assume (without loss of generality) that $\lim_{k\to\infty} \rho(\mu_k,\mu) = 0$ for some $\mu \in D$. By virtue of Lemma 7.2 and since $g_N(u,y)$ is bounded, the latter leads to the equality $\lim_{k\to\infty} \int g_N(u,y)\mu_k(du,dy) = \int g_N(u,y)\mu(du,dy)$, which, in turn, implies that

$$\begin{aligned}
&\lim_{k\to\infty} \left\| \int g(u,y)\mu_k(du,dy) - \int g(u,y)\mu(du,dy) \right\| \\
&\le \limsup_{k\to\infty} \left\| \int g(u,y)\mu_k(du,dy) - \int g_N(u,y)\mu_k(du,dy) \right\| \\
&\quad + \lim_{k\to\infty} \left\| \int g_N(u,y)\mu_k(du,dy) - \int g_N(u,y)\mu(du,dy) \right\| \\
&\quad + \left\| \int g_N(u,y)\mu(du,dy) - \int g(u,y)\mu(du,dy) \right\| \\
&\le \frac{2a_3}{N^{\alpha-1}} \quad \Rightarrow \quad \lim_{k\to\infty} \int g(u,y)\mu_k(du,dy) = \int g(u,y)\mu(du,dy) \in V_g.
\end{aligned}$$

This proves that V_g is compact.

Let $V_{g_N}(S,y_0)$ and V_{g_N} be defined by (4.10) and (4.11) with the replacement of $g(\cdot)$ by $g_N(\cdot)$. By (7.14),

$$d_H(V_{g_N}, V_g) \le \frac{a_3}{N^{\alpha-1}}. \tag{7.15}$$

Similarly to (7.12), from Assumption 1 it follows that, for any y_0 having a probability distribution from the class (C,α),

$$E\left[\int_{||y||\ge N} ||y||\mu(du,dy) \right] \le \frac{a_4}{N^{\alpha-1}} \qquad \forall \mu \in \mathcal{M}(S,y_0) \tag{7.16}$$

$$\Rightarrow \ E\left[\left\| \int g(u,y)\mu(du,dy) - \int g_N(u,y)\mu(du,dy) \right\| \right] \le \frac{a_5}{N^{\alpha-1}} \ \ \forall \mu \in \mathcal{M}(S,y_0) \tag{7.17}$$

$$\Rightarrow \qquad d_H^E(V_{g_N}(S,y_0), V_g(S,y_0)) \le \frac{a_5}{N^{\alpha-1}}, \tag{7.18}$$

where a_4, a_5 are positive constants. From (7.15) and (7.18) it follows that, to prove (4.12), it is enough to prove that

$$\sup_{v\in V_{g_N}(S,y_0)} E[d(v,V_{g_N})] \le \nu_{g_N}(S), \qquad \lim_{S\to\infty} \nu_{g_N}(S) = 0. \tag{7.19}$$

Assume it is not true. Then there exists $\delta > 0$ and sequences $S_i, \lim_{i\to\infty} S_i = 0$, and $\mu_i \in \mathcal{M}(S_i, y_0$ such that

$$E\left[d\left(\int g_N(u,y)\mu_i(du,dy), V_{g_N}\right)\right] \geq \delta, \tag{7.20}$$

with $\mu_i \to \mu^*$ in law as $i \to \infty$. Like in the proof of Theorem 4.2, let $\tilde{\mu}_i$ and $\tilde{\mu}^*$ be $\mathcal{P}(U \times R^m)$-valued random variables defined on a common probability space such that they agree in law with μ_i and μ^*, respectively, and such that (7.2) is satisfied. From (7.2) and Lemma 7.2 it follows that

$$\lim_{i\to\infty} \int g_N(u,y)\tilde{\mu}_i(du,dy) = \int g_N(u,y)\tilde{\mu}^*(du,dy) \in V_{g_N} \quad a.s.,$$

the last inclusion being implied by the fact that $\tilde{\mu}^* \in D$ (which is established similarly to the proof of Theorem 4.2). Hence, $E[d(\int g_N(u,y)\tilde{\mu}^*(du,dy), V_{g_N})] = 0$. This contradicts the following inequalities resulting from (7.20) and the fact that $\tilde{\mu}_i$ and μ_i agree in law:

$$E\left[d\left(\int g_N(u,y)\tilde{\mu}^*(du,dy),\ V_{g_N}\right)\right] = \lim_{i\to\infty} E\left[d\left(\int g_N(u,y)\tilde{\mu}_i(du,dy), V_{g_N}\right)\right] \geq \delta$$

Thus Corollary 4.4 is proved. □

Proof of Corollary 4.5. By (7.15) and (7.18), to prove (4.12), it is sufficient to prove that

$$\sup_{v\in V_{g_N}} d^E(v, V_{g_N}(S, y_0)) \leq \ \nu_{g_N}(S), \qquad \lim_{S\to\infty} \nu_{g_N}(S) = 0. \tag{7.21}$$

Assume it is not true. Then there exist a number $\delta > 0$ and sequences $\mu_i \in D$ and $S_i, i = 1, 2, \ldots, (S_i \to \infty$ as $i \to \infty)$ such that

$$E\left[\left\|\int g_N(u,y)\mu_i(du,dy) - \int g_N(u,y)\mu(du,dy)\right\|\right] \geq \delta \qquad \forall \mu \in \mathcal{M}(S_i, y_0). \tag{7.22}$$

From Theorem 3.4(ii) (see (3.19)) it follows that there exists $\mu^{S_i} \in \mathcal{M}(S_i, y_0)$, such that

$$E[\rho(\mu_i, \mu^{S_i})] \leq 2\tilde{\nu}^{(C,\alpha)}(S_i), \qquad \lim_{i\to\infty} \tilde{\nu}^{(C,\alpha)}(S_i) = 0. \tag{7.23}$$

Without loss of generality, one may assume that $\mu_i \to \mu^*$ and $\mu^{S_i} \to \mu^{**}$ in law. Also, using Skorohod's theorem, one can verify (similar to the way it is done in the proof of Theorem 4.2) that there exist $\mathcal{P}(U \times R^m)$-valued random variables $\tilde{\mu}_i, \tilde{\tilde{\mu}}_i, \tilde{\mu}^*$ and $\tilde{\tilde{\mu}}^{**}$ defined on a common probability space such that they agree in law with μ_i, μ^{S_i}, μ^* and μ^{**} and such that, almost sure, $\tilde{\mu}_i \to \tilde{\mu}^*$, $\tilde{\tilde{\mu}}_i \to \tilde{\tilde{\mu}}^{**}$. Note that $E[\rho(\tilde{\mu}_i, \tilde{\tilde{\mu}}_i)] = E[\rho(\mu_i, \mu^{S_i})]$ and, hence, by (7.23),

$$E[\rho(\tilde{\mu}^*, \tilde{\tilde{\mu}}^{**})] \leq \lim_{i\to\infty} E[\rho(\tilde{\mu}^*, \tilde{\mu}_i)] + \limsup_{i\to\infty} E[\rho(\tilde{\mu}_i, \tilde{\tilde{\mu}}_i)] + \lim_{i\to\infty} E[\rho(\tilde{\tilde{\mu}}_i, \tilde{\tilde{\mu}}^{**})]$$

$$\leq \lim_{i\to\infty} 2\tilde{\nu}^{(C,\alpha)}(S_i) = 0 \qquad \Rightarrow \qquad E[\rho(\tilde{\mu}^*, \tilde{\tilde{\mu}}^{**})] = 0 \qquad \Rightarrow \qquad \tilde{\mu}^* = \tilde{\tilde{\mu}}^{**} \ \ a.s.$$

The latter implies (by virtue of Lemma 7.2) that

$$\lim_{i\to\infty}\int g_N(u,y)\tilde{\mu}_i(du,dy)=\lim_{i\to\infty}\int g_N(u,y)\tilde{\tilde{\mu}}_i(du,dy)=\int g_N(u,y)\tilde{\mu}^*(du,dy)\quad a.s.$$
$$\Rightarrow\quad 0=\lim_{i\to\infty}E\left[\left\|\int g_N(u,y)\tilde{\mu}_i(du,dy)-\int g_N(u,y)\tilde{\tilde{\mu}}_i(du,dy)\right\|\right]$$
$$=\lim_{i\to\infty}E\left[\left\|\int g_N(u,y)\mu_i(du,dy)-\int g_N(u,y)\mu^{S_i}(du,dy)\right\|\right].$$

These equalities contradict (7.22) and, thus, prove the corollary. □

8. Proofs for section 5.

LEMMA 8.1. *Let the assumptions of Theorem* 5.1 *be satisfied. Then any admissible solution* $(y^\epsilon(\cdot),z^\epsilon(\cdot))$ *of the singularly perturbed CSDE* (5.1) *and* (5.2) *and any admissible solution* $z(\cdot)$ *of the averaged CSDE* (5.4) *satisfy the inequalities*

$$E[||y^\epsilon(t)||^4]\le L\qquad E[||z^\epsilon(t)||^4]\le L\qquad E[||z(t)||^4]\le L\quad\forall t\in[0,T];\tag{8.1}$$

$$E[||z^\epsilon(t)-z^\epsilon(\theta)||^2]\le L|t-\theta|\qquad E[||z(t)-z(\theta)||^2]\le L|t-\theta|\quad\forall t,\theta\in[0,T],\tag{8.2}$$

where L *is a positive constant.*

Proof. The proof follows a standard argument based on Lemma 4.12 in [40, p. 125] and an application of the Gronwall–Bellman lemma. □

Proof of Theorem 5.1(i). Let $u^\epsilon(t)$ be an admissible control and $(y^\epsilon(t),z^\epsilon(t))$ be the solution of the singularly perturbed CSDE (5.1) and (5.2) obtained with this control. Divide the interval $[0,T]$ by the points $t_l\overset{def}{=}l\Delta(\epsilon)$, $l=0,1,\dots,N_\epsilon$, where $\Delta(\epsilon)$ is a function of ϵ such that

$$\lim_{\epsilon\to0}\Delta(\epsilon)=0,\qquad\lim_{\epsilon\to0}\frac{\Delta(\epsilon)}{\epsilon}=\infty\tag{8.3}$$

and N_ϵ is the integer part of $\frac{T}{\Delta(\epsilon)}$. For $l=1,\dots,N_\epsilon$, define a $\mathcal{P}(U\times\bar{R}^m)$-valued random variable $\bar{\mu}_l$ by

$$\int f_i(u,y)\bar{\mu}_l(du,dy)=\frac{1}{\Delta(\epsilon)}\int_{t_{l-1}}^{t_l}f_i(u^\epsilon(t),y^\epsilon(t))dt\ =\frac{1}{S_\epsilon}\int_0^{S_\epsilon}f_i(\bar{u}(\tau),\bar{y}(\tau))d\tau,\tag{8.4}$$

where $f_i(u,y)$, $i=1,2,\dots$, are as in (2.7) and

$$(\bar{u}(\tau),\bar{y}(\tau))\overset{\text{def}}{=}(u(t_{l-1}+\epsilon\tau),y(t_{l-1}+\epsilon\tau)),\qquad S_\epsilon\overset{\text{def}}{=}\frac{\Delta(\epsilon)}{\epsilon}.$$

The equations in (8.4) imply that $\bar{\mu}_l$ is the occupational measure generated on the interval $[0,S_\epsilon]$ by the control $\bar{u}(\tau)$ and the corresponding solution $\bar{y}(\tau)$ of the associated system (2.4). Hence, by Theorem 4.2,

$$E[\rho(\bar{\mu}_l,D)]\le\bar{\nu}^{(C,\alpha)}(S_\epsilon)\overset{\text{def}}{=}\nu(\epsilon),\qquad\lim_{\epsilon\to0}\nu(\epsilon)=0.\tag{8.5}$$

For any $\mu',\mu''\in\mathcal{P}(U\times\bar{R}^m)$, let

$$\hat{\rho}(\mu',\mu'')\overset{\text{def}}{=}\rho(\mu',\mu'')+\left(\sum_{i=1}^{\infty}2^{-2i}\left|\int f_i(u,y)\mu'(du,dy)-\int f_i(u,y)\mu''(du,dy)\right|^2\right)^{\frac{1}{2}}\tag{8.6}$$

It is easy to see that $\hat{\rho}(\cdot,\cdot)$ is a metric on $\mathcal{P}(U \times \bar{R}^m)$ and that

$$\rho(\mu',\mu'') \le \hat{\rho}(\mu',\mu'') \le 2\rho(\mu',\mu'') \qquad \forall \mu',\mu'' \in \mathcal{P}(U \times \bar{R}^m). \tag{8.7}$$

The advantage of using $\hat{\rho}(\cdot,\cdot)$ instead of $\rho(\cdot,\cdot)$ is that, for any μ, the solution of the problem $\min_{\mu' \in D} \hat{\rho}(\mu,\mu')$, called the projection of μ onto D, is unique (this being easily verifiable on the basis of the inequality $\hat{\rho}(\mu,(1-\lambda)\mu'+\lambda\mu'') < (1-\lambda)\hat{\rho}(\mu,\mu') + \lambda\hat{\rho}(\mu,\mu'')$ $\forall \lambda \in (0,1)$).

Let μ_l stand for the projection of $\bar{\mu}_l$ onto D. By (8.5) and (8.7),

$$E[\hat{\rho}(\bar{\mu}_l,\mu_l)] = E[\hat{\rho}(\bar{\mu}_l,D)] \le 2\nu(\epsilon) \tag{8.8}$$

Using (8.8) and slightly extending arguments in the proof of Corollary 4.5 (to take into account the dependence on z), one can verify that, for any $N > 0$,

$$E[||\tilde{g}(\bar{\mu}_l,z) - \tilde{g}(\mu_l,z)||] \le \nu_N(\epsilon) \quad \forall\, z : ||z|| \le N, \qquad \lim_{\epsilon \to 0} \nu_N(\epsilon) = 0. \tag{8.9}$$

Also, it can be verified that the estimates (8.1) and Assumption 2 (with $\alpha = 4$) imply that

$$E[||\tilde{g}(\bar{\mu}_l,z^\epsilon(t_l))||^4] \le L_1, \quad E[||\tilde{g}(\mu_l,z^\epsilon(t_l))||^4] \le L_1, \quad L_1 = \text{const.} \tag{8.10}$$

as well as that

$$E[||\tilde{g}(\bar{\mu}_l,z^\epsilon(t_l))||\chi_N] \le \kappa(N), \quad E[||\tilde{g}(\mu_l,z^\epsilon(t_l))||\chi_N] \le \kappa(N), \quad \lim_{N \to \infty} \kappa(N) = 0, \tag{8.11}$$

where χ_N is the indicator function of the event: $||z^\epsilon(t_l)|| > N$ ($\bar{\chi}_N$ below will stand for the indicator function of $||z^\epsilon(t_l)|| \le N$). From (8.9) and (8.11) it follows that

$$E[||\tilde{g}(\bar{\mu}_l,z^\epsilon(t_l)) - \tilde{g}(\mu_l,z^\epsilon(t_l))||] \le E[||\tilde{g}(\bar{\mu}_l,z^\epsilon(t_l)) - \tilde{g}(\mu_l,z^\epsilon(t_l))||\bar{\chi}_N]$$

$$+ E[||\tilde{g}(\bar{\mu}_l,z^\epsilon(t_l))||\chi_N] + E[||\tilde{g}(\mu_l,z^\epsilon(t_l))||\chi_N] \;\le\; \nu_N(\epsilon) + 2\kappa(N),$$

which implies that there exists $\hat{\nu}(\epsilon)$, $\lim_{\epsilon \to 0} \hat{\nu}(\epsilon) = 0$, such that

$$E[||\tilde{g}(\bar{\mu}_l,z^\epsilon(t_l)) - \tilde{g}(\mu_l,z^\epsilon(t_l))||] \le \hat{\nu}(\epsilon). \tag{8.12}$$

This estimate and (8.10) imply, in turn, that

$$E[||\tilde{g}(\mu_l,z^\epsilon(t_l)) - \tilde{g}(\bar{\mu}_l,z^\epsilon(t_l))||^2]$$

$$\le \sqrt{E[||\tilde{g}(\mu_l,z^\epsilon(t_l)) - \tilde{g}(\bar{\mu}_l,z^\epsilon(t_l))||]}\sqrt{E[||\tilde{g}(\mu_l,z^\epsilon(t_l)) - \tilde{g}(\bar{\mu}_l,z^\epsilon(t_l))||^3]}$$

$$\le L_2\sqrt{\hat{\nu}(\epsilon)}, \qquad L_2 = \text{const.}$$

Thus, denoting $\bar{\nu}(\epsilon) \stackrel{\text{def}}{=} L_2\sqrt{\hat{\nu}(\epsilon)}$, one obtains

$$E[||\tilde{g}(\mu_l,z^\epsilon(t_l)) - \tilde{g}(\bar{\mu}_l,z^\epsilon(t_l))||^2] \le \bar{\nu}(\epsilon), \qquad \lim_{\epsilon \to 0} \bar{\nu}(\epsilon) = 0. \tag{8.13}$$

Now define the admissible control $\mu(t)$ of the averaged system as follows. On the intervals $[t_0,t_1)$ and $[t_{N_\epsilon},T]$, take $\mu(t) = \mu$ (an arbitrary element of D). On any other

interval $[t_l, t_{l+1}), l = 1, 2, \dots, N_\epsilon - 1$, take $\mu(t) = \mu_l$. Let $z(t)$ be the solution of the averaged system (5.4) obtained with the control $\mu(t)$. By definition, it satisfies

$$z(t) = z_0 + \int_0^t \tilde{g}(\mu(t'), z(t'))dt' + \int_0^t \sigma(z(t'))dB_2(t').$$

Subtracting this from

$$z^\epsilon(t) = z_0 + \int_0^t g(u^\epsilon(t'), y^\epsilon(t'), z^\epsilon(t'))dt' + \int_0^t \sigma(z^\epsilon(t'))dB_2(t'),$$

one can obtain that

(8.14)

$$E[||z^\epsilon(t) - z(t)||^2] \le K\left\{E\left[\left\|\int_0^t g(u^\epsilon(t'), y^\epsilon(t'), z^\epsilon(t'))dt' - \int_0^t \tilde{g}(\mu(t'), z(t'))dt'\right\|^2\right] + \int_0^t E[||z^\epsilon(t') - z(t')||]^2 dt'\right\},$$

where K is a positive constant. Let us evaluate the first term on the right-hand side of (8.14). Let k_t stand for the integer part of $\frac{t}{\Delta(\epsilon)} (k_t\Delta(\epsilon) \le t \le T)$. Then

$$\begin{aligned}
&E\left[\left\|\int_0^t g(u^\epsilon(t'), y^\epsilon(t'), z^\epsilon(t'))dt' - \int_0^t \tilde{g}(\mu(t'), z(t'))dt'\right\|^2\right] \\
&\le K_1\Bigg\{ E\Bigg[\Bigg\|\sum_{l=1}^{k_t}\int_{t_{l-1}}^{t_l}(g(u^\epsilon(t'), y^\epsilon(t'), z^\epsilon(t')) - g(u^\epsilon(t'), y^\epsilon(t'), z^\epsilon(t_l)))dt' \\
&+\sum_{l=1}^{k_t}(\tilde{g}(\bar{\mu}_l, z^\epsilon(t_l)) - \tilde{g}(\mu_l, z^\epsilon(t_l)))\Delta(\epsilon) + \sum_{l=1}^{k_t-1}(\tilde{g}(\mu_l, z^\epsilon(t_l)) - \tilde{g}(\mu_l, z(t_l)))\Delta(\epsilon) \\
&+\sum_{l=1}^{k_t-1}\int_{t_l}^{t_{l+1}}(\tilde{g}(\mu(t'), z(t_l)) - \tilde{g}(\mu(t'), z(t')))dt'\Bigg\|^2\Bigg] + \Delta(\epsilon)\Bigg\} \\
&\le K_2\Bigg\{E\Bigg[\Bigg\|\sum_{l=1}^{k_t}\int_{t_{l-1}}^{t_l}(g(u^\epsilon(t'), y^\epsilon(t'), z^\epsilon(t')) - g(u^\epsilon(t'), y^\epsilon(t'), z^\epsilon(t_l)))dt'\Bigg\|^2\Bigg] \\
&+E\Bigg[\Bigg\|\sum_{l=1}^{k_t}(\tilde{g}(\bar{\mu}_l, z^\epsilon(t_l)) - \tilde{g}(\mu_l, z^\epsilon(t_l)))\Bigg\|^2 + \Bigg\|\sum_{l=1}^{k_t-1}(\tilde{g}(\mu_l, z^\epsilon(t_l)) - \tilde{g}(\mu_l, z(t_l)))\Bigg\|^2\Bigg]\Delta^2(\epsilon) \\
&+E\Bigg[\Bigg\|\sum_{l=1}^{k_t-1}\int_{t_l}^{t_{l+1}}(\tilde{g}(\mu(t'), z(t_l)) - \tilde{g}(\mu(t'), z(t')))dt'\Bigg\|^2\Bigg] + \Delta(\epsilon)\Bigg\}, \quad K_1, K_2 = \text{const.}
\end{aligned}$$

(8.15)

Using Cauchy–Schwarz inequality (two times), one can obtain that

$$E\left[\left\|\sum_{l=1}^{k_t}\int_{t_{l-1}}^{t_l}(g(u^\epsilon(t'), y^\epsilon(t'), z^\epsilon(t')) - g(u^\epsilon(t'), y^\epsilon(t'), z^\epsilon(t_l)))dt'\right\|^2\right]$$

$$\leq\ k_t \sum_{l=1}^{k_t} E\left[\left\| \int_{t_{l-1}}^{t_l} (g(u^\epsilon(t'), y^\epsilon(t'), z^\epsilon(t')) - g(u^\epsilon(t'), y^\epsilon(t'), z^\epsilon(t_l)))dt' \right\|^2\right]$$

$$\leq\ k_t \Delta(\epsilon) \sum_{l=1}^{k_t} \int_{t_{l-1}}^{t_l} E\left[\| g(u^\epsilon(t'), y^\epsilon(t'), z^\epsilon(t')) - g(u^\epsilon(t'), y^\epsilon(t'), z^\epsilon(t_l))\|^2\right] dt'$$

$$\leq K_3 \Delta(\epsilon), \qquad K_3 = \text{const}, \tag{8.16}$$

where, to obtain the last inequality, it has been taken into account that $g(u, y, z)$ satisfies Lipschitz conditions in z and also that, by (8.2), $E[||z^\epsilon(t') - z^\epsilon(t_l)||^2] \leq L\Delta(\epsilon)\ \forall t' \in [t_{l-1}, t_l]$. Similarly, using Cauchy–Schwarz inequality and the fact that $\tilde{g}(\mu, z)$ satisfies Lipschitz conditions in z as well as that $E[||z(t') - z(t_l)||^2] \leq L\Delta(\epsilon)\ \forall t' \in [t_l, t_{l+1}]$, one can obtain that

$$E\left[\left\| \sum_{l=1}^{k_t - 1} \int_{t_l}^{t_{l+1}} (\tilde{g}(\mu(t'), z(t_l)) - \tilde{g}(\mu(t'), z(t')))dt' \right\|^2\right] \leq\ K_4 \Delta(\epsilon),\ \ K_4 = \text{const}. \tag{8.17}$$

Also, by (8.13),

$$E\left[\left\| \sum_{l=1}^{k_t} (\tilde{g}(\bar{\mu}_l, z^\epsilon(t_l)) - \tilde{g}(\mu_l, z^\epsilon(t_l))) \right\|^2\right] \Delta^2(\epsilon)$$

$$\leq k_t \sum_{l=1}^{k_t} E[||\tilde{g}(\bar{\mu}_l, z^\epsilon(t_l)) - \tilde{g}(\mu_l, z^\epsilon(t_l))||^2] \Delta^2(\epsilon)\ \leq\ K_5 \bar{\nu}(\epsilon), \quad K_5 = \text{const}, \tag{8.18}$$

and, by (8.1) and (8.2),

$$E\left[\left\| \sum_{l=1}^{k_t - 1} (\tilde{g}(\mu_l, z^\epsilon(t_l)) - \tilde{g}(\mu_l, z(t_l))) \right\|^2\right] \Delta^2(\epsilon)$$

$$\leq k_t \sum_{l=1}^{k_t - 1} E\left[\|\tilde{g}(\mu_l, z^\epsilon(t_l)) - \tilde{g}(\mu_l, z(t_l))\|^2\right] \Delta^2(\epsilon) \leq K_6 \sum_{l=1}^{k_t - 1} E[||z^\epsilon(t_l) - z(t_l)||^2] \Delta(\epsilon)$$

$$\leq K_7 \left(\int_0^t E[||z^\epsilon(t') - z(t')||^2] dt' + \Delta^{\frac{1}{2}}(\epsilon) \right), \qquad K_6, K_7 = \text{const}. \tag{8.19}$$

Substitution of (8.16)–(8.19) into (8.15) leads to

$$E\left[\left\| \int_0^t g(u^\epsilon(t'), y^\epsilon(t'), z^\epsilon(t'))dt' - \int_0^t \tilde{g}(\mu(t'), z(t'))dt' \right\|^2\right]$$

$$\leq\ K_8 \left(\int_0^t E[||z^\epsilon(t') - z(t')||^2] dt' + \bar{\nu}(\epsilon) + \Delta^{\frac{1}{2}}(\epsilon) \right), \qquad K_8 = \text{const}. \tag{8.20}$$

The latter, in turn, being substituted into (8.14) implies (with the help of a Gronwall–Bellman lemma) the validity of (5.7) with $\tilde{\nu}(\epsilon) = K_9(\bar{\nu}(\epsilon) + \Delta^{\frac{1}{2}}(\epsilon)), K_9 = \text{const}$. The validity of (5.8) follows from the Lipschitz continuity of $G(z)$. □

Proof of Theorem 5.1(ii) *(Outline).* Let $S_\epsilon = \frac{\Delta(\epsilon)}{\epsilon}$ (as in the proof of Theorem 5.1(i)) and let

$$\nu(\epsilon) \stackrel{\text{def}}{=} \tilde{\nu}^{(C,\alpha)}(S_\epsilon), \tag{8.21}$$

where $\tilde{\nu}^{(C,\alpha)}(\cdot)$ is the function from the estimate (4.8). Note that $\lim_{\epsilon\to 0}\nu(\epsilon) = 0$. Let J^ϵ be the integer part of $\nu^{-\frac{1}{2}}(\epsilon)$, which implies, in particular, that

$$\lim_{\epsilon\to 0} J^\epsilon = \infty, \qquad \lim_{\epsilon\to 0}(\nu(\epsilon)J^\epsilon) = 0. \tag{8.22}$$

Using the fact that D is compact, one can show that, for any $\epsilon > 0$, there exists a finite subset $D^\epsilon \stackrel{\text{def}}{=} \{\Upsilon^\epsilon_1, \dots, \Upsilon^\epsilon_{J^\epsilon}\}$ of D such that $\rho_H(D^\epsilon, D) \le \delta(\epsilon)$, where $\delta(\epsilon)$ is some function tending to zero as ϵ tends to zero.

It can be verified (by standard applications of a Gronwall–Bellman lemma) that, given a solution $z'(t)$ of the averaged system (5.4) obtained with an arbitrary admissible control $\mu'(t)$, there exists a piecewise constant admissible control $\mu(t)$:

$$\mu(t) = \mu_l \in D^\epsilon \quad \forall t \in [t_{l-1}, t_l), \qquad l = 1, \dots, N_\epsilon \tag{8.23}$$

such that the solution $z(t)$ of the averaged system (5.4), obtained with the use of this control, satisfies the inequality

$$\max_{t\in[0,T]} E[||z'(t) - z(t)||^2] \le \kappa(\epsilon), \qquad \lim_{\epsilon\to 0}\kappa(\epsilon) = 0.$$

Let us show that corresponding to any solution $z(t)$ of the averaged system (5.4) obtained with a control (8.23), there exists an admissible control $u^\epsilon(t)$ the use of which in the singularly perturbed CSDE (5.1) and (5.2) leads to the solution $(y^\epsilon(t), z^\epsilon(t))$ satisfying (5.7).

Take $u^\epsilon(t) = u$ (an arbitrary element of U) on the intervals $[0, t_1)$ and $[t_{N_\epsilon}, T]$ and denote by $(y^\epsilon(t), z^\epsilon(t))$ the solution of (5.1) and (5.2) on the interval $[0, t_1]$ obtained with this control.

From Corollary 4.3 (see (4.8) and the notation (8.21)) it follows that there exist random variables $\bar{\Upsilon}^\epsilon_j \in \mathcal{M}(S_\epsilon, y^\epsilon(t_1))$ such that

$$\frac{E[\rho^2(\bar{\Upsilon}^\epsilon_j, \Upsilon^\epsilon_j)]}{2} \le E[\rho(\bar{\Upsilon}^\epsilon_j, \Upsilon^\epsilon_j)] \le \nu(\epsilon), \quad j = 1, \dots, J^\epsilon, \tag{8.24}$$

where the left inequality is obtained by taking into account that $\frac{\rho(\cdot,\cdot)}{2} \le 1$ (see (2.1)) and, hence, $\frac{\rho^2(\cdot,\cdot)}{2} \le \rho(\cdot,\cdot)$. Define $\bar{\mu}_1$ by

$$\bar{\mu}_1 \stackrel{\text{def}}{=} \sum_{j=1}^{J^\epsilon} \bar{\Upsilon}^\epsilon_j \chi(\mu_1 = \Upsilon^\epsilon_j), \tag{8.25}$$

where $\chi(A)$ is the indicator function of the "event A." By (8.24),

$$E[\rho(\bar{\mu}_1, \mu_1)] = \sum_{j=1}^{J^\epsilon} E[\rho(\bar{\Upsilon}^\epsilon_j, \Upsilon^\epsilon_j)\chi(\mu_1 = \Upsilon^\epsilon_j)]$$

$$\le \sum_{j=1}^{J^\epsilon} \sqrt{E[\rho^2(\bar{\Upsilon}^\epsilon_j, \Upsilon^\epsilon_j)]}\sqrt{E[\chi(\mu_1 = \Upsilon^\epsilon_j)]} \le \sqrt{2\nu(\epsilon)}\sum_{j=1}^{J^\epsilon}\sqrt{E[\chi(\mu_1 = \Upsilon^\epsilon_j)]}$$

$$\leq \sqrt{2\nu(\epsilon)}\sqrt{J^\epsilon}\sqrt{\sum_{j=1}^{J^\epsilon} E[\chi(\mu_1 = \Upsilon_j^\epsilon)]} = \sqrt{2\nu(\epsilon)}\sqrt{J^\epsilon} \stackrel{\text{def}}{=} \nu^*(\epsilon). \tag{8.26}$$

Note that, as follows from (8.22), $\lim_{\epsilon\to 0}\nu^*(\epsilon) = 0$.

The fact that $\bar\Upsilon_j^\epsilon$ is an element of $\mathcal{M}(S_\epsilon, y^\epsilon(t_1))$ implies that there exists an admissible control $\bar u_j^\epsilon(\tau)$ and the corresponding solution $\bar y_j^\epsilon(\tau)$ of the associated system with $\bar y_j^\epsilon(0) = y^\epsilon(t_1)$ such that the occupational measure generated by this pair on the interval $[0, S_\epsilon]$ coincides with $\bar\Upsilon_j$. That is,

$$\frac{1}{S_\epsilon}\int_0^{S_\epsilon} f_i(\bar u_j^\epsilon(\tau), \bar y_j^\epsilon(\tau))d\tau = \int f_i(u,y)\bar\Upsilon_j(du,dy) \quad \forall i = 1,2,\dots. \tag{8.27}$$

Now take

$$u^\epsilon(t) \stackrel{\text{def}}{=} \sum_{j=1}^{J^\epsilon} \bar u_j^\epsilon\left(\frac{t-t_1}{\epsilon}\right)\chi(\mu_1 = \Upsilon_j^\epsilon) \quad \forall t \in [t_1, t_2)$$

and, using this control, extend the solution $(y^\epsilon(t), z^\epsilon(t))$ of the CSDE (5.1) and (5.2) to the interval $[t_1, t_2]$. By construction, $y^\epsilon(t) = \sum_{j=1}^{J^\epsilon}\bar y_j^\epsilon(\frac{t-t_1}{\epsilon})\chi(\mu_1 = \Upsilon_j^\epsilon)$ and also (see (8.25) and (8.27))

$$\frac{1}{\Delta(\epsilon)}\int_{t_1}^{t_2} f_i(u^\epsilon(t), y^\epsilon(t))dt = \sum_{j=1}^{J^\epsilon}\left(\frac{1}{S_\epsilon}\int_0^{S_\epsilon} f_i(\bar u_j^\epsilon(\tau), \bar y_j^\epsilon(\tau))d\tau\right)\chi(\mu_1 = \Upsilon_j^\epsilon)$$

$$= \sum_{j=1}^{J^\epsilon}\left(\int f_i(u,y)\bar\Upsilon_j^\epsilon(du,dy)\right)\chi(\mu_1 = \Upsilon_j^\epsilon) = \int f_i(u,y)\bar\mu_1(du,dy)\ \forall i = 1,2,\dots.$$

Continuing in a similar fashion, one can define an admissible control $u^\epsilon(t)$ and the corresponding solution $(y^\epsilon(t), z^\epsilon(t))$ of the CSDE (5.1) and (5.2) such that, on any interval $[t_l, t_{l+1}), l = 1, \dots, N_\epsilon - 1$,

$$\frac{1}{\Delta(\epsilon)}\int_{t_l}^{t_{l+1}} f_i(u^\epsilon(t), y^\epsilon(t))dt = \int f_i(u,y)\bar\mu_l(du,dy) \quad \forall i = 1,2,\dots, \tag{8.28}$$

where $\bar\mu_l$ satisfy the inequalities

$$E[\rho(\bar\mu_l, \mu_l)] \leq \nu^*(\epsilon) \qquad \forall l = 1, \dots, N_\epsilon - 1, \tag{8.29}$$

with $\nu^*(\epsilon)$ being as in (8.26).

Using arguments similar to the proof of Corollary 4.5, one can verify that (8.29) implies the validity of (8.9) which, in turn, implies the validity of (8.12) and (8.13). The latter leads to the estimate similar to (8.20), which, being substituted into (8.14), leads to (5.7). Due to the Lipschitz continuity of $G(z)$, one can easily derive now that $\limsup_{\epsilon\to 0} G_\epsilon^* \leq G_{av}^*$, which, along with (5.8), implies (5.9). □

Acknowledgment. The authors express their gratitude to Prof. H. Kushner for useful discussions at early stages of this work.

REFERENCES

[1] E. Altman and V. Gaitsgory, *Asymptotic optimization of a nonlinear hybrid system governed by a Markov decision process*, SIAM J. Control Optim., 35 (1997), pp. 2070–2085.

[2] O. Alvarez and M. Bardi, *Viscosity solutions methods for singular perturbations in deterministic and stochastic control*, SIAM J. Control Optim., 40 (2001/2002), pp. 1159–1188.

[3] O. Alvarez and M. Bardi, *Singular perturbations of nonlinear degenerate parabolic PDEs: A general convergence result*, Arch. Ration. Mech. Anal., 170 (2003), pp. 17–61.

[4] Z. Artstein, *Relaxation of singularly perturbed control systems*, Proceedings of CDC-2002, Control and Decision Conference, Las Vegas, 2002.

[5] Z. Artstein, *An occupational measure solution to a singularly perturbed optimal control problem*, Control Cybernet., 31 (2002), pp. 623–642.

[6] Z. Artstein and V. Gaitsgory, *Tracking fast trajectories along a slow dynamics: A singular perturbations approach*, SIAM J. Control Optim., 35 (1997), pp. 1487–1507.

[7] Z. Artstein and V. Gaitsgory, *Convergence to convex compact sets in infinite dimensions*, J. Math. Anal. Appl., 284 (2003), pp. 471–480.

[8] Z. Artstein and A. Leizarowitz, *Singularly perturbed control systems with one-dimensional fast dynamics*, SIAM J. Control Optim., 41 (2002), pp. 641–658.

[9] K. E. Avrachenkov, J. Filar, and M. Haviv, *Singular perturbations of Markov chains and decision processes*, in Handbook of Markov Decision Processes. Methods and Applications, Kluwer Academic Publishers, Boston, 2002, pp. 113–150.

[10] F. Bagagiolo and M. Bardi, *Singular perturbation of a finite horizon problem with state-space constraints*, SIAM J. Control Optim., 36 (1998), pp. 2040–2060.

[11] G. K. Basak, V. S. Borkar, and M. K. Ghosh, *Ergodic control of degenerate diffusions*, Stochastic Anal. Appl., 15 (1997), pp. 1–17.

[12] A. Bensoussan, *Perturbation Methods in Optimal Control*, John Wiley, Chichester, 1988.

[13] A. G. Bhatt and V. S. Borkar, *Occupation measures for controlled Markov processes: Characterization and optimality*, Ann. Probab., 24 (1996), pp. 1531–1562.

[14] V. S. Borkar, *Optimal control of diffusion processes*, Pitman Research Notes in Mathematics Series 203, Longman Scientific and Technical, Harlow, UK, 1989.

[15] V. S. Borkar, *Probability Theory: An Advanced Course*, Springer-Verlag, New York, 1995.

[16] V. S. Borkar, *Stability of annealing schemes and related processes*, Systems Control Lett., 41 (2000), pp. 325–331.

[17] F. Colonius and R. Fabri, *Controllability for systems with slowly varying parameters*, ESAIM Control Optim. Calc. Var., 9 (2003), pp. 207–216.

[18] T. D. Donchev and I. Slavov, *Averaging method for one-sided Lipschitz differential inclusions with generalized solutions*, SIAM J. Control Optim., 37 (1999), pp. 1600–1613.

[19] T. D. Donchev and A. L. Dontchev, *Singular perturbations in infinite-dimensional control systems*, SIAM J. Control Optim., 42 (2003), pp. 1795–1812.

[20] A. L. Dontchev, T. D. Donchev, and I. Slavov, *A Tikhonov-type theorem for singularly perturbed differential inclusions*, Nonlinear Anal., 26 (1996), pp. 1547–1554.

[21] J. A. Filar, V. Gaitsgory, and A. Haurie, *Control of singularly perturbed hybrid stochastic systems*, IEEE Trans. Automat. Control, 46 (2001), pp. 179–190.

[22] O. P. Filatov and M. M. Hapaev, *Averaging of Systems of Differential Inclusions*, Moscow University Publishing House, Moscow, 1998 (in Russian).

[23] G. B. Folland, *Real Analysis, Modern Techniques and their Applications*, John Wiley, New York, 1984.

[24] V. Gaitsgory, *Control of Systems with Slow and Fast Motions*, Nauka, Moscow, 1991 (in Russian).

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

[26] V. Gaitsgory, *On a representation of the limit occupational measures set of a control system with applications to singularly perturbed control systems*, SIAM J. Control Optim., 43 (2004), pp. 325–340.

[27] V. Gaitsgory and A. Leizarowitz, *Limit occupational measures set for a control system and averaging of singularly perturbed control systems*, J. Math. Anal. Appl., 233 (1999), pp. 461–475.

[28] V. Gaitsgory and M. T. Nguyen, *Multiscale singularly perturbed control systems: Limit occupational measures sets and averaging*, SIAM J. Control Optim., 41 (2002), pp. 954–974.

[29] V. Gaitsgory and S. Rossomakhine, *Linear programming approach to deterministic long run average problems of optimal control*, SIAM J. Control Optim., accepted.

[30] G. Grammel, *Averaging of singularly perturbed systems*, Nonlinear Anal., 28 (1997), pp. 1851–1865.
[31] G. Grammel, *On nonlinear control systems with multiple time scales*, J. Dynam. Control Systems, 10 (2004), pp. 11–28.
[32] Y. Kabanov and S. Pergamenshchikov, *On convergence of attainability sets for controlled two-scale stochastic linear systems*, SIAM J. Control Optim., 35 (1997), pp. 134–159.
[33] Y. Kabanov and S. Pergamenshchikov, *Two-Scale Stochastic Systems, Asymptotic Analysis and Control*, Springer-Verlag, Berlin, 2003.
[34] R. Z. Khasminskii, *Stochastic Stability of Differential Equations*, Sijthoff and Noordhoff, Alphen aan den Rijn, Germantown, MD, 1980.
[35] R. Z. Khasminskii and G. Yin, *On averaging principles: An asymptotic expansion approach*, SIAM J. Math. Anal. 35, (2004), pp. 1534–1560.
[36] P. V. Kokotovic, *Applications of singular perturbation techniques to control problems*, SIAM Rev., 26 (1984), pp. 501–550.
[37] P. V. Kokotovic, H. K. Khalil, and J. O'Reilly, *Singular Perturbation Methods in Control: Analysis and Design*, Academic Press, London, 1986.
[38] H. J. Kushner, *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*, Birkhauser, Boston, 1990.
[39] A. Leizarowitz, *Order reduction is invalid for singularly perturbed control problems with a vector fast variable*, Math. Control Signals Systems, 15 (2002), pp. 101–119.
[40] R. S. Liptser and A. N. Shiryayev, *Statistics of Random Processes* I: *General Theory*, Springer-Verlag, New York, 1977.
[41] S. D. Naidu, *Singular perturbations and time scales in control theory and applications: An overview*, Dyn. Contin. Discrete Impuls. Syst. Ser. B Appl. Algorithms, 9 (2002), pp. 233–278.
[42] R. E. O'Malley, Jr., *Singular perturbations and optimal control*, in Mathematical Control Theory, W. A. Copel, ed., Lecture Notes in Math. 680, Springer-Verlag, Berlin, 1978.
[43] A. A. Pervozvanskii and V. Gaitsgory, *Theory of Suboptimal Decisions*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1988.
[44] V. A. Plotnikov, A. V. Plotnikov, and A. N. Vituk, *Differential Equations with Multivalued Right-Hand Sides: Asymptotic Methods*, AstroPrint, Odessa, Ukraine, 1999 (in Russian).
[45] M. Quincampoix and H. Zhang, *Singular perturbations in non-linear optimal control systems*, Differential Integral Equations, 8 (1995), pp. 931–944.
[46] M. Quincampoix and F. Watbled, *Averaging method for discontinuous Mayer's problem of singularly perturbed control systems*, Nonlinear Anal., 54 (2003), pp. 819–837.
[47] R. T. Rockafellar, *Convex Analysis*, Princeton Mathematical Series 5, Princeton University Press, Princeton, NJ, 1970.
[48] S. P. Sethi and Q. Zhang, *Hierarchical Decision Making in Stochastic Manufacturing Systems*, Birkhauser, Boston, 1994
[49] R. H. Stockbridge, *Time-average control of a martingale problem: Existence of a stationary solution*, Ann. Probab., 18 (1990), pp. 190–205.
[50] V. Veliov, *A generalization of Tikhonov theorem for singularly perturbed differential inclusions*, J. Dynam. Control Systems, 3 (1997), pp. 291–319.
[51] A. Vigodner, *Limits of singularly perturbed control problems with statistical dynamics of fast motions*, SIAM J. Control Optim., 35 (1997), pp. 1–28.
[52] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
[53] G. G. Yin and Q. Zhang, *Continuous-Time Markov Chains and Applications. A Singular Perturbation Approach*, Springer-Verlag, New York, 1998.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 4, pp. 1474–1528

STRUCTURED NONCOMMUTATIVE MULTIDIMENSIONAL LINEAR SYSTEMS*

JOSEPH A. BALL†, GILBERT GROENEWALD‡, AND TANIT MALAKORN§

Abstract. We introduce a class of multidimensional linear systems with evolution along a free semigroup. The transfer function for such a system is a formal power series in noncommuting indeterminates. Standard system-theoretic properties (the operations of cascade/parallel connection and inversion, controllability, observability, Kalman decomposition, state-space similarity theorem, minimal state-space realizations, Hankel operators, realization theory) are developed for this class of systems. We also draw out the connections with the much earlier studied theory of rational and recognizable formal power series. Applications include linear-fractional models for classical discrete-time systems with structured, time-varying uncertainty, dimensionless formulas in robust control, multiscale systems and automata theory, and the theory of formal languages.

Key words. multidimensional linear systems, free semigroup, controllability, observability, minimality, realization, formal power series, noncommuting indeterminates

AMS subject classifications. 93B10, 13F25, 47A56, 93B28

DOI. 10.1137/S0363012904443750

1. Introduction.

This paper considers extensions of standard system-theoretic ideas for classical, discrete-time, input/state/output linear systems to the case of certain types of generalized i/s/o systems having evolution along a free semigroup (in place of evolution along the nonnegative integers, as in the classical case). One can introduce formal frequency-domain techniques and arrive at a transfer function for such a system which is a formal power series in noncommuting variables; such objects have occurred in the context of the theory of formal languages and automata theory as well as in connection with realization theory for bilinear systems in the work of Schützenberger and Fliess (see [37, 38, 39, 20, 21, 22, 23] and the book [15] for a good survey).

We first review those aspects of the classical theory which we here generalize to the setting of systems evolving on a free semigroup; this material can be found in many books on linear system and control theory (see, e.g., [32, 16]). By a classical, discrete-time, i/s/o linear system (referred to here simply as a *linear system* for short) we mean a system Σ of linear equations of the form

$$\begin{aligned} x(n+1) &= Ax(n) + Bu(n), \\ y(n) &= Cx(n) + Du(n) \end{aligned} \tag{1.1}$$

*Received by the editors May 11, 2004; accepted for publication (in revised form) February 26, 2005; published electronically November 4, 2005. Portions of this paper are based on the dissertation [34] of the third author, written under the direction of the first author.
http://www.siam.org/journals/sicon/44-4/44375.html

†Department of Mathematics, Virginia Tech, Blacksburg, VA 24061-0123 (ball@math.vt.edu). This author was supported by the U.S. National Science Foundation under grant DMS-9987636.

‡Department of Mathematics, North West University, Potchefstroom 2520, South Africa (wskgjg@puknet.puk.ac.za). This author is supported by the National Research Foundation of South Africa under grant 2053733.

§Department of Electrical and Computer Engineering, Naresuan University, Phitsanulok 65000, Thailand (tanitm@nu.ac.th). This author was supported by a grant from Naresuan University, Thailand.

(where n takes values in the integers $\mathbb{Z}$), with $x(n)$ taking values in the *state-space* $\mathcal{H}$, $u(n)$ taking values in the *input-space* $\mathcal{U}$, and $y(n)$ taking values in the *output-space* $\mathcal{Y}$, where here we assume that $\mathcal{H}$, $\mathcal{U}$, and $\mathcal{Y}$ are *finite-dimensional* linear spaces over the field of complex numbers $\mathbb{C}$. It is convenient to identify the operator

$$U = \begin{bmatrix} A & B \\ C & D \end{bmatrix} : \begin{bmatrix} \mathcal{H} \\ \mathcal{U} \end{bmatrix} \to \begin{bmatrix} \mathcal{H} \\ \mathcal{Y} \end{bmatrix}$$

as the *connection matrix* or *colligation* of the system Σ. Given such a system Σ, if one initializes the state $x(0)$ at time 0 and feeds in an input string $\{u(n)\}_{n\in\mathbb{Z}_+}$, one can use the system equations (1.1) to uniquely determine the state $x(n)$ for all future times $n > 0$ and the output $y(n)$ for the present and all future times $n \geq 0$; the result is

$$\begin{aligned} x(n) &= A^n x(0) + \sum_{k=0}^{n-1} A^{n-1-k} B u(k), \\ y(n) &= CA^n x(0) + \sum_{k=0}^{n-1} CA^{n-1-k} B u(k) + Du(n). \end{aligned} \tag{1.2}$$

Application of the *Z-transform*

$$\{x(n)\}_{n\in\mathbb{Z}_+} \mapsto \sum_{n=0}^{\infty} x(n) z^n$$

to the system equations (1.1) converts the expressions (1.2) to the so-called frequency-domain formulas

$$\begin{aligned} \widehat{x}(z) &= (I - zA)^{-1} x(0) + z(I - zA)^{-1} B \widehat{u}(z), \\ \widehat{y}(z) &= C(I - zA)^{-1} x(0) + T_\Sigma(z) \widehat{u}(z), \end{aligned} \tag{1.3}$$

where

$$T_\Sigma(z) = D + zC(I - zA)^{-1} B \tag{1.4}$$

is a rational $\mathcal{L}(\mathcal{U}, \mathcal{Y})$-valued function analytic at the origin called the *transfer function* of the system Σ. Standard system-theoretic ideas in this context are *controllability* and *observability*. The system is said to be *controllable* if for every h in the state-space $\mathcal{H}$ there is an $N < 0$ and an input string $\{u(n)\}_{n=N,N+1,\ldots,-1}$ so that h is achievable as $h = x(0)$ if the system is run with initialization $x(N) = 0$ and input string $\{u(n)\}_{n=N,N+1,\ldots,-1}$. It works out that the system Σ is controllable if and only if the *controllability operator*

$$\mathcal{C} = \begin{bmatrix} B & AB & A^2B & \cdots \end{bmatrix} : \ell_{\text{fin}}(\mathbb{Z}_-, \mathcal{U}) \to \mathcal{H}$$

has full rank (equal to $\dim \mathcal{H}$).[1] Here $\ell_{\text{fin}}(\mathbb{Z}_-, \mathcal{U})$ denotes the linear space of all $\mathcal{U}$-valued summable sequences on $\mathbb{Z}_-$ with finite support. Similarly, the system is said to be *observable* if the state-vector $h \in \mathcal{H}$ can be uniquely recovered from the output string $\{y(n)\}_{n\geq 0}$ generated by running the system with initial condition $x(0) = h$ and

[1] By the Cayley–Hamilton theorem, it suffices to consider only the finite matrix $\mathcal{C}_n = \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}$, where $n = \dim \mathcal{H}$ in place of $\mathcal{C}$.

zero input string $u(n) = 0$ for $n \geq 0$; this in turn is equivalent to the *observability operator*

$$\mathcal{O} = \mathrm{col}_{n\geq 0}[CA^n] \colon \mathcal{H} \to \ell(\mathbb{Z}_+, \mathcal{Y})$$

being injective.[2] Here and in what follows, we often use the following notation. If $\mathcal{H}_i$, $\widetilde{\mathcal{H}}_j$, $\mathcal{U}$, and $\mathcal{Y}$ are finite-dimensional linear spaces (for each index i in an index set S and index j in an index set $\widetilde{S}$), and if we are given linear operators $B_j \colon \mathcal{U} \to \widetilde{\mathcal{H}}_j$ and $C_i \colon \mathcal{H}_i \to \mathcal{Y}$, then $\mathrm{col}_{j\in\widetilde{S}}\, B_j$ denotes the block-operator *column* matrix representing a linear operator from $\mathcal{U}$ into $\oplus_{j\in\widetilde{S}}\widetilde{\mathcal{H}}_j$ given by

$$\mathrm{col}_{j\in\widetilde{S}}\, B_j \colon u \to \oplus_{j\in\widetilde{S}} B_j u, \tag{1.5}$$

while $\mathrm{row}_{i\in S}\, C_i$ denotes the block-operator *row* matrix representing a linear operator from $\oplus_{i\in S}\mathcal{H}_i$ into $\mathcal{Y}$ given by

$$\mathrm{row}_{i\in S}\, C_i \colon \oplus_{i\in S}\, h_i \mapsto \sum_{i\in S} C_i h_i. \tag{1.6}$$

We say that the system $\Sigma = (U \colon (\mathcal{H}\oplus\mathcal{U}) \to (\mathcal{H}\oplus\mathcal{Y}))$ is a *realization* of the $\mathcal{L}(\mathcal{U},\mathcal{Y})$-valued function $T(z)$ if $T(z) = T_\Sigma(z)$. There is a theory of *minimality* of a realization Σ of a given matrix-valued function $T(z)$: we say that the realization $\Sigma = (U \colon (\mathcal{H}\oplus\mathcal{U}) \to (\mathcal{H}\oplus\mathcal{Y}))$ of $T(z)$ is a *minimal realization* if, whenever $\Sigma' = (U' \colon (\mathcal{H}'\oplus\mathcal{U}) \to (\mathcal{H}'\oplus\mathcal{Y}))$ is another realization of $T(z)$, it is the case that $\dim\mathcal{H} \leq \dim\mathcal{H}'$. It is well known that Σ is a minimal realization of $T_\Sigma(z)$ if and only if Σ is both controllable and observable; moreover, given a realization $\Sigma' = (U' \colon (\mathcal{H}' \oplus \mathcal{U}) \to (\mathcal{H}' \oplus \mathcal{Y}))$ of $T(z)$ which is not controllable and/or not observable, the *Kalman decomposition* of the system leads to a procedure for cutting down the system to a controllable and observable (and therefore minimal) realization $\Sigma'_{c/o} = (U_{c/o} \colon \mathcal{H}'_{c/o} \oplus \mathcal{U} \to \mathcal{H}'_{c/o} \oplus \mathcal{Y})$ for $T(z)$ ($T_{\Sigma'_{c/o}}(z) = T_\Sigma(z) = T(z)$). Moreover, the *Hankel operator* $\mathbb{H} = \mathcal{O}\cdot\mathcal{C} \colon \ell_{\mathrm{fin}}(\mathbb{Z}_-,\mathcal{U}) \to \ell(\mathbb{Z}_+,\mathcal{Y})$, the map of a past input signal to the future output signal generated by the system (under the assumption that the state is initialized to be zero sufficiently far in the past and if the input string is taken to be zero on the present and future), plays a prominent role in realization theory, since $\mathbb{H} = \mathbb{H}^T$ is also completely determined by the Taylor coefficients of the transfer function $T(z)$ of Σ. Indeed, a given $\mathcal{L}(\mathcal{U},\mathcal{Y})$-valued function $T(z)$ analytic at the origin can be realized as the transfer function $T(z) = T_\Sigma(z)$ for some finite-dimensional system Σ (1.1) if and only if the Hankel operator $\mathbb{H}^T$ constructed from $T(z)$ has finite rank; in this case there is a canonical construction (the *shift realization*) of a minimal realization $\Sigma_{\mathbb{H}^T} = (U_{\mathbb{H}^T} \colon \mathcal{H}_{\mathbb{H}^T} \oplus \mathcal{U} \to \mathcal{H}_{\mathbb{H}^T} \oplus \mathcal{Y})$ for $T(z)$ with $\dim\mathcal{H}_{\mathbb{H}^T} = \mathrm{rank}\,\mathbb{H}^T$.

The purpose of this paper is to extend these ideas to various classes of systems with evolution along a free semigroup rather than along $\mathbb{Z}_+$ or $\mathbb{Z}$. We consider three main classes of such systems, which we refer to as (1) *noncommutative Fornasini–Marchesini systems*, (2) *noncommutative Givone–Roesser systems*, and (3) *noncommutative full-structured systems*. In all these examples, application of a formal Z-transform to the system equations, under the assumption that the state-vector is initialized to be zero, gives rise to the input-output map for the system being given by

[2] Again by the Cayley–Hamilton theorem, for the present classical case one can replace $\mathcal{O}$ by the finite matrix $\mathcal{O}_n = \mathrm{col}_{j=0,1,\ldots,n-1}[CA^j]$, where n is the dimension of the state-space $\mathcal{H}$.

multiplication by a formal power series in noncommuting indeterminates (the *transfer function* of the system) of the form

$$T_\Sigma(z) = D + C(I - Z(z)A)^{-1}Z(z)B, \tag{1.7}$$

where $Z(z)$ is a linear pencil in noncommuting indeterminates $z = (z_1, \dots, z_d)$. The particular form of the linear pencil $Z(z)$ is determined by the particular form of the state equations. For the reader's convenience, section 2 states the main results in explicit, concrete form for these particular classes of examples. In section 3 the added formalism is introduced to describe a general structured noncommutative multidimensional linear system (SNMLS; see Definition 3.7 below). In section 4 we show that such standard system-theoretic operations as cascade connection, parallel connection, and system inversion can be carried out in this context. With the formalism from section 3 in hand, unified proofs are given of the results on controllability, observability, Kalman decomposition, state-space similarity, minimality of realizations, Hankel operators, and construction of minimal realizations in sections 5, 6, 7, 8, 9, 10, and 11, respectively. The final section 12 makes connections of our framework with the theory of recognizable formal power series presented in [15], developed in the work of Schützenberger [37, 38, 39] and Fliess [20, 21].

In applications it is sometimes convenient to view the indeterminates as noncommuting variables, and a formal power series $T(z) = \sum_{w \in \mathcal{F}_d} T_w z^w$ (where $\mathcal{F}_d$ is the set of all words in the d letters $1, 2, \dots, d$ and where $z^w = z_{i_N} \cdots z_{i_1}$ if $w = i_N \cdots i_1$) as a function $\delta \mapsto T(\delta) = \sum_{w \in \mathcal{F}_d} T_w \otimes \delta^w$ defined on some domain of noncommuting operator-tuples $\delta = (\delta_1, \dots, \delta_d)$ (where $\delta^w = \delta_{i_N} \cdots \delta_{i_1}$, multiplication here given by operator composition); this calculus of operator-substitution is important for several of the applications listed below.

Now we mention several areas for applications of the results of this paper.

1. *Robust control theory.* Formal power series and their realizations appear prominently in the theory of robust control of classical 1-D (one-dimensional) systems subject to structured possibly time-varying uncertainty (see [33, 13, 12, 10, 11]). A commonly used model for structured uncertainty in a classical linear, finite-dimensional, feedback-control system is a so-called linear-fractional model, whereby the uncertainty is assumed to have a certain block structure which then enters the nominal plant through a feedback loop. In the case where one considers time-varying uncertainty, the time-varying input-output operator for the disturbed plant can be identified with the evaluation of the transfer function $T_\Sigma(z)$ at $z = \delta$, where $\delta = (\delta_1, \dots, \delta_d)$ is a d-tuple of time-varying operators on ℓ^2 parametrizing the time-varying structured uncertainty. Questions concerning minimality, realizability, and reduction which we explore here have direct relevance for this application. In a companion paper [5], we impose an energy balance law on an SNMLS to define the notion of a *conservative* SNMLS and obtain a realization theorem for this class of noncommutative systems; such conservative (or more generally dissipative) SNMLSs are directly relevant to the robust H^∞-control problems discussed in [33]. In the followup paper [6], we make more explicit the connections of this paper and [5] with linear-fractional models for structured uncertainty and μ-analysis in the presence of structured time-varying uncertainty. Conservative SNMLSs of noncommutative Fornasini–Marchesini type appear also in [7] and [8] in connection with other kinds of problems from multivariable operator theory. Recent closely related work of Alpay and Kalyuzhnyĭ-Verbovetzkiĭ [1] uses the state-space similarity theorem for noncommutative Givone–Roesser systems to develop a realization theory for noncommutative rational J-unitary formal

power series, including connections with noncommutative formal reproducing kernel Pontryagin spaces.

2. *Dimensionless linear matrix inequalities.* As pointed out in [28, 30], many formulas occurring in engineering involving matrix quantities have the same form independent of the size of the matrices. This motivates the study of rational functions in noncommuting variables and the study of noncommutative positivity domains associated with such rational expressions. Realizations such as (1.7) are exactly what is needed to convert (numerically unmanageable) rational matrix inequalities into (highly manageable) linear matrix inequalities (see [30]). Here one substitutes d-tuples of symmetric matrices of variable common size for the indeterminates in the noncommutative rational expression.

3. *Wavelet analysis/multiscale systems.* There have been some attempts in the literature (see [14, 2]) to attach a system evolution to multiresolution structure and multiscale modeling. We expect the setting and results of this paper to have some connections with the work in [14, 2], but details remain to be worked out.

4. *Automata and the theory of formal languages.* It has been known for some time (see [15] and the references there) that formal power series in noncommuting variables, particularly, recognizable and rational formal power series (see section 12 below), arise naturally in connection with the theory of automata and formal languages. In this context, the coefficients of the formal power series may come from a semiring (a ring without subtraction such as the nonnegative integers or the nonnegative rational numbers) rather than operators between two Hilbert spaces, and the free semigroup may be only a monoid. Roughly, a formal power series is said to be recognizable if the support set of its coefficients is recognizable. A subset of a free semigroup (or, more generally, of a monoid) is said to be *recognizable*, in turn, if it can be identified with the set of successful paths (from an initial state to a final state) generated by a finite automaton. Recognizability of a formal power series turns out to be equivalent to existence of a certain type of realization (see section 12 below). Many of the familiar results (e.g., realization through a Hankel-matrix construction, equivalence of minimality of realization with simultaneous controllability and observability, and a state-space similarity theorem) have been worked out in this automaton context. Further details can be found in [15, 19, 32]. Our results give a broader perspective in which to view recognizable formal power series.

5. *Commutative multidimensional system theory.* We view the "noncommutative Fornasini–Marchesini systems" introduced here as noncommutative analogues of the (commutative) Fornasini–Marchesini systems introduced by Fornasini and Marchesini [24] in the multidimensional system theory literature, while the "noncommutative Givone–Roesser systems" are noncommutative analogues of the (commutative) multidimensional Givone–Roesser systems appearing in [26, 27, 36]. In what we call the *commutative case* (evolution along an integer lattice rather than along a free semigroup), the theory of controllability, observability, state-space similarity, and reduction to and construction of a minimal realization of a transfer function is problematic (see, e.g., [31, 25]). By the results here, however, the situation in the noncommutative case is much more like the classical 1-D case. A possible direction for future work is the application of the noncommutative theory as a vehicle for deeper understanding of the commutative case; indeed, the realization theorem in [24] for commutative Fornasini–Marchesini systems is based on the noncommutative realization theorem from [20].

In other directions the commutative theory is ahead of the noncommutative theory. We mention the recent work of Ambrozie and Timotin [3] and of the first author

and Bolotnikov [4], which studies classes of functions with a realization similar to the type of realizations discussed here (see (3.19)) in a commutative (and conservative) setting but with resolvent containing a certain polynomial in the frequency variables rather than just a linear term. In particular, [4] contains a realization result which generalizes the commutative analogue of the main result of [5]. A nonlinear analogue of the realization results of [4] would probably demand a nonlinear version of the Taylor functional calculus (see [41] for a start in this direction). Results on minimality, controllability, and observability obtained in the present paper for this case of higher-degree polynomial in the resolvent of the realization could be obtained by first finding an equivalent system representation having a linear resolvent (or first-order system equations), or, more directly, by developing a more coordinate-free behavioral framework for noncommutative system theory (see [35] for the commutative case).

2. Three classes of examples of structured noncommutative multidimensional linear systems. In this section we introduce and state the main results for the three main examples of SNMLSs. Here the reader can understand the examples and statements of all the main results without having to confront the added formalism of the general definition involving an "admissible graph" (see Definition 3.7 below).

2.1. Noncommutative Fornasini–Marchesini systems. For d a positive integer, let $\mathcal{F}_d$ be the free semigroup generated by the set of d letters $\{1, 2, \dots, d\}$. Elements of $\mathcal{F}_d$ are words w of the form $w = i_N i_{N-1} \cdots i_1$, where $i_k \in \{1, 2, \dots, d\}$ for each $k = 1, \dots, N$. We include the empty word $\emptyset$ as an element of $\mathcal{F}_d$. The semigroup operation is concatenation: $w \cdot w' = i_N i_{N-1} \cdots i_1 i'_{N'} i'_{N'-1} \cdots i'_1$ if $w = i_N i_{N-1} \cdots i_1$ and $w' = i'_{N'} i'_{N'-1} \cdots i'_1$; the empty word $\emptyset$ serves as the identity element of the semigroup $\mathcal{F}_d$. A Fornasini–Marchesini connection matrix U^{FM} is a matrix of the form

$$U^{FM} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A_1 & B_1 \\ \vdots & \vdots \\ A_d & B_d \\ C & D \end{bmatrix} : \begin{bmatrix} \mathcal{H} \\ \mathcal{U} \end{bmatrix} \to \begin{bmatrix} \oplus_{i=1}^{d} \mathcal{H} \\ \mathcal{Y} \end{bmatrix}.$$

The associated system equations are

$$\Sigma^{FM}: \begin{cases} x(1w) = A_1 x(w) + B_1 u(w), \\ \qquad \vdots \\ x(dw) = A_d x(w) + B_d u(w), \\ y(w) \quad = Cx(w) + Du(w) \quad \text{for } w \in \mathcal{F}_d, \end{cases} \tag{2.1}$$

where the state $x(w)$ takes values in the state-space $\mathcal{H}$ and consists of only one component, $u(w)$ takes values in the input-space $\mathcal{U}$, and $y(w)$ takes values in the output-space $\mathcal{Y}$. We consider this type of system as a noncommutative analogue of the (commutative) multidimensional linear systems studied by Fornasini and Marchesini (see, e.g., [24]). We let $z = (z_1, \dots, z_d)$ be a collection of d formal noncommuting variables and consider the formal noncommutative multivariable Z-transform

$$\{x(w)\}_{w \in \mathcal{F}_d} \mapsto \widehat{x}(z) := \sum_{w \in \mathcal{F}_d} x(w) z^w, \tag{2.2}$$

where $z^w = z_{i_N} z_{i_{N-1}} \cdots z_{i_1}$ if $w = i_N i_{N-1} \cdots i_1$. Then, as will be seen in more generality in section 3 (see Example 3.8 and formula (3.20) below), application of the formal Z-transform to the system (2.1) on $\mathcal{T}_{\text{future}}$ leads to the representation

$$\widehat{x}(z) = (I - (Z_{\text{row}}(z) \otimes I_{\mathcal{H}})A)^{-1} x(\emptyset) + (I - (Z_{\text{row}}(z) \otimes I_{\mathcal{H}})A)^{-1} (Z_{\text{row}}(z) \otimes I_{\mathcal{H}}) \cdot B\widehat{u}(z),$$
$$\widehat{y}(z) = C(I - (Z_{\text{row}}(z) \otimes I_{\mathcal{H}})A)^{-1} x(\emptyset) + T_{\Sigma^{FM}}(z)\widehat{u}(z), \tag{2.3}$$

where the formal power series $T_{\Sigma}^{FM}(z)$ (the *transfer function* of the noncommutative Fornasini–Marchesini system Σ^{FM}) is given by

$$\begin{aligned} T_{\Sigma^{FM}}(z) &= D + C(I - (Z_{\text{row}}(z) \otimes I_{\mathcal{H}})A)^{-1}(Z_{\text{row}}(z) \otimes I_{\mathcal{H}})B \\ &= D + C(I - z_1 A_1 - \cdots - z_d A_d)^{-1}(z_1 B_1 + \cdots + z_d B_d) \\ &= D + \sum_{v \in \mathcal{F}_d} \sum_{j=1}^{d} C A^v B_j z^v z_j, \end{aligned} \tag{2.4}$$

where we have used the conventions

$$\begin{aligned} Z_{\text{row}}(z) \otimes I_{\mathcal{H}} &= \begin{bmatrix} z_1 I_{\mathcal{H}} & \cdots & z_d I_{\mathcal{H}} \end{bmatrix}, \\ A^v &= A_{i_N} A_{i_{N-1}} \cdots A_{i_1} \text{ if } v = i_N i_{N-1} \cdots i_1. \end{aligned}$$

We now also consider the associated backward system equations

$$\Sigma_{\text{past}}^{FM}\colon \begin{cases} x(w) = \sum_{i=1}^{d} A_i x(wi) + \sum_{i=1}^{d} B_i u(wi), \\ y(w) = Cx(w) + Du(w) \quad \text{for } w \in \mathcal{F}_d. \end{cases} \tag{2.5}$$

We view the system as running on both the present and future $\mathcal{T}_{\text{future}} := \mathcal{F}_d$ and on the past $\mathcal{T}_{\text{past}} := \mathcal{F}_d \setminus \emptyset$ (where we think of the two appearances of $\mathcal{F}_d$ here as two distinct copies of $\mathcal{F}_d$). The forward equations (2.1) apply for $w \in \mathcal{T}_{future}$, while the backward equations (2.5) apply for $wi \in \mathcal{T}_{past}$. The noncommutative Fornasini–Marchesini system Σ^{FM} is said to be *FM-controllable* if any state-vector $h \in \mathcal{H}$ can be achieved as $h = x(\emptyset)$ by running the system on the past $\mathcal{T}_{\text{past}}$ with state-initialization equal to zero on all locations $w \in \mathcal{T}_{\text{past}}$ of sufficiently long length with some input string $\{u(w)\}_{w \in \mathcal{T}_{\text{past}}}$ having finite support on the past; this condition turns out to be equivalent to the *Fornasini–Marchesini controllability matrix* $\mathcal{C}^{FM}$ given by

$$\mathcal{C}^{FM} = \text{row}_{N=1,2,\ldots} \, \text{row}_{i_1, i_2, \ldots, i_N \in \{1,\ldots,d\}} [A_{i_N} A_{i_{N-1}} \cdots A_{i_2} B_{i_1}] \tag{2.6}$$

having full rank, i.e., having $\operatorname{im} \mathcal{C}^{FM} = \mathcal{H}$. This fact amounts to the specialization of the analysis in section 5 to Example 3.8; a direct analysis can be found in [34].

Dually, we say that the noncommutative Fornasini–Marchesini system Σ^{FM} is *FM-observable* if the state-vector $h \in \mathcal{H}$ can be uniquely recovered from the present and future output string $\{y_i(w)\}_{w \in \mathcal{T}_{\text{future}}}$ generated by running the forward system equations (2.1) of Σ^{FM} with the state initialized by $x(\emptyset) = h$ and with zero input string on the future ($u(w) = 0$ for $w \in \mathcal{T}_{\text{future}} = \mathcal{F}_d$). In terms of the system operators, FM-observability of Σ^{FM} is equivalent to the *Fornasini–Marchesini observability operator* $\mathcal{O}^{FM}$ being injective, where

$$\mathcal{O}^{FM} = \text{col}_{N=0,1,2,\ldots} \, \text{col}_{i_1, i_2, \ldots, i_N \in \{1,\ldots,d\}} [C A_{i_N} A_{i_{N-1}} \cdots A_{i_1}]. \tag{2.7}$$

(Here and elsewhere we interpret $A_{i_N}A_{i_{N-1}}\cdots A_{i_1}$ to be equal to the identity operator $I_{\mathcal{H}}$ in case $N = 0$.) This fact follows from specializing the results of section 6 to Example 3.8 below; again a direct discussion can be found in [34].

The Hankel operator $\mathbb{H}^{FM}$ of the noncommutative Fornasini–Marchesini system Σ^{FM} is the composition $\mathbb{H}^{FM} = \mathcal{O}^{FM}\mathcal{C}^{FM} : \ell_{\text{fin}}(\mathcal{T}_{\text{past}}, \mathcal{U}) \to \ell(\mathcal{T}_{\text{future}}, \mathcal{Y})$; the Hankel operator has the same physical interpretation as in the classical case; $\mathbb{H}^{FM}$ maps a past input to the corresponding future output of a given system trajectory, under the assumption that the state has been initialized to zero in the distant past. Matrix entries of $\mathbb{H}^{FM}$ are given by

$$\mathbb{H}^{FM}_{i_N i_{N-1}\cdots i_1; i'_{N'} i'_{N'-1}\cdots i'_1} = CA_{i_N}A_{i_{N-1}}\cdots A_{i_1}A_{i'_{N'}}A_{i'_{N'-1}}\cdots A_{i'_2}B_{i'_1}, \tag{2.8}$$

where $N = 0, 1, 2, \ldots$, $N' = 1, 2, \ldots$, and $i_k, i'_{k'} \in \{1, \ldots, d\}$ for all k, k'. From the factorization $\mathbb{H}^{FM} = \mathcal{O}^{FM}\mathcal{C}^{FM}$ we see that $\mathbb{H}^{FM}$ has finite rank for any (finite-dimensional) noncommutative Fornasini–Marchesini system. The matrix entries of $\mathbb{H}^{FM}$ can also be expressed directly in terms of the Taylor coefficients (sometimes also called Markov parameters) of the transfer function $T_{\Sigma^{FM}}(z) = \sum_{w\in\mathcal{F}_d} T_w z^w$:

$$\mathbb{H}^{FM}_{v,w} = T_{vw}. \tag{2.9}$$

This type of Hankel operator is obtained by specializing the Hankel operator discussed in section 10 to Example 3.8 below; an explicit discussion of this (Fornasini–Marchesini) case is given in [34].

Given a formal power series $T(z) = \sum_{w\in\mathcal{F}_d} T_w z^w$ in d noncommuting variables $z = (z_1, \ldots, z_d)$ (where $z^w = z_{i_N}\cdots z_{i_1}$ if $w = i_N\cdots i_1$ and where $z^\emptyset = 1$) with operator-valued coefficients $T_w \in \mathcal{L}(\mathcal{U}, \mathcal{Y})$, we say that the noncommutative Fornasini–Marchesini system Σ^{FM} is a (noncommutative Fornasini–Marchesini) realization of $T(z)$ if $T(z) = T_{\Sigma^{FM}}(z)$. A given (noncommutative Fornasini–Marchesini) realization Σ^{FM} of $T(z)$ with state-space $\mathcal{H}$ is said to be *FM-minimal* if, whenever $\Sigma^{FM\prime}$ is another noncommutative Fornasini–Marchesini realization of $T(z)$ with state-space $\mathcal{H}'$, then $\dim\mathcal{H} \le \dim\mathcal{H}'$. Two noncommutative Fornasini–Marchesini systems Σ^{FM} and $\Sigma^{FM\prime}$ with the same input- and output-spaces and connection matrices

$$U^{FM} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A_1 & B_1 \\ \vdots & \vdots \\ A_d & B_d \\ C & D \end{bmatrix} : \begin{bmatrix} \mathcal{H} \\ \mathcal{U} \end{bmatrix} \to \begin{bmatrix} \oplus_{i=1}^d \mathcal{H} \\ \mathcal{Y} \end{bmatrix},$$

$$U^{FM\prime} = \begin{bmatrix} A' & B' \\ C' & D' \end{bmatrix} = \begin{bmatrix} A'_1 & B'_1 \\ \vdots & \vdots \\ A'_d & B'_d \\ C' & D' \end{bmatrix} : \begin{bmatrix} \mathcal{H}' \\ \mathcal{U} \end{bmatrix} \to \begin{bmatrix} \oplus_{i=1}^d \mathcal{H}' \\ \mathcal{Y} \end{bmatrix}$$

are said to be *FM-similar* if there is a bijective linear operator $\Gamma : \mathcal{H} \to \mathcal{H}'$ such that

$$\begin{bmatrix} \Gamma & & & \\ & \ddots & & \\ & & \Gamma & \\ & & & I_{\mathcal{Y}} \end{bmatrix} \begin{bmatrix} A_1 & B_1 \\ \vdots & \vdots \\ A_d & B_d \\ C & D \end{bmatrix} = \begin{bmatrix} A'_1 & B'_1 \\ \vdots & \vdots \\ A'_d & B'_d \\ C' & D' \end{bmatrix} \begin{bmatrix} \Gamma & 0 \\ 0 & I_{\mathcal{U}} \end{bmatrix}.$$

The following theorem summarizes the results of Theorems 8.2, 9.1, and 11.1 when specialized to the case of noncommutative Fornasini–Marchesini systems (Example 3.8).

THEOREM 2.1.

(1) *Suppose that* Σ^{FM} *and* $\Sigma^{FM\prime}$ *are two noncommutative Fornasini–Marchesini systems which are both FM-controllable and FM-observable. Then* Σ^{FM} *and* $\Sigma^{FM\prime}$ *are FM-similar if and only if they realize the same transfer function:*

$$T_{\Sigma^{FM}}(z) = T_{\Sigma^{FM\prime}}(z).$$

(2) *The noncommutative Fornasini–Marchesini system* Σ^{FM} *is an FM-minimal realization of its transfer function* $T_{\Sigma^{FM}}(z)$ *if and only if* Σ^{FM} *is both FM-controllable and FM-observable.*

(3) *Suppose that* $T(z) = \sum_{w\in\mathcal{F}_d} T_w z^w$ *is a formal power series in* d *noncommuting variables* $z = (z_1, \dots, z_d)$ *with matrix coefficients* $T_w \in \mathcal{L}(\mathcal{U}, \mathcal{Y})$. *Then* $T(z)$ *can be realized as the transfer function* $T(z) = T_{\Sigma^{FM}}(z)$ *of a finite-dimensional noncommutative Fornasini–Marchesini system* Σ^{FM} *if and only if the associated Hankel matrix*

$$\mathbb{H}^T = [T_{vw}]_{v\in\mathcal{F}_d, w\in\mathcal{F}_d\setminus\{\emptyset\}}$$

has finite rank. In this case there is a canonical construction (shift realization) of a minimal realization with state-space $\mathcal{H}$ *having* $\dim\mathcal{H} = \operatorname{rank}\mathbb{H}^T$.

2.2. Noncommutative Givone–Roesser systems. Just as was done above for the case of noncommutative Fornasini–Marchesini systems, the domain evolution for a *noncommutative Givone–Roesser system* which we discuss now is the free semigroup $\mathcal{F}_d$ on the set of d letters $\{1, 2, \dots, d\}$ (for d a positive integer). We take the associated Givone–Roesser connection matrix U^{GR}, however, to have the form

$$U^{GR} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A_{11} & \dots & A_{1d} & B_1 \\ \vdots & & \vdots & \vdots \\ A_{d1} & \dots & A_{dd} & B_d \\ C_1 & \dots & C_d & D \end{bmatrix} : \begin{bmatrix} \mathcal{H}_1 \\ \vdots \\ \mathcal{H}_d \\ \mathcal{U} \end{bmatrix} \to \begin{bmatrix} \mathcal{H}_1 \\ \vdots \\ \mathcal{H}_d \\ \mathcal{Y} \end{bmatrix}$$

for *auxiliary state-spaces* $\mathcal{H}_1, \dots, \mathcal{H}_d$, an input-space $\mathcal{U}$, and an output-space $\mathcal{Y}$ (all finite-dimensional linear spaces for our discussion here). The associated system equations then are

$$\Sigma^{GR}: \begin{cases} x_1(1w) = A_{11}x_1(w) + \cdots + A_{1d}x_d(w) + B_1u(w), \\ \qquad\vdots \\ x_d(dw) = A_{d1}x_1(w) + \cdots + A_{dd}x_d(w) + B_du(w), \\ y(w) \quad = C_1x_1(w) + \cdots + C_dx_d(w) + Du(w), \end{cases} \text{for } w \in \mathcal{F}_d, \tag{2.10}$$

where the state $x(w) = \operatorname{col}_{j=1,\dots,d} x_j(w)$ at position $w \in \mathcal{F}_d$ consists of d components $x_1(w), \dots, x_d(w)$ with $x_j(w)$ taking values in the auxiliary state-space $\mathcal{H}_j$ for $j = 1, \dots, d$; $u(w)$ takes values in the input-space $\mathcal{U}$; and $y(w)$ takes values in the output-space $\mathcal{Y}$. In case $i, j \in \{1, \dots, d\}$ with $i \neq j$ we set $x_i(jw) = 0$. We consider this type of system as a noncommutative analogue of the (commutative) multidimensional linear systems introduced by Givone and Roesser (see, e.g., [26, 27, 36]). If we apply

the noncommutative formal Z-transform (2.2) to the system equations (2.10) and solve, we get

$$\begin{aligned}\widehat{x}(z) &= (I - (Z_{\mathrm{diag}}(z) \otimes I_{\mathcal{H}})A)^{-1}x(\emptyset) + (I - (Z_{\mathrm{diag}}(z) \otimes I_{\mathcal{H}})A)^{-1}(Z_{\mathrm{diag}}(z) \otimes I_{\mathcal{H}}) \\ &\qquad \cdot B\widehat{u}(z), \\ \widehat{y}(z) &= C(I - (Z_{\mathrm{diag}}(z) \otimes I_{\mathcal{H}})A)^{-1}x(\emptyset) + T_{\Sigma^{GR}}(z)\widehat{u}(z),\end{aligned} \tag{2.11}$$

where the formal power series $T_{\Sigma^{GR}}(z)$ (the *transfer function* of the noncommutative Givone–Roesser system Σ^{GR}) is given by

$$\begin{aligned}T_{\Sigma^{GR}}(z) &= D + C(I - (Z_{\mathrm{diag}}(z) \otimes I_{\mathcal{H}})A)^{-1}(Z_{\mathrm{diag}}(z) \otimes I_{\mathcal{H}})B \\ &= D + \begin{bmatrix} C_1 & \cdots & C_d \end{bmatrix} \left(\begin{bmatrix} I_{\mathcal{H}_1} & & \\ & \ddots & \\ & & I_{\mathcal{H}_d} \end{bmatrix} - \begin{bmatrix} z_1A_{11} & \cdots & z_1A_{1d} \\ \vdots & & \vdots \\ z_dA_{d1} & \cdots & z_dA_{dd} \end{bmatrix} \right)^{-1} \begin{bmatrix} z_1B_1 \\ \vdots \\ z_dB_d \end{bmatrix} \\ &= D + \sum_{N=1}^{\infty} \sum_{i_1,\ldots,i_N \in \{1,\ldots,d\}} C_{i_N} A_{i_N,i_{N-1}} A_{i_{N-1},i_{N-2}} \cdots A_{i_2,i_1} B_{i_1} z_{i_N} z_{i_{N-1}} \cdots z_{i_2} z_{i_1},\end{aligned} \tag{2.12}$$

where we have used the convention

$$Z_{\mathrm{diag}}(z) \otimes I_{\mathcal{H}} = \begin{bmatrix} z_1 I_{\mathcal{H}_1} & & \\ & \ddots & \\ & & z_d I_{\mathcal{H}_d} \end{bmatrix}.$$

We now also consider the associated backward system equations

$$\Sigma^{GR}_{\mathrm{past}} : \begin{cases} x_1(w) = \sum_{i=1}^d A_{1i}x_i(w1) + B_1u(w1), \\ \qquad \vdots \\ x_d(w) = \sum_{i=1}^d A_{di}x_i(wd) + B_du(wd), \\ y(w) \;\; = \sum_{i=1}^d C_ix(w) + Du(w) \quad \text{for } w \in \mathcal{F}_d. \end{cases} \tag{2.13}$$

We follow the same convention as explained above for noncommutative Fornasini–Marchesini systems and view the system Σ^{GR} as running on both the present and future $\mathcal{T}_{\mathrm{future}} := \mathcal{F}_d$ and on the past $\mathcal{T}_{\mathrm{past}} := \mathcal{F}_d \setminus \emptyset$, with the forward equations (2.10) applying for $w \in \mathcal{T}_{future}$ and the backward equations (2.13) applying for $wi \in \mathcal{T}_{past}$. The noncommutative Givone–Roesser system Σ^{GR} is said to be *GR-controllable* if, for each $i \in \{1, \ldots, d\}$, any state-vector $h_i \in \mathcal{H}_i$ can be achieved as the ith component $h_i = x_i(\emptyset)$ of the state-vector $x(\emptyset)$ at the empty-set location by running the system on the past $\mathcal{T}_{\mathrm{past}}$ with state-initialization equal to zero on all locations $w \in \mathcal{T}_{\mathrm{past}}$ of sufficiently long length with some input string $\{u(w)\}_{w \in \mathcal{T}_{\mathrm{past}}}$ having finite support on the past; this condition turns out to be equivalent to the ith *Givone–Roesser controllability matrix* $\mathcal{C}_i^{GR}$ given by

$$\mathcal{C}_i^{GR} = \mathrm{row}_{N=0,1,\ldots} \, \mathrm{row}_{i_1,i_2,\ldots,i_N \in \{1,\ldots,d\}} [A_{i,i_N} A_{i_N,i_{N-1}} \cdots A_{i_2,i_1} B_{i_1}] \tag{2.14}$$

(where the $N = 0$ term is to be interpreted as simply B_i) having full rank, i.e., having $\mathrm{im}\, \mathcal{C}_i^{GR} = \mathcal{H}_i$ for each $i = 1, \ldots, d$. This fact follows by specializing the analysis in section 5 to Example 3.9 below; a direct discussion is in [34].

Dually, we say that the noncommutative Givone–Roesser system Σ^{GR} is *GR-observable* if, for each $i = 1, \dots, d$, the state-vector $h_i \in \mathcal{H}_i$ can be uniquely recovered from the present and future output string $\{y(w)\}_{w \in \mathcal{T}_{\text{future}}}$ generated by running the forward system equations (2.10) of Σ^{GR} with the state initialized by $x_i(\emptyset) = h_i$ and $x_{i'}(\emptyset) = 0$ for $i' \neq i$, and with zero input string on the future ($u(w) = 0$ for $w \in \mathcal{T}_{\text{future}} = \mathcal{F}_d$). In terms of the system operators, GR-observability of Σ^{GR} is equivalent to the ith *Givone–Roesser observability operator* $\mathcal{O}_i^{GR}$ being injective for each $i = 1, \dots, d$, where

$$\mathcal{O}_i^{GR} = \operatorname{col}_{N=0,1,2,\dots} \operatorname{col}_{i_1,i_2,\dots,i_N \in \{1,\dots,d\}} [C_{i_N} A_{i_N,i_{N-1}} A_{i_{N-1},i_{N-2}} \cdots A_{i_1,i}]. \tag{2.15}$$

Here the $N = 0$ term is to be interpreted as simply C_i. All these matters follow upon specialization of the analysis in section 6 to Example 3.9 below; again, a direct discussion is in [34].

There are d Hankel operators $\mathbb{H}^{GR,1}, \dots, \mathbb{H}^{GR,d}$ for a noncommutative Givone–Roesser system Σ^{GR}; namely, for each $i = 1, \dots, d$,

$$\mathbb{H}^{GR,i} = \mathcal{O}_i^{GR} \mathcal{C}_i^{GR} \colon \ell_{\text{fin}}(\mathcal{T}_{\text{past}}, \mathcal{U}) \to \ell(\mathcal{T}_{\text{future}}, \mathcal{Y}).$$

Each Hankel operator $\mathbb{H}^{GR,i}$ again has a physical interpretation as mapping a past input to the corresponding future output of a given system trajectory under the assumption that the state has been initialized to zero in the distant past, but where the observations are taken only with respect to the ith component $x_i(\emptyset)$ of the state at position $\emptyset$. Matrix entries of $\mathbb{H}^{GR,i}$ are given by

$$\begin{aligned} &\mathbb{H}^{GR,i}_{i_N i_{N-1} \cdots i_1;\, i'_{N'} i'_{N'-1} \cdots i'_1} \\ &\qquad = C_{i_N} A_{i_N,i_{N-1}} A_{i_{N-1},i_{N-2}} \cdots A_{i_1,i} A_{i,i'_{N'}} A_{i'_{N'},i'_{N'-1}} \cdots A_{i'_2,i'_1} B_{i'_1}, \end{aligned} \tag{2.16}$$

where $N' = 0, 1, 2, \dots$, $N = 0, 1, 2, \dots$, and $i_k, i'_{k'} \in \{1, \dots, d\}$ for all k, k'. Some small values of N and N' in formula (2.16) require special interpretation; for example, for case $N = 0$ and $N' = 0$ we interpret (2.16) as giving

$$\mathbb{H}^{GR,i}_{\emptyset;\emptyset} = C_i B_i.$$

From the factorization $\mathbb{H}^{GR,i} = \mathcal{O}_i^{GR} \mathcal{C}_i^{GR}$ we see that $\mathbb{H}^{GR,i}$ has finite rank for each $i = 1, \dots, d$ for any (finite-dimensional) noncommutative Givone–Roesser system. The matrix entries of $\mathbb{H}^{GR,i}$ can also be expressed directly in terms of the Taylor coefficients (sometimes also called Markov parameters) of the transfer function $T_{\Sigma^{GR}}(z) = \sum_{w \in \mathcal{F}_d} T_w z^w$: indeed,

$$\mathbb{H}^{GR,i}_{v,w} = T_{viw} \quad \text{for } v, w \in \mathcal{F}_d,\ i \in \{1, \dots, d\}. \tag{2.17}$$

These details amount to the specialization of section 10 to Example 3.9 below, and also can be found (in explicit form) in [34].

Given a formal power series $T(z) = \sum_{w \in \mathcal{F}_d} T_w z^w$ in d noncommuting variables $z = (z_1, \dots, z_d)$ (where $z^w = z_{i_N} \cdots z_{i_1}$ if $w = i_N \cdots i_1$ and where $z^\emptyset = 1$) with operator-valued coefficients $T_w \in \mathcal{L}(\mathcal{U}, \mathcal{Y})$, we say that the noncommutative Givone–Roesser system Σ^{GR} is a (noncommutative Givone–Roesser) realization of $T(z)$ if $T(z) = T_{\Sigma^{GR}}(z)$. A given (noncommutative Givone–Roesser) realization Σ^{GR} of $T(z)$ with auxiliary state-spaces $\mathcal{H}_1, \dots, \mathcal{H}_d$ is said to be *GR-minimal* if, whenever $\Sigma^{GR\prime}$

is another noncommutative Givone–Roesser realization of $T(z)$ with auxiliary state-spaces $\mathcal{H}'_1, \dots, \mathcal{H}'_d$, then $\dim \mathcal{H}_i \le \dim \mathcal{H}'_i$ for each $i = 1, \dots, d$. Two noncommutative Givone–Roesser systems Σ^{GR} and $\Sigma^{GR\prime}$ with the same input- and output-spaces and connection matrices

$$U^{GR} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A_{11} & \dots & A_{1d} & B_1 \\ \vdots & & \vdots & \vdots \\ A_{d1} & \dots & A_{dd} & B_d \\ C_1 & \dots & C_d & D \end{bmatrix} : \begin{bmatrix} \mathcal{H}_1 \\ \vdots \\ \mathcal{H}_d \\ \mathcal{U} \end{bmatrix} \to \begin{bmatrix} \mathcal{H}_1 \\ \vdots \\ \mathcal{H}_d \\ \mathcal{Y} \end{bmatrix},$$

$$U^{GR\prime} = \begin{bmatrix} A' & B' \\ C' & D' \end{bmatrix} = \begin{bmatrix} A'_{11} & \dots & A'_{1d} & B'_1 \\ \vdots & & \vdots & \vdots \\ A'_{d1} & \dots & A'_{dd} & B'_d \\ C'_1 & \dots & C'_d & D' \end{bmatrix} : \begin{bmatrix} \mathcal{H}'_1 \\ \vdots \\ \mathcal{H}'_d \\ \mathcal{U} \end{bmatrix} \to \begin{bmatrix} \mathcal{H}'_1 \\ \vdots \\ \mathcal{H}'_d \\ \mathcal{Y} \end{bmatrix}$$

are said to be *GR-similar* if, for each $i = 1, \dots, d$, there is a bijective linear operator $\Gamma_i \colon \mathcal{H}_i \to \mathcal{H}'_i$ such that

$$\begin{bmatrix} \Gamma_1 & & & \\ & \ddots & & \\ & & \Gamma_d & \\ & & & I_{\mathcal{Y}} \end{bmatrix} \begin{bmatrix} A_{11} & \cdots A_{1d} & B_1 \\ \vdots & \vdots & \vdots \\ A_{d1} & \cdots A_{dd} & B_d \\ C_1 & \cdots C_d & D \end{bmatrix} = \begin{bmatrix} A'_{11} & \cdots & A'_{1d} & B'_1 \\ \vdots & & \vdots & \vdots \\ A'_{d1} & \cdots & A'_{dd} & B'_d \\ C'_1 & \cdots & C'_d & D' \end{bmatrix} \begin{bmatrix} \Gamma_1 & & & \\ & \ddots & & \\ & & \Gamma_d & \\ & & & I_{\mathcal{U}} \end{bmatrix}.$$

The following theorem summarizes the results of Theorems 8.2, 9.1, and 11.1 when specialized to the case of noncommutative Givone–Roesser systems (Example 3.9).

THEOREM 2.2.

(1) *Suppose that Σ^{GR} and $\Sigma^{GR\prime}$ are two noncommutative Givone–Roesser systems which are both GR-controllable and GR-observable. Then Σ^{GR} and $\Sigma^{GR\prime}$ are GR-similar if and only if they realize the same transfer function:*

$$T_{\Sigma^{GR}}(z) = T_{\Sigma^{GR\prime}}(z).$$

(2) *The noncommutative Givone–Roesser system Σ^{GR} is a GR-minimal realization of its transfer function $T_{\Sigma^{GR}}(z)$ if and only if Σ^{GR} is both GR-controllable and GR-observable.*

(3) *Suppose that $T(z) = \sum_{w \in \mathcal{F}_d} T_w z^w$ is a formal power series in d noncommuting variables $z = (z_1, \dots, z_d)$ with matrix coefficients $T_w \in \mathcal{L}(\mathcal{U}, \mathcal{Y})$. Then $T(z)$ can be realized as the transfer function $T(z) = T_{\Sigma^{GR}}(z)$ of a finite-dimensional noncommutative Givone–Roesser system Σ^{GR} if and only if the associated Hankel matrices*

$$\mathbb{H}^{T,i} = [T_{viw}]_{v \in \mathcal{F}_d, w \in \mathcal{F}_d}$$

have finite rank for $i = 1, \dots, d$. In this case there is a canonical construction (shift realization) of a minimal realization with auxiliary state-space $\mathcal{H}_i$ having $\dim \mathcal{H}_i = \operatorname{rank} \mathbb{H}^{T,i}$ for $i = 1, \dots, d$.

In fact, it can be shown that a formal power series $T(z) = \sum_{w \in \mathcal{F}_d} T_w z^w$ in d noncommuting variables $z = (z_1, \ldots, z_d)$ has an FM-realization if and only if it has a GR-realization if and only if it is *rational* in the sense of Schützenberger (see [15]). One of the points of part (3) in Theorems 2.1 and 2.2 is that they identify the precise Hankel matrices with rank(s) equal to the state-space dimension(s) in a minimal realization of Fornasini–Marchesini or Givone–Roesser type. We discuss these connections between various types of noncommutative realizations in section 12.

2.3. Noncommutative full-structured systems. Our last concrete example of a structured noncommutative system is what we call a "full-structured" system. For this case it is convenient to assume that the evolution of the system takes place on the free semigroup generated by a certain Cartesian product set. Denote by $\mathcal{F}_{n,m}$ the free semigroup generated by the set $E = \{1, \ldots, n\} \times \{1, \ldots, m\}$. Thus elements of $\mathcal{F}_{n,m}$ are words w of the form $(i_N, j_N)(i_{N-1}, j_{N-1}) \cdots (i_1, j_1)$, where $i_k \in \{1, \ldots, n\}$ and $j_k \in \{1, \ldots, m\}$ for all $k = 1, \ldots, N$. Again we let $\emptyset$ denote the empty word which serves as the identity for the semigroup $\mathcal{F}_{n,m}$. By a *full-structured connection matrix* U^{full} we mean a block-operator matrix of the form

$$U^{\text{full}} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A_{11} & \cdots & A_{1n} & B_1 \\ \vdots & & \vdots & \vdots \\ A_{m1} & \cdots & A_{mn} & B_m \\ C_1 & \cdots & C_n & D \end{bmatrix} : \begin{bmatrix} \oplus_{i=1}^n \mathcal{H} \\ \mathcal{U} \end{bmatrix} \to \begin{bmatrix} \oplus_{j=1}^m \mathcal{H} \\ \mathcal{Y} \end{bmatrix},$$

where $\mathcal{H}$ (the *state-space*), $\mathcal{U}$ (the *input-space*), and $\mathcal{Y}$ (the *output-space*) are finite-dimensional linear spaces. The associated system equations are

(2.18)
$$\Sigma^{\text{full}}: \begin{cases} x_1((1,j)\cdot w) = A_{j1}x_1(w) + \cdots + A_{jn}x_n(w) + B_j u(w) & \text{for } j = 1, \ldots, m, \\ \qquad\qquad \vdots \\ x_n((n,j)\cdot w) = A_{j1}x_1(w) + \cdots + A_{jn}x_n(w) + B_j u(w) & \text{for } j = 1, \ldots, m, \\ x_i((i',j)\cdot w) = 0 \quad \text{if } i \neq i', \\ y(w) = C_1 x_1(w) + \cdots + C_n x_n(w) + Du(w). \end{cases}$$

Here the state-vector $x(w) = \text{col}_{i=1,\ldots,n}\, x_i(w) \in \oplus_{i=1}^n \mathcal{H}$ consists of n components $x_i(w)$ for $i = 1, \ldots, n$ with each $x_i(w)$ in the auxiliary state-space $\mathcal{H}$, while $u(w)$ assumes values in the input-space $\mathcal{U}$ and $y(w)$ assumes values in the output-space $\mathcal{Y}$. Note that the state trajectory $\{x(w)\}_{w \in \mathcal{F}_{n,m}}$ incorporates some redundancy; namely, *if* $\{x(w)\}_{w \in \mathcal{F}_{n,m}} = \{\text{col}_{i=1,\ldots,n}[x_i(w)]\}_{w \in \mathcal{F}_{n,m}}$ *is the state trajectory satisfying the state-update equation in* (2.18) *for some choice of input signal* $\{u(w)\}_{w \in \mathcal{F}_{n,m}}$*, then, for each fixed* $j \in \{1, \ldots, m\}$ *and* $w \in \mathcal{F}_{n,m}$,

(2.19) $$x_i((i,j)\cdot w) \text{ is independent of } i \in \{1, \ldots, n\}.$$

We shall work with the redundant form (2.18) of the system equations rather than rewriting them in a more economical form.

We let $z = (z_{11}, \ldots, z_{1m}; z_{21}, \ldots, z_{2m}; \ldots; z_{n1}, \ldots, z_{nm})$ be a collection of nm noncommuting variables indexed by $\{1, \ldots, n\} \times \{1, \ldots, m\}$. Application of the noncommutative Z-transform (2.2) (with respect to $\mathcal{F}_{n,m}$ rather than with respect to $\mathcal{F}_d$)

converts the system equations to

$$\widehat{x}(z) = (I - (Z_{\text{full}}(z) \otimes I_{\mathcal{H}})A)^{-1}x(\emptyset) + (I - (Z_{\text{full}}(z) \otimes I_{\mathcal{H}})A)^{-1}(Z_{\text{full}}(z) \otimes I_{\mathcal{H}}) \cdot B\widehat{u}(z),$$
$$\widehat{y}(z) = C(I - (Z_{\text{full}}(z) \otimes I_{\mathcal{H}})A)^{-1}x(\emptyset) + T_{\Sigma^{\text{full}}}(z)\widehat{u}(z), \tag{2.20}$$

where $T_{\Sigma^{\text{full}}}(z)$ (the *transfer function* of the noncommutative full-structured system Σ^{full}) is given by

$$\begin{aligned}
T_{\Sigma^{\text{full}}}(z) &= D + C(I - (Z_{\text{full}}(z) \otimes I_{\mathcal{H}})A)^{-1}(Z_{\text{full}}(z) \otimes I_{\mathcal{H}})B \\
&= D + \begin{bmatrix} C_1 & \cdots & C_n \end{bmatrix} \\
&\cdot \left(\begin{bmatrix} I_{\mathcal{H}} & & \\ & \ddots & \\ & & I_{\mathcal{H}} \end{bmatrix} - \begin{bmatrix} \sum_{j=1}^m z_{1j}A_{j1} & \cdots & \sum_{j=1}^m z_{1j}A_{jn} \\ \vdots & & \vdots \\ \sum_{j=1}^m z_{nj}A_{j1} & \cdots & \sum_{j=1}^m z_{nj}A_{jn} \end{bmatrix} \right)^{-1} \begin{bmatrix} \sum_{j=1}^m z_{1j}B_j \\ \vdots \\ \sum_{j=1}^m z_{nj}B_j \end{bmatrix} \\
&= D + \sum_{N=1}^{\infty} \sum_{i_1,\ldots,i_N \in \{1,\ldots,n\}} \sum_{j_1,\ldots,j_N \in \{1,\ldots,m\}} C_{i_N}A_{j_N,i_{N-1}}A_{j_{N-1},i_{N-2}} \cdots A_{j_2,i_1}B_{j_1} \\
&\qquad \cdot z_{i_N,j_N}z_{i_{N-1},j_{N-1}} \cdots z_{i_2,j_2}z_{i_1,j_1},
\end{aligned} \tag{2.21}$$

and where $Z_{\text{full}}(z) \otimes I_{\mathcal{H}}$ is given by

$$Z_{\text{full}}(z) \otimes I_{\mathcal{H}} = \begin{bmatrix} z_{1,1}I_{\mathcal{H}} & \cdots & z_{1,m}I_{\mathcal{H}} \\ \vdots & & \vdots \\ z_{n,1}I_{\mathcal{H}} & \cdots & z_{n,m}I_{\mathcal{H}} \end{bmatrix}.$$

The backward full-structured system equations have the form

$$\Sigma^{\text{full}}_{past}\colon \begin{cases} x_1(w) = \sum_{j=1}^m \sum_{i'=1}^n A_{j,i'}x_{i'}(w \cdot (1,j)) + \sum_{j=1}^m B_j u(w \cdot (1,j)), \\ \qquad \vdots \\ x_n(w) = \sum_{j=1}^m \sum_{i'=1}^n A_{j,i'}x_{i'}(w \cdot (n,j)) + \sum_{j=1}^m B_j u(w \cdot (n,j)), \\ y(w) \quad = \sum_{i=1}^n C_i x_i(w) + Du(w), \end{cases} \tag{2.22}$$

and are to be interpreted as the evolution of the system on the past $\mathcal{T}_{\text{past}} = \mathcal{F}_{n,m} \setminus \{\emptyset\}$. The noncommutative full-structured system Σ^{full} is said to be *full-controllable* if, for each $i \in \{1,\ldots,n\}$, any state-vector $h \in \mathcal{H}$ can be achieved as the ith component $h_i = x_i(\emptyset)$ (for some, or equivalently for any $i \in \{1,\ldots,n\}$) of the state-vector $x(\emptyset)$ at the empty-set location by running the system on the past $\mathcal{T}_{\text{past}}$ with state-initialization equal to zero on all locations $w \in \mathcal{T}_{\text{past}}$ of sufficiently long length with some input string $\{u(w)\}_{w \in \mathcal{T}_{\text{past}}}$ having finite support on the past; this condition turns out to be equivalent to the *full-structured controllability matrix* $\mathcal{C}_1^{\text{full}}$ given by

$$\begin{aligned}
\mathcal{C}_1^{\text{full}} = \text{row}_{N=1,2,\ldots} \text{row}_{(1,j_N)(i_{N-1},j_{N-1})\cdots(i_1,j_1)\colon\, i_1,i_2,\ldots,i_{N-1} \in \{1,\ldots,n\};j_1,\ldots,j_N \in \{1,\ldots,m\}} \\
[A_{j_N,i_{N-1}}A_{j_{N-1},i_{N-2}} \cdots A_{j_2,i_1}B_{j_1}]
\end{aligned} \tag{2.23}$$

having full rank, i.e., having $\text{im}\,\mathcal{C}_1^{\text{full}} = \mathcal{H}$. These facts amount to the specialization of the results of section 5 to Example 3.10 below.

Dually, we say that the noncommutative full-structured system $\Sigma^{\rm full}$ is *full-observable* if the state-vector $h \in \mathcal{H}$ can be uniquely recovered from the n-tuple of present and future output strings $\{y_i(w)\}_{i=1,\dots,n;w\in\mathcal{T}_{\rm future}}$ (with $\mathcal{T}_{\rm future} = \mathcal{F}_{n,m}$). Here, for each $i = 1,\dots,n$, the ith output string $\{y_i(w)\}_{w\in\mathcal{F}_{n,m}}$ is generated by running the forward system equations (2.18) of $\Sigma^{\rm full}$ with the state initialized by $x_i(\emptyset) = h_i$ and $x_{i'}(\emptyset) = 0$ for $i' \neq i$ and with zero input string on the future ($u(w) = 0$ for $w \in \mathcal{T}_{\rm future} = \mathcal{F}_{n,m}$). In terms of the system operators, full-observability of $\Sigma^{\rm full}$ is equivalent to the *full observability operator* $\mathcal{O}^{\rm full}\colon \mathcal{H} \to \oplus_{i=1}^n \ell(\mathcal{F}_{n,m},\mathcal{Y})$ being injective, where

$$\begin{aligned}\mathcal{O}^{\rm full} &= \operatorname{col}_{i=1,\dots,n} \operatorname{col}_{N=0,1,2,\dots} \operatorname{col}_{(i_N,j_N)\cdots(i_1,j_1)\colon\, i_1,\dots,i_N\in\{1,\dots,n\};j_1,\dots,j_N\in\{1,\dots,m\}} \\ &\qquad [C_{i_N} A_{j_N,i_{N-1}} A_{j_{N-1},i_{N-2}} \cdots A_{j_1,i}]. \end{aligned} \tag{2.24}$$

Here the $N = 0$ term is to be interpreted as simply C_i. These matters amount to specialization of the results of section 6 to Example 3.10 below.

We define the Hankel operator $\mathbb{H}^{\rm full}$ for a noncommutative full-structured system $\Sigma^{\rm full}$ as the composition $\mathbb{H}^{\rm full} = \mathcal{O}^{\rm full}\mathcal{C}_1^{\rm full}\colon \ell_{\rm fin}(\mathcal{T}_{\rm past}^1,\mathcal{U}) \to \oplus_{i=1}^n \ell(\mathcal{T}_{\rm future},\mathcal{Y})$; here $\mathcal{T}_{\rm past}^1$ denotes a certain subset of the past $\mathcal{T}_{\rm past}$, namely, the set of all nonempty words $(i_1,j_1)\cdot(i_2,j_2)\cdots(i_N,j_N)$ for which the leading letter (i_1,j_1) has first component i_1 equal to 1. Again the Hankel operator $\mathbb{H}^{\rm full}$ has a physical interpretation as mapping a past input to the corresponding future output of a given system trajectory (in this case an n-tuple of future outputs) under the assumption that the state has been initialized to zero in the distant past. Matrix entries of $\mathbb{H}^{\rm full}$ are given by

$$\begin{aligned}&\mathbb{H}^{\rm full}_{i,(i_N,j_N)\cdots(i_1,j_1);(i'_{N'},j'_{N'}),(i'_{N'-1},j'_{N'-1})\cdots(i'_1,j'_1)} \\ &\qquad = C_{i_N} A_{j_N,i_{N-1}} A_{j_{N-1},i_{N-2}} \cdots A_{j_1,i} A_{j'_{N'},i'_{N'-1}} \cdots A_{j'_2,i'_1} B_{j'_1}, \end{aligned} \tag{2.25}$$

where $N = 0,1,2,\dots$, $N' = 1,2,\dots$, and $i_k, i'_{k'} \in \{1,\dots,n\}$ and $j_k, j'_{k'} \in \{1,\dots,m\}$ for all k and k'; some small values of N and N' in formula (2.25) require special interpretation; for example, for case $N = 0$ and $N' = 1$ we interpret (2.25) as giving

$$\mathbb{H}^{\rm full}_{i,\emptyset;(1,j)} = C_i B_j.$$

From the factorization $\mathbb{H}^{\rm full} = \mathcal{O}^{\rm full}\mathcal{C}_1^{\rm full}$ we see that $\mathbb{H}^{\rm full}$ has finite rank equal to the dimension of the state-space in a minimal realization for any (finite-dimensional) noncommutative full-structured system. The matrix entries of $\mathbb{H}^{\rm full}$ can also be expressed directly in terms of the Taylor coefficients of the transfer function $T_{\Sigma^{\rm full}}(z) = \sum_{w\in\mathcal{F}_{n,m}} T_w z^w$: indeed

$$\mathbb{H}^{\rm full}_{i,v;(1,j_N)w'} = T_{v\cdot(i,j_N)\cdot w'} \quad \text{for } v,w' \in \mathcal{F}_{n,m},\ i \in \{1,\dots,n\}. \tag{2.26}$$

These results all fall out of specializing the results of section 10 to Example 3.10 below.

Given a formal power series $T(z) = \sum_{w\in\mathcal{F}_{n,m}} T_w z^w$ in $n\cdot m$ noncommuting variables $z = (z_{11},\dots,z_{1m};\cdots;z_{n1},\cdots,z_{nm})$ (where $z^w = z_{i_N,j_N}\cdots z_{i_1,j_1}$ if $w = (i_N,j_N)\cdots(i_1,j_1)$ and where $z^\emptyset = 1$) with $\mathcal{L}(\mathcal{U},\mathcal{Y})$-valued coefficients T_w, we say that the noncommutative full-structured system $\Sigma^{\rm full}$ is a (noncommutative full) realization of $T(z)$ if $T(z) = T_{\Sigma^{\rm full}}(z)$. A given (noncommutative full) realization $\Sigma^{\rm full}$ of $T(z)$ with auxiliary state-space $\mathcal{H}$ is said to be *full-minimal* if, whenever $\Sigma^{\rm full\,\prime}$ is another noncommutative full realization of $T(z)$ with auxiliary state-space $\mathcal{H}$, then

$\dim \mathcal{H} \le \dim \mathcal{H}'$. Two noncommutative full-structured systems Σ^{full} and $\Sigma^{\text{full}\prime}$ with the same input- and output-spaces and connection matrices

$$U^{\text{full}} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A_{11} & \cdots & A_{1n} & B_1 \\ \vdots & & \vdots & \vdots \\ A_{m1} & \cdots & A_{mn} & B_m \\ C_1 & \cdots & C_n & D \end{bmatrix} : \begin{bmatrix} \oplus_{i=1}^n \mathcal{H} \\ \mathcal{U} \end{bmatrix} \to \begin{bmatrix} \oplus_{j=1}^m \mathcal{H} \\ \mathcal{Y} \end{bmatrix},$$

$$U^{\text{full}\prime} = \begin{bmatrix} A' & B' \\ C' & D' \end{bmatrix} = \begin{bmatrix} A'_{11} & \cdots & A'_{1n} & B'_1 \\ \vdots & & \vdots & \vdots \\ A'_{m1} & \cdots & A'_{mn} & B'_m \\ C'_1 & \cdots & C'_n & D' \end{bmatrix} : \begin{bmatrix} \oplus_{i=1}^n \mathcal{H}' \\ \mathcal{U} \end{bmatrix} \to \begin{bmatrix} \oplus_{j=1}^m \mathcal{H}' \\ \mathcal{Y} \end{bmatrix}$$

are said to be *full-similar* if there is a bijective linear operator $\Gamma : \mathcal{H} \to \mathcal{H}'$ such that

$$\begin{bmatrix} \Gamma & & & \\ & \ddots & & \\ & & \Gamma & \\ & & & I_{\mathcal{Y}} \end{bmatrix} \begin{bmatrix} A_{11} & \cdots A_{1n} & B_1 \\ \vdots & \vdots & \vdots \\ A_{m1} & \cdots A_{mn} & B_m \\ C_1 & \cdots C_n & D \end{bmatrix} = \begin{bmatrix} A'_{11} & \cdots & A'_{1n} & B'_1 \\ \vdots & & \vdots & \vdots \\ A'_{m1} & \cdots & A'_{mn} & B'_m \\ C'_1 & \cdots & C'_n & D' \end{bmatrix} \begin{bmatrix} \Gamma & & & \\ & \ddots & & \\ & & \Gamma & \\ & & & I_{\mathcal{U}} \end{bmatrix}.$$

The following theorem summarizes the results of Theorems 8.2, 9.1, and 11.1 when specialized to the case of noncommutative full-structured systems (Example 3.10).

THEOREM 2.3.

(1) *Suppose that Σ^{full} and $\Sigma^{full\prime}$ are two noncommutative full-structured systems which are both full-controllable and full-observable. Then Σ^{full} and $\Sigma^{full\prime}$ are full-similar if and only if they realize the same transfer function:*

$$T_{\Sigma^{full}}(z) = T_{\Sigma^{full\prime}}(z).$$

(2) *The noncommutative full-structured system Σ^{full} is a full-minimal realization of its transfer function $T_{\Sigma^{full}}(z)$ if and only if Σ^{full} is both full-controllable and full-observable.*

(3) *Suppose that $T(z) = \sum_{w \in \mathcal{F}_{n,m}} T_w z^w$ is a formal power series in $n \cdot m$ noncommuting variables $z = (z_{11}, \ldots, z_{1m}; \ldots; z_{n1}, \ldots, z_{nm})$ with matrix coefficients $T_w \in \mathcal{L}(\mathcal{U}, \mathcal{Y})$. Then $T(z)$ can be realized as the transfer function $T(z) = T_{\Sigma^{full}}(z)$ of a finite-dimensional noncommutative full-structured system Σ^{full} if and only if the associated Hankel matrix*

$$\mathbb{H}^T = [T_{v \cdot (i,i_N) \cdot w}]_{i \in \{1,\ldots,n\}, v \in \mathcal{F}_{n,m}; (1,j_N) \cdot w \in \mathcal{F}_{n,m} \setminus \{\emptyset\}}$$

has finite rank for $i = 1, \ldots, n$. In this case there is a canonical construction (shift realization) of a minimal realization with state-space $\mathcal{H}$ having $\dim \mathcal{H} = \operatorname{rank} \mathbb{H}^T$.

3. Structured noncommutative multidimensional linear systems: Definition and basic properties. Our general notion of structured noncommutative

multidimensional linear system (SNMLS) will be associated with a graph G. As is standard, a graph G consists of a set of vertices V together with a set of edges E. Each edge $e \in E$ connects a source vertex $\mathbf{s}(e)$ (where $\mathbf{s}\colon E \to V$ is the *source map*) to a range vertex $\mathbf{r}(e)$ (where $\mathbf{r}\colon E \to V$ is the *range map*). We assume throughout that V and E are *finite* sets. For our application to SNMLSs, we require a few additional properties, encoded in the following definition of an *admissible graph*.

DEFINITION 3.1. *We say that the graph $G = (V, E, \mathbf{s}\colon E \to V, \mathbf{r}\colon E \to V)$ is an admissible graph if*

(1) *the set of vertices V of G has a disjoint partitioning $V = S \dot\cup R$ into two subsets S and R such that each edge e of G has source vertex $\mathbf{s}(e) \in S$ and range vertex $\mathbf{r}(e) \in R$;*
(2) *for a given $s \in S$ and $r \in R$ there is at most one edge $e \in E$ connecting s to r (i.e., at most one edge e with $\mathbf{s}(e) = s$ and $\mathbf{r}(e) = r$);*
(3) *each pathwise-connected component G_k of G is a nondegenerate complete bipartite graph; i.e., the vertices of G_k have a partitioning $V(G_k) = S_k \dot\cup R_k$ (with $S_k \subset S$, $R_k \subset R$ and both $S_k \neq \emptyset$ and $R_k \neq \emptyset$) such that for each pair (s, r) with $s \in S_k$ and $r \in R_k$ there is exactly one edge $e \in E$ with $\mathbf{s}(e) = s$ and $\mathbf{r}(e) = r$.*

In other words, conditions (1) and (2) say that G is a *bipartite* graph. Thus admissible graphs amount to bipartite graphs having connected path components which are complete bipartite subgraphs. Thus the set of edges E can be identified with a subset of the Cartesian product $S \times R$, where S and R are called the source vertices and range vertices, respectively.

Admissible graphs G have the following intrinsic characterization.

THEOREM 3.2. *Suppose that we are given finite disjoint sets S, R, and E together with mappings $\mathbf{s}\colon E \to S$ and $\mathbf{r}\colon E \to R$. Associated with these data is a graph G defined as follows: the vertex set of G is $V := S \cup R$, and there exists an edge connecting v to v' if and only if there is an $e \in E$ either with $v = \mathbf{s}(e), v' = \mathbf{r}(e)$ or with $v' = \mathbf{s}(e), v = \mathbf{r}(e)$. Then G is admissible in the sense of Definition* 3.1 *if and only if the following conditions hold:*

(1) *$\mathbf{s}\colon E \to S$ is surjective.*
(2) *$\mathbf{r}\colon E \to R$ is surjective.*
(3) *The map $\mathbf{s} \times \mathbf{r}\colon E \to S \times R$ given by*

$$\mathbf{s} \times \mathbf{r}\colon e \mapsto (\mathbf{s}(e), \mathbf{r}(e))$$

is injective.
(4) *Whenever e_1, e_2, and e_3 are elements of E with $\mathbf{r}(e_1) = \mathbf{r}(e_2)$ and $\mathbf{s}(e_1) = \mathbf{s}(e_3)$, then there is an edge e_4 in E, with $\mathbf{s}(e_4) = \mathbf{s}(e_2)$, and $\mathbf{r}(e_4) = \mathbf{r}(e_3)$.*

Proof. Let G be an admissible graph with pathwise-connected components equal to the subgraphs $G_1, \dots, G_K$. Since each G_k is a complete bipartite graph by assumption, we have that the vertex set $V(G_k)$ has a disjoint partitioning $V(G_k) = S_k \dot\cup R_k$ for nonempty subsets $S_k \subset S$ and $R_k \subset R$, and the edge set $E(G_k)$ of G_k can be identified with the Cartesian product $S_k \times R_k$ (with $\mathbf{s}(s, r) = s$ and $\mathbf{r}(s, r) = r$ for $s \in S_k$ and $r \in R_k$). As $\mathbf{s}$ maps $E(G_k)$ onto S_k and $\mathbf{r}$ maps $E(G_k)$ onto R_k for each $k = 1, \dots, K$, we see that $\mathbf{s}$ maps E onto S and $\mathbf{r}$ maps E onto R. Condition (2) in Definition 3.1 says that $\mathbf{s} \times \mathbf{r}$ is injective on E. Finally, suppose that $e_1, e_2, e_3 \in E$ as in condition (4). Then $\mathbf{r}(e_1) = \mathbf{r}(e_2) = r$ implies that $\mathbf{s}(e_1)$ and $\mathbf{s}(e_2)$ are in the same pathwise-connected component S_i of G. On the other hand, $\mathbf{s}(e_1) = \mathbf{s}(e_3)$ implies that $\mathbf{s}(e_3)$ is also in S_i and $\mathbf{r}(e_3) \in R_i$. The assumption that the pathwise-connected

component G_i is a complete bipartite graph implies that there is an edge e_4 connecting $\mathbf{s}(e_2)$ to $\mathbf{r}(e_3)$.

Conversely, suppose that G arises from source vertex function $\mathbf{s}\colon E \to S$ and range vertex function $\mathbf{r}\colon E \to R$ satisfying conditions (1)–(4) as in the statement of the theorem. By definition, the vertex set V is partitioned into two disjoint subsets S and R such that each edge of G connects an element of S with an element of R or vice versa; i.e., Definition 3.1(1) holds. Condition (3) in Theorem 3.2 gives Definition 3.1(2). Suppose that $s \in S$ and $r \in R$ are in the same pathwise-connected component of the graph G. By the bipartite structure of G, this means that there is a path $e_1 e_2 \cdots e_{2N-1}$ (necessarily of odd length) connecting s to r:

$$\mathbf{s}(e_1) = s,\ \mathbf{r}(e_1) = \mathbf{s}(e_2),\ \mathbf{r}(e_2) = \mathbf{s}(e_3),\ \ldots,\ \mathbf{r}(e_{2N-2}) = \mathbf{s}(e_{2N-1}),\ \mathbf{r}(e_{2N-1}) = r.$$

Without loss of generality we may suppose that we have chosen the shortest such path. If $N > 1$, we may use condition (4) to produce a shorter path connecting s to r. Hence it must be the case that $N = 1$ and the path consists of a single edge $e \in E$ connecting s to r and hence (s, r). Thus if $s \in S$ and $r \in R$ are connected by a path of G, then they are connected by a path of length 1. Condition (1) in the theorem implies that every $s \in S$ is connected to some $r \in R$. We conclude that each pathwise-connected component of G is a complete bipartite graph; i.e., Definition 3.1(3) is satisfied, and the theorem follows. □

If e is an edge in the admissible graph G, then we have the notation $\mathbf{s}(e)$ for the source vertex of e, and $\mathbf{r}(e)$ for the range vertex of e. Conversely, given an $s \in S$ and a $r \in R$, there is an edge e connecting s to r (i.e., $e \in E$ with $\mathbf{s}(e) = s$ and $\mathbf{r}(e) = r$); exactly one s and r are in the same path-connected component p of G. For v any vertex of G (either a source vertex or a range vertex) we shall let $[v]$ denote the path-connected component containing v. Thus s and r are in the same path-connected component exactly when $[s] = [r]$. When this is the case, by the admissibility axioms the edge e connecting s to r is unique. We shall denote this edge by $e_{s,r}$:

$$e_{s,r} \text{ determined by } \mathbf{s}(e_{s,r}) = s \text{ and } \mathbf{r}(e_{s,r}) = r. \tag{3.1}$$

Note that $e_{s,r}$ is defined for $s \in S$ and $r \in R$ exactly when $[s] = [r]$.

We associate with each admissible graph G a linear form in noncommuting indeterminates $z = (z_e : e \in E)$ indexed by the edge set E of G, as follows. For each $e \in E$, define a matrix $I_{G,e} = [I_{G,e;s,r}]_{s\in S, r\in R}$ (with rows indexed by S and columns indexed by R) with matrix entries given by

$$I_{G,e;s,r} = \begin{cases} 1 & \text{if } (s,r) = (\mathbf{s}(e), \mathbf{r}(e)), \\ 0 & \text{otherwise.} \end{cases} \tag{3.2}$$

We then define the *structure matrix* $Z_G(z)$ associated with each admissible graph G to be the linear form in the noncommuting indeterminates $z = (z_e : e \in E)$ given by

$$Z_G(z) = \sum_{e\in E} I_{G,e} z_e.$$

We are now ready to give examples of admissible graphs with their associated structure matrices in connection with certain well-known noncommutative multidimensional linear models. We refer to [34] for further details on the motivation and construction of these models.

Example 3.3 (noncommutative Fornasini–Marchesini structure matrix). In this case, we take the admissible graph G^{FM} to be a complete bipartite graph having only one source vertex. Thus we take $S^{FM} = \{1\}$, and $R^{FM} = E^{FM} = \{1, \dots, d\}$ with $\mathbf{s}^{FM}(i) = 1$, $\mathbf{r}^{FM}(i) = i$, i.e., $n = 1, m = d$. Thus we have

$$I_{G^{FM},i} = \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{bmatrix},$$

where 1 is located in the ith slot. Thus, the structure matrix for the noncommutative Fornasini–Marchesini case is simply given by

$$Z_{G^{FM}}(z) = \sum_{i=1}^{d} I_{G^{FM},i} z_i = \begin{bmatrix} z_1 & \cdots & z_d \end{bmatrix} =: Z_{\mathrm{row}}(z).$$

Example 3.4 (noncommutative Givone–Roesser structure matrix). In this case, we take the admissible graph G^{GR} to have d path-connected components, with each path-connected component containing only one source and one range vertex. Thus, we take $S^{GR} = R^{GR} = E^{GR} = \{1, \dots, d\}$ with $\mathbf{s}^{GR}(i) = i$, $\mathbf{r}^{GR}(i) = i$, and thus $n = d = m$. We then have

$$I_{G^{GR},i} = \begin{bmatrix} 0 & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & 1 & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \end{bmatrix},$$

where 1 is located at the (i, i)th entry. Therefore, the structure matrix for the noncommutative Givone–Roesser case has the diagonal form

$$Z_{G^{GR}}(z) = \sum_{i=1}^{d} z_i I_{G^{GR},i} = \begin{bmatrix} z_1 & & \\ & \ddots & \\ & & z_d \end{bmatrix} := Z_{\mathrm{diag}}(z).$$

Example 3.5 (full matrix block structure matrix). In this case, we take G^{full} to be a general finite, complete bipartite graph. Thus we take $S = \{1, \dots, n\}$, $R = \{1, \dots, m\}$, and $E = \{(i, j) : i \in S, j \in R\}$ with $\mathbf{s}^{\mathrm{full}}(i, j) = i$, $\mathbf{r}^{\mathrm{full}}(i, j) = j$, where $d = nm$. Then we have

$$I_{G^{\mathrm{full}},(i,j)} = \begin{bmatrix} 0 & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & 1 & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \end{bmatrix},$$

where 1 is located at the (i, j)th entry. Thus the structure matrix for this case has the full-block structure

$$Z_{G^{\mathrm{full}}}(z) = \begin{bmatrix} z_{1,1} & \cdots & z_{1,m} \\ \vdots & & \vdots \\ z_{n,1} & \cdots & z_{n,m} \end{bmatrix} =: Z_{\mathrm{full}}(z).$$

Note that Example 3.3 amounts to the special case of this example where $n = 1$.

Example 3.6 (the general structure matrix). Suppose that the admissible graph G has path-connected components G_k with source vertices $S_k = \{(k,1),\dots,(k,n_k)\}$, range vertices $R_k = \{(k,1),\dots,(k,m_k)\}$, and edge sets $E_k = \{(k,i,j) : 1 \le i \le n_k, 1 \le j \le m_k\}$ for $k = 1,\dots,K$. Define a graph G to have source vertex set

$$S = \cup_{k=1}^{K} S_k = \{(k,i) : 1 \le k \le K, 1 \le i \le n_k\},$$

range vertex set

$$R = \cup_{k=1}^{K} R_k = \{(k,j) : 1 \le k \le K, 1 \le j \le m_k\},$$

and edge set

$$E = \cup_{k=1}^{K} E_k = \{(k,i,j) : 1 \le k \le K, 1 \le i \le n_k, 1 \le j \le m_k\},$$

with $\mathbf{s}(k,i,j) = (k,i)$, $\mathbf{r}(k,i,j) = (k,j)$ for $(k,i,j) \in E$. Then the associated structure matrix $Z_G(z)$ is given by

$$Z_G(z) = \begin{bmatrix} Z_{\text{full},1}(z^1) & & \\ & \ddots & \\ & & Z_{\text{full},K}(z^K) \end{bmatrix},$$

where we let z^k denote the $(n_k \cdot m_k)$-tuple of variables $z^k = (z_{k,i,j} : 1 \le i \le n_k; 1 \le j \le m_k)$ and where

$$Z_{\text{full},k}(z^k) = \begin{bmatrix} z_{k,1,1} & \cdots & z_{k,1,m_k} \\ \vdots & & \vdots \\ z_{k,n_k,1} & \cdots & z_{k,n_k,m_k} \end{bmatrix}$$

is as in Example 3.5 for $k = 1,\dots,K$. By the definition of an admissible graph as a graph with path-connected components equal to complete bipartite graphs, we see that this example amounts to the general case.

To define an SNMLS, in addition to an admissible graph we require a collection of finite-dimensional linear spaces $\mathcal{H}_p$ indexed by each path-connected component p of G. We often abbreviate the whole collection simply by

$$\mathcal{H} = \{\mathcal{H}_p : p \in P(G)\},$$

where $P(G)$ denotes the set of path-connected components of G. In general, for $v \in V$ we use the notation $[v]$ to denote the path-connected component of G containing v (whether v be in S or in R). Thus, for each $s \in S$ and $r \in R$ we have associated finite-dimensional linear spaces $\mathcal{H}_{[s]}$ and $\mathcal{H}_{[r]}$, which are distinct only for s and r in distinct path-connected components of G. In addition, we need a *connection matrix* or *colligation*

(3.3)
$$U = \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} [A_{r,s}]_{r\in R, s\in S} & \operatorname{col}_{r\in R}[B_r] \\ \operatorname{row}_{s\in S}[C_s] & D \end{bmatrix} : \begin{bmatrix} \operatorname{col}_{s\in S}\mathcal{H}_{[s]} \\ \mathcal{U} \end{bmatrix} \to \begin{bmatrix} \operatorname{col}_{r\in R}\mathcal{H}_{[r]} \\ \mathcal{Y} \end{bmatrix},$$

where $\mathcal{U}$ and $\mathcal{Y}$ are linear spaces, here taken also to be finite-dimensional, called the input-space and output-space, respectively.

We now introduce our notion SNMLS.

DEFINITION 3.7. *By an* SNMLS*, we mean a collection of objects*

$$\Sigma = (G,\ \mathcal{H},\ U), \tag{3.4}$$

where

(1) *G is an admissible graph (called the* structure graph *of Σ),*
(2) *$\mathcal{H} = \{\mathcal{H}_p : p \in P(G)\}$ is a collection of finite-dimensional spaces $\mathcal{H}_p$ (called the* auxiliary state-spaces *of Σ), and*
(3) *U is a matrix of the form* (3.3) *(called the* connection matrix *or* colligation *of Σ).*

With any SNMLS we associate an i/s/o linear system with evolution along a free semigroup as follows. We denote by $\mathcal{F}_E$ the free semigroup generated by the edge set E. An element of $\mathcal{F}_E$ is then a word w of the form $w = e_N \cdots e_1$, where each e_r is an edge of G for $r = 1, \dots, N$. We denote the empty word (consisting of no letters) by $\emptyset$. The semigroup operation is concatenation: if $w = e_N \cdots e_1$ and $w' = e'_{N'} \cdots e'_1$, then ww' is defined to be

$$ww' = e_N \cdots e_1 e'_{N'} \cdots e'_1.$$

Note that the empty word $\emptyset$ acts as the identity element for this semigroup. Equivalently, we may view $\mathcal{F}_E$ as a homogeneous tree of degree $\#E + 1$ (where $\#E$ is the number of edges of G) with root $\emptyset$; this point of view appears in the "multiscale system theory" in [14].

If $\Sigma = (G,\ \mathcal{H},\ U)$ is an SNMLS, we associate the system equations (with evolution along $\mathcal{F}_E$)

$$\Sigma\colon \begin{cases} x_{\mathbf{s}(e)}(ew) &= \Sigma_{s\in S} A_{\mathbf{r}(e),s} x_s(w) + B_{\mathbf{r}(e)} u(w), \\ x_{s'}(ew) &= 0 \quad \text{if } s' \neq \mathbf{s}(e), \\ y(w) &= \Sigma_{s\in S} C_s x_s(w) + Du(w). \end{cases} \tag{3.5}$$

Here the *state-vector* $x(w)$ at position w (for $w \in \mathcal{F}_E$) has the form of a column vector

$$x(w) = \operatorname{col}_{s\in S}\, x_s(w),$$

with column entries indexed by the source vertices $s \in S$ and with column entry $x_s(w) \in \mathcal{H}_{[s]}$ (thus $x(w) \in \oplus_{s\in S}\mathcal{H}_{[s]}$), while $u(w) \in \mathcal{U}$ denotes the *input* at position w and $y(w) \in \mathcal{Y}$ denotes the *output* at position w. Just as in the classical case, if we specify an initial condition $x(\emptyset) \in \oplus_{s\in S}\mathcal{H}_{[s]}$ and feed in an input string $\{u(w)\}_{w\in\mathcal{F}_E}$, then (3.5) enables us to recursively compute $x(w)$ for all $w \in \mathcal{F}_E \setminus \{\emptyset\}$ and $y(w)$ for all $w \in \mathcal{F}_E$.

As these systems include the full-structured case discussed in section 2.3 as a special case (see Example 3.10 below) where some redundancy occurs in the state-vector of a system trajectory (see (2.18)), in general some redundancy in the state-vector occurs for trajectories of a general SNMLS Σ as well. Indeed, the analogue of (2.19) for this more general setting is the following: *if $\{x(w)\}_{w\in\mathcal{F}_E} = \{\operatorname{col}_{s\in S}[x_s(w)]\}_{w\in\mathcal{F}_E}$ is the state trajectory solving the state-update equation in* (3.5) *for some choice of input signal $\{u(w)\}_{w\in\mathcal{F}_E}$, then necessarily, for each fixed $r \in R$ and $w \in \mathcal{F}_E$,*

$$x_s(e_{s,r}w) \text{ is independent of } s \text{ for all } s \text{ with } [s] = [r]. \tag{3.6}$$

It will be convenient for purposes of the matrix manipulations to come that we maintain the form (3.5) of the system equations rather than rewriting them in a more economical form.

The solution of these recursions can be made more explicit as follows. Note first of all that a consequence of the system equations is that

$$x(ew) \in \mathcal{H}_{\mathbf{s}(e)} := \mathrm{col}_{s\in S}[\boldsymbol{\delta}_{s,\mathbf{s}(e)}\mathcal{H}_{[\mathbf{s}(e)]}] \quad \text{for all } e \in E \text{ and } w \in \mathcal{F}_E$$

(where $\boldsymbol{\delta}_{s,s'}$ is the Kronecker delta function). Given $x(\emptyset)$ and $\{u(w)\}_{w\in\mathcal{F}_E}$, we can solve the system equations (3.5) or (3.10) uniquely for $\{x(w)\}_{w\in\mathcal{F}_E\setminus\{\emptyset\}}$ and $\{y(w)\}_{w\in\mathcal{F}_E}$ as follows:

$$\begin{aligned} x_{\mathbf{s}(e_N)}(e_N\cdots e_1) &= \sum_{s\in S} A_{\mathbf{r}(e_N),\mathbf{s}(e_{N-1})}A_{\mathbf{r}(e_{N-1}),\mathbf{s}(e_{N-2})}\cdots A_{\mathbf{r}(e_1),s}x_s(\emptyset) \\ &\quad + \sum_{r=1}^{N} A_{\mathbf{r}(e_N),\mathbf{s}(e_{N-1})}\cdots A_{\mathbf{r}(e_{r+1}),\mathbf{s}(e_r)}B_{\mathbf{r}(e_r)}u(e_{r-1}\cdots e_1), \end{aligned} \tag{3.7}$$

where we interpret $u(e_{r-1}\cdots e_1)$ to be $u(\emptyset)$ when $r = 1$, and where we set

$$x_s(e_Ne_{N-1}\cdots e_1) = 0 \quad \text{if } s \neq \mathbf{s}(e_N).$$

Also,

$$\begin{aligned} y(e_N\cdots e_1) &= \sum_{s\in S} C_{\mathbf{s}(e_N)}A_{\mathbf{r}(e_N),\mathbf{s}(e_{N-1})}A_{\mathbf{r}(e_{N-1}),\mathbf{s}(e_{N-2})}\cdots A_{\mathbf{r}(e_1),s}x_s(\emptyset) \\ &\quad + \sum_{r=1}^{N} C_{\mathbf{s}(e_N)}A_{\mathbf{r}(e_N),\mathbf{s}(e_{N-1})}\cdots A_{\mathbf{r}(e_{r+1}),\mathbf{s}(e_r)}B_{\mathbf{r}(e_r)}u(e_{r-1}\cdots e_1) + Du(e_N\cdots e_1). \end{aligned} \tag{3.8}$$

This formula must be interpreted appropriately for special cases. As examples, for the particular cases $r = 1$ and $r = N$ we have the interpretations

$$\begin{aligned} &A_{\mathbf{r}(e_N),\mathbf{s}(e_{N-1})}\cdots A_{\mathbf{r}(e_{r+1}),\mathbf{s}(e_r)}B_{\mathbf{r}(e_r)}u(e_{r-1}\cdots e_1)|_{r=1} \\ &\quad = A_{\mathbf{r}(e_N),\mathbf{s}(e_{N-1})}\cdots A_{\mathbf{r}(e_2),\mathbf{s}(e_1)}B_{\mathbf{r}(e_1)}u(\emptyset), \\ &A_{\mathbf{r}(e_N),\mathbf{s}(e_{N-1})}\cdots A_{\mathbf{r}(e_{r+1}),\mathbf{s}(e_r)}B_{\mathbf{r}(e_r)}u(e_{r-1}\cdots e_1)|_{r=N} = B_{\mathbf{r}(e_N)}u(e_{N-1}\cdots e_1). \end{aligned}$$

If we set

$$\Delta_e = i_{\mathbf{s}(e)}A_{\mathbf{r}(e),\cdot}\colon \oplus_{s\in S}\mathcal{H}_{[s]} \to \oplus_{s\in S}\mathcal{H}_{[s]},$$

where i_s denotes the natural injection $h \mapsto \mathrm{col}_{s'\in S}[\delta_{s',s}h]$ of $\mathcal{H}_{[s]}$ into $\oplus_{s'\in S}\mathcal{H}_{[s']}$, and if we use our assumption that $x_{s'}(ew) = 0$ if $s' \neq \mathbf{s}(e)$, then (3.7) and (3.8) can be rewritten as

$$\begin{aligned} x(w) &= \Delta^w x(\emptyset) + \sum_{w',w''\in\mathcal{F}_E,e\in E:\ w'ew''=w} \Delta^{w'}i_{\mathbf{s}(e)}B_{\mathbf{r}(e)}u(w''), \\ y(w) &= C\Delta^w x(\emptyset) + \sum_{w',w''\in\mathcal{F}_E,e\in E:\ w'ew''=w} C\Delta^{w'}i_{\mathbf{s}(e)}B_{\mathbf{r}(e)}u(w'') + Du(w), \end{aligned} \tag{3.9}$$

where we use the noncommutative functional calculus

$$\Delta^v = \Delta_{e_N}\Delta_{e_{N-1}}\cdots\Delta_{e_1} \quad \text{if } v = e_N e_{N-1}\cdots e_1 \in \mathcal{F}_E, \qquad \Delta^{\emptyset} = I_{\mathcal{H}}.$$

The system equations (3.5) can also be written more compactly in operator-theoretic form as

$$\Sigma\colon \begin{cases} x(ew) &= I_{\Sigma,e}Ax(w) + I_{\Sigma,e}Bu(w), \\ y(w) &= Cx(w) + Du(w), \end{cases} \tag{3.10}$$

where $I_{\Sigma;e}$ is a higher-multiplicity version of the coefficient matrices $I_{G,e}$ appearing in (3.2):

$$I_{\Sigma;e}\colon \oplus_{r\in R}\mathcal{H}_{[r]} \to \oplus_{s\in S}\mathcal{H}_{[s]}$$

with matrix entries $[I_{\Sigma;e}]_{s\in S, r\in R}$ given by

$$[I_{\Sigma;e}]_{s,r} = \begin{cases} I_{\mathcal{H}_{[\mathbf{s}(e)]}} = I_{\mathcal{H}_{[\mathbf{r}(e)]}} & \text{if } s = \mathbf{s}(e) \text{and } r = \mathbf{r}(e), \\ 0 & \text{otherwise.} \end{cases} \tag{3.11}$$

Also, just as in the classical case, it is convenient to introduce "frequency-domain" notation for explicit representation of system trajectories. For any linear space $\mathcal{H}$, we define the formal noncommutative Z-transform of a sequence of $\mathcal{H}$-valued functions as a formal power series in several noncommuting indeterminates $z = (z_e : e \in E)$ as follows:

$$\{h(w)\}_{w\in\mathcal{F}_E} \mapsto \widehat{h}(z) = \sum_{w\in\mathcal{F}_E} h(w)z^w, \tag{3.12}$$

where $z^{\emptyset} = 1$, $z^w = z_{e_N}z_{e_{N-1}}\cdots z_{e_1}$ if $w = e_N e_{N-1}\cdots e_1$. Then, applying the Z-transform to (3.10) gives

$$\sum_{w\in\mathcal{F}_E} x(ew)z^w = I_{\Sigma,e}A\widehat{x}(z) + I_{\Sigma,e}B\widehat{u}(z). \tag{3.13}$$

Multiply (3.13) on the left by z_e to get

$$\sum_{w\in\mathcal{F}_E} x(ew)z^{ew} = z_e I_{\Sigma,e}A\widehat{x}(z) + z_e I_{\Sigma,e}B\widehat{u}(z). \tag{3.14}$$

Summing (3.14) over all edges $e \in E$, we get

$$\sum_{e\in E}\sum_{w\in\mathcal{F}_E} x(ew)z^{ew} = Z_\Sigma(z)A\widehat{x}(z) + Z_\Sigma(z)B\widehat{u}(z), \tag{3.15}$$

where we have set

$$Z_\Sigma(z) = \sum_{e\in E} z_e I_{\Sigma,e}. \tag{3.16}$$

Note that the definition of the formal Z-transform yields

$$\sum_{e\in E}\sum_{w\in\mathcal{F}_E} x(ew)z^{ew} = \widehat{x}(z) - x(\emptyset).$$

Thus (3.15) becomes

$$\widehat{x}(z) = x(\emptyset) + Z_\Sigma(z)A\widehat{x}(z) + Z_\Sigma(z)B\widehat{u}(z). \tag{3.17}$$

Solving (3.17) for $\widehat{x}(z)$, we obtain

$$\widehat{x}(z) = (I - Z_\Sigma(z)A)^{-1}\, x(\emptyset) + (I - Z_\Sigma(z)A)^{-1}\, Z_\Sigma(z)B\widehat{u}(z). \tag{3.18}$$

Substitution of (3.17) into the formal Z-transform of the output equation of (3.10) then gives

$$\begin{aligned} \widehat{y}(z) &= C\widehat{x}(z) + D\widehat{u}(z) \\ &= C\,(I - Z_\Sigma(z)A)^{-1}\, x(\emptyset) + T_\Sigma(z)\widehat{u}(z), \end{aligned} \tag{3.19}$$

where we have set

$$T_\Sigma(z) = D + C(I - Z_\Sigma(z)A)^{-1}Z_\Sigma(z)B \tag{3.20}$$

equal to the *transfer function* of the SNMLS Σ, where the inverse is taken in the algebra $\mathcal{L}(\oplus_{s\in S}\mathcal{H}_{[s]})\langle\langle z\rangle\rangle$ of formal power series with operator coefficients in the noncommuting variables $z = (z_e : e \in E)$. We can write $T_\Sigma(z)$ explicitly as a formal power series in the form

$$\begin{aligned} T_\Sigma(z) = T_\emptyset + \sum_{N=1}^{\infty} \sum_{e_1,\dots,e_N\in E} & C_{\mathbf{s}(e_N)} \\ & \cdot A_{\mathbf{r}(e_N),\mathbf{s}(e_{N-1})} \cdots A_{\mathbf{r}(e_2),\mathbf{s}(e_1)} B_{\mathbf{r}(e_1)} z_{e_N} z_{e_{N-1}} \cdots z_{e_2} z_{e_1}. \end{aligned} \tag{3.21}$$

Example 3.8 (noncommutative Fornasini–Marchesini system). Here we continue Example 3.3. As the structure graph G is connected in this case, we assume that we are given a single finite-dimensional linear space $\mathcal{H}$ together with an input-space $\mathcal{U}$ and an output-space $\mathcal{Y}$. Then the structure matrix (3.16) $Z_{FM}(z)$ is the row matrix

$$Z_{\Sigma^{FM}}(z) = \sum_{j=1}^{d} z_j I_{\Sigma^{FM},j} = \begin{bmatrix} z_1 I_\mathcal{H} & \cdots & z_d I_\mathcal{H} \end{bmatrix} =: Z_{\mathrm{row}}(z) \otimes I_\mathcal{H},$$

where

$$I_{\Sigma^{FM},j} = \begin{bmatrix} 0 & \cdots & 0 & I_\mathcal{H} & 0 & \cdots & 0 \end{bmatrix}$$

(with nonzero entry in the jth column), and the connection matrix U^{FM} has the form

$$U^{FM} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} \mathrm{col}_{j=1,\dots,d}[A_j] & \mathrm{col}_{j=1,\dots,d}[B_j] \\ C & D \end{bmatrix} : \begin{bmatrix} \mathcal{H} \\ \mathcal{U} \end{bmatrix} \to \begin{bmatrix} \oplus_{j=1}^d \mathcal{H} \\ \mathcal{Y} \end{bmatrix}. \tag{3.22}$$

Thus, $I_{\Sigma^{FM},j}A = A_j$, $I_{\Sigma^{FM},j}B = B_j$, and therefore the associated noncommutative Fornasini–Marchesini system is given by

$$\Sigma^{FM}: \begin{cases} x(1w) = A_1 x(w) + B_1 u(w), \\ \qquad\vdots \\ x(dw) = A_d x(w) + B_d u(w), \\ y(w) \;\; = Cx(w) + Du(w), \end{cases} \tag{3.23}$$

i.e., we are in the setting of the noncommutative Fornasini–Marchesini systems discussed in section 2.1. Since in this case $Z_{\Sigma^{FM}}(z)A = \sum_{i=1}^{d} z_i A_i$ and similarly $Z_{\Sigma^{FM}}(z)B = \sum_{i=1}^{d} z_i B_i$, the transfer function $T_{\Sigma^{FM}}(z)$ in (3.20) for the noncommutative Fornasini–Marchesini system has the form given in (2.4).

We remark that any SNMLS can be embedded into a noncommutative Fornasini–Marchesini system having a certain special form as follows. Given a general SNMLS $\Sigma = (G, \mathcal{H}, U)$, we associate a Fornasini–Marchesini system

$$\Sigma^{FM} = (G^{FM}, \mathcal{H}^{FM}, U^{FM})$$

as follows. We let G^{FM} be the unique Fornasini–Marchesini graph having the same edge set as G: $E^{FM} = E$. Thus we take the source-vertex set S^{FM} to be $S^{FM} = \{1\}$, and the range-vertex set R^{FM} to be $R^{FM} = E$, with associated source and range vertex maps $\mathbf{s}^{FM}$ and $\mathbf{r}^{FM}$ given by $\mathbf{s}^{FM}(e) = 1$ and $\mathbf{r}^{FM}(e) = e$ for $e \in E$. We let $\mathcal{H}^{FM} = \oplus_{s \in S} \mathcal{H}$, and we define the connection matrix $U^{FM} = \left[\begin{smallmatrix} A^{FM} & B^{FM} \\ C^{FM} & D^{FM} \end{smallmatrix}\right]$ by

$$\begin{bmatrix} A^{FM} & B^{FM} \\ C^{FM} & D^{FM} \end{bmatrix} = \begin{bmatrix} \mathrm{col}_{e \in E}[A_e^{FM}] & \mathrm{col}_{e \in E}[B_e^{FM}] \\ C & D \end{bmatrix} : \begin{bmatrix} \mathcal{H}^{FM} \\ \mathcal{U} \end{bmatrix} \to \begin{bmatrix} \oplus_{e \in E} \mathcal{H}^{FM} \\ \mathcal{Y} \end{bmatrix}$$

and by

$$\begin{aligned} A_e^{FM} &= i_{\mathbf{s}(e)} A_{\mathbf{r}(e),\cdot} \colon \mathcal{H}^{FM} \to \mathcal{H}^{FM}, \\ B_e^{FM} &= i_{\mathbf{s}(e)} B_{\mathbf{r}(e)} \colon \mathcal{U} \to \mathcal{H}^{FM}, \\ C^{FM} &= C \colon \mathcal{H}^{FM} \to \mathcal{Y}, \\ D^{FM} &= D \colon \mathcal{U} \to \mathcal{Y}, \end{aligned}$$

where $i_{\mathbf{s}(e)} \colon \mathcal{H}_{[s]} \to \mathrm{col}_{s' \in S} \mathcal{H}_{[s']}$ is the natural injection $h \mapsto \mathrm{col}_{s' \in S} \delta_{s',s} h$. A consequence of formula (3.9) is that Σ *and* Σ^{FM} *associated in this way have the same system trajectories.*

Example 3.9 (noncommutative Givone–Roesser system). Here we continue Example 3.4. In this case the structure graph G has d connected components, so we assume that we give d auxiliary state-spaces $\mathcal{H}_1, \dots, \mathcal{H}_d$. The structure matrix (3.16) then has the diagonal form

$$Z_{\Sigma^{GR}}(z) = \sum_{j=1}^{d} I_{\Sigma^{GR},j} z_j = \begin{bmatrix} z_1 I_{\mathcal{H}_1} & & \\ & \ddots & \\ & & z_d I_{\mathcal{H}_d} \end{bmatrix} =: Z_{\mathrm{diag}}(z) \otimes I_{\mathcal{H}},$$

where $I_{\Sigma^{GR},j}$ is a $d \times d$ matrix with zero entries except at the (j,j)th entry, where $[I_{\Sigma^{GR},j}]_{j,j} = I_{\mathcal{H}_j}$, and the connecting matrix U^{GR} is of the form

(3.24)
$$U^{GR} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} [A_{j,i}]_{j,i=1,\dots,d} & \mathrm{col}_{j=1,\dots,d}[B_j] \\ \mathrm{row}_{i=1,\dots,d}[C_i] & D \end{bmatrix} : \begin{bmatrix} \oplus_{i=1}^{d} \mathcal{H}_i \\ \mathcal{U} \end{bmatrix} \to \begin{bmatrix} \oplus_{j=1}^{d} \mathcal{H}_j \\ \mathcal{Y} \end{bmatrix}.$$

Thus,

(3.25)
$$I_{\Sigma^{GR},i} A = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ A_{i,1} & \cdots & A_{i,d} \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix} \quad \text{and} \quad I_{\Sigma^{GR},i} B = \begin{bmatrix} 0 \\ \vdots \\ B_i \\ \vdots \\ 0 \end{bmatrix}$$

(where the nonzero row is row i in both expressions), and therefore the noncommutative Givone–Roesser system is given by

$$\Sigma^{GR}\colon \begin{cases} x_i(iw) &= \sum_{i'\in S} A_{i,i'}x_{i'}(w) + B_i u(w) \quad \text{for } e\in E,\\ x_{i''}(iw) &= 0 \quad \text{if } i''\neq i,\\ y(w) &= \sum_{i'=1}^{d} C_{i'}x_{i'}(w) + Du(w), \end{cases} \tag{3.26}$$

as stated in section 2.2. Here $x_i(iw)\in\mathcal{H}_i$ for $i=1,\dots,d$. The transfer function $T_{\Sigma^{GR}}(z)$ for the noncommutative Givone–Roesser system then has the form as given in (2.12).

Example 3.10 (noncommutative full-structured system). Here we continue Example 3.5. We assume that the structure matrix G has the form G^{full}, as in Example 3.5. As the structure graph G^{full} has only one connected component, we need specify only one auxiliary state-space $\mathcal{H}$ for an SNMLS $\Sigma=(G^{\text{full}},\mathcal{H},U)$ with structure graph G^{full}. The structure matrix (3.16) is the full-block operator matrix with each matrix entry containing one of the variables

$$Z_{\Sigma^{\text{full}}}(z)=\sum_{i=1}^{n}\sum_{j=1}^{m} I_{\Sigma^{\text{full}},(i,j)}z_{i,j}=\begin{bmatrix} z_{1,1}I_{\mathcal{H}} & \cdots & z_{1,m}I_{\mathcal{H}}\\ \vdots & & \vdots\\ z_{n,1}I_{\mathcal{H}} & \cdots & z_{n,m}I_{\mathcal{H}}\end{bmatrix} =: Z_{\text{full}}(z)\otimes I_{\mathcal{H}},$$

where $I_{\Sigma^{\text{full}},(i,j)}$ is an $n\times m$ matrix with zero entries except at the (i,j)th entry, where $[I_{\Sigma^{\text{full}},(i,j)}]_{i,j}=I_{\mathcal{H}}$. The connecting operator U^{full} in this case is given by

$$U^{\text{full}}=\begin{bmatrix} A & B\\ C & D\end{bmatrix}:\begin{bmatrix}\oplus_1^n\mathcal{H}\\ \mathcal{U}\end{bmatrix}\to\begin{bmatrix}\oplus_1^m\mathcal{H}\\ \mathcal{Y}\end{bmatrix},$$

where

$$A=\begin{bmatrix} A_{1,1} & \cdots & A_{1,n}\\ \vdots & & \vdots\\ A_{m,1} & \cdots & A_{m,n}\end{bmatrix},\quad B=\begin{bmatrix}B_1\\ \vdots\\ B_m\end{bmatrix},\quad C=\begin{bmatrix}C_1 & \cdots & C_n\end{bmatrix},$$

and the system equations (3.5) assume the form

$$\Sigma^{\text{full}}\colon \begin{cases} x_i((i,j)\cdot w) &= \sum_{i'=1}^{n} A_{j,i'}x_{i'}(w) + B_j u(w),\\ x_{i''}((i,j)\cdot w) &= 0 \quad \text{if } i''\neq i,\\ y(w) &= \sum_{i'=1}^{n} C_{i'}x_{i'}(w) + Du(w), \end{cases} \tag{3.27}$$

and we are in the setting of the noncommutative full-structured systems discussed in section 2.3. The transfer function $T_{\Sigma^{\text{full}}}(z)$ for the full-block operator matrix case then has the form as in (2.21).

Example 3.11 (the general SNMLS system). Here we continue Example 3.6. Suppose that the admissible graph G is the union of complete bipartite graphs G_k with source-vertex set $S_k=\{(k,i)\colon 1\le i\le n_k\}$, range-vertex set $R_k=\{(k,j)\colon 1\le j\le m_k\}$, and edge set $E_k=\{(k,i,j)\colon 1\le i\le n_k; 1\le j\le m_k\}$ for $k=1,\dots,K$. Note that $k=1,\dots,K$ labels the set P of path-connected components of G. Let $\mathcal{H}=\{\mathcal{H}_k\colon k=1,\dots,K\}$ denote a specification of a finite-dimensional linear space for each path-connected component $k=1,\dots,K$, and suppose that $\Sigma=(G,\mathcal{H},U)$ is an SNMLS with structure graph G. Then the connection matrix U has the form

$$U=\begin{bmatrix}[A_{k',k}] & [B_{k'}]\\ [C_k] & D\end{bmatrix}:\begin{bmatrix}\oplus_{k=1}^{K}[\oplus_{i=1}^{n_k}\mathcal{H}_k]\\ \mathcal{U}\end{bmatrix}\to\begin{bmatrix}\oplus_{k'=1}^{K}[\oplus_{j=1}^{m_{k'}}\mathcal{H}_{k'}]\\ \mathcal{Y}\end{bmatrix},$$

where each $A_{k',k}$, $B_{k'}$, and C_k in turn has the form

$$\begin{aligned}
&A_{k',k} = [A_{k',k;j,i}]_{j=1,\dots,m_{k'};i=1,\dots,n_k}, \quad \text{where } A_{k',k;j,i}\colon \mathcal{H}_k \to \mathcal{H}_{k'},\\
&B_{k'} = \operatorname{col}_{j=1,\dots,m_{k'}}[B_{k',j}], \quad \text{where } B_{k',j}\colon \mathcal{U} \to \mathcal{H}_{k'},\\
&C_k = \operatorname{row}_{i=1,\dots,n_k}[C_{k,i}], \quad \text{where } C_{k,i}\colon \mathcal{H}_k \to \mathcal{Y}.
\end{aligned}$$

The structure matrix $Z_\Sigma(z)$ has the block-diagonal form

$$Z_\Sigma(z) = \begin{bmatrix} Z_{\text{full},1}(z^1)\otimes I_{\mathcal{H}_1} & & \\ & \ddots & \\ & & Z_{\text{full},K}(z^K)\otimes I_{\mathcal{H}_K} \end{bmatrix},$$

where z^k is the collection of variables $z^k = \{z_{k,i,j} : i = 1,\dots,n_k; j = 1,\dots,m_k\}$ and each $Z_{\text{full},k}(z^k)\otimes I_{\mathcal{H}_k}$ is a full-block structure matrix (of block size $n_k \times m_k$), as in Example 3.10. While the structure matrix splits as the direct sum, the system trajectories for the whole system Σ in general can be quite complicated since there is no corresponding splitting for the A matrix generating the system dynamics.

If one substitutes general noncommuting operators $\delta = (\delta_{k,i,j}\colon k = 1,\dots,K; i = 1,\dots,n_k; j = 1,\dots,m_k)$ for the noncommuting formal variables $z_{k,i,j}$, then $Z_\Sigma(\delta)$ is the most general structure matrix coming up in μ-synthesis analysis (see [33]). Part of the advantage of the notion of SNMLS introduced here is the setting thereby given for proving results in the theory of μ-synthesis in a unified way for a general structure. We refer to [6] for further details.

4. System operations: Cascade/parallel connection and inversion. Suppose that we are given two SNMLSs

$$\Sigma'' = (G, \mathcal{H}'', U''), \qquad \Sigma' = (G, \mathcal{H}', U')$$

with the same structure graph G and with connection matrices

$$\begin{aligned}
U'' &= \begin{bmatrix} A'' & B'' \\ C'' & D'' \end{bmatrix} : \begin{bmatrix} \operatorname{col}_{s\in S}\mathcal{H}''_{[s]} \\ \mathcal{U}'' \end{bmatrix} \to \begin{bmatrix} \operatorname{col}_{r\in R}\mathcal{H}''_{[r]} \\ \mathcal{Y}'' \end{bmatrix},\\
U' &= \begin{bmatrix} A' & B' \\ C' & D' \end{bmatrix} : \begin{bmatrix} \operatorname{col}_{s\in S}\mathcal{H}'_{[s]} \\ \mathcal{U}' \end{bmatrix} \to \begin{bmatrix} \operatorname{col}_{r\in R}\mathcal{H}'_{[r]} \\ \mathcal{Y}' \end{bmatrix},
\end{aligned}$$

with the property that the output-space for U' coincides with the input-space for U'':

$$\mathcal{Y}' = \mathcal{U}''.$$

We then define the *cascade connection* $\Sigma = \Sigma''\circ\Sigma'$ of Σ'' with Σ' to be the SNMLS $\Sigma = (G, \mathcal{H}, U)$ with auxiliary state-spaces $\mathcal{H}_p$ given by $\mathcal{H}_p = \begin{bmatrix} \mathcal{H}''_p \\ \mathcal{H}'_p \end{bmatrix}$ and with colligation U given by

$$U = \begin{bmatrix} A & B \\ C & D \end{bmatrix} := \begin{bmatrix} A'' & B''C' & B''D' \\ 0 & A' & B' \\ C'' & D''C' & D''D' \end{bmatrix} : \begin{bmatrix} \operatorname{col}_{s\in S}\mathcal{H}''_{[s]} \\ \operatorname{col}_{s\in S}\mathcal{H}'_{[s]} \\ \mathcal{U}' \end{bmatrix} \to \begin{bmatrix} \operatorname{col}_{r\in R}\mathcal{H}''_{[r]} \\ \operatorname{col}_{r\in R}\mathcal{H}'_{[r]} \\ \mathcal{Y}'' \end{bmatrix}.$$

Here we have identified the space $\operatorname{col}_{s\in S}\begin{bmatrix} \mathcal{H}''_{[s]} \\ \mathcal{H}'_{[s]} \end{bmatrix}$ with $\begin{bmatrix} \operatorname{col}_{s\in S}\mathcal{H}''_{[s]} \\ \operatorname{col}_{s\in S}\mathcal{H}'_{[s]} \end{bmatrix}$ as well as $\operatorname{col}_{r\in R}\begin{bmatrix} \mathcal{H}''_{[r]} \\ \mathcal{H}'_{[r]} \end{bmatrix}$ with $\begin{bmatrix} \operatorname{col}_{r\in R}\mathcal{H}''_{[r]} \\ \operatorname{col}_{r\in R}\mathcal{H}'_{[r]} \end{bmatrix}$ in the natural way. In more detail, the colligation coefficients A, B,

C, D are given by

$$A_{r,s} = \begin{bmatrix} A''_{r,s} & B''_r C'_s \\ 0 & A'_{r,s} \end{bmatrix} : \begin{bmatrix} \mathcal{H}''_{[s]} \\ \mathcal{H}'_{[s]} \end{bmatrix} \to \begin{bmatrix} \mathcal{H}''_{[r]} \\ \mathcal{H}'_{[r]} \end{bmatrix}, \qquad B_r = \begin{bmatrix} B''_r D' \\ B'_r \end{bmatrix} : \mathcal{U}' \to \begin{bmatrix} \mathcal{H}''_{[r]} \\ \mathcal{H}'_{[r]} \end{bmatrix},$$

$$C_s = \begin{bmatrix} C''_s & D''C'_s \end{bmatrix} : \begin{bmatrix} \mathcal{H}''_{[s]} \\ \mathcal{H}'_{[s]} \end{bmatrix} \to \mathcal{Y}'', \qquad D = D''D' : \mathcal{U}' \to \mathcal{Y}''.$$

We note that the cascade connection $\Sigma = \Sigma'' \circ \Sigma'$ has the following interpretation. Suppose that we are given an initial condition $x'(\emptyset) = x'_0$ and an input string $\{u'(w)\}_{w \in \mathcal{F}_E}$ to generate a trajectory $\{u'(w), x'(w), y'(w)\}_{w \in \mathcal{F}_E}$ of Σ' via the system equations

$$\Sigma' : \begin{cases} x'_{\mathbf{s}(e)}(ew) &= \Sigma_{s \in S} A'_{\mathbf{r}(e),s} x'_s(w) + B'_{\mathbf{r}(e)} u'(w), \\ x'_{s'}(ew) &= 0 \quad \text{if } s' \neq \mathbf{s}(e), \\ y'(w) &= \Sigma_{s \in S} C'_s x'_s(w) + D' u'(w). \end{cases} \tag{4.1}$$

We then let $x''(\emptyset) = x''_0 \in \mathcal{H}''$ be arbitrary and set $u''(w) = y'(w)$ to generate a system trajectory $\{u''(w), x''(w), y''(w)\}_{w \in \mathcal{F}_E}$ of Σ'', via the system equations

$$\Sigma'' : \begin{cases} x''_{\mathbf{s}(e)}(ew) &= \Sigma_{s \in S} A''_{\mathbf{r}(e),s} x''_s(w) + B''_{\mathbf{r}(e)} u''(w), \\ x''_{s'}(ew) &= 0 \quad \text{if } s' \neq \mathbf{s}(e), \\ y''(w) &= \Sigma_{s \in S} C''_s x''_s(w) + D'' u''(w). \end{cases} \tag{4.2}$$

The resulting triple $\left\{u'(w), \left[\begin{smallmatrix} x''(w) \\ x'(w) \end{smallmatrix}\right], y''(w)\right\}_{w \in \mathcal{F}_E}$ then is a system trajectory of $\Sigma = \Sigma'' \circ \Sigma'$, and every system trajectory of $\Sigma'' \circ \Sigma'$ arises in this way.

The main result concerning cascade connection is that this is the state-space operation corresponding to multiplication of the corresponding transfer functions.

THEOREM 4.1. *Let Σ'' and Σ' be SNMLSs for which the cascade connection $\Sigma := \Sigma'' \circ \Sigma'$ is defined as above. Then the transfer function $T_\Sigma(z)$ for Σ is the product of the transfer functions $T_{\Sigma''}(z)$ and $T_{\Sigma'}(z)$ for Σ'' and Σ':*

$$T_{\Sigma'' \circ \Sigma'}(z) = T_{\Sigma''}(z) \cdot T_{\Sigma'}(z). \tag{4.3}$$

Proof. We have seen (see (3.19)) that the transfer function $T_\Sigma(z)$ is characterized by the property that

$$\widehat{y}(z) = T_\Sigma(z) \widehat{u}(z)$$

whenever $\{u(w), x(w), y(w)\}_{w \in \mathcal{F}_E}$ is a trajectory of Σ with $x(\emptyset) = 0$. By the interpretation for the cascade connection $\Sigma'' \circ \Sigma'$ given in the preceding paragraph, we know that $\{u(w), x(w), y(w)\}_{w \in \mathcal{F}_E}$ has the form $\left\{u'(w), \left[\begin{smallmatrix} x''(w) \\ x'(w) \end{smallmatrix}\right], y''(w)\right\}_{w \in \mathcal{F}_E}$, where

$$\{u'(w), x'(w), y'(w)\}_{w \in \mathcal{F}_E}$$

is a trajectory of Σ' with $x'(\emptyset) = 0$, where $\{u''(w), x''(w), y''(w)\}_{w \in \mathcal{F}_E}$ is a trajectory of Σ'' with $x''(\emptyset) = 0$, and where we impose the interconnection law $y'(w) = u''(w)$. It therefore follows that

$$\begin{aligned} \widehat{y}(z) = \widehat{y''}(z) &= T_{\Sigma''}(z) \widehat{y'}(z) \\ &= T_{\Sigma''}(z) \left(T_{\Sigma'}(z) \widehat{u'}(z) \right) \\ &= \left(T_{\Sigma''}(z) T_{\Sigma'}(z) \right) \widehat{u}(z), \end{aligned}$$

and we conclude that it must be the case that $T_\Sigma(z) = T_{\Sigma''}(z)T_{\Sigma'}(z)$, as asserted. Of course the result can also be verified by direct computation using the formula (3.20) for the transfer function in terms of A, B, C, D. ∎

We next define the *parallel connection* of two SNMLSs as follows. We suppose that we are given two SNMLSs

$$\Sigma'' = (G, \mathcal{H}'', U''), \qquad \Sigma' = (G, \mathcal{H}', U')$$

with the same structure graph G and with the same input-space $\mathcal{U}$ and the same output-space $\mathcal{Y}$:

$$U'' = \begin{bmatrix} A'' & B'' \\ C'' & D'' \end{bmatrix} : \begin{bmatrix} \operatorname{col}_{s\in S} \mathcal{H}''_{[s]} \\ \mathcal{U} \end{bmatrix} \to \begin{bmatrix} \operatorname{col}_{r\in R} \mathcal{H}''_{[r]} \\ \mathcal{Y} \end{bmatrix},$$
$$U' = \begin{bmatrix} A' & B' \\ C' & D' \end{bmatrix} : \begin{bmatrix} \operatorname{col}_{s\in S} \mathcal{H}'_{[s]} \\ \mathcal{U} \end{bmatrix} \to \begin{bmatrix} \operatorname{col}_{r\in R} \mathcal{H}'_{[r]} \\ \mathcal{Y} \end{bmatrix}.$$

We then define the *parallel sum* $\Sigma = \Sigma''[+]\Sigma'$ of Σ'' and Σ' to be $\Sigma = (G, \mathcal{H}, U)$ with auxiliary state-spaces $\mathcal{H}_p$ again equal to the direct sums $\mathcal{H}_p = \begin{bmatrix} \mathcal{H}''_p \\ \mathcal{H}'_p \end{bmatrix}$ and with connection matrix U given by

$$U = \begin{bmatrix} A'' & 0 & B'' \\ 0 & A' & B' \\ C'' & C' & D'' + D' \end{bmatrix} : \begin{bmatrix} \operatorname{col}_{s\in S} \mathcal{H}''_{[s]} \\ \operatorname{col}_{s\in S} \mathcal{H}'_{[s]} \\ \mathcal{U} \end{bmatrix} \to \begin{bmatrix} \operatorname{col}_{r\in R} \mathcal{H}''_{[r]} \\ \operatorname{col}_{r\in R} \mathcal{H}'_{[r]} \\ \mathcal{Y} \end{bmatrix}.$$

Here again we identify $\operatorname{col}_{s\in S} \begin{bmatrix} \mathcal{H}''_{[s]} \\ \mathcal{H}'_{[s]} \end{bmatrix}$ with $\begin{bmatrix} \operatorname{col}_{s\in S} \mathcal{H}''_{[s]} \\ \operatorname{col}_{s\in S} \mathcal{H}'_{[s]} \end{bmatrix}$ and $\operatorname{col}_{r\in R} \begin{bmatrix} \mathcal{H}''_{[r]} \\ \mathcal{H}'_{[r]} \end{bmatrix}$ with $\begin{bmatrix} \operatorname{col}_{r\in R} \mathcal{H}''_{[r]} \\ \operatorname{col}_{r\in R} \mathcal{H}'_{[r]} \end{bmatrix}$ in the natural way. In this case the physical interpretation is that we feed an initial state $x'(\emptyset) = x'_0 \in \operatorname{col}_{s\in S} \mathcal{H}'_{[s]}$ and an input string $\{u(w)\}_{w\in\mathcal{F}_E}$ into Σ' to generate a trajectory $\{u(w), x'(w), y'(w)\}_{w\in\mathcal{F}_E}$ of Σ' along with an initial state $x''(\emptyset) = x''_0 \in \operatorname{col}_{s\in S} \mathcal{H}''_{[s]}$ and the same input string $(u(w))_{w\in\mathcal{F}_E}$ to generate a trajectory $\{u(w), x''(w), y''(w)\}$ of Σ''. We then set $y(w) = y'(w) + y''(w)$. Then $\left\{u(w), \begin{bmatrix} x''(w) \\ x'(w) \end{bmatrix}, y(w)\right\}_{w\in\mathcal{F}_E}$ is a system trajectory of $\Sigma = \Sigma''[+]\Sigma'$, and every trajectory of $\Sigma''[+]\Sigma'$ is of this form. With this system interpretation, the following result follows easily along the same lines as the proof of Theorem 4.1.

THEOREM 4.2. *Suppose that Σ'' and Σ' are two SNMLSs for which the parallel sum $\Sigma := \Sigma''[+]\Sigma'$ is defined as above. Then the transfer function $T_\Sigma(z)$ for Σ is the sum of the transfer functions $T_{\Sigma''}(z)$ and $T_{\Sigma'}(z)$ for Σ'' and Σ':*

$$T_{\Sigma''[+]\Sigma'}(z) = T_{\Sigma''}(z) + T_{\Sigma'}(z). \tag{4.4}$$

Our final system operation is inversion. We suppose that we are given an SNMLS $\Sigma = (G, \mathcal{H}, U)$ for which the colligation

$$U = \begin{bmatrix} A & B \\ C & D \end{bmatrix} : \begin{bmatrix} \operatorname{col}_{s\in S} \mathcal{H}_{[s]} \\ \mathcal{U} \end{bmatrix} \to \begin{bmatrix} \operatorname{col}_{r\in R} \mathcal{H}_{[r]} \\ \mathcal{Y} \end{bmatrix}$$

is such that the feedthrough operator $D : \mathcal{U} \to \mathcal{Y}$ is invertible. We then define the inverse colligation

$$\Sigma^\times = (G, \mathcal{H}, U^\times)$$

with the same structure graph G and auxiliary state-spaces $\mathcal{H} = \{\mathcal{H}_p : p \in P(G)\}$ but with colligation $U^\times$ given by

$$U^\times = \begin{bmatrix} A^\times & B^\times \\ C^\times & D^\times \end{bmatrix} = \begin{bmatrix} A - BD^{-1}C & BD^{-1} \\ -D^{-1}C & D^{-1} \end{bmatrix} : \begin{bmatrix} \mathrm{col}_{s\in S}\, \mathcal{H}_{[s]} \\ \mathcal{Y} \end{bmatrix} \to \begin{bmatrix} \mathrm{col}_{r\in R}\, \mathcal{H}_{[r]} \\ \mathcal{U} \end{bmatrix}.$$

The point here is that $\{y(w), x(w), u(w)\}_{w\in\mathcal{F}_E}$ is a system trajectory of $U^\times$ if and only if $\{u(w), x(w), y(w)\}_{w\in\mathcal{F}_E}$ is a system trajectory of U; i.e., system-inversion amounts to interchange of inputs and outputs. If we then work with system trajectories having $x(\emptyset) = 0$, we see that $\widehat{y}(z) = T_\Sigma(z)\widehat{u}(z)$ is equivalent to $\widehat{u}(z) = T_{\Sigma^\times}(z)\widehat{y}(z)$. Of course it is also possible to verify the formal power series identities

$$T_{\Sigma^\times}(z) \cdot T_\Sigma(z) = I_\mathcal{U}, \qquad T_\Sigma(z) \cdot T_{\Sigma^\times}(z) = I_\mathcal{Y}$$

directly by use of the explicit formula (3.20) for the transfer function. In any case, we record this observation in the following theorem.

THEOREM 4.3. *Suppose that $\Sigma = (G, \mathcal{H}, U)$ is an SNMLS with colligation*

$$U = \begin{bmatrix} A & B \\ C & D \end{bmatrix} : \begin{bmatrix} \mathrm{col}_{s\in S}\, \mathcal{H}_{[s]} \\ \mathcal{U} \end{bmatrix} \to \begin{bmatrix} \mathrm{col}_{r\in R}\, \mathcal{H}_{[r]} \\ \mathcal{Y} \end{bmatrix}$$

having invertible feedthrough operator $D : \mathcal{U} \to \mathcal{Y}$. Then

$$T_\Sigma(z) = D + C(I - Z_\Sigma(z)A)^{-1}Z_\Sigma(z)B$$

is invertible in the space $\mathcal{L}(\mathcal{U}, \mathcal{Y})\langle\langle z\rangle\rangle$ (formal power series in the noncommuting variables $z = (z_e)_{e\in E}$ with coefficients in the space $\mathcal{L}(\mathcal{U}, \mathcal{Y})$ of operators from $\mathcal{U}$ to $\mathcal{Y}$), with inverse $T_\Sigma^{-1}(z) \in \mathcal{L}(\mathcal{Y}, \mathcal{U})\langle\langle z\rangle\rangle$ given by

$$T_\Sigma^{-1}(z) = T_{\Sigma^\times}(z) := D^{-1} - D^{-1}C(I - Z_\Sigma(z)[A - BD^{-1}C])^{-1}Z_\Sigma(z)BD^{-1}. \tag{4.5}$$

Remark 4.4. For the classical case, there exists a converse to Theorem 4.1; i.e., given Σ, it is possible to describe geometrically all possible nontrivial decompositions of Σ as $\Sigma = \Sigma'' \circ \Sigma'$ (see, e.g., [9]). These results can also be extended to more general linear-fractional decompositions (see [29] and [18]). Presumably such results can also be worked out for SNMLSs, but we leave this project to another occasion.

5. Reachability and controllability. The building blocks for reachability and controllability operators are certain operators $\Psi_w : \mathcal{U} \to \mathcal{H}_s$ associated with any word w,

$$\Psi_w = A_{\mathbf{r}(e_N),\mathbf{s}(e_{N-1})} \cdots A_{\mathbf{r}(e_2),\mathbf{s}(e_1)} B_{\mathbf{r}(e_1)} \quad \text{if } w = e_N \cdots e_1. \tag{5.1}$$

Note that the word $w = e_N e_{N-1} \cdots e_2 e_1$ can be written, for each $r = 1, 2, \ldots, N$, as the concatenation

$$w = w'_r w''_{r-1},$$

where we have set

$$w'_r = e_N e_{N-1} \cdots e_r \text{ for } r = 1, \ldots, N, \quad w''_{r-1} = e_{r-1} \cdots e_1 \text{ for } r = 2, \ldots, N, \quad w''_0 = \emptyset.$$

From formula (3.7), we see that the $\mathbf{s}(e_N)$th component of the state trajectory at location $w = e_N \cdots e_1$ for Σ generated by input string $\{u(v)\}_{v \in \mathcal{F}_E}$ with zero initial condition $x(\emptyset) = 0$ is given by

$$\begin{aligned} x_{\mathbf{s}(e_N)}(w) &= \sum_{r=1}^{N} A_{\mathbf{r}(e_N),\mathbf{s}(e_{N-1})} \cdots A_{\mathbf{r}(e_{r+1}),\mathbf{s}(e_r)} B_{\mathbf{r}(e_r)} u(w''_{r-1}) \\ &= \sum_{r=1}^{N} \Psi_{w'_r} u(w''_{r-1}). \end{aligned}$$

Just as in the classical case, the indexing is a little more natural if we consider controllability operators instead. Up to this point we have been considering the system evolution only on the "future time" $\mathcal{T}_{\text{future}} := \mathcal{F}_E$. We now define the "past time" $\mathcal{T}_{\text{past}}$ to be a second copy of $\mathcal{F}_E$ but with the empty word deleted: $\mathcal{T}_{\text{past}} := \mathcal{F}_E \setminus \{\emptyset\}$. We emphasize that $\mathcal{T}_{\text{future}}$ and $\mathcal{T}_{\text{past}}$ are considered to be disjoint sets; given a nonempty word w in $\mathcal{F}_E$, we will specify in the particular context whether it is to be considered as an element of $\mathcal{T}_{\text{future}}$ or of $\mathcal{T}_{\text{past}}$.

Let us now introduce the system evolution on the past, which is given by

$$\Sigma_{\text{past}}\colon \begin{cases} x_s(w) = \sum_{e\colon \mathbf{s}(e)=s} \sum_{s' \in S} A_{\mathbf{r}(e),s'} x_{s'}(we) + \sum_{e\colon \mathbf{s}(e)=s} B_{\mathbf{r}(e)} u(we), \\ y(w) \;\; = \sum_{s \in S} C_s x_s(w) + Du(w), \end{cases} \tag{5.2}$$

or, in aggregate form,

$$\Sigma_{\text{past}}\colon \begin{cases} x(w) = \sum_{e \in E} I_{\Sigma,e} A x(we) + \sum_{e \in E} I_{\Sigma,e} B u(we), \\ y(w) = Cx(w) + Du(w). \end{cases} \tag{5.3}$$

This evolution can actually be derived from the forward evolution by doing the change of "time" variable $w''_{r-1} \mapsto w'_r$ along each finite path w (where the initial segment w''_{r-1} is viewed as a point in the future $\mathcal{T}_{\text{future}}$, while the corresponding final segment w'_r is viewed as a position in the past $\mathcal{T}_{\text{past}}$), and then taking a span over paths as was done above. In this way, the span of all vectors generated at some finite position in the future from zero initial condition on the state at $\emptyset$ over all possible input strings on $\mathcal{T}_{\text{future}}$ is transformed into the set of all possible states achieved at time $\emptyset$ (the final point for the past) over all possible finitely supported input strings on the past with zero state initialization in the distant past.

More precisely, fix a finite word $w = e_N \cdots e_1$, and assume that we run the system in the past $\mathcal{T}_{\text{past}}$ using the system equations (5.2) or (5.3) under the assumption that $x(v) = 0$ for all $v \in \mathcal{T}_{\text{past}}$ with $|v| \geq N$, where N is an arbitrary length, and that $u(v) = 0$ for all $v \in \mathcal{T}_{\text{past}}$ except for those of the form $v = w'_r = e_N \cdots e_r$ for some r with $1 \leq r \leq N$. Then the $\mathbf{s}(e_N)$th component of the resulting state trajectory $x(\cdot)$ at the location $\emptyset$ is

$$\begin{aligned} x_{\mathbf{s}(e_N)}(\emptyset) &= \sum_{r=1}^{N} A_{\mathbf{r}(e_N),\mathbf{s}(e_{N-1})} \cdots A_{\mathbf{r}(e_{r+1}),\mathbf{s}(e_r)} B_{\mathbf{r}(e_r)} u(w'_r) \\ &= \sum_{r=1}^{N} \Psi_{w'_r} u(w'_r). \end{aligned}$$

Then the linear space $\boldsymbol{\mathcal{C}}_w$ consisting of all vectors $x_s \in \mathcal{H}_s$ achievable as $x_s(\emptyset)$ when the system is run with state set equal to zero in the distant past and with input taken to be equal to zero except along some left segment of the word w is characterized as

$$\boldsymbol{\mathcal{C}}_w = \operatorname{im} \mathcal{C}_w,$$

where the *controllability operator* associated with the word w is given by

$$\mathcal{C}_w = \operatorname*{row}_{r=1,\dots,N} \Psi_{w'_r} : \ell_{\mathrm{fin}}(\mathcal{T}^w_{\mathrm{past}}, \mathcal{U}) \to \mathcal{H}_{[s]}, \tag{5.4}$$

where $\mathcal{T}^w_{\mathrm{past}} = \{w'_r : r = 1, \dots, N\} \subset \mathcal{T}_{\mathrm{past}}$.

More generally, we denote by $\mathcal{F}_E^{\infty_R}$ the set of all nonempty words which have a beginning on the left but are infinite to the right:

$$\mathcal{F}_E^{\infty_R} = \{e_1 e_2 \cdots e_N \cdots : e_j \in E \text{ for } j = 1, 2, 3, \dots\}.$$

Fix an infinite word $w = e_1 e_2 \cdots e_N \cdots \in \mathcal{F}_E^{\infty_R}$. Set $w^N = e_1 e_2 \cdots e_N$ equal to the finite word obtained as the truncation of w after N letters, and define

$$\mathcal{C}_w = \operatorname{row}_{N=1,2,3,\dots} \Psi_{w^N} : \ell_{\mathrm{fin}}(\mathcal{T}^w_{\mathrm{past}}, \mathcal{U}) \to \mathcal{H}_{[\mathbf{s}(LL[w])]},$$

where $LL[w]$ (for w a finite or infinite word) denotes the *leading letter* of w,

$$LL[e_1 e_2 \cdots e_N \cdots] = e_1,$$

and where $\mathcal{T}^w_{\mathrm{past}} = \cup\{w^N : N = 1, 2, 3, \dots\}$. Then the image of $\mathcal{C}_w$ (as an operator on $\ell_{\mathrm{fin}}(\mathcal{T}^w_{\mathrm{past}}, \mathcal{U})$) is the linear space of all possible states $x_{\mathbf{s}(e_1)} \in \mathcal{H}_{[\mathbf{s}(e_1)]}$ ($e_1 = LL[w]$) arising in the form $x_{\mathbf{s}(e_1)}(\emptyset)$ from a system trajectory (5.2) under the assumptions that $x(w) = 0$ for all words $w \in \mathcal{T}_{\mathrm{past}}$ of sufficiently large length and that the input string $\{u(w)\}_{w \in \mathcal{T}_{\mathrm{past}}}$ is supported on $w^1, \dots, w^N$ for some finite N.

It is natural to initialize the state to be zero in the far past but to allow input strings of arbitrary finite support. Given $s \in S$, we define the controllability operator $\mathcal{C}_s$ as the block row matrix

$$\mathcal{C}_s = \operatorname*{row}_{w \in \mathcal{T}_{\mathrm{past}} \text{ with } \mathbf{s}(LL[w]) = s} \Psi_w : \ell_{\mathrm{fin}}(\mathcal{T}^s_{\mathrm{past}}, \mathcal{U}) \to \mathcal{H}_{[s]}. \tag{5.5}$$

Here we set

$$\mathcal{T}^s_{\mathrm{past}} = \bigcup_{w \in \mathcal{T}_{\mathrm{past}} \text{ with } s = \mathbf{s}(LL[w])} \mathcal{T}^w_{\mathrm{past}}. \tag{5.6}$$

If we define $\boldsymbol{\mathcal{C}}_s$ to be the linear space of all vectors $x_s \in \mathcal{H}_s$ achievable as $x_s = x_s(\emptyset)$ when we run the system on $\mathcal{T}_{\mathrm{past}}$ with an input string of finite support and with state initialization set equal to zero at all positions $v \in \mathcal{T}_{\mathrm{past}}$ with $|v|$ sufficiently large, then we have

$$\boldsymbol{\mathcal{C}}_s = \operatorname{im} \mathcal{C}_s.$$

Remark 5.1. More generally, we may define an apparently more general controllability operator as follows. For $p \in P$ (the set of path-connected components of the structure graph G associated with the SNMLS Σ (see Definition 3.7)), set

$\mathcal{T}^p_{\rm past} = \bigcup_{s:[s]=p} \mathcal{T}^s_{\rm past}$. We define the controllability operator $\mathcal{C}_p$ as the block row matrix

$$\mathcal{C}_p = \operatorname*{row}_{s:[s]=p} \mathcal{C}_s \colon \ell_{\rm fin}(\mathcal{T}^p_{\rm past}, \mathcal{U}) \to \mathcal{H}_p. \tag{5.7}$$

Then the image of $\mathcal{C}_p$ consists of the linear span of all vectors $x_p \in \mathcal{H}_p$ expressible as $x_s(\emptyset)$ (for some s with $[s] = p$) when the system Σ is run over the past $\mathcal{T}^p_{\rm past}$ with some input string on $\mathcal{T}^p_{\rm past}$ of finite support and with state-vector initialized to be zero at all positions v sufficiently far in the past.

Note, however, from the formula (5.1) for Ψ_w that Ψ_w is independent of the value of $\mathbf{s}(LL[w])$; i.e., if $w = e_N e_{N-1} \cdots e_2 e_1$ and $w' = e'_N e_{N-1} \cdots e_2 e_1$, then $\Psi_{w'} = \Psi_w$, as long as $\mathbf{r}(e'_N) = \mathbf{r}(e_N)$. Thus $\operatorname{im} \mathcal{C}_s = \operatorname{im} \mathcal{C}_p$ for any $s \in S$ with $[s] = p$.

It will be convenient to make this invariance property more explicit. We define a bijection $w \mapsto w^{\wedge s}$ from $\mathcal{T}^{s'}_{\rm past}$ to $\mathcal{T}^s_{\rm past}$ by

$$w^{\wedge s} = e_{s,\mathbf{r}(e_N)} e_{N-1} \cdots e_1 \quad \text{if } w = e_N e_{N-1} \cdots e_1. \tag{5.8}$$

Note that $e_{s,\mathbf{r}(e_N)}$ is well defined as (3.1) whenever it is the case that $[s] = [\mathbf{s}(e_N)] (= [\mathbf{r}(e_N)])$. As observed in the previous paragraph, the controllability-operator building blocks Ψ_w given by (5.1) are invariant under this transformation:

$$\text{for } s, s' \in S \text{ with } [s] = [s'], \quad \Psi_w = \Psi_{w^{\wedge s'}} \text{ for } w \in \mathcal{T}^s_{\rm past}. \tag{5.9}$$

For each of the three choices of controllability operator $\mathcal{C}_w$, $\mathcal{C}_s$, and $\mathcal{C}_p$ (where $\mathcal{C}_p$ is as in Remark 5.1), we have a corresponding notion of controllability, namely, the system Σ is X-controllable (where $X = \mathcal{F}_E^{\infty_R}$ (the set of words which are infinite to the right), $X = S$ or $X = P$) if the operator $\mathcal{C}_x$ is surjective for all $x \in X$. A consequence of Remark 5.1, however, is that S-controllability and P-controllability are equivalent. The notion of controllability most convenient for our purposes here is the weakest of these, namely P-controllability (or equivalently, S-controllability). We therefore make the following definition.

DEFINITION 5.2. *We say that the SNMLS Σ is* structured-controllable *or simply* controllable *if the operator*

$$\mathcal{C}_p \colon \ell_{fin}(\mathcal{T}^p_{past}, \mathcal{U}) \to \mathcal{H}_p$$

given by (5.7) *is surjective for each path-connected component p of the admissible graph G associated with Σ, or equivalently (by Remark* 5.1*), if the operator*

$$\mathcal{C}_s \colon \ell_{fin}(\mathcal{T}^s_{past}, \mathcal{U}) \to \mathcal{H}_{[s]}$$

given by (5.5) *is surjective for each $s \in S$ (or equivalently, for some s with $[s] = p$ for each $p \in P$).*

6. Observability. Analogously, we have a dual array of observability operators, but with one additional parameter (roughly due to the fact that $\mathcal{T}_{\rm future}$ includes the empty word $\emptyset$, while $\mathcal{T}_{\rm past}$ does not), namely $\mathcal{O}_{s,w}$ for each $s \in S$ and infinite word $w = e_1 e_2 \cdots e_N \cdots \in \mathcal{F}_E^{\infty_R}$, $\mathcal{O}_s$ for each $s \in S$, and $\mathcal{O}_p$ for each $p \in P$. For $w = e_1 e_2 \cdots e_N \cdots \in \mathcal{F}_E^{\infty_R}$ and $s \in S$, we define $\mathcal{O}_{s,w}$ as the block-operator column matrix

$$\mathcal{O}_{s,w} = \operatorname*{col}_{N=0,1,2,\ldots} [C_{\mathbf{s}(e_N)} A_{\mathbf{r}(e_N),\mathbf{s}(e_{N-1})} \cdots A_{\mathbf{r}(e_1),s}] \colon \mathcal{H}_{[s]} \to \ell(\mathcal{T}^w_{\rm future}, \mathcal{Y}), \tag{6.1}$$

where we interpret the formula for the case $N = 0$ to mean

$$[\mathcal{O}_{s,w}]_0 = C_s \tag{6.2}$$

and where we put $\mathcal{T}^w_{\text{future}} = \{(w^N)^\top = e_N e_{N-1} \cdots e_1 \colon N = 0, 1, 2, \dots\} \subset \mathcal{T}_{\text{future}}$. For any $s \in S$, we define an associated observability operator $\mathcal{O}_s$ as the column matrix

$$\mathcal{O}_s = \operatorname*{col}_{v = e_N e_{N-1} \cdots e_1 \in \mathcal{T}_{\text{future}}} [C_{\mathbf{s}(e_N)} A_{\mathbf{r}(e_N), \mathbf{s}(e_{N-1})} \cdots A_{\mathbf{r}(e_1), s}] \colon \mathcal{H}_{[s]} \to \ell(\mathcal{T}_{\text{future}}, \mathcal{Y}), \tag{6.3}$$

with again the interpretation (6.2) for the case $v = \emptyset$ column entry. Finally, for path-connected component $p \in P$ we define an associated observability operator $\mathcal{O}_p$ by

$$\mathcal{O}_p = \operatorname*{col}_{s \in S \colon [s] = p} \mathcal{O}_s \colon \mathcal{H}_p \to \operatorname*{col}_{s \in S \colon [s] = p} \ell(\mathcal{T}_{\text{future}}, \mathcal{Y}). \tag{6.4}$$

Clearly, for each infinite word $w \in \mathcal{F}_E^{\infty_R}$, index $s \in S$, and path-connected component $p \in P$ with $[s] = p$, we have the subspace inclusions

$$\ker \mathcal{O}_p \subset \ker \mathcal{O}_s \subset \ker \mathcal{O}_{s,w}. \tag{6.5}$$

For each of the cases $X = S \times \mathcal{F}_E^{\infty_R}$, $X = S$, and $X = P$, we have a notion of X-observability: Σ is X-observable if the operator $\mathcal{O}_x$ is injective for all $x \in X$. By the set of inclusions (6.5) we see that we have the chain of implications: $S \times \mathcal{F}_E^{\infty_R}$-observability implies S-observability, which in turn implies P-observability. Note that each of these observability notions has a system-theoretic interpretation, as follows:

1. $S \times \mathcal{F}_E^{\infty_R}$-observability means that, for each fixed infinite word $w \in \mathcal{F}_E^{\infty_R}$, an initial state $x_s \in \mathcal{H}_{[s]}$ is uniquely determined from the observations $y((w^N)^\top)$ (for $N = 0, 1, 2, \dots$) obtained by letting the system drift with initial condition $x_s(\emptyset) = x_s$ and $x_{s'}(\emptyset) = 0$ for $s' \neq s$ and with zero input string $u(w) = 0$ for all $w \in \mathcal{F}_E$.

2. S-observability means again that, for each $s \in S$, one can detect an initial state $x_s \in \mathcal{H}_{[s]}$ by the same experiment, but with additional observations, namely $y(v)$ for all $v \in \mathcal{F}_E$.

3. P-observability means again that one can detect an initial state $x_p \in \mathcal{H}_p$ but one must do the experiment described above for S-observability with initial condition $x_s(\emptyset) = x_p$ and $x_{s'}(\emptyset) = 0$ for $s' \neq s$ for each $s \in S$ with $[s] = p$.

For our notion of observability here, we take the weakest of these notions and make the following definition.

DEFINITION 6.1. *We say that the SNMLS* $\Sigma = (G, \mathcal{H}, U)$ *is* structured-observable *(or simply* observable*) if the operator* $\mathcal{O}_p \colon \mathcal{H}_p \to col_{s \in S \colon [s] = p} \ell(\mathcal{T}_{future}, \mathcal{Y})$ *given by* (6.4) *is injective for each* $p \in P$.

7. Kalman decomposition. In this section we obtain a Kalman-type decomposition for SNMLSs; for a good summary of these results for the classical case, we refer to [16].

Let $\Sigma = (G, \mathcal{H}, U)$ be an SNMLS as in Definition 3.7. For each $p \in P$ (the set of path-connected components of the admissible graph G), we let $\mathcal{C}_p$ be the controllability operator defined by (5.7) and $\mathcal{O}_p$ be the observability operator defined by (6.4).[3] From

[3] As it is only the images $\operatorname{im} \mathcal{C}_p$ of the controllability operators $\mathcal{C}_p$ which enter in here, by Remark 5.1 without loss of generality one can in all the discussion below replace $\mathcal{C}_p$ with $\mathcal{C}_{s_p}$ for any fixed choice of $s_p \in S$ with $[s_p] = p$.

the definitions we see that

$$A_{r,s}\colon \operatorname{im}\mathcal{C}_{[s]} \to \operatorname{im}\mathcal{C}_{[r]}, \tag{7.1}$$

$$A_{r,s}\colon \ker\mathcal{O}_{[s]} \to \ker\mathcal{O}_{[r]}, \tag{7.2}$$

$$\ker\mathcal{O}_{[s]} \subset \ker C_s, \tag{7.3}$$

$$\operatorname{im} B_r \subset \operatorname{im}\mathcal{C}_{[r]} \tag{7.4}$$

for all $r \in R$ and $s \in S$. We introduce a direct-sum decomposition

$$\mathcal{H}_p = \mathcal{H}_{p,c/o} \oplus \mathcal{H}_{p,c/no} \oplus \mathcal{H}_{p,nc/o} \oplus \mathcal{H}_{p,nc/no} \tag{7.5}$$

according to the following recipe:

1. Set $\mathcal{H}_{p,c/no} = \operatorname{im}\mathcal{C}_p \cap \ker\mathcal{O}_p$.
2. Choose $\mathcal{H}_{p,c/o}$ so that $\mathcal{H}_{p,c/no} \oplus \mathcal{H}_{p,c/o} = \operatorname{im}\mathcal{C}_p$.
3. Choose $\mathcal{H}_{p,nc/no}$ such that $\mathcal{H}_{p,c/no} \oplus \mathcal{H}_{p,nc/no} = \ker\mathcal{O}_p$.
4. Choose $\mathcal{H}_{p,nc/o}$ such that $\mathcal{H}_p = \mathcal{H}_{p,c/o} \oplus \mathcal{H}_{p,c/no} \oplus \mathcal{H}_{p,nc/o} \oplus \mathcal{H}_{p,nc/no}$.

Fix an $r \in R$ and an $s \in S$. Note that $A_{r,s}\colon \mathcal{H}_{[s]} \to \mathcal{H}_{[r]}$, $B_r\colon \mathcal{U} \to \mathcal{H}_{[r]}$, and $C_s\colon \mathcal{H}_{[s]} \to \mathcal{Y}$, while $\mathcal{H}_{[s]}$, and $\mathcal{H}_{[r]}$ have the direct-sum decompositions

$$\begin{aligned}\mathcal{H}_{[s]} &= \mathcal{H}_{[s],c/o} \oplus \mathcal{H}_{[s],c/no} \oplus \mathcal{H}_{[s],nc/o} \oplus \mathcal{H}_{[s],nc/no},\\ \mathcal{H}_{[r]} &= \mathcal{H}_{[r],c/o} \oplus \mathcal{H}_{[r],c/no} \oplus \mathcal{H}_{[r],nc/o} \oplus \mathcal{H}_{[r],nc/no}.\end{aligned}$$

We may therefore represent $A_{r,s}$, B_r, and C_s as matrices with respect to these direct-sum decompositions of $\mathcal{H}_{[s]}$ and $\mathcal{H}_{[r]}$:

$$A_{r,s} = \begin{bmatrix} A_{r,s;c/o,c/o} & A_{r,s;c/o,c/no} & A_{r,s;c/o,nc/o} & A_{r,s;c/o,nc/no} \\ A_{r,s;c/no,c/o} & A_{r,s;c/no,c/no} & A_{r,s;c/no,nc/o} & A_{r,s;c/no,nc/no} \\ A_{r,s;nc/o,c/o} & A_{r,s;nc/o,c/no} & A_{r,s;nc/o,nc/o} & A_{r,s;nc/o,nc/no} \\ A_{r,s;nc/no,c/o} & A_{r,s;nc/no,c/no} & A_{r,s;nc/no,nc/o} & A_{r,s;nc/no,nc/no} \end{bmatrix},$$

$$B_r = \begin{bmatrix} B_{r,c/o} \\ B_{r,c/no} \\ B_{r,nc/o} \\ B_{r,nc/no} \end{bmatrix}, \qquad C_s = \begin{bmatrix} C_{s,c/o} & C_{s,c/no} & C_{s,nc/o} & C_{s,nc/no} \end{bmatrix}.$$

From (7.1) we see that

$$A_{r,s;nc/o,c/o} = 0, \quad A_{r,s;nc/o,c/no} = 0, \quad A_{r,s;nc/no,c/o} = 0, \quad A_{r,s;nc/no,c/no} = 0.$$

From (7.2) we see that

$$A_{r,s;c/o,c/no} = 0, \quad A_{r,s;c/o,nc/no} = 0, \quad A_{r,s;nc/o,c/no} = 0, \quad A_{r,s;nc/o,nc/no} = 0.$$

From (7.4) we see that

$$B_{r,nc/o} = 0, \qquad B_{r,nc/no} = 0.$$

From (7.3) we see that

$$C_{s,c/no} = 0, \qquad C_{s,nc/no} = 0.$$

We are therefore left with

$$A_{r,s} = \begin{bmatrix} A_{r,s;c/o,c/o} & 0 & A_{r,s;c/o,nc/o} & 0 \\ A_{r,s;c/no,c/o} & A_{r,s;c/no,c/no} & A_{r,s;c/no,nc/o} & A_{r,s;c/no,nc/no} \\ 0 & 0 & A_{r,s;nc/o,nc/o} & 0 \\ 0 & 0 & A_{r,s;nc/no,nc/o} & A_{r,s;nc/no,nc/no} \end{bmatrix},$$

$$B_r = \begin{bmatrix} B_{r,c/o} \\ B_{r,c/no} \\ 0 \\ 0 \end{bmatrix}, \qquad C_s = \begin{bmatrix} C_{s,c/o} & 0 & C_{s,nc/o} & 0 \end{bmatrix}. \tag{7.6}$$

This analysis leads us to the following result.

THEOREM 7.1. *Let $\Sigma = (G, \mathcal{H}, U)$ be an SNMLS. Decompose each $\mathcal{H}_p$ as in* (7.5) *with resulting decompositions* (7.6) *for the system matrices $A_{r,s}$, B_r, and C_s.*

(1) *Define a reduced SNMLS $\Sigma_{c/o} = (G, \mathcal{H}_{c/o}, U_{c/o})$ with auxiliary state-spaces $(\mathcal{H}_{c/o})_p = \mathcal{H}_{p,c/o}$ as in* (7.5) *and with connection matrix*

$$U_{c/o} = \begin{bmatrix} A_{c/o} & B_{c/o} \\ C_{c/o} & D_{c/o} \end{bmatrix} : \begin{bmatrix} \oplus_{s\in S}\mathcal{H}_{[s],c/o} \\ \mathcal{U} \end{bmatrix} \to \begin{bmatrix} \oplus_{r\in R}\mathcal{H}_{[r],c/o} \\ \mathcal{Y} \end{bmatrix}$$

given by

$$[A_{c/o}]_{r,s} = A_{r,s;c/o,c/o}, \quad [B_{c/o}]_r = B_{r,c/o}, \quad [C_{c/o}]_s = C_{s,c/o}, \quad D_{c/o} = D$$

determined as in (7.6). *Then the SNMLS $\Sigma_{c/o}$ is structured-controllable and structured-observable and has the same transfer function as Σ:*

$$D + C(I - Z_\Sigma(z)A)^{-1}Z_\Sigma(z)B = D_{c/o} + C_{c/o}(I - Z_{\Sigma_{c/o}}(z)A_{c/o})^{-1}Z_{\Sigma_{c/o}}(z)B_{c/o}.$$

(2) *Define a reduced system $\Sigma_c = (G, \mathcal{H}_c, U_c)$ with auxiliary state-spaces*

$$(\mathcal{H}_c)_p = \mathcal{H}_{p,c/o} \oplus \mathcal{H}_{p,c/no}$$

with components determined as in (7.5) *and with connection matrix*

$$U_c = \begin{bmatrix} A_c & B_c \\ C_c & D_c \end{bmatrix} : \begin{bmatrix} \oplus_{s\in S}\mathcal{H}_{[s],c} \\ \mathcal{U} \end{bmatrix} \to \begin{bmatrix} \oplus_{r\in R}\mathcal{H}_{[r],c} \\ \mathcal{Y} \end{bmatrix}$$

given by

$$[A_c]_{r,s} = \begin{bmatrix} A_{r,s;c/o,c/o} & 0 \\ A_{r,s;c/no,c/o} & A_{r,s;c/no,c/no} \end{bmatrix}, \qquad [B_c]_r = \begin{bmatrix} B_{r,c/o} \\ B_{r,c/no} \end{bmatrix},$$
$$[C_c]_s = \begin{bmatrix} C_{s,c/o} & 0 \end{bmatrix}, \qquad D_c = D,$$

with matrix entries determined as in (7.6). *Then the SNMLS Σ_c is structured-controllable and has the same transfer function as Σ:*

$$D + C(I - Z_\Sigma(z)A)^{-1}Z_\Sigma(z)B = D_c + C_c(I - Z_{\Sigma_c}(z)A_c)^{-1}Z_{\Sigma_c}(z)B_c.$$

(3) *Define a reduced system $\Sigma_o = (G, \mathcal{H}_o, U_o)$ with auxiliary state-spaces $(\mathcal{H}_o)_p = \mathcal{H}_{p,c/o} \oplus \mathcal{H}_{p,nc/o}$ with components determined as in* (7.5) *and with connection matrix*

$$U_o = \begin{bmatrix} A_o & B_o \\ C_o & D_o \end{bmatrix} : \begin{bmatrix} \oplus_{s\in S}\mathcal{H}_{[s],o} \\ \mathcal{U} \end{bmatrix} \to \begin{bmatrix} \oplus_{r\in R}\mathcal{H}_{[r],o} \\ \mathcal{Y} \end{bmatrix}$$

given by

$$[A_o]_{r,s} = \begin{bmatrix} A_{r,s;c/o,c/o} & A_{r,s;c/o,nc/o} \\ 0 & A_{r,s;nc/o,nc/o} \end{bmatrix}, \qquad [B_o]_r = \begin{bmatrix} B_{r,c/o} \\ 0 \end{bmatrix},$$
$$[C_o]_s = \begin{bmatrix} C_{s,c/o} & C_{s,nc/o} \end{bmatrix}, \qquad D_o = D,$$

with matrix entries determined as in (7.6). *Then the SNMLS* Σ_o *is structured-observable and has the same transfer function as* Σ*:*

$$D + C(I - Z_\Sigma(z)A)^{-1}Z_\Sigma(z)B = D_o + C_o(I - Z_{\Sigma_o}(z)A_o)^{-1}Z_{\Sigma_o}(z)B_o.$$

8. State-space similarity theorem. We begin with a definition.

DEFINITION 8.1. *Given two SNMLSs* $\Sigma = (G, \mathcal{H}, U)$ *and* $\Sigma' = (G, \mathcal{H}', U')$ *with a common structure graph* G *and with common input- and output-spaces, so that*

$$U = \begin{bmatrix} A & B \\ C & D \end{bmatrix} : \begin{bmatrix} \oplus_{s\in S}\mathcal{H}_{[s]} \\ \mathcal{U} \end{bmatrix} \to \begin{bmatrix} \oplus_{r\in R}\mathcal{H}_{[r]} \\ \mathcal{Y} \end{bmatrix},$$
$$U' = \begin{bmatrix} A' & B' \\ C' & D' \end{bmatrix} : \begin{bmatrix} \oplus_{s\in S}\mathcal{H}'_{[s]} \\ \mathcal{U} \end{bmatrix} \to \begin{bmatrix} \oplus_{r\in R}\mathcal{H}'_{[r]} \\ \mathcal{Y} \end{bmatrix},$$

we say that Σ *and* Σ' *are* similar *(via a state-space similarity) if there is a collection* $\Gamma = \{\Gamma_p : p \in P\}$ *of bijective linear operators* $\Gamma_p : \mathcal{H}_p \to \mathcal{H}'_p$ *(for each path-connected component* p *of* G*) such that*

$$\begin{bmatrix} (\oplus_{r\in R}\Gamma_{[r]}) & 0 \\ 0 & I_{\mathcal{Y}} \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A' & B' \\ C' & D' \end{bmatrix} \begin{bmatrix} (\oplus_{s\in S}\Gamma_{[s]}) & 0 \\ 0 & I_{\mathcal{U}} \end{bmatrix}. \tag{8.1}$$

It is an easy computation to see that two systems Σ and Σ' have the same transfer functions if they are similar. On the other hand, Theorem 7.1 is not true in general, since an SNMLS Σ which is not already structured-controllable and structured-observable cannot be similar to its structured-controllable/structured-observable part, as in this case necessarily $\dim \mathcal{H}_{p,c/o} < \dim \mathcal{H}_p$ for some p. The next theorem gives the converse under a controllability/observability hypothesis.

THEOREM 8.2. *Suppose that* $\Sigma = (G, \mathcal{H}, U)$ *and* $\Sigma' = (G, \mathcal{H}', U')$ *are two SNMLSs with a common structure graph* G *and common input- and output-spaces* $\mathcal{U}$ *and* $\mathcal{Y}$*. Assume that both* Σ *and* Σ' *are structured-controllable and structured-observable. Then* Σ *and* Σ' *are similar; i.e., there are bijective linear maps* $\Gamma_p : \mathcal{H}_p \to \mathcal{H}'_p$ *for each path-connected component* p *of* G *such that* (8.1) *holds if and only if* Σ *and* Σ' *have the same transfer function*

$$T_\Sigma(z) = T_{\Sigma'}(z).$$

Moreover, in this situation the collection of state-space similarity operators

$$\Gamma_p : \mathcal{H}_{[p]} \to \mathcal{H}'_{[p]}$$

implementing the similarity between Σ *and* Σ' *is unique.*

Proof. We have already observed that in general two systems which are similar have the same transfer function. It remains to show the following: *under the assumption that* Σ *and* Σ' *are structured-controllable and structured-observable, if*

$T_\Sigma(z) = T_{\Sigma'}(z)$, then Σ and Σ' are similar. From the expression (3.21) for the transfer function, we see that the hypothesis that $T_\Sigma(z) = T_{\Sigma'}(z)$ amounts to the assertion that

$$\begin{aligned} &C_{\mathbf{s}(e_N)} A_{\mathbf{r}(e_N),\mathbf{s}(e_{N-1})} A_{\mathbf{r}(e_{N-1}),\mathbf{s}(e_{N-2})} \cdots A_{\mathbf{r}(e_2),\mathbf{s}(e_1)} B_{\mathbf{r}(e_1)} \\ &\quad = C'_{\mathbf{s}(e_N)} A'_{\mathbf{r}(e_N),\mathbf{s}(e_{N-1})} A'_{\mathbf{r}(e_{N-1}),\mathbf{s}(e_{N-2})} \cdots A'_{\mathbf{r}(e_2),\mathbf{s}(e_1)} B'_{\mathbf{r}(e_1)} \end{aligned} \tag{8.2}$$

for all nonempty words $w = e_N e_{N-1} \cdots e_1 \in \mathcal{F}_E$, with the interpretation

$$C_{\mathbf{s}(e_1)} B_{\mathbf{r}(e_1)} = C'_{\mathbf{s}(e_1)} B'_{\mathbf{r}(e_1)} \tag{8.3}$$

in case $w = e_1$ has length 1 together with

$$D = D' \tag{8.4}$$

corresponding to the case $w = \emptyset$. Recalling the definitions (6.3) and (5.1), we see immediately from (8.2) and (8.3) that

$$[\mathcal{O}_s]_v \mathcal{C}_w = [\mathcal{O}'_s]_v \mathcal{C}'_w \tag{8.5}$$

whenever $s \in S$, $v \in \mathcal{T}_{\text{future}}$, and $w \in \mathcal{T}^s_{\text{past}}$. By the same type of argument as that appearing in Remark 5.1, in fact (8.5) holds for each $s \in S$, $v \in \mathcal{T}_{\text{future}}$, and $w \in \mathcal{T}^{s'}_{\text{past}}$ for any $s' \in S$ in the same path-connected component as s (i.e., with $[s'] = [s]$); indeed, if $w = ew' \in \mathcal{T}^s_{\text{past}}$, there is a unique adjustment $e' \in E$ of e so that $w' = e'w' \in \mathcal{T}^{s'}_{\text{past}}$, $\mathcal{C}_{w'} = \mathcal{C}_w$, and also $\mathcal{C}'_{w'} = \mathcal{C}'_w$. Hence the equality (8.5) with $w \in \mathcal{T}^s_{\text{past}}$ implies the equality (8.5) with $w \in \mathcal{T}^{s'}_{\text{past}}$ for any s' with $[s'] = [s]$ as well.

We attempt to define $\Gamma_p \colon \mathcal{H}_p \to \mathcal{H}'_p$ by

$$\Gamma_p \colon \Psi_w u \mapsto \Psi'_w u \quad \text{for } u \in \mathcal{U} \text{ and } w \in \mathcal{F}_E \text{ with } [\mathbf{r}(LL[w])] = s_p, \tag{8.6}$$

where Ψ_w and Ψ'_w are given by (5.1) and where $s_p \in S$ is any choice of source vertex with $[s_p] = p$. Note that a consequence of Remark 5.1 is that we can always adjust $LL[w]$ to achieve $\mathbf{s}(LL[w]) = s_p$ (for any fixed choice of $s_p \in S$ with $[s_p] = p$) without affecting $\operatorname{im} \Psi_w$ and $\operatorname{im} \Psi'_w$. Explicitly, we have

$$\begin{aligned} \Gamma_p \colon\ & A_{\mathbf{r}(e_N),\mathbf{s}(e_{N-1})} A_{\mathbf{r}(e_{N-1}),\mathbf{s}(e_{N-2})} \cdots A_{\mathbf{r}(e_2),\mathbf{s}(e_1)} B_{\mathbf{r}(e_1)} u \\ &\mapsto A'_{\mathbf{r}(e_N),\mathbf{s}(e_{N-1})} A'_{\mathbf{r}(e_{N-1}),\mathbf{s}(e_{N-2})} \cdots A'_{\mathbf{r}(e_2),\mathbf{s}(e_1)} B'_{\mathbf{r}(e_1)} u, \end{aligned} \tag{8.7}$$

where $w = e_N e_{N-1} \cdots e_1 \in \mathcal{F}_E$ and where e_N is normalized so that $\mathbf{s}(e_N) = s_p$ with the interpretation

$$\Gamma_p \colon B_{\mathbf{r}(e_1)} u \mapsto B'_{\mathbf{r}(e_1)} u \tag{8.8}$$

in case $w = e_1$ (with $\mathbf{s}(e_1) = s_p$) has length 1. We then extend Γ_p to

$$\mathcal{D}_{s_p} = \operatorname{span}\{\Psi_w u \colon w \in \mathcal{T}^{s_p}_{\text{future}} \text{ with } \mathbf{s}(LL[w]) = s_p,\ u \in \mathcal{U}\} \tag{8.9}$$

by linearity, where we set

$$\mathcal{T}^{s_p}_{\text{future}} = \{w \in \mathcal{F}_E \setminus \{\emptyset\} \colon \mathbf{s}(LL[w]) = s_p\}.$$

We first wish to check that Γ_p is well defined. We must therefore show the following: *given a map $w \mapsto u_w$ from $\mathcal{T}^{s_p}_{future}$ to $\mathcal{U}$ with finite support (so $u_w = 0$ for*

all but finitely many words $w \in \mathcal{T}_{\text{future}}^{s_p}$) *such that* $\sum_{w\in\mathcal{T}_{\text{future}}^{s_p}} \Psi_w u_w = 0$, *it follows that* $\sum_{w\in\mathcal{T}_{\text{future}}^{s_p}} \Psi'_w u_w = 0$. Since $\sum_{w\in\mathcal{T}_{\text{future}}^{s_p}} \Psi_w u_w = 0$, we then also have

$$\mathcal{O}_p \cdot \sum_{w\in\mathcal{T}_{\text{future}}^{s_p}} \Psi_w u_w = 0. \tag{8.10}$$

From the definition of $\mathcal{O}_p$, equation (8.10) in turn means that

$$\mathcal{O}_s \cdot \sum_{w\in\mathcal{T}_{\text{future}}^{s_p}} \Psi_w u_w = 0 \quad \text{for each } s \in S \text{ with } [s] = p. \tag{8.11}$$

From the extended domain of validity of (8.5) explained above, (8.11) immediately implies

$$\mathcal{O}'_s \cdot \sum_{w\in\mathcal{T}_{\text{future}}^{s_p}} \Psi'_w u_w = \quad \text{for each } s \in S \text{ with } [s] = p. \tag{8.12}$$

By the assumption that Σ' is structured-observable, we know that $\mathcal{O}'_p$ is injective. Hence we see from (8.12) that

$$\sum_{w\in\mathcal{T}_{\text{future}}^{s_p}} \Psi'_w u_w = 0.$$

We conclude that Γ_p is well-defined on its domain $\mathcal{D}_{s_p}$ (see (8.9)), as wanted.

Since Σ by hypothesis is structured-controllable, we see that in fact $\mathcal{D}_{s_p} = \mathcal{H}_p$, and hence Γ_p is defined on all of $\mathcal{H}_p$. Similarly, since Σ' is structured-controllable, we see that $\Gamma_p(\mathcal{H}_p)$ is equal to all of $\mathcal{H}'_p$, i.e., that Γ_p is surjective.

It remains to see that Γ_p is injective; i.e., *given a map* $w \mapsto u_w$ *from* $\mathcal{T}_{future}^{s_p}$ *to* $\mathcal{U}$ *with finite support such that* $\sum_{w\in\mathcal{T}_{future}^{s_p}} \Psi'_w u_w = 0$, *it follows that* $\sum_{w\in\mathcal{T}_{future}^{s_p}} \Psi_w u_w = 0$. This follows by the same argument as in the proof that Γ_p is well defined, with the roles of Σ and Σ' interchanged. We conclude that (8.6) extends by linearity to define a bijective linear transformation from $\mathcal{H}_p$ onto $\mathcal{H}'_p$.

It remains now only to check that $\Gamma = \{\Gamma_p \colon p \in P\}$ satisfies (8.1). This amounts to verifying

$$\Gamma_{[r]} A_{r,s} = A'_{r,s} \Gamma_{[s]}, \tag{8.13}$$

$$\Gamma_{[r]} B_r = B'_r, \tag{8.14}$$

$$C_s = C'_s \Gamma_{[s]}, \tag{8.15}$$

$$D = D'. \tag{8.16}$$

Note that (8.16) follows immediately from (8.4), while (8.14) follows from (8.8). By the structured-controllability hypothesis on Σ, to show (8.13) and (8.15) it suffices to show

$$\begin{aligned} &\Gamma_{[r]} A_{r,s} A_{\mathbf{r}(e_N),\mathbf{s}(e_{N-1})} A_{\mathbf{r}(e_{N-1}),\mathbf{s}(e_{N-2})} \cdots A_{\mathbf{r}(e_2),\mathbf{s}(e_1)} B_{\mathbf{r}(e_1)} \\ &\quad = A'_{r,s} \Gamma_{[s]} A_{\mathbf{r}(e_N),\mathbf{s}(e_{N-1})} A_{\mathbf{r}(e_{N-1}),\mathbf{s}(e_{N-2})} \cdots A_{\mathbf{r}(e_2),\mathbf{s}(e_1)} B_{\mathbf{r}(e_1)}, \end{aligned} \tag{8.17}$$

$$\begin{aligned} &C_s A_{\mathbf{r}(e_N),\mathbf{s}(e_{N-1})} A_{\mathbf{r}(e_{N-1}),\mathbf{s}(e_{N-2})} \cdots A_{\mathbf{r}(e_2),\mathbf{s}(e_1)} B_{\mathbf{r}(e_1)} \\ &\quad = C'_s \Gamma_{[s]} A_{\mathbf{r}(e_N),\mathbf{s}(e_{N-1})} A_{\mathbf{r}(e_{N-1}),\mathbf{s}(e_{N-2})} \cdots A_{\mathbf{r}(e_2),\mathbf{s}(e_1)} B_{\mathbf{r}(e_1)} \quad \text{if } [s] = [\mathbf{r}(e_N)] \end{aligned} \tag{8.18}$$

for all words $w = e_N e_{N-1} \cdots e_1 \in \mathcal{T}^{s_p}_{\text{future}}$ (with proper interpretation for $N = 1$) for each $p \in P$. Note that (8.17) is an immediate consequence of the definition (8.6) of Γ_p together with the completeness of the path-connected components of G, while (8.18) follows from the definition (8.6) combined with the completeness of the path-connected components of G and the equality of moments (8.2) and (8.3).

As for the last statement in Theorem 8.2, suppose that $\Gamma'_p \colon \mathcal{H}_p \to \mathcal{H}'_p$ is any other linear isomorphism between $\mathcal{H}_p$ and $\mathcal{H}'_p$ so that (8.1) is satisfied. Then a consequence of (8.1) is that necessarily Γ'_p must also satisfy (8.6) (with Γ'_p in place of Γ_p). By the first part of the proof, $\Gamma'_p = \Gamma_p$ for all $p \in P$, and the uniqueness assertion in Theorem 8.2 follows as well. This completes the proof of Theorem 8.2. ∎

9. Minimal state-space realizations. Suppose that we are given an admissible graph G together with a formal power series

$$T(z) = \sum_{w \in \mathcal{F}_E} T_w z^w$$

in the noncommuting variables $z = \{z_e \colon e \in E\}$ (where E is the edge set of G) with coefficients T_w in the space $\mathcal{L}(\mathcal{U}, \mathcal{Y})$ of linear operators between the (finite-dimensional) linear spaces $\mathcal{U}$ and $\mathcal{Y}$. We say that the SNMLS $\Sigma = (G, \mathcal{H}, U)$ (with structure graph equal to G) is a *G-structured realization* for $T(z)$ if $T(z)$ is equal to the transfer function of Σ, i.e., if

$$T_\emptyset = D, \qquad T_{e_N e_{N-1} \cdots e_1} = C_{\mathbf{s}(e_N)} A_{\mathbf{r}(e_N), \mathbf{s}(e_{N-1})} A_{\mathbf{r}(e_{N-1}), \mathbf{s}(e_{N-2})} \cdots A_{\mathbf{r}(e_2), \mathbf{s}(e_1)} B_{\mathbf{r}(e_1)},$$

where the connection matrix U for Σ has the form

$$U = \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} [A_{r,s}] & [B_r] \\ [C_s] & D \end{bmatrix} : \begin{bmatrix} \oplus_{s \in S} \mathcal{H}_{[s]} \\ \mathcal{U} \end{bmatrix} \to \begin{bmatrix} \oplus_{r \in R} \mathcal{H}_{[r]} \\ \mathcal{Y} \end{bmatrix}.$$

We say that the SNMLS Σ is a *structured-minimal realization* for $T(z)$ if $\dim \mathcal{H}'_p \geq \dim \mathcal{H}_p$ for each path-connected component p of G whenever $\Sigma' = (G, \mathcal{H}', U')$ is another G-structured realization for $T(z)$. The following theorem establishes the equivalence of structured-minimality with simultaneous structured-controllability and structured-observability for G-structured realizations of a given formal power series $T(z)$.

THEOREM 9.1. *Suppose that $\Sigma = (G, \mathcal{H}, U)$ is a G-structured realization for the formal power series $T(z) = \sum_{w \in \mathcal{F}_E} T_w z^w$. Then Σ is a G-structured minimal realization for $T(z)$ if and only if Σ is both structured-controllable and structured-observable (with structure graph G).*

Proof. Suppose first that $\Sigma = (G, \mathcal{H}, U)$ is a structured-controllable and structured-observable realization of $T(z)$ and that $\Sigma' = (G, \mathcal{H}', U')$ is another structured realization of $T(z)$ (with the same structure graph G). By part (1) of Theorem 7.1, we may cut the realization Σ' down to a structured-controllable and structured-observable realization $\Sigma'_{c/o} = (G, \mathcal{H}'_{c/o}, U'_{c/o})$ for $T(z)$; as part of the construction we have $\dim \mathcal{H}'_p \geq \dim \mathcal{H}'_{p,c/o}$ for each $p \in P$. We now have that $\Sigma = (G, \mathcal{H}, U)$ and $\Sigma'_{c/o} = (G, \mathcal{H}'_{c/o}, U'_{c/o})$ are both structured-controllable and structured-observable realizations of the same formal power series $T(z)$. By the state-space-similarity theorem (Theorem 8.2), it follows that Σ and $\Sigma'_{c/o}$ are similarvia a state-space similarity

$\Gamma = \{\Gamma_p \colon \mathcal{H}_p \to \mathcal{H}'_{p,c/o} \colon p \in P\}$. In particular,

$$\dim \mathcal{H}_p = \dim \mathcal{H}'_{p,c/o} \le \dim \mathcal{H}'_p.$$

As Σ' was any other G-structured realization of $T(z)$, it follows that Σ is a G-structured minimal realization, as wanted.

Conversely, suppose that Σ is G-structured minimal. By part (1) of Theorem 7.1, we may cut Σ down to a structured-controllable and structured-observable realization $\Sigma_{c/o} = (G, \mathcal{H}_{c/o}, U_{c/o})$ of the same formal power series $T(z)$. By the construction in Theorem 7.1, $\mathcal{H}_{p,c/o} \subset \mathcal{H}_p$. On the other hand, by the assumption that Σ is G-structured minimal, we must also have $\dim \mathcal{H}_p \le \dim \mathcal{H}_{p,c/o}$, and hence we must have the equality $\mathcal{H}_p = \mathcal{H}_{p,c/o}$ for each $p \in P$. From the construction in Theorem 7.1, this means that the realization Σ is itself structured-controllable and structured-observable. This completes the proof of Theorem 9.1. ∎

10. Hankel operators. The notion of a Hankel operator $\mathbb{H}$ for a classical (1-D) linear system is the map which maps a past input sequence to the future output sequence, under the assumptions that the state has been initialized to be zero at $-\infty$ (roughly speaking) and that the future input string is set equal to zero. Since the controllability operator $\mathcal{C}$ maps the past history to the state at time zero, also under the assumption that the state has been initialized to be zero at $-\infty$, while the observability operator $\mathcal{O}$ maps a given state at time 0 into the future output sequence (under the assumption that the future input string is set equal to zero), we see immediately from the definitions that the Hankel operator $\mathbb{H}$ has the factorization $\mathbb{H} = \mathcal{O} \cdot \mathcal{C}$. For the case of SNMLSs, we have three notions ($\mathcal{C}_w$ for $w \in \mathcal{F}_E^{\infty_R}$, $\mathcal{C}_s$ for $s \in S$, and $\mathcal{C}_p$ for $p \in P$) of controllability operators which map some version of the past ($\mathcal{T}^w_{\text{past}}$, $\mathcal{T}^s_{\text{past}}$, or $\mathcal{T}^p_{\text{past}}$) to a state at the "present" position $\emptyset$, and three notions of observability operator ($\mathcal{O}_{s,w}$, $\mathcal{O}_s$, and $\mathcal{O}_p$ for $(s,w) \in S \times \mathcal{F}_E^{\infty_R}$, $s \in S$, and $p \in P$) mapping some state at the present position $\emptyset$ to outputs supported on some version of the future ($\mathcal{T}^w_{\text{future}}$, $\mathcal{T}_{\text{future}}$, or $\cup_{s\colon [s]=p} \mathcal{T}_{\text{future}}$. Thus a priori we have nine distinct possible notions of a Hankel operator. However, for purposes of the realization theory to be presented in section 11 below, only some of these are of interest for our purposes here, so we focus on them.

Let $\Sigma = (G, \mathcal{H}, U)$ be an SNMLS as in Definition 3.7. In this section we shall fix a cross section $p \mapsto s_p \in S$ of the map $[\cdot]\colon S \to P$ mapping a source vertex s to its associated path-connected component $[s] \in P$; i.e., for each $p \in P$, we let s_p be a fixed choice of element of S such that $[s_p] = p$. Consider any past input string $\{u(v)\}_{v \in \mathcal{T}^{s_p}_{\text{past}}}$. Run the system with this input string $u(w)$ for $w \in \mathcal{T}^{s_p}_{\text{past}}$ and with the state initialized to be zero in the distant past to generate a state $x(\emptyset)$ with s_pth component x_{s_p} equal to, say, $x_p \in \mathcal{H}_p$. For each $s \in S$ with $[s] = p$, we next run the system with zero inputs $u(w)$ for $w \in \mathcal{T}_{\text{future}}$ and with initial condition $x_s(\emptyset) = x_p$, $x_{s'}(\emptyset) = 0$ for $s' \ne s$. The result is an output sequence $\{y_s(w)\}_{w \in \mathcal{T}_{\text{future}}}$. The resulting composite map defined as taking the input string $\{u(v)\}_{v \in \mathcal{T}^{s_p}_{\text{past}}}$ to the output string $\{y_s(w)\}_{s\colon [s]=p; w \in \mathcal{T}_{\text{future}}}$ we define to be the Hankel operator $\mathbb{H}^p$:

$$\mathbb{H}^p = \mathcal{O}_p \mathcal{C}_{s_p} \colon \ell_{\text{fin}}(\mathcal{T}^{s_p}_{\text{past}}, \mathcal{U}) \to \oplus_{s\colon [s]=p} \ell(\mathcal{T}_{\text{future}}, \mathcal{Y}). \tag{10.1}$$

Explicitly, $\mathbb{H}^p$ is given as a bi-infinite matrix $[\mathbb{H}^p_{(s,w),v}]$ with rows indexed by pairs (s,w) with $s \in S$ with $[s] = p$ and with $w \in \mathcal{T}_{\text{future}}$, and with columns indexed by words $v \in \mathcal{T}^{s_p}_{\text{past}}$. In terms of the connecting operator U for Σ, the matrix entries are

given explicitly as

$$\begin{aligned}\mathbb{H}^p_{(s,w),w'} &= C_{\mathbf{s}(e_N)} A_{\mathbf{r}(e_N),\mathbf{s}(e_{N-1})} A_{\mathbf{r}(e_{N-1}),\mathbf{s}(e_{N-2})} \cdots A_{\mathbf{r}(e_2),\mathbf{s}(e_1)} \\ &\qquad \cdot A_{\mathbf{r}(e_1),s} A_{\mathbf{r}(e'_{N'}),\mathbf{s}(e'_{N'-1})} A_{\mathbf{r}(e'_{N'-1}),\mathbf{s}(e'_{N'-2})} \cdots A_{\mathbf{r}(e'_2),\mathbf{s}(e'_1)} B_{\mathbf{s}(e'_1)} \\ &= C_{\mathbf{s}(e_N)} A_{\mathbf{r}(e_N),\mathbf{s}(e_{N-1})} A_{\mathbf{r}(e_{N-1}),\mathbf{s}(e_{N-2})} \cdots A_{\mathbf{r}(e_2),\mathbf{s}(e_1)} A_{\mathbf{r}(e_1),s} \Psi_{w'} \qquad (10.2)\end{aligned}$$

if $w = e_N e_{N-1} \cdots e_2 e_1$ and $w' = e'_{N'} e'_{N'-1} \cdots e'_2 e'_1$, where $e'_{N'}$ is constrained to satisfy $\mathbf{s}(e'_{N'}) = s_p$ and where we use (5.1) to define $\Psi_{w'}$. (We leave it to the reader to give the appropriate interpretations for these formulas in case $N = 1$ and/or $N' = 0$.) As explained in the context of Remark 5.1, if we replace w' by w'' of the form

$$w'' = e_{s,\mathbf{r}(e'_{N'})} e'_{N'-1} \cdots e'_2 e'_1$$

for any s with $[s] = [\mathbf{s}(e'_{N'})] = [\mathbf{r}(e'_{N'})]$, then $\Psi_{w''} = \Psi_{w'}$. Since $v \in \mathcal{T}^{s_p}_{\text{past}}$, where $[s_p] = p$, we may therefore rewrite the Hankel matrix entry as a moment of the transfer function $T_\Sigma(z) = \sum_{w \in \mathcal{F}_E} T_w z^w$, namely,

$$\begin{aligned}\mathbb{H}^p_{(s,w),v} &= C_{\mathbf{s}(e_N)} A_{\mathbf{r}(e_N),\mathbf{s}(e_{N-1})} A_{\mathbf{r}(e_{N-1}),\mathbf{s}(e_{N-2})} \cdots A_{\mathbf{r}(e_2),\mathbf{s}(e_1)} \\ &\qquad \cdot A_{\mathbf{r}(e_1),\mathbf{s}(e_{s,\mathbf{r}(e'_{N'})})} A_{\mathbf{r}(e_{s,\mathbf{r}(e'_{N'})}),\mathbf{s}(e'_{N'-1})} A_{\mathbf{r}(e'_{N'-1}),\mathbf{s}(e'_{N'-2})} \cdots A_{\mathbf{r}(e'_2),\mathbf{s}(e'_1)} B_{\mathbf{s}(e'_1)} \\ &= T_{e_N e_{N-1} \cdots e_1 e_{s,\mathbf{r}(e'_{N'})} e'_{N'-1} \cdots e'_2 e'_1}, \qquad (10.3)\end{aligned}$$

or, more compactly,

$$\mathbb{H}^p_{(s,w),ev'} = T_{w e_{s,\mathbf{r}(e)} v'} \qquad (10.4)$$

for $s \in S$, $w \in \mathcal{T}_{\text{future}}$, and ev' (with $e \in E$ with $\mathbf{s}(e) = s_p$ and $v' \in \mathcal{F}_E$) the generic form of an element in $\mathcal{T}^{s_p}_{\text{past}}$.

From the factorization (10.1) and the definitions, it is easy to see the following result; we shall obtain a converse in section 11 below.

THEOREM 10.1. *Suppose that the SNMLS Σ (see Definition 3.7) is structured-controllable and structured-observable. Then the dimension of the auxiliary state-space $\mathcal{H}_p$ (for a given path-connected component $p \in P$ of the structure graph) is given by*

$$\dim \mathcal{H}_p = \operatorname{rank} \mathbb{H}^p.$$

Proof. By definition, $\mathcal{C}_{s_p}$ is a surjective map to $\mathcal{H}_p$ if Σ is structured-controllable, and $\mathcal{O}_p$ is an injective map if Σ is structured-observable. Hence the result is immediate from the factorization (10.1). ∎

COROLLARY 10.2. *If $T(z)$ is the transfer function of an SNMLS Σ having structure graph G, then, for each path-connected component $p \in P$, the Hankel operator $\mathbb{H}^p$ formed from G and $T(z)$ according to the formula* (10.4) *has finite rank.*

We shall obtain a converse of Corollary 10.2 in section 11 below.

11. Realization theory for structured noncommutative linear systems. Suppose that we are given an admissible graph G together with a formal power series $T(z) = \sum_{v \in \mathcal{F}_E} T_v z^v$ in noncommuting variables $z = (z_e : e \in E)$ indexed by the edge set E of G and with coefficients T_v equal to linear operators between the finite-dimensional linear spaces $\mathcal{U}$ and $\mathcal{Y}$. The realization problem associated with the data set $\mathbb{D} := (G, T(z))$ then is the following: *construct a finite-dimensional SNMLS $\Sigma = (G, \mathcal{H}, U)$ having G as its structure graph and $T(z)$ as its transfer function.*

A necessary condition for the problem to have a solution was formulated in Corollary 10.2. The content of the following theorem is the converse. We shall need the following conventions. Let G be an admissible graph. As in section 10, we assume that we have specified a cross section $p \mapsto s_p$ of the map $[\cdot]\colon S \to P$, so $s_p \in S$ with $[s_p] = p$ for each $p \in P$. For $v \in \mathcal{T}^s_{\rm past}$ (where $\mathcal{T}^s_{\rm past}$ is defined as in (5.6)), we let δ_v be the Kronecker delta function on $\mathcal{T}^s_{\rm past}$:

$$\delta_v(v') = \begin{cases} 1 & \text{if } v' = v, \\ 0 & \text{if } v' \neq v, \end{cases} \qquad \text{for } v' \in \mathcal{T}^s_{\rm past}.$$

Then $\{\delta_v u\colon v \in \mathcal{T}^s_{\rm past},\, u \in \mathcal{U}\}$ is a spanning set for the linear space $\ell_{\rm fin}(\mathcal{T}^s_{\rm past}, \mathcal{U})$. Recall the notation $e_{s,r}$ as in (3.1) for the unique edge connecting $s \in S$ to $r \in R$, defined whenever $[s] = [r]$, and the notation $w^{\wedge s}$ introduced in (5.8).

THEOREM 11.1. *Suppose that we are given the data set $\mathbb{D} = (G, T(z))$ for a realization problem as above. For each path-connected component $p \in P$ of G, associate the Hankel matrix $\mathbb{H}^p$ as in* (10.4). *Then the realization problem for the data set $\mathbb{D}$ is solvable if and only if*

$$\operatorname{rank} \mathbb{H}^p < \infty \quad \textit{for each } p \in P. \tag{11.1}$$

When the condition (11.1) *holds, a structured-minimal realization of $T(z)$ can be constructed as follows.*

For each $p \in P$, let $\mathcal{H}_p$ be the linear space

$$\mathcal{H}_p = \ell_{fin}(\mathcal{T}^{s_p}_{past}, \mathcal{U}) / \ker \mathbb{H}^p, \tag{11.2}$$

and set $\mathcal{H}$ equal to the collection

$$\mathcal{H} = \{\mathcal{H}_p\colon p \in P\}.$$

For each source vertex $s \in S$ and range vertex $r \in R$ of G, define linear operators $A_{r,s}\colon \mathcal{H}_{[s]} \to \mathcal{H}_{[r]}$, $B\colon \mathcal{U} \to \mathcal{H}_{[r]}$, $C_s\colon \mathcal{H}_{[s]} \to \mathcal{Y}$, and $D\colon \mathcal{U} \to \mathcal{Y}$ by

$$A_{r,s}\colon \left[\{u(v)\}_{v \in \mathcal{T}^{s_{[s]}}_{past}}\right]_{\mathcal{H}_{[s]}} \mapsto \left[\{u'(v)\}_{v \in \mathcal{T}^{s_{[r]}}_{past}}\right]_{\mathcal{H}_{[r]}}, \ \textit{where}$$

$$u'(v) = \begin{cases} u((v')^{\wedge s_{[s]}}) & \textit{if } v \textit{ has the form } v = e_{s_{[r]},r} v' \textit{ with } v' \in \mathcal{T}^s_{past}, \\ 0 & \textit{otherwise,} \end{cases}$$

$$B_r\colon u \mapsto [\delta_{e_{s_{[r]},r}} u]_{\mathcal{H}_{[r]}},$$

$$C_s\colon \left[\{u(v)\}_{v \in \mathcal{T}^{s_{[s]}}_{past}}\right]_{\mathcal{H}_{[s]}} \mapsto \mathbb{H}^{[s]}_{(s,\emptyset),\cdot}\left(\{u(v)\}_{v \in \mathcal{T}^{s_{[s]}}_{past}}\right)$$

$$= \sum_{v \in \mathcal{T}^{s_{[s]}}_{past}} T_{v^{\wedge s}} u(v),$$

$$D = T_\emptyset. \tag{11.3}$$

Use (11.3) *to define a connection matrix U by*

$$U = \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} [A_{r,s}] & [B_r] \\ [C_s] & D \end{bmatrix} : \begin{bmatrix} \oplus_{s \in S} \mathcal{H}_{[s]} \\ \mathcal{U} \end{bmatrix} \to \begin{bmatrix} \oplus_{r \in R} \mathcal{H}_{[r]} \\ \mathcal{Y} \end{bmatrix}.$$

Then the collection $\Sigma = (G, \mathcal{H}, U)$ is a structured-minimal SNMLS with structure graph G having $T(z)$ as its transfer function.

Proof. We have already observed in Corollary 10.2 the necessity of the condition (11.1) for the realization problem to have a solution. It remains to prove the sufficiency. This follows if we can verify that the formulas (11.2) and (11.3) provide a structured-minimal realization of $T(z)$ (with structure matrix G).

As a preliminary step, we note that the formula for $A_{r,s}$ in (11.3) when specialized to elements of $\mathcal{H}_{[s]}$ of the form $[\delta_v u]_{\mathcal{H}_{[s]}}$ (where $v \in \mathcal{T}_{\text{past}}^{s_{[s]}}$) assumes the form

$$A_{r,s}\colon [\delta_v u]_{\mathcal{H}_{[s]}} \mapsto [\delta_{e_{s_{[r]},r}(v^{\wedge s})}]_{\mathcal{H}_{[r]}}. \tag{11.4}$$

Note also that the set $\{[\delta_v u]_{\mathcal{H}_{[s]}}\colon v \in \mathcal{T}_{\text{past}}^{s_{[s]}},\ u \in \mathcal{U}\}$ is a spanning set for $\mathcal{H}_{s_{[s]}}$ since $\{\delta_v u\colon v \in \mathcal{T}_{\text{past}}^{s_{[s]}},\ u \in \mathcal{U}\}$ is a spanning set for $\ell_{\text{fin}}(\mathcal{T}_{\text{past}}^{s_{[s]}}, \mathcal{U})$. Similarly, the action of C_s in (11.3) on delta functions can be written as

$$C_s\colon [\delta_v u]_{\mathcal{H}_{[s]}} \mapsto T_{v^{\wedge s}} u \quad \text{for } v \in \mathcal{T}_{\text{past}}^{s_{[s]}}. \tag{11.5}$$

The verification proceeds via a number of steps.

Step 1: *Verification that $A_{r,s}$ is well defined.* Suppose that $\{u(v)\}_{v \in \mathcal{T}_{\text{past}}^{s_{[s]}}}$ represents the zero element of $\mathcal{H}_{[s]}$; thus $\mathbb{H}^{[s]}(\{u(v)\}_{v \in \mathcal{T}_{\text{past}}^{s_{[s]}}}) = 0$. Explicitly, this means

$$\sum_{v \in \mathcal{T}_{\text{past}}^{s_{[s]}}} T_{wv^{\wedge s'}} u(v) = 0 \quad \text{for all } w \in \mathcal{F}_E \text{ and } s' \in S \text{ with } [s'] = [s]. \tag{11.6}$$

View $\{u(v)\}_{v \in \mathcal{T}_{\text{past}}^{s_{[s]}}}$ as equal to $\sum_{v \in \mathcal{T}_{\text{past}}^{s_{[s]}}} \delta_v u(v)$, and use the formula (11.4) combined with linearity: the result is

$$A_{r,s}\colon \sum_{v \in \mathcal{T}_{\text{past}}^{s_{[s]}}} \delta_v u(v) \mapsto \sum_{v \in \mathcal{T}_{\text{past}}^{s_{[s]}}} \delta_{e_{s_{[r]},r} v^{\wedge s}} u(v) \in \ell(\mathcal{T}_{\text{past}}^{s_{[r]}}, \mathcal{U}).$$

For the right-hand side of this formula to represent the zero element of $\mathcal{H}_{[r]}$ we need to have $\mathbb{H}^{[r]}(\sum_{v \in \mathcal{T}_{\text{past}}^{s_{[s]}}} \delta_{e_{s_{[r]},r} v^{\wedge s}} u(v)) = 0$, which is to say

$$\sum_{v \in \mathcal{T}_{\text{past}}^{s_{[s]}}} T_{w'(e_{s_{[r]},r} v^{\wedge s})^{\wedge s''}} u(v) = 0 \quad \text{for all } w' \in \mathcal{F}_E, s'' \in S \text{ with } [s''] = [r]. \tag{11.7}$$

However, it is easily verified that

$$(e_{s_{[r]},r} v^{\wedge s})^{\wedge s''} = e_{s'',r} v^{\wedge s}.$$

Hence the condition (11.7) amounts to the known condition (11.6) for the special case $w = w' e_{s'',r}$ and $s' = s$. We conclude that the formula for $A_{r,s}$ in (11.3), or equivalently the formula (11.4) for $A_{r,s}$ on a spanning subset of $\mathcal{H}_{[r]}$, is well defined.

Step 2: *Verification that C_s is well defined.* We again suppose that $\{u(v)\}_{v \in \mathcal{T}_{\text{past}}^{s_{[s]}}}$ represents the zero element of $\mathcal{H}_{[s]}$, i.e., that (11.6) holds. Then $C_s(\{u(v)\}_{v \in \mathcal{T}_{\text{past}}^{[s]}})$ by definition is the left-hand side of (11.6) for the special case $w = \emptyset$ and $s' = s$. Hence C_s is well defined, as wanted.

Step 3: *Verification that* $T_\Sigma(z) = T(z)$. Let $e \in E$ be an edge of G. Use the formula for B_r in (11.3) and the formula (11.5) for the action of C_s on delta functions to compute

$$\begin{aligned} C_{\mathbf{s}(e)}B_{\mathbf{r}(e)}u &= C_{\mathbf{s}(e)}\left(\left[\delta_{e_{s_{[\mathbf{r}(e)]},\mathbf{r}(e)}}u\right]_{\mathcal{H}_{[\mathbf{r}(e)]}}\right) = T_{(e_{s_{[\mathbf{r}(e)]},\mathbf{r}(e)})^{\wedge\mathbf{s}(e)}u} \\ &= T_e u, \end{aligned} \tag{11.8}$$

where the equality $(e_{s_{[\mathbf{r}(e)]},\mathbf{r}(e)})^{\wedge\mathbf{s}(e)} = e_{\mathbf{s}(e),\mathbf{r}(e)} = e$ follows from the uniqueness condition (3) in the admissibility conditions (see Definition 3.1) for the graph G. Similarly, by using the formula for B_r in (11.3) combined with (11.4), a straightforward induction argument gives that, for any word $w = e_N e_{N-1}\cdots e_2 e_1$ of length at least 2,

$$A_{\mathbf{r}(e_N),\mathbf{s}(e_{N-1})}A_{\mathbf{r}(e_{N-1}),\mathbf{s}(e_{N-2})}\cdots A_{\mathbf{r}(e_2),\mathbf{s}(e_1)}B_{\mathbf{r}(e_1)}u = \left[\delta_{w^{\wedge s_{[\mathbf{r}(e_N)]}}}u\right]_{\mathcal{H}_{[\mathbf{r}(e_N)]}}. \tag{11.9}$$

From the uniqueness axiom in Definition 3.1 we have

$$(w^{\wedge[\mathbf{r}(e_N)]})^{\wedge\mathbf{s}(e_N)} = w \quad \text{if } e_N = LL[w]. \tag{11.10}$$

Applying the formula (11.5) to (11.9) and using (11.10), we get

$$\begin{aligned} &C_{\mathbf{s}(e_N)}A_{\mathbf{r}(e_N),\mathbf{s}(e_{N-1})}A_{\mathbf{r}(e_{N-1}),\mathbf{s}(e_{N-2})}\cdots A_{\mathbf{r}(e_2),\mathbf{s}(e_1)}B_{\mathbf{r}(e_1)}u = T_{(w^{\wedge[\mathbf{r}(e_N)]})^{\wedge\mathbf{s}(e_N)}u} \\ &= T_w u \text{ for } w = e_N e_{N-1}\cdots e_2 e_1. \end{aligned} \tag{11.11}$$

Combining (11.8) and (11) with the definition $D = T_\emptyset$ in (11.3), we see that $T_\Sigma(z) = T(z)$, as wanted.

Step 4: *Verification that* Σ *is structured-controllable.* By formula (11.9) we have

$$\Psi_w u = [\delta_w u]_{\mathcal{H}_p} \quad \text{for } w \in \mathcal{T}_{\text{past}}^{s_p} \text{ and } u \in \mathcal{U}.$$

As the set $\{[\delta_w u]_{\mathcal{H}_p} : w \in \mathcal{T}_{\text{past}}^{s_p} \text{ and } u \in \mathcal{U}\}$ is spanning for the space

$$\mathcal{H}_p = \ell_{\text{fin}}(\mathcal{T}_{\text{past}}^{s_p}, \mathcal{U})/\ker \mathbb{H}^p,$$

we conclude that Σ is structured-controllable, as wanted.

Step 5: *Verification that* Σ *is structured-observable.* From the various definitions it is easy to verify that

$$\mathcal{O}_s\left(\left[\{u(v)\}_{v\in\mathcal{T}_{\text{past}}^{s_{[s]}}}\right]_{\mathcal{H}_{[s]}}\right) = \mathbb{H}^{[s]}_{(s,\cdot),\cdot}\left(\{u(v)\}_{v\in\mathcal{T}_{\text{past}}^{s_{[s]}}}\right) \in \ell(\mathcal{T}_{\text{future}}, \mathcal{Y})$$

for each source vertex $s \in S$. Since, by definition, $\mathcal{O}_p = \text{col}_{s:\,[s]=p}\,\mathcal{O}_s$ for each $p \in P$, we can then make the identification

$$\mathcal{O}_p\left(\left[\{u(v)\}_{v\in\mathcal{T}_{\text{past}}^{s_p}}\right]_{\mathcal{H}_p}\right) = \mathbb{H}^p\left(\{u(v)\}_{v\in\mathcal{T}_{\text{past}}^{s_p}}\right) \in \oplus_{s:\,[s]=p}\ell(\mathcal{T}_{\text{future}}, \mathcal{Y}).$$

In this way we see that $[\{u(v)\}_{v\in\mathcal{T}_{\text{past}}^{s_p}}]_{\mathcal{H}_p} \in \ker \mathcal{O}_p$ if and only if $\{u(v)\}_{v\in\mathcal{T}_{\text{past}}^{s_p}} \in \ker \mathbb{H}^p$, i.e., if and only if $[\{u(v)\}_{v\in\mathcal{T}_{\text{past}}^{s_p}}]_{\mathcal{H}_p}$ is the zero equivalence class in $\mathcal{H}_p$. We conclude that Σ is structured-observable as wanted, and the proof of Theorem 11.1 is now complete. ∎

We now consider the situation where the formal power series $T(z) = \sum_{v \in \mathcal{F}_d} T_v z^v$ is given but the admissible graph G is not specified. By comparing the various Hankel operators involved, we have the following result.

THEOREM 11.2. *Suppose that we are given the formal power series in the d noncommuting variables $z = (z_1, \dots, z_d)$, and let G and G' be two admissible graphs with edge sets E and E' of the same cardinality. Then $T(z)$ has a G-structured realization $\Sigma = \{G, \mathcal{H}, U\}$ if and only if $T(z)$ has a G'-structured realization $\Sigma' = \{G', \mathcal{H}', U'\}$.*

Proof. Let G be any admissible graph with edge set E labeled as $E = \{1, \dots, d\}$, and let G^{FM} be the Fornasini–Marchesini admissible graph with source-vertex set $S = \{1\}$, range-vertex set $R = \{1, \dots, d\}$, and edge set $E = \{1, \dots, d\}$, with $\mathbf{s}(j) = 1$ and $\mathbf{r}(j) = j$ for $j = 1, \dots, d$. We show that $T(z)$ has a G-structured realization $\Sigma = (G, \mathcal{H}, U)$ if and only if $T(z)$ has a G^{FM}-structured realization $\Sigma^{FM} = (G^{FM}, \mathcal{H}^{FM}, U^{FM})$. For s in S define the Hankel operator $\mathbb{H}^s \colon \ell_{\mathrm{fin}}(\mathcal{T}^s_{\mathrm{past}}, \mathcal{U}) \to \ell(\mathcal{T}_{\mathrm{future}}, \mathcal{Y})$ by

$$\mathbb{H}^s \colon \{u(v)\}_{v \in \mathcal{T}^s_{\mathrm{past}}} \mapsto \mathbb{H}^{[s]}_{(s,\cdot),\cdot} \left(\{u(v^{\wedge s})\}_{v \in \mathcal{T}^{s_{[s]}}_{\mathrm{past}}} \right).$$

As the map $v \mapsto v^{s_{[s]}}$ is a bijection between $\mathcal{T}^s_{\mathrm{past}}$ and $\mathcal{T}^{s_{[s]}}_{\mathrm{past}}$, we see that $\mathbb{H}^s$ is similar to $\mathbb{H}^{[s]}_{(s,\cdot),\cdot}$. By definition,

$$\mathbb{H}^p = \mathrm{col}_{s \colon [s]=p} \left[\mathbb{H}^p_{(s,\cdot),\cdot} \right]$$

from which we get the estimates

$$\max_{s \colon [s]=p} \operatorname{rank} \mathbb{H}^p_{(s,\cdot),\cdot} \le \operatorname{rank} \mathbb{H}^p \le \sum_{s \colon [s]=p} \operatorname{rank} \mathbb{H}^p_{(s,\cdot),\cdot}. \tag{11.12}$$

As we observed above that $\mathbb{H}^s$ and $\mathbb{H}^{[s]}_{(s,\cdot),\cdot}$ have the same rank, we can rewrite (11.12) as

$$\max_{s \colon [s]=p} \operatorname{rank} \mathbb{H}^s \le \operatorname{rank} \mathbb{H}^p \le \sum_{s \colon [s]=p} \operatorname{rank} \mathbb{H}^s. \tag{11.13}$$

From the characterization (10.4) of $\mathbb{H}^p$ we see that

$$\mathbb{H}^{FM} = \mathrm{col}_{p \in P} \, \mathrm{col}_{s \colon [s]=p} [\mathbb{H}^s] = \mathrm{col}_{s \in S} [\mathbb{H}^s]. \tag{11.14}$$

By combining (10.4) with the estimates (11.13), we see that $\mathbb{H}^{FM}$ has finite rank if and only if $\mathbb{H}^p$ has finite rank for each $p \in P$.

Now suppose that G and G' are two admissible graphs with the same edge set E and that $T(z)$ is a given formal power series in the noncommuting variables $z = (z_e \colon e \in E)$. By the first part of the proof, realizability of $T(z)$ as the transfer function of an SNMLS with structure graph G and realizability of $T(z)$ as the transfer function of an SNMLS with structure graph G' are each equivalent to realizability of $T(z)$ as the transfer function of a noncommutative Fornasini–Marchesini system with structure graph G^{FM} having edge set E. Hence G-realizability and G'-realizability are equivalent to each other. This completes the proof of Theorem 11.2. ∎

12. Recognizable and rational formal power series. Formal power series in noncommuting variables of the form arising here have come up in the theory of formal languages as studied in computer science [15]. For the sake of concreteness we index the noncommuting variables simply by $\{1,\dots,d\}$ and work with the semigroup $\mathcal{F}_d$ generated by the concrete set of letters $\{1,\dots,d\}$, as was done in sections 2.1 and 2.2 in the setting of noncommutative Fornasini–Marchesini and Givone–Roesser systems. We specialize the discussion in [15] to the setting here, where we take the scalars to be the field $\mathbb{C}$ of complex numbers rather than a general semiring, i.e., a "ring without subtraction." A formal power series $\sum_{v\in\mathcal{F}_d} T_v z^v$ (with coefficients T_v equal to linear operators acting between the finite-dimensional linear spaces $\mathcal{U}$ and $\mathcal{Y}$) is said to be *recognizable* if there are finite-dimensional linear space $\mathcal{H}$ and operators $A_1,\dots,A_d\colon \mathcal{H}\to\mathcal{H}$, $B\colon \mathcal{U}\to\mathcal{H}$, and $C\colon \mathcal{H}\to\mathcal{Y}$ such that

$$T_v = CA^vB \quad \text{for } v\in\mathcal{F}_d.$$

In terms of the linear systems discussed here, one can view a recognizable series $T(z)=\sum_{v\in\mathcal{F}_d}(CA^vB)z^v$ as the transfer function of a noncommutative Fornasini–Marchesini system

$$\Sigma^{FM}\colon \begin{cases} x(jw) &= A_jx(w)+B_ju(w) \quad \text{for } j=1,\dots,d,\\ y(w) &= Cx(w)+Du(w), \end{cases}$$

with the special structure that

$$B_j =: B \text{ is independent of } j \text{ and } D=CB.$$

More economical is to consider the recognizable series as the transfer function of a system of the form

$$\Sigma^{\mathrm{rec}}\colon \begin{cases} x(1w) &= A_1x(w)+Bu(1w),\\ &\vdots\\ x(dw) &= A_dx(w)+Bu(dw),\\ y(w) &= Cx(w). \end{cases} \tag{12.1}$$

One can check that application of the formal noncommutative Z-transform (2.2) to the system equations Σ^{rec} yields the frequency-domain formulas

$$\begin{aligned} \widehat{x}(z) &= (I-Z_{\mathrm{row}}(z)A)^{-1}(x(\emptyset)-Bu(\emptyset))+(I-Z_{\mathrm{row}}(z)A)^{-1}B\widehat{u}(z),\\ \widehat{y}(z) &= C(I-Z_{\mathrm{row}}(z)A)^{-1}(x(\emptyset)-Bu(\emptyset))+T_{\Sigma^{\mathrm{rec}}}(z)\cdot\widehat{u}(z), \end{aligned} \tag{12.2}$$

where the *transfer function* $T_{\Sigma^{\mathrm{rec}}}(z)$ for the recognizable system Σ^{rec} given by

$$T_{\Sigma^{\mathrm{rec}}}(z)=\sum_{v\in\mathcal{F}_d} CA^vBz^v \tag{12.3}$$

has the form of a recognizable formal series. In particular, if the initial condition is given by the input-injection $x(\emptyset)=Bu(\emptyset)$, then multiplication by the transfer function $T_{\Sigma^{\mathrm{rec}}}(z)$ provides the input-output map in the frequency domain $\widehat{y}(z)=T_{\Sigma^{\mathrm{rec}}}(z)\widehat{u}(z)$.

All the results in sections 5, 6, 8, and 11 (notions of controllability and observability, equivalence of controllability and observability with minimality, state-space similarity theorem, realization theorem) have parallels for the case of recognizable

systems in place of general SNMLSs; in fact, as surveyed nicely in Chapters 1 and 2 of [15], all these results, with the exception of the identification of a recognizable series $T(z) = \sum_{v \in \mathcal{F}_d} (CA^v B) z^v$ as the transfer function of a noncommutative linear system of the form (12.1), already appear in the literature—even in the more general setting where the scalars are taken to be a general semiring rather than the field $\mathbb{C}$ of complex numbers as is done here (see [37, 38, 17, 39, 20, 21, 22, 23]). We now survey these results from our system-theoretic perspective.

To obtain a physical interpretation for the recognizable controllability operator $\mathcal{C}^{\mathrm{rec}}$ introduced below, it is natural to define the backward system equations giving the evolution on the past $\mathcal{T}^{\mathrm{rec}}_{\mathrm{past}} = \mathcal{F}_d$ to be

$$\Sigma^{\mathrm{rec}}_{\mathrm{backward}}\colon \begin{cases} x(w) &= \sum_{i=1}^{d} A_i x(wi) + Bu(w), \\ y(w) &= Cx(w). \end{cases} \tag{12.4}$$

If we run the backward system equations on the past and present $\mathcal{T}^{\mathrm{rec}}_{\mathrm{past}} := \mathcal{F}_d$ with the state initialized to be zero sufficiently far in the past and with an input string $\{u(w)\}_{w \in \mathcal{T}^{\mathrm{rec}}_{\mathrm{past}}}$ with finite support on $\mathcal{T}^{\mathrm{rec}}_{\mathrm{past}}$ to compute the state $x(\emptyset)$ at location $\emptyset$, the result is

$$x(\emptyset) = \mathcal{C}^{\mathrm{rec}}(\{u(w)\}_{w \in \mathcal{T}^{\mathrm{rec}}_{\mathrm{past}}}),$$

where the *recognizable controllability operator* $\mathcal{C}^{\mathrm{rec}}$ is given by

$$\mathcal{C}^{\mathrm{rec}} = \mathrm{row}_{w \in \mathcal{T}^{\mathrm{rec}}_{\mathrm{past}}} A^w B, \tag{12.5}$$

where we set $A^w = A_{i_N} A_{i_{N-1}} \cdots A_{i_1}$ if $w = i_N i_{N-1} \cdots i_1 \in \mathcal{T}^{\mathrm{rec}}_{\mathrm{past}}$ (with $A^{\emptyset} = I_{\mathcal{H}}$). Note that this controllability operator has close to the same form as the Fornasini–Marchesini controllability operator $\mathcal{C}^{FM}$ (2.6); the difference is that a recognizable system has only one input operator B and that the columns of $\mathcal{C}^{\mathrm{rec}}$ are indexed by $\mathcal{T}^{\mathrm{rec}}_{\mathrm{past}}$ which includes the empty word, with $[\mathcal{C}^{\mathrm{rec}}]_{\emptyset} = B$.

We say that the system Σ^{rec} is *recognizable-controllable* if the image $\operatorname{im} \mathcal{C}^{\mathrm{rec}}$ of the recognizable-controllability operator $\mathcal{C}^{\mathrm{rec}}$ is the whole state-space $\mathcal{H}$.

The observability operator $\mathcal{O}^{\mathrm{rec}}\colon \mathcal{H} \to \ell(\mathcal{T}^{\mathrm{rec}}_{\mathrm{future}}, \mathcal{Y})$ produces the future output $\{y(v)\}_{v \in \mathcal{T}^{\mathrm{rec}}_{\mathrm{future}}}$ generated by the system for a given prescribed initial condition $x(\emptyset) \in \mathcal{H}$ under the assumption that the zero input string $\{u(v)\}_{v \in \mathcal{T}^{\mathrm{rec}}_{\mathrm{future}}}$ is fed into the system; explicitly, we have[4]

$$\mathcal{O}^{\mathrm{rec}} = \operatorname*{row}_{v \in \mathcal{F}_d} CA^v. \tag{12.6}$$

Note that $\mathcal{O}^{\mathrm{rec}}$ has exactly the same form as the Fornasini–Marchesini observability operator $\mathcal{O}^{FM}$ from (2.7). We say that the system Σ^{rec} is *recognizable-observable* if the recognizable-observability operator $\mathcal{O}^{\mathrm{rec}}$ is injective on $\mathcal{H}$.

We can now obtain a *recognizable Kalman decomposition* of the state-space $\mathcal{H}$,

$$\mathcal{H} = \mathcal{H}_{c/o} \oplus \mathcal{H}_{c/no} \oplus \mathcal{H}_{nc/o} \oplus \mathcal{H}_{nc/no},$$

[4]Here $\mathcal{T}^{\mathrm{rec}}_{\mathrm{future}}$ is taken to be $\mathcal{F}_d$; the location $\emptyset$ in $\mathcal{T}^{\mathrm{rec}}_{\mathrm{future}}$ is identified with the location $\emptyset$ in $\mathcal{T}^{\mathrm{rec}}_{\mathrm{past}}$ (i.e., both $\mathcal{T}^{\mathrm{rec}}_{\mathrm{future}}$ and $\mathcal{T}^{\mathrm{rec}}_{\mathrm{past}}$ contain the "present"), but a given nonempty word w as an element of the future $\mathcal{T}^{\mathrm{rec}}_{\mathrm{future}}$ is to be considered distinct from the same word w considered as an element of the past $\mathcal{T}^{\mathrm{rec}}_{\mathrm{past}}$.

by the same recipe used in section 7 (by using $\mathcal{C}^{\rm rec}$ in place of $\mathcal{C}_{s_p}$ and $\mathcal{O}^{\rm rec}$ in place of $\mathcal{O}_p$). We then obtain the decompositions

$$A_j = \begin{bmatrix} A_{j;c/o,c/o} & 0 & A_{j;c/o,nc/o} & 0 \\ A_{j;c/no,c/o} & A_{j;c/no,c/no} & A_{j;c/no,nc/o} & A_{j;c/no,nc/no} \\ 0 & 0 & A_{j;nc/o,nc/o} & 0 \\ 0 & 0 & A_{j;nc/no,nc/o} & A_{j;nc/no,nc/no} \end{bmatrix},$$

$$B = \begin{bmatrix} B_{c/o} \\ B_{c/no} \\ 0 \\ 0 \end{bmatrix}, \qquad C = \begin{bmatrix} C_{c/o} & 0 & C_{nc/o} & 0 \end{bmatrix} \tag{12.7}$$

for the system matrices $A_1, \dots, A_d, B, C$ of $\Sigma^{\rm rec}$. It is then easily verified that the reduced recognizable system $\Sigma^{\rm rec}_{c/o}$ with system matrices

$$A_{1;c/o,c/o}, \dots, A_{d;c/o,c/o}, B_{c/o}, C_{c/o}$$

is both recognizable-controllable and recognizable-observable and produces the same transfer function: $T_{\Sigma^{\rm rec}}(z) = T_{\Sigma^{\rm rec}_{c/o}}(z)$. Given two recognizable systems $\Sigma^{\rm rec}$ with system matrices $A_1, \dots, A_d, B, C$ and $\Sigma^{\rm rec\prime}$ with system matrices $A'_1, \dots, A'_d, B', C'$, let us say that $\Sigma^{\rm rec}$ and $\Sigma^{\rm rec\prime}$ are *recognizable-similar* if there is a bijective linear map $\Gamma\colon \mathcal{H} \to \mathcal{H}'$ so that $\Gamma A_j = A'_j \Gamma$ for $j = 1, \dots, d$, $\Gamma B = B'$, and $C' = C\Gamma$. Following the same argument as in section 8, we have the *state-space similarity theorem for recognizable systems*: *given two recognizable systems* $\Sigma^{rec} = (A_1 \dots, A_d, B, C)$ *and* $\Sigma^{rec\prime} = (A'_1, \dots, A'_d, B', C')$ *with the same input-space* $\mathcal{U}$ *and output-space* $\mathcal{Y}$*, which are both recognizable-controllable and recognizable-observable, then* Σ^{rec} *and* $\Sigma^{rec\prime}$ *have the same transfer function*

$$T_{\Sigma^{rec}}(z) = T_{\Sigma^{rec\prime}}(z)$$

if and only if Σ^{rec} *and* $\Sigma^{rec\prime}$ *are recognizable-similar.* Furthermore, one can say that the recognizable system $\Sigma^{\rm rec}$ with state-space $\mathcal{H}$ is a *recognizable-minimal realization* for its transfer function $T(z) = T_{\Sigma^{\rm rec}}(z)$ if, whenever $\Sigma^{\rm rec\prime}$ with state-space $\mathcal{H}'$ is any other recognizable realization for the same $T(z)$, then $\dim \mathcal{H} \le \dim \mathcal{H}'$. Following the same line of argument as in section 9, one can show the following: *the recognizable system* Σ^{rec} *is a recognizable-minimal realization of its transfer function* $T_{\Sigma^{rec}}(z)$ *if and only if* Σ^{rec} *is recognizable-controllable and recognizable-observable.*

We next define the *recognizable Hankel operator* by

$$\mathbb{H}^{\rm rec} = \mathcal{O}^{\rm rec} \cdot \mathcal{C}^{\rm rec}\colon \ell_{\rm fin}(\mathcal{T}^{\rm rec}_{\rm past}, \mathcal{U}) \to \ell(\mathcal{T}^{\rm rec}_{\rm future}, \mathcal{Y}). \tag{12.8}$$

The matrix entries of $\mathbb{H}^{\rm rec}$ are then given by

$$\mathbb{H}^{\rm rec}_{w,v} = CA^{wv}B \quad \text{for } w, v \in \mathcal{F}_d \tag{12.9}$$

or directly in terms of the Taylor coefficients of the transfer function $T_{\Sigma^{\rm rec}}(z) = \sum_{v \in \mathcal{F}_d} T_v z^v$ as

$$\mathbb{H}^{\rm rec}_{w,v} = T_{wv} \quad \text{for } w, v \in \mathcal{F}_d. \tag{12.10}$$

In the case that $\Sigma^{\rm rec}$ is both recognizable-controllable and recognizable-observable, we see from the factorization (12.8) that

$$\operatorname{rank} \mathbb{H}^{\rm rec} = \dim \mathcal{H}.$$

In particular, $\operatorname{rank}\mathbb{H}^{\mathrm{rec}} < \infty$, where we now use (12.10) to define $\mathbb{H}^{\mathrm{rec}}$ directly in terms of the formal power series $T(z) = \sum_{v\in\mathcal{F}_d} T_v z^v$, which is a necessary condition for $T(z)$ to have a recognizable realization $T(z) = \sum_{v\in\mathcal{F}_d}(CA^vB)z^v$. For the converse, we have the following realization theorem.

THEOREM 12.1. *Let the formal power series $T(z) = \sum_{v\in\mathcal{F}_d} T_v z^v$ in d noncommuting indeterminates $z = (z_1, \dots, z_d)$, with coefficients T_v equal to linear operators between the linear spaces $\mathcal{U}$ and $\mathcal{Y}$, be given. Then a necessary and sufficient condition for $T(z)$ to be recognizable, i.e., for the existence of a linear space $\mathcal{H}$ and operators $A_1, \dots, A_d$ on $\mathcal{H}$, $B\colon \mathcal{U} \to \mathcal{H}$, and $C\colon \mathcal{H} \to \mathcal{Y}$ with $T_v = CA^vB$ for $v \in \mathcal{F}_d$, is that*

$$\operatorname{rank}\mathbb{H}^{rec} < \infty. \tag{12.11}$$

When this holds, a recognizable-minimal realization $(A_1, \dots, A_d, B, C)$ can be constructed as follows: set

$$\mathcal{H} = \ell_{fin}(\mathcal{T}^{rec}_{past}, \mathcal{U})/\ker\mathbb{H}^{rec} \tag{12.12}$$

and define operators $A_j\colon \mathcal{H} \to \mathcal{H}$ (for $j = 1, \dots, d$), $B\colon \mathcal{U} \to \mathcal{H}$, and $C\colon \mathcal{H} \to \mathcal{Y}$:

$$A_j\colon [\delta_v]_{\mathcal{H}} \mapsto [\delta_{jv}]_{\mathcal{H}} \quad \text{for } v \in \mathcal{T}^{rec}_{past},$$

$$B\colon u \mapsto [\delta_\emptyset]_{\mathcal{H}}, \tag{12.13}$$

$$C\colon [\{u(v)\}_{v\in\mathcal{T}^{rec}_{past}}]_{\mathcal{H}} \mapsto \sum_{v\in\mathcal{T}^{rec}_{past}} T_v u(v). \tag{12.14}$$

Proof. The proof parallels the ideas in the proof of Theorem 11.1, so we omit the details. The result is also essentially contained in Theorem 1.5 of [15] (without any system-theoretic interpretation using the system equations (12.1) and (12.4)), where it is attributed to [17] and [20]. □

Note that the recognizable Hankel $\mathbb{H}^{\mathrm{rec}}$ is almost the same as the Fornasini–Marchesini Hankel $\mathbb{H}^{\mathrm{FM}}$; namely, we have

$$\mathbb{H}^{\mathrm{rec}} = \begin{bmatrix}\operatorname{col}_{v\in\mathcal{F}_d}[T_v] & \mathbb{H}^{FM}\end{bmatrix}. \tag{12.15}$$

In particular, we see that

$$\operatorname{rank}\mathbb{H}^{FM} \le \operatorname{rank}\mathbb{H}^{\mathrm{rec}} \le \dim\mathcal{U} + \operatorname{rank}\mathbb{H}^{FM},$$

and hence $\mathbb{H}^{\mathrm{FM}}$ has finite rank if and only if $\mathbb{H}^{\mathrm{rec}}$ has finite rank. Combining this observation with Theorems 12.1, 11.1, and 11.2, we arrive at the following result.

COROLLARY 12.2. *Let a formal power series $T(z) = \sum_{v\in\mathcal{F}_d} T_v z^v$ in d noncommuting variables $z = (z_1, \dots, z_d)$ and an admissible graph G with edge set E labeled as $E = \{1, \dots, d\}$ be given. Then T has a realization of the form $T(z) = D + C(I - Z_\Sigma(z)A)^{-1}Z_\Sigma(z)B$ for an SNMLS $\Sigma = (G, \mathcal{H}, U)$ if and only if $T(z) = C(I - z_1A_1 - \cdots - z_dA_d)^{-1}B$ is recognizable.*

A related notion arising in the theory of formal languages, particularly in the work of Schützenberger, is that of rationality. We say that a formal power series $T(z) = \sum_{v\in\mathcal{F}_d} T_v z^v \in \mathbb{C}\langle\langle z\rangle\rangle$ in noncommuting variables $z = (z_1, \dots, z_d)$ with scalar coefficients $T_v \in \mathbb{C}$ is *rational* if it is in the smallest subalgebra of $\mathbb{C}\langle\langle z\rangle\rangle$ which contains the polynomials and is invariant under the operator $R(z) \mapsto R^*(z) = \sum_{n=0}^{\infty}(R(z))^n$ defined on proper formal power series $R(z) = \sum_{v\in\mathcal{F}_d\setminus\{\emptyset\}} R_v z^v$. The demand here that the constant term $R_\emptyset$ vanish guarantees that, for each word w, the w-coefficient

of $R(z)^n$ vanishes for all $n \geq N_w$ for some $N_w < \infty$, and hence that the infinite series expression for $R^*(z)$ is convergent in the topology of coefficientwise convergence. The $*$-operation also makes sense in the setting where the scalars are taken from a general semiring K; in case K is a field (as we assume), the $*$-operation $R(z) \mapsto R^*(z)$ can be identified as $R^*(z) = (I - R(z))^{-1}$. In case that $T(z) = \sum_{v \in \mathcal{F}_d} T_v z^v \in \mathcal{L}(\mathcal{U}, \mathcal{Y})\langle\langle z \rangle\rangle$ has coefficients T_v equal to operators between finite-dimensional linear spaces $\mathcal{U}$ and $\mathcal{Y}$, we say that $T(z)$ is rational if each of its matrix entries (with respect to some bases for $\mathcal{U}$ and $\mathcal{Y}$) is rational. In case $\mathcal{U} = \mathcal{Y}$ and $T_\emptyset = 0$, we can define

$$T^*(z) := \sum_{n=0}^{\infty} (T(z))^n = (I - T(z))^{-1} \tag{12.16}$$

just as in the scalar case. The following lemma assures us that $T^*(z)$ is again rational if $T(z)$ is rational. This result is actually a special case of Lemma I.6.3 in [15], but we include a proof for the sake of completeness.

LEMMA 12.3. *Suppose that $T(z) = [T_{ij}(z)]_{i,j=1}^N \in \mathcal{L}(\mathbb{C}^N)\langle\langle z \rangle\rangle$ is a formal power series in the noncommuting variables $z = (z_1, \ldots, z_d)$ with matrix entries $T_{ij}(z) \in \mathbb{C}\langle\langle z \rangle\rangle$ all rational such that $T_\emptyset = [T_{\emptyset,ij}]_{i,j=1}^N = 0$. Then all matrix entries of the formal power series $T^*(z)$ given by* (12.16) *are also rational.*

Proof. If $N = 1$, the result is clear. By induction we assume that the result is true for all $N < N_0$ and seek to prove the result for $N = N_0$. Given $T(z) \in \mathcal{L}(\mathbb{C}^{N_0})\langle\langle z \rangle\rangle$ with $T_\emptyset = 0$, consider a block decomposition of $T(z)$,

$$T(z) = \begin{bmatrix} a(z) & b(z) \\ c(z) & d(z) \end{bmatrix},$$

and a corresponding block decomposition of $T^*(z) = (I_{N_0} - T(z))^{-1}$,

$$(I_{N_0} - T(z))^{-1} = \begin{bmatrix} \alpha(z) & \beta(z) \\ \gamma(z) & \delta(z) \end{bmatrix},$$

where $a(z)$ and $\alpha(z)$ are both of size $K \times K$ for some K with $1 \leq K < N_0$. From the identity

$$(I_{N_0} - T(z))^{-1} = I_{N_0} + T(z)(I_{N_0} - T(z))^{-1}$$

we get the collection of identities

$$\begin{aligned} \alpha(z) &= I_K + a(z)\alpha(z) + b(z)\gamma(z), \\ \beta(z) &= a(z)\beta(z) + b(z)\delta(z), \\ \gamma(z) &= c(z)\alpha(z) + d(z)\gamma(z), \\ \delta(z) &= I_{N_0-K} + c(z)\beta(z) + d(z)\delta(z). \end{aligned} \tag{12.17}$$

We may then solve the second and third equations in (12.17) for $\beta(z)$ and $\gamma(z)$, respectively, to get

$$\beta(z) = (I_K - a(z))^{-1} b(z)\delta(z), \tag{12.18}$$

$$\gamma(z) = (I_{N_0-K} - d(z))^{-1} c(z)\alpha(z). \tag{12.19}$$

By the induction assumption we see immediately from (12.18) and (12.19) that $\beta(z)$ and $\gamma(z)$ are rational. Plugging back into the first and fourth identities in (12.17)

then gives

$$\alpha(z) = I_K + a(z)\alpha(z) + b(z)(I_{N_0-K} - d(z))^{-1}c(z)\alpha(z),$$
$$\delta(z) = I_{N_0-K} + c(z)(I_K - a(z))^{-1}b(z)\delta(z) + d(z)\delta(z).$$

We may then solve these equations for $\alpha(z)$ and $\delta(z)$ to get

$$\alpha(z) = \big(I_K - [a(z) + b(z)(I_{N_0-K} - d(z))^{-1}c(z)]\big)^{-1}, \tag{12.20}$$

$$\delta(z) = \big(I_{N_0-K} - [c(z)(I_K - a(z))^{-1}b(z) + d(z)]\big)^{-1}. \tag{12.21}$$

Again as a consequence of the induction assumption, (12.20) and (12.21) imply that $\alpha(z)$ and $\delta(z)$ are rational as well, and the lemma follows. □

The following characterization of rational formal power series can be seen as a corollary of the results of this paper.

COROLLARY 12.4. *Let a formal power series $T(z) = \sum_{v\in\mathcal{F}_d} T_v z^v$ in d noncommuting variables $z = (z_1, \dots, z_d)$ and an admissible graph G with edge set $E = \{1, \dots, d\}$ be given. Then the following are equivalent:*

1. *$T(z)$ is rational.*
2. *For each path-connected component p of G, the Hankel operator $\mathbb{H}^p$ given by (10.4) has finite rank.*
3. *$T(z)$ has a realization $T(z) = D + C(I - Z_\Sigma(z)A)^{-1}Z_\Sigma(z)B$ for an SNMLS $\Sigma = (G, \mathcal{H}, U)$ having structure graph G.*

Proof. We first show that (1) $\Longrightarrow$ (3). Note first that any scalar constant D (considered as a formal power series in noncommuting variables $z = (z_1, \dots, z_d)$) is realizable (with zero auxiliary state-spaces $\mathcal{H}_p$).

We next note that any monomial z_e is realizable for each edge $e = 1, \dots, d$. Indeed, set $\mathcal{H}_{[\mathbf{s}(e)]} = \mathbb{C}$ and $\mathcal{H}_p = \{0\}$ for $p \neq [\mathbf{s}(e)]$ and set

$$A = [A_{r,s}]_{r\in R, s\in S} \quad \text{with } A_{r,s} = 0,$$
$$B = \mathrm{col}_{r\in R}[B_r] \quad \text{with } B_r = \begin{cases} 1 & \text{if } r = \mathbf{r}(e), \\ 0 & \text{otherwise,} \end{cases}$$
$$C = \mathrm{row}_{s\in S}[C_s] \quad \text{with } C_s = \begin{cases} 1 & \text{if } s = \mathbf{s}(e), \\ 0 & \text{otherwise,} \end{cases}$$
$$D = 0.$$

Then the associated transfer function is given by

$$\begin{aligned} T_\Sigma(z) &= D + C(I - Z_\Sigma(z)A)^{-1}Z_\Sigma(z)B \\ &= 0 + CZ_\Sigma(z)B \\ &= \sum_{s\in S}\sum_{r\in R} C_s[Z_\Sigma(z)]_{s,r}B_r \\ &= \sum_{s\in S}\sum_{r\in R}\sum_{e'\in E} C_s I_{\Sigma,e';s,r}B_r z_{e'} \\ &= \sum_{e'\in E} C_{\mathbf{s}(e')}B_{\mathbf{r}(e')}z_{e'} \\ &= z_e. \end{aligned}$$

We conclude that each monomial z_e has a realization as asserted.

By Theorems 4.1, 4.2, and 4.3, products, sums, and inverses of invertible formal power series which are realizable (as the transfer function of an SNMLS Σ with structure graph G) are again realizable. By the inductive definition of rational formal power series given above, we may now conclude that any scalar rational formal power series $T(z)$ has the form of a transfer function $T(z) = T_\Sigma(z)$ for an SNMLS Σ with given admissible graph G as structure graph.

If each scalar entry $[T(z)]_{i,j}$ of a matrix of formal power series is realizable, it is easy to construct a realization (not necessarily minimal) for the formal power series $T(z)$ with matrix coefficients. This concludes the proof of (1) $\Longrightarrow$ (3).

We next verify (3) $\Longrightarrow$ (1). Assume that the formal power series $T(z)$ has a realization of the form $T(z) = D + C(I - Z_\Sigma(z)A)^{-1}Z_\Sigma(z)B$ for a finite-dimensional SNMLS $\Sigma = (G, \mathcal{H}, \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right])$. By Lemma 12.3 it follows that $(I - Z_\Sigma(z)A)^{-1}$ is rational. As products and sums of rational matrix functions are rational, it then follows that $T(z)$ is rational, as wanted.

The equivalence of (2) and (3) is just a restatement of Theorem 11.1. ∎

Remark 12.5. We note that the equivalence (1) $\Longleftrightarrow$ (2) between rationality and finiteness of the rank of an associated Hankel operator is known as Kronecker's theorem in the classical case.

Remark 12.6. Combining (1) $\Longleftrightarrow$ (3) in Corollary 12.4 with Corollary 12.2, we see that a formal power series is recognizable if and only if it is rational; this result goes back to Schützenberger (see Theorem I.6.1 in [15]).

Remark 12.7. In [22] Fliess gives an alternative system interpretation of a recognizable formal power series in terms of a homogeneous bilinear system with evolution along the nonnegative integers $\mathbb{Z}_+$ but with state-update equation of the form

$$x(n+1) = \left[\sum_{j=0}^{d} u_j(n)A_j\right] x(n),$$

with $A_0, \ldots, A_d$ linear operators on the state-space $\mathcal{H}$ and with $u_0(n), \ldots, u_d(n)$ equal to $d+1$ scalar-valued controls. The input-output operator for the system is obtained as

$$(x_0, (u_0(n), \ldots, u_d(n))_{n\in\mathbb{Z}_+}) \mapsto T_\Sigma(u)x_0,$$

where $T_\Sigma(z)$ is the recognizable formal power series $T_\Sigma(z) = C(I - z_0A_0 - z_1A_1 - \cdots - z_dA_d)^{-1}$ and where $T_\Sigma(u)$ is defined via the substitution

$$z_{i_N} z_{i_{N-1}} \ldots z_{i_0} \mapsto u_{i_N}(N)u_{i_{N-1}}(N-1)\cdots u_{i_0}(0).$$

Multidimensional versions of such bilinear systems, including connections with formal power series in this more general setting, are given in [23]. Sontag [40] used a variation of Fliess's Hankel-matrix construction to solve the following related moment problem connected with an alternative formulation of a bilinear system realization problem: *given operators $T_w \in \mathcal{L}(\mathcal{U}, \mathcal{Y})$ for $w \in \mathcal{F}_E$ ($E = \{1, \ldots, d\}$), find operators $C_1, \ldots, C_d\colon \mathcal{H} \to \mathcal{Y}$, $A_1, \ldots, A_d\colon \mathcal{H} \to \mathcal{H}$, and $B_1, \ldots, B_d\colon \mathcal{U} \to \mathcal{H}$ so that $T_{i_N i_{N-1}\cdots i_2 i_1} = C_{i_N}A_{i_{N-1}}\cdots A_{i_1}B_{i_1}$.*

Our discussion here gives a linear (rather than bilinear) system interpretation for a formal power series, but with evolution along a free semigroup rather than along $\mathbb{Z}_+$ and with a somewhat contrived input-injection for the initial condition on the

state required to recover the precise form of a recognizable series. The awkwardness of these various system interpretations for a recognizable formal power series gives some explanation as to why system operations work out well for transfer functions of SNMLSs (see section 4) but not so well for recognizable series—a point discussed in [30].

Acknowledgments. We thank the referees for useful suggestions which led to improvements in the paper; in particular, the material of section 4 appears as a result of a suggestion of one of the referees.

REFERENCES

[1] D. Alpay and D.S. Kalyuzhnyĭ-Verbovetzkiĭ, *Matrix-J-unitary noncommutative rational formal power series*, in The State Space Method, D. Alpay and I. Gohberg, eds., Oper. Theory Adv. Appl. 161, Birkhäuser-Verlag, Basel, Switzerland, to appear.

[2] D. Alpay and D. Volok, *Point evaluation and Hardy space on an homogeneous tree*, Integral Equations Operator Theory, to appear.

[3] C.-G. Ambrozie and D. Timotin, *A von Neumann type inequality for certain domains in* $\mathbb{C}^n$, Proc. Amer. Math. Soc., 131 (2003), pp. 859–869.

[4] J.A. Ball and V. Bolotnikov, *Realization and interpolation for Schur–Agler-class functions on domains with matrix polynomial defining function in* $\mathbb{C}^n$, J. Funct. Anal., 213 (2004), pp. 45–87.

[5] J.A. Ball, G. Groenewald, and T. Malakorn, *Conservative structured noncommutative multidimensional linear systems*, in The State Space Method, D. Alpay and I. Gohberg, eds., Oper. Theory Adv. Appl. 161, Birkhäuser-Verlag, Basel, Switzerland, to appear.

[6] J.A. Ball, G. Groenewald, and T. Malakorn, *Bounded Real Lemma for Structured Noncommutative Multidimensional Linear Systems and Robust Control*, manuscript.

[7] J.A. Ball and V. Vinnikov, *Functional models for representations of the Cuntz algebra*, in Operator Theory, Systems Theory and Scattering Theory: Multidimensional Generalizations, Oper. Theory Adv. Appl. 157, Birkhäuser-Verlag, Basel, Switzerland, 2005, pp. 1–60.

[8] J.A. Ball and V. Vinnikov, *Lax-Phillips scattering and conservative linear systems: A Cuntz-algebra multidimensional setting*, Mem. Amer. Math. Soc., to appear.

[9] H. Bart, I. Gohberg, and M.A. Kaashoek, *Minimal Factorization of Matrix and Operator Functions*, OT1, Birkhäuser-Verlag, Basel, Boston, 1979.

[10] C.L. Beck, *On formal power series representations for uncertain systems*, IEEE Trans. Automat. Control, 46 (2001), pp. 314–319.

[11] C.L. Beck and R. D'Andrea, *Noncommuting multidimensional realization theory: Minimality, reachability, and observability*, IEEE Trans. Automat. Control, 49 (2004), pp. 1815–1820.

[12] C.L. Beck and J.C. Doyle, *A necessary and sufficient minimality condition for uncertain systems*, IEEE Trans. Automat. Control, 44 (1999), pp. 1802–1813.

[13] C.L. Beck, J.C. Doyle, and K. Glover, *Model reduction of multidimensional and uncertain systems*, IEEE Trans. Automat. Control, 41 (1996), pp. 1466–1477.

[14] A. Benveniste, R. Nikoukhah, and A.S. Willsky, *Multiscale system theory*, IEEE Trans. Circuits Systems I Fund. Theory Appl., 41 (1994), pp. 2–15.

[15] J. Berstel and C. Reutenauer, *Rational Series and Their Languages*, EATCS Monogr. Theoret. Comput. Sci., Springer-Verlag, Berlin, New York, 1984.

[16] F.M. Callier and C.A. Desoer, *Linear System Theory*, Springer-Verlag, Berlin, 1991.

[17] J.W. Carlyle and A. Paz, *Realizations by stochastic finite automaton*, J. Comput. System Sci., 5 (1971), pp. 26–40.

[18] N. Cohen, *Decoupling of transfer functions*, Integral Equations Operator Theory, 50 (2004), pp. 317–322.

[19] S. Eilenberg, *Automata, Languages, and Machines, Volume A*, Academic Press, New York, 1974.

[20] M. Fliess, *Matrices de Hankel*, J. Math. Pures Appl., 53 (1974), pp. 197–222; Erratum, 54 (1975), p. 481.

[21] M. Fliess, *Matrices de Hankel*, *Sur divers produits de séries formelles*, Bull. Soc. Math. France, 102 (1974), pp. 181–191.

[22] M. Fliess, *Matrices de Hankel*, *Un codage non commutatif pour certains systèmes échantillonnés non linéaires*, Inform. and Control, 38 (1978), pp. 264–287.

[23] M. Fliess, *Matrices de Hankel, Une théorie fonctionnelle de la réalisation en filtrage multidimensionnel, échantillonné, récurrent*, Inform. and Control, 43 (1979), pp. 338–355.
[24] E. Fornasini and G. Marchesini, *Doubly-indexed dynamical systems: State space models and structural properties*, Math. System Theory, 12 (1978), pp. 59–72.
[25] K. Gałkowski, *Minimal state-space realization for a class of linear, discrete, nD, SISO systems*, Internat. J. Control, 74 (2001), pp. 1279–1294.
[26] D.D. Givone and R.P. Roesser, *Multidimensional linear iterative circuits—General properties*, IEEE Trans. Comput., C-21 (1972), pp. 1067–1073.
[27] D.D. Givone and R.P. Roesser, *Minimization of multidimensional linear iterative circuits*, IEEE Trans. Comput., C-22 (1973), pp. 673–678.
[28] J.W. Helton, *Manipulating matrix inequalities automatically*, in Mathematical Systems Theory in Biology, Communication, Computation, and Finance, D. Gilliam and J. Rosenthal, eds., IMA Vol. Math. Appl. 134, Springer-Verlag, New York, 2003.
[29] J.W. Helton and J.A. Ball, *The cascade decompositions of a given system vs. the linear fractional decompositions of its transfer function*, Integral Equations Operator Theory, 5 (1982), pp. 341–385.
[30] J.W. Helton, S.A. McCullough, and V. Vinnikov, *Noncommutative Convexity Arises from Linear Matrix Inequalities*, manuscript.
[31] T. Kaczorek, *Two Dimensional Linear Systems*, Lecture Notes in Control and Inform. Sci. 68, Springer-Verlag, Berlin, 1985.
[32] R.F. Kalman, P.L. Falb, and M.A. Arbib, *Topics in Mathematical System Theory*, McGraw–Hill, New York, 1969.
[33] W.-M. Lu, K. Zhou, and J.C. Doyle, *Stabilization of uncertain linear systems: An LFT approach*, IEEE Trans. Automat. Control, 41 (1996), pp. 50–65.
[34] T. Malakorn, *Multidimensional Linear Systems and Robust Control*, Dissertation, Department of Electrical and Computer Engineering, Virginia Tech, Blackburg, VA, 2003; available online at http://scholar.lib.vt.edu/theses/available/etd-04142003-144447/.
[35] J.W. Polderman and J.C. Willems, *Introduction to Mathematical Systems Theory*, Springer-Verlag, Berlin, 1998.
[36] R.P. Roesser, *A concrete state-space model for linear image processing*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 1–10.
[37] M.P. Schützenberger, *On the definition of a family of automata*, Inform. and Control, 4 (1961), pp. 245–270.
[38] M.P. Schützenberger, *On a theorem of R. Jungen*, Proc. Amer. Math. Soc., 13 (1962), pp. 885–889.
[39] M.P. Schützenberger, *Certain elementary families of automata*, in Proceedings of the Symposium on Mathematical Theory of Automata, Polytechnic Institute Brooklyn, Brooklyn, NY, 1962, pp. 139–153.
[40] E. Sontag, *Realization theory of discrete-time nonlinear systems: Part* I—*The bounded case*, IEEE Trans. Circuits Systems, CAS-26 (1979), pp. 342–356.
[41] J.L. Taylor, *Functions of several noncommuting variables*, Bull. Amer. Math. Soc., 79 (1973), pp. 1–34.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 4, pp. 1529–1541

OPTIMAL CONTROL UNDER A DYNAMIC FUEL CONSTRAINT*

PETER BANK†

Abstract. We present a new approach to solve optimal control problems of the monotone follower type. The key feature of our approach is that it allows us to include an arbitrary dynamic fuel constraint. Instead of dynamic programming, we use the convexity of our cost functional to derive a first order characterization of optimal policies based on the Snell envelope of the objective functional's gradient at the optimum. The optimal control policy is constructed explicitly in terms of the solution to a representation theorem for stochastic processes obtained in Bank and El Karoui (2004), *Ann. Probab.*, 32, pp. 1030–1067. As an illustration, we show how our methodology allows us to extend the scope of the explicit solutions obtained for the classical monotone follower problem and for an irreversible investment problem arising in economics.

Key words. monotone follower, Snell envelope, dynamic fuel constraint

AMS subject classifications. 49J55, 93E20, 60H30, 91B28

DOI. 10.1137/040616966

Introduction. Many optimization problems involve so-called finite fuel constraints on the allowable control policies, i.e., upper bounds on the resources a control policy can use. The usual methodology to address these optimization problems is to specify a Markovian framework and to compute the problem's value function either by PDE methods based on the problem's Hamilton–Jacobi–Bellman equation or by probabilistic methods and the variational method of switching paths. In some special cases this leads to a more or less explicit solution to the optimization problem.

In any case, the constraint has so far only been specified by a *constant* upper bound for the overall amount of "fuel" a control policy is allowed to use. *Dynamic* upper bounds, by contrast, are difficult to take into account as their introduction increases the dimensionality of the problem, making it typically impossible to solve the Hamilton–Jacobi–Bellman equation explicitly. On the other hand, it is well known that in some problems the optimal policy for a (constant) finite fuel constraint can be derived from the optimal policy obtained when disregarding the fuel constraint completely: one just has to follow the unconstraint policy up to the moment when all fuel has been spent; see, e.g., Chow, Menaldi, and Robin (1985), Karatzas (1985), and Fleming and Soner (1993). It is thus natural to conjecture that a suitable variant of this principle should hold true for situations where a dynamic finite fuel constraint is specified by an increasing adapted process. The corroboration of this conjecture and the description of a general framework where it holds true constitute the main results of the present paper.

We consider a convex minimization problem in which a policy θ incurs the costs

$$C(\theta) = \mathbb{E}\int_0^\infty c(t,\theta_t)\,\mu(dt) + \mathbb{E}\int_0^\infty k_t\,d\theta_t \quad (\theta \in \mathscr{A})\,,$$

*Received by the editors October 13, 2004; accepted for publication (in revised form) April 24, 2005; published electronically November 14, 2005. This work was supported by Deutsche Forschungsgemeinschaft through DFG-Research Center "Mathematics for Key Technologies" (FZT 86) and grant BA 2276/1-1.

http://www.siam.org/journals/sicon/44-4/61696.html

†Department of Mathematics, Columbia University in the City of New York, 2990 Broadway, Mail Code 4433, New York, NY 10027 (pbank@math.columbia.edu, www.math.columbia.edu/∼pbank).

where $c(t,.)$ describes the (convex) running costs and k_t the control costs at time $t \geq 0$. Our approach is based on a characterization of optimal policies in terms of first order conditions. More specifically, Theorem 2.2 shows that an optimal control policy will exercise control whenever its impact is maximal as measured by the Snell envelope of the cost functional's subgradient at the optimum; it also shows that, actually, all available fuel should be spent whenever this Snell envelope tends to decrease. The occurrence of Snell envelopes in this characterization highlights the intimate relationship between singular control and optimal stopping problems which has already been observed in Karatzas and Shreve (1984, 1985) and El Karoui and Karatzas (1988, 1991).

The construction of an optimal policy is achieved in Theorem 3.1, which relates the dynamic finite fuel problem with a stochastic representation theorem obtained in Bank and El Karoui (2004). This representation theorem has found a number of other applications ranging from utility maximization to optimal stopping; we refer the reader to Bank and Föllmer (2003) for an overview. Here it provides us with a lower bound which the optimal control policy has to respect if enough fuel is available to do so. This lower bound turns out to be independent of the fuel constraint, thus providing a universal signal process which allows one to construct optimal policies for a whole class of finite fuel problem at the same time.

As an application we provide an explicit solution to the monotone follower problem for Lévy processes with quadratic cost functional in the spirit of Beneš, Shepp, and Witsenhausen (1980/81). We also illustrate how explicit solutions obtained for singular control problems without any fuel constraint, as obtained, e.g., in Kobila (1993), can actually be used to describe optimal policies for problems with a dynamic fuel constraint.

Notation and conventions. All (in)equalities between random variables are meant to hold true in the $\mathbb{P}$–a.s. sense. We shall let $\mathscr{T}$ denote the set of all stopping times, and we use $\mathscr{T}(I)$ to denote the class of stopping times almost surely taking values in a given random set I, such as, e.g., $I = [S, +\infty]$ with $S \in \mathscr{T}$. A supremum over an empty set is defined to be $\sup \varnothing \stackrel{\Delta}{=} -\infty$. Intervals $[a,b]$ with $b < a$ are interpreted as the empty set. We also put $x^+ \stackrel{\Delta}{=} x \vee 0 = \max\{x, 0\}$ and $x^- \stackrel{\Delta}{=} (-x)^+$.

1. The general control problem. A well-known problem in stochastic optimization is the problem of controlling the motion of a particle so as to keep it as close to the origin as possible over some period of time. In the formulation as a monotone follower problem suggested and analyzed by Karatzas and Shreve (1984), one considers a model where the dynamics of the uncontrolled particle is given by a standard Brownian motion W on the real line and where the control θ is an increasing adapted process θ which specifies the downward displacement of the particle caused by the control. Hence, in this case, the controlled particle would follow the dynamics $W_t - \theta_t$ $(t \geq 0)$. The cost incurred by a control policy θ can, for instance, be described as

$$C(\theta) = \mathbb{E} \int_0^\infty \delta e^{-\delta t} \frac{1}{2} \left(W_t - \theta_t\right)^2 \, dt,$$

and one could start studying the optimization problem to minimize C subject to, e.g., a finite fuel constraint on the control θ.

More generally, let $(\Omega, \mathscr{F}_\infty, (\mathscr{F}_t)_{t\geq 0}, \mathbb{P})$ be a filtered probability space satisfying the usual conditions of right-continuity and completeness. Controls θ are given by increasing, left-continuous adapted processes starting at $\theta_0 = \underline{\vartheta} \in \mathbb{R}$. We shall impose

a dynamic finite fuel constraint, specified by an increasing adapted process $\overline{\vartheta}$ with left-continuous paths and values in $[\underline{\vartheta}, +\infty]$. The class of admissible controls is therefore

$$\mathscr{A} \triangleq \{\theta \text{ incr., left-cont., adapted with } \underline{\vartheta} = \theta_0 \leq \theta_t \leq \overline{\vartheta}_t \text{ for all } t \geq 0 \ \mathbb{P}\text{–a.s.}\}\,.$$

REMARK 1.1. *Note that the lower bound $\underline{\vartheta}$ is assumed to be a real constant, not a process. Assuming a dynamic lower bound would mean that a minimum amount of control must have been exercised up to each point in time, a natural, yet much more demanding, extension which is beyond the scope of the present paper.*

The costs incurred by a control policy will be composed of running costs and control costs. The running costs are described by a measurable random field

$$c : \Omega \times [0, \infty) \times \mathbb{R} \to \mathbb{R}$$

and a positive optional random measure $\mu = \mu(\omega, dt)$ on the time axis satisfying the following convexity and regularity assumption.

ASSUMPTION 1.

(i) *The measure μ is atomless and has full support* $\operatorname{supp}\mu = [0, +\infty)$ *almost surely.*

(ii) *For any $(\omega, t) \in \Omega \times [0, +\infty)$, the mapping $\vartheta \mapsto c(\omega, t, \vartheta)$ is strictly convex with continuous derivative $c'(\omega, t, \vartheta) = \frac{\partial}{\partial \vartheta} c(\omega, t, \vartheta)$ increasing from $c'(\omega, t, -\infty) = -\infty$ to $c'(\omega, t, +\infty) = +\infty$.*

(iii) *For $\vartheta \in \mathbb{R}$ fixed, $(\omega, t) \mapsto c(\omega, t, \vartheta)$ is progressively measurable and $\mathbb{P} \otimes \mu$-integrable.*

(iv) *The process $(\omega, t) \mapsto \inf_{\vartheta \in [\underline{\vartheta}, \overline{\vartheta}_t(\omega)]} c(\omega, t, \vartheta)$ is $\mathbb{P} \otimes \mu$-integrable.*

The control costs are described by a stochastic process k with the following properties.

ASSUMPTION 2. *The process k is optional, of class (D), and continuous in expectation with $k_\infty = 0$. Moreover, the family of random variables $(\int_0^\infty k_t^-\, d\theta_t,\ \theta \in \mathscr{A})$ is bounded in $L^1(\mathbb{P})$.*

REMARK 1.2. *Recall that an optional process $k = (k_t)_{t \geq 0}$ is of class (D) if the family of random variables $(k_T,\ T \in \mathscr{T})$ is uniformly integrable. Such a process is continuous in expectation provided $\lim_n \mathbb{E}[k_{T_n}] = \mathbb{E}[k_T]$ for any monotone sequence of stopping times (T_n) with limit $T = \lim_n T_n$.*

These assumptions allow us to consider the cost functional

$$C(\theta) = \mathbb{E} \int_0^\infty c(t, \theta_t)\, \mu(dt) + \mathbb{E} \int_0^\infty k_t\, d\theta_t \quad (\theta \in \mathscr{A})\,.$$

The general optimization problem we shall be concerned with in this paper can now be stated as follows:

$$\text{Minimize } C(\theta) \text{ over } \theta \in \mathscr{A}\,. \tag{1}$$

REMARK 1.3.

(i) *Full support of μ ensures that strict convexity transfers from the random field c to our cost functional C. Our assumptions on the derivative of c will be used when applying a representation theorem obtained in Bank and El Karoui* (2004)*; see section* 3. *Integrability of $c(., \vartheta)$ for $\vartheta \in \mathbb{R}$ means that the decision not to intervene at all will not cause infinite costs. Integrability of $\inf_{\vartheta \in [\underline{\vartheta}, \overline{\vartheta}_\cdot]} c(., \vartheta)$ is assumed to ensure that our minimization problem* (1) *has a finite value. For the same reason we assume L^1-boundedness of*

($\int_0^\infty k_t^-\, d\theta_t,\ \theta \in \mathscr{A}$), *which amounts to requiring that the negative (!) "costs" of exercising control must not be "too large."*

(ii) *Observe that our introductory example would be accommodated in this setting by choosing*

$$c(\omega, t, \vartheta) \triangleq \frac{1}{2}\left(W_t(\omega) - \vartheta\right)^2, \quad \mu(dt) = \delta e^{-\delta t}\, dt \quad \textit{and} \quad k \equiv 0\,.$$

Observe furthermore that this setting can also accommodate the monotone follower problems studied in Chow, Menaldi, and Robin (1985)*, Karatzas* (1985)*, Karatzas and Shreve* (1984)*, as well as the irreversible investment problems solved in Kobila* (1993) *(see Criterion* (3.2) *and Condition* (5.1)*), Scheinkman and Zariphopoulou* (2001) *(see section* 4.2*), and Baldursson and Karatzas* (1997)*. Settings* not *covered by our framework include Chiarolla* (1997) *and Jacka* (1999, 2002) *since their cost functional is specified in terms of the controlled system instead of the cumulatively exercised control. We also do* not *cover the "cheap monotone follower" of Chiarolla and Haussmann* (1994) *as they allow for two-dimensional controls. Also uncovered remains the finite fuel problem of Beneš, Shepp, and Witsenhausen* (1980/81) *and Karatzas and Shreve* (1985)*, where two-sided controls are considered.*

2. First order conditions for optimality. In this section, we are going to provide a first order characterization of optimal control policies for problem (1). While the main avenue of approach to achieve this characterization is classical, we need to be a little bit careful to ensure that our Assumptions 1 and 2 suffice to deduce all the integrability requirements we need along the way.

Our first step is to note that the convex functional C is supported by the subgradients

$$\nabla C(\theta)_S \triangleq \mathbb{E}\left[\int_S^\infty c'(t, \theta_t)\, \mu(dt) \,\middle|\, \mathscr{F}_S\right] + k_S \quad (S \in \mathscr{T}) \tag{2}$$

in the following sense.

LEMMA 2.1. *For any $\theta \in \mathscr{A}$, the optional process $\nabla C(\theta)$ of* (2) *is well defined and $\nabla C(\theta)^-$ is $\mathbb{P}\otimes d\theta$-integrable. If also $\nabla C(\theta)^+$ is $\mathbb{P}\otimes d\theta$-integrable, then $C(\theta) < \infty$ and $\nabla C(\theta)$ satisfies the subgradient property*

$$C(\theta') - C(\theta) \geq \mathbb{E}\int_0^\infty \nabla C(\theta)_s\, d(\theta'_s - \theta_s) \quad \textit{for any} \quad \theta' \in \mathscr{A} \textit{ with } C(\theta') < +\infty\,.$$

Proof.

(i) As $c'(t, \theta_t) \geq c'(t, \underline{\vartheta}) \in L^1(\mathbb{P}\otimes\mu)$ by convexity and $\mathbb{P}\otimes\mu$-integrability of $c(t, \vartheta)$ for $\vartheta \in \mathbb{R}$, the conditional expectation appearing in (2) is well defined as a random variable taking values in $(-\infty, +\infty]$. As for $\mathbb{P}\otimes d\theta$-integrability of $\nabla C(\theta)^-$, we note that

$$c'(t, \theta_t)^-(\theta_t - \underline{\vartheta}) \leq c(t, \underline{\vartheta}) - \inf_{\vartheta \in [\underline{\vartheta}, \overline{\vartheta}_t]} c(t, \vartheta) \in L^1(\mathbb{P}\otimes\mu)$$

by Assumption 1(iii) and (iv), and this yields

$$\mathbb{E}\int_0^\infty \int_s^\infty c'(t, \theta_t)^-\, \mu(dt)\, d\theta_s = \mathbb{E}\int_0^\infty c'(t, \theta_t)^-(\theta_t - \underline{\vartheta})\, \mu(dt) < +\infty$$

by Fubini's theorem. By Assumption 2,

$$\mathbb{E}\int_0^\infty k_s^-\, d\theta_s < +\infty\,,$$

and it follows that

$$\nabla C(\theta)_s^- \le \mathbb{E}\left[\int_s^\infty c'(t,\theta_t)^-\,\mu(dt)\,\middle|\,\mathscr{F}_s\right] + k_s^-$$

is $\mathbb{P}\otimes d\theta_s$-integrable.

(ii) Let us now assume $\mathbb{P}\otimes d\theta$-integrability of $\nabla C(\theta)^+$ and show that both $\int_\cdot^\infty c'(t,\theta_t)^+\,\mu(dt)$ and k^+ inherit this integrability property. In addition, we shall see that $C(\theta)<\infty$.
To wit, note that in (i) we actually obtained $\int_\cdot^\infty c'(t,\theta)^-\,\mu(dt)\in L^1(\mathbb{P}\otimes d\theta)$ and that $k^-\in L^1(\mathbb{P}\otimes d\theta)$ by Assumption 2. So we can write

$$\mathbb{E}\left[\int_s^\infty c'(t,\theta_t)^+\,\mu(dt)\,\middle|\,\mathscr{F}_s\right] + k_s^+ = \nabla C(\theta)_s + \mathbb{E}\left[\int_s^\infty c'(t,\theta_t)^-\,\mu(dt)\,\middle|\,\mathscr{F}_s\right] + k_s^-$$

to deduce that also both summands on the left side are $\mathbb{P}\otimes d\theta$-integrable if, as we assume, $\nabla C(\theta)$ is.
To obtain $C(\theta)<\infty$, note that by convexity of $c(t,.)$ we have

$$c(t,\theta_t) \le c(t,\underline{\vartheta}) + c'(t,\theta_t)(\theta_t-\underline{\vartheta})\,.$$

By Assumption 1 the $c(t,\underline{\vartheta})$-term on the right side of the above estimate is $\mathbb{P}\otimes\mu$-integrable. Moreover, the $\mathbb{P}\otimes d\theta$-integrability of $\int_\cdot^\infty c'(t,\theta_t)^\pm\,\mu(dt)$ established before entails that also $c'(t,\theta_t)(\theta_t-\vartheta)$ is $\mathbb{P}\otimes\mu$-integrable by Fubini's theorem. Hence $c(t,\theta_t)^+\in L^1(\mathbb{P}\otimes\mu)$, which in conjunction with $k\in L^1(\mathbb{P}\otimes d\theta)$ yields $C(\theta)<\mathbb{R}$.

(iii) We finally can prove our subgradient estimate for θ' with $C(\theta')<+\infty$, assuming as in (ii) that $\nabla C(\theta)^+$ (and thus $\nabla C(\theta)$) is $\mathbb{P}\otimes d\theta$-integrable.
We start from the convexity estimate

$$(3)\quad c(t,\theta_t') - c(t,\theta_t) \ge c'(t,\theta_t)(\theta_t'-\theta_t) = c'(t,\theta_t)(\theta_t'-\underline{\vartheta}) - c'(t,\theta_t)(\theta_t-\underline{\vartheta})\,.$$

Since $C(\theta'), C(\theta)<\infty$, the cost process k is both $\mathbb{P}\otimes d\theta'$- and $\mathbb{P}\otimes d\theta$-integrable and the left side of (3) is $\mathbb{P}\otimes\mu$-integrable. In (ii) we have shown that also the last term in (3), $c'(t,\theta_t)(\theta_t-\underline{\vartheta})$, is $\mathbb{P}\otimes\mu$-integrable. It thus follows that the positive part of the remaining $c'(t,\theta_t)(\theta_t'-\underline{\vartheta})$-term on the right side of (3) is $\mathbb{P}\otimes\mu$-integrable or, equivalently by Fubini's theorem, that $\int_\cdot^\infty c'(t,\theta_t)^+\,\mu(dt)$ is $\mathbb{P}\otimes d\theta'$-integrable. As a consequence, the expectation

$$\mathbb{E}\int_0^\infty\left\{\int_s^\infty c'(t,\theta_t)\,\mu(dt)+k_s\right\}d(\theta_s'-\theta_s) = \mathbb{E}\int_0^\infty \nabla C(\theta)_s\,d(\theta_s'-\theta_s)\in[-\infty,+\infty)$$

is well defined and indeed not larger than $C(\theta')-C(\theta)$. □

Let us denote by $\mathbb{S}(\theta)$ the lower Snell envelope of $\nabla C(\theta)$ as follows:

$$\mathbb{S}(\theta)_S = \operatorname*{ess\,inf}_{T\in\mathscr{T}([S,\infty])} \mathbb{E}\left[\nabla C(\theta)_T \,|\, \mathscr{F}_S\right] \quad (S\in\mathscr{T}).$$

We refer the reader to El Karoui (1981) for a comprehensive account on Snell envelopes. Let us just note here that $S(\theta)$ is a submartingale taking values in $(-\infty, 0]$. Indeed, we can choose $T=\infty$ to see that $\mathbb{S}(\theta)_S \le \mathbb{E}\left[\nabla C(\theta)_\infty \,|\, \mathscr{F}_S\right] = 0$ by definition of $\nabla C(\theta)$ and our assumption that $k_\infty = 0$. Moreover, using that $c'(t,\underline{\vartheta}) \ge c(t,\underline{\vartheta}-1) - c(t,\underline{\vartheta}) \in L^1(\mathbb{P}\otimes\mu)$ by Assumption 1 and that k is of class (D) by Assumption 2, we obtain

$$\inf_{T\in\mathscr{T}} \mathbb{E}\nabla C(\theta)_T \ge \mathbb{E}\int_0^\infty c'(t,\underline{\vartheta})\wedge 0\,\mu(dt) + \inf_{T\in\mathscr{T}} \mathbb{E}k_T > -\infty\,.$$

This entails that almost surely $\mathbb{S}(\theta)$ does not take the value $-\infty$.

We shall use $M(\theta)$ and $A(\theta)$ to denote the martingale and predictable increasing part in the Doob–Meyer decomposition

$$\mathbb{S}(\theta) = M(\theta) + A(\theta)$$

of the submartingale $\mathbb{S}(\theta)$.

After these preliminaries, we can now give the following characterization of optimal policies in terms of first order conditions.

THEOREM 2.2. *Under Assumptions* 1 *and* 2, *a control policy* $\theta^*\in\mathscr{A}$ *is optimal for problem* (1) *if*

(i) θ^* *is flat off* $\{\nabla C(\theta^*) = \mathbb{S}(\theta^*)\}$ *and*
(ii) $A(\theta^*)$ *is flat off* $\{\theta^* = \overline{\vartheta}\}$.

REMARK 2.3.

(i) *An increasing process* θ *is said to be flat off a set* $A\in\mathscr{F}_\infty\otimes\mathscr{B}([0,\infty))$ *if the induced measure* $d\theta$ *almost surely does not charge the set* A: $\mathbb{E}\int_0^\infty 1_A\,d\theta = 0$.
(ii) *Condition* (i) *requires that control should be exercised only when its marginal impact on future costs is maximal. Condition* (ii) *reveals that all fuel should be spent at moments when the maximal expected marginal impact tends to decrease.*
(iii) *In fact, conditions* (i) *and* (ii) *are also* necessary *for optimality of* $\theta^*\in\mathscr{A}$. *This result could be derived using arguments from the calculus of variations. It is much easier, however, to deduce this observation directly from our construction of the unique optimal policy in section* 3.

The proof of this theorem uses the following two lemmata and will be given at the end of this section.

LEMMA 2.4. *A plan* $\theta^*\in\mathscr{A}$ *is optimal for problem* (1) *if for any* $\theta\in\mathscr{A}$ *the process* $\nabla C(\theta^*)^-$ *is* $\mathbb{P}\otimes d\theta$*-integrable and we have*

$$\mathbb{E}\int_0^\infty \nabla C(\theta^*)_s\,d\theta^*_s \le \mathbb{E}\int_0^\infty \nabla C(\theta^*)_s\,d\theta_s\,. \tag{4}$$

Proof. For $\theta\equiv\underline{\vartheta}$ the right side in (4) vanishes and therefore $\nabla C(\theta^*)^+$ must be $\mathbb{P}\otimes d\theta^*$-integrable. This allows us to use the subgradient estimate of Lemma 2.1 to obtain

$$C(\theta) - C(\theta^*) \ge \mathbb{E}\int_0^\infty \nabla C(\theta^*)_s\,d(\theta-\theta^*)_s$$

for any $\theta \in \mathscr{A}$ with $C(\theta) < \infty$. By (4), the right side in this estimate is nonnegative and therefore θ^* attains the minimum of $C(.)$ over $\mathscr{A}$ as claimed. □

The preceding lemma suggests that an optimal policy for our convex optimization problem (1) should also be a solution to some linear minimization problem. Solutions to this kind of linear minimization problem are characterized by the following result.

LEMMA 2.5. *Let $\phi \leq 0$ be an optional process of class (D) which is continuous in expectation with $\phi_\infty = 0$. Let ψ denote its lower Snell envelope*

$$\psi_S = \operatorname*{ess\,inf}_{T \in \mathscr{T}([S,\infty])} \mathbb{E}\,[\,\phi_T \,|\, \mathscr{F}_S\,] \quad (S \in \mathscr{T}),$$

and consider the corresponding Doob–Meyer decomposition $\psi = M + A$ into a uniformly integrable martingale M and an increasing, predictable process A with $A_0 = 0$.

Then θ^* solves the linear optimization problem

$$\textit{Minimize } \mathbb{E}\int_0^\infty \phi_s\, d\theta_s \textit{ subject to } \theta \in \mathscr{A}$$

if it satisfies

(i) *θ^* is flat off $\{\phi = \psi\}$ and*

(ii) *A is flat off $\{\theta^* = \overline{\vartheta}\}$.*

If the value of the above minimization problem is finite, these two conditions are also necessary for optimality of $\theta^* \in \mathscr{A}$.

Proof. As ϕ is of class (D) and continuous in expectation, so is its Snell envelope ψ. In particular, ψ is a right-continuous process with left limits and its predictable compensator A has continuous paths almost surely increasing to $A_\infty \in L^1(\mathbb{P})$. Moreover, $0 = \psi_\infty = M_\infty + A_\infty$ implies that $\psi_S = \mathbb{E}\,[\,A_T - A_\infty \,|\, \mathscr{F}_T\,]$ for $T \in \mathscr{T}$. For any control policy $\theta \in \mathscr{A}$, this allows us to derive the following estimate:

$$\begin{aligned}(5)\qquad \mathbb{E}\int_0^\infty \phi_t\, d\theta_t &\geq \mathbb{E}\int_0^\infty \psi_t\, d\theta_t = \mathbb{E}\int_0^\infty (A_t - A_\infty)\, d\theta_t = -\mathbb{E}\int_0^\infty (\theta_t - \underline{\vartheta})\, dA_t \\ &\geq -\mathbb{E}\int_0^\infty (\overline{\vartheta}_t - \underline{\vartheta})\, dA_t\,.\end{aligned}$$

Indeed, the first estimate is due to $\phi \geq \psi$, the second equality follows by partial integration, and the last estimate holds true because $dA \geq 0$ and $\theta \leq \overline{\vartheta}$ by admissibility of θ.

It is now easy to see that any θ^* satisfying (i) and (ii) will yield equality everywhere in (5). On the other hand, these two conditions are also necessary for a plan θ^* to minimize $\mathbb{E}\int_0^\infty \phi_s\, d\theta_s$ over $\theta \in \mathscr{A}$, provided the value of our linear minimization problem is finite. This follows readily from (5) in conjunction with the identity

$$(6)\qquad \inf_{\theta \in \mathscr{A}} \mathbb{E}\int_0^\infty \phi_s\, d\theta_s = -\mathbb{E}\int_0^\infty (\overline{\vartheta}_t - \underline{\vartheta})\, dA_t\,.$$

To prove this identity, we introduce for $n = 1, 2, \ldots$ the sequence of stopping times

$$\begin{aligned}T_0^n &\triangleq \inf\{t \geq 0 \mid \phi_t = \psi_t\},\\ T_j^n &\triangleq \inf\{t \geq T_{j-1}^n \mid \phi_t = \psi_t,\ \overline{\vartheta}_t > \overline{\vartheta}_{T_{j-1}^n} + 1/n\} \quad (j = 1, 2, \ldots)\end{aligned}$$

and consider the admissible control policy

$$\theta_t^n \triangleq \sum_{j=0}^\infty \overline{\vartheta}_{T_j^n + 1}(T_j^n, T_{j+1}^n](t) \quad (t \geq 0)\,.$$

For $\theta = \theta^n$ we have equality in the first part of estimate (5), and so we obtain

$$\mathbb{E}\int_0^\infty \phi_s \, d\theta_s^n = -\mathbb{E}\int_0^\infty (\theta_t^n - \underline{\vartheta}) \, dA_t \, .$$

It follows from general results on optimal stopping that dA is supported by the set $\{t \geq 0 \mid \phi_t = \psi_t\}$ almost surely. By definition of the stopping times T_j^n $(j = 0, 1, \ldots)$ this entails that almost surely $\overline{\vartheta}_t \leq \theta_t^n + 1/n$ for dA-a.e. t. We can thus conclude that

$$\mathbb{E}\int_0^\infty \phi_s \, d\theta_s^n = -\mathbb{E}\int_0^\infty (\theta_t^n - \underline{\vartheta}) \, dA_t \leq -\mathbb{E}\int_0^\infty (\overline{\vartheta}_t - \underline{\vartheta}) \, dA_t + \frac{1}{n}\mathbb{E}A_\infty \, .$$

For $n \uparrow \infty$ this establishes the desired identity (6), accomplishing our proof. □

It is now easy to give the following proof.

Proof of Theorem 2.2. Let $\theta^* \in \mathscr{A}$ be a policy such that θ^* is flat off $\{\nabla C(\theta^*) = \psi(\theta^*)\}$ and $A(\theta^*)$ is flat off $\{\theta^* = \overline{\vartheta}\}$. Since $k_\infty = 0$ entails $\nabla C(\theta^*)_\infty = 0$, the process $\psi \triangleq \psi(\theta^*) \leq 0$ is actually the Snell envelope of both $\nabla C(\theta^*)$ and $\phi \triangleq \nabla C(\theta^*) \wedge 0$, an optional process of class (D) which is continuous in expectation due to our assumptions on c, k, and μ. As a consequence, the above flat off conditions entail that θ^* actually satisfies both optimality conditions of Lemma 2.5 for this choice of $\phi \leq 0$. It follows that

$$\mathbb{E}\int_0^\infty \nabla C(\theta^*) \, d\theta^* = \mathbb{E}\int_0^\infty \phi \, d\theta^* \leq \mathbb{E}\int_0^\infty \phi \, d\theta = \mathbb{E}\int_0^\infty \nabla C(\theta^*) \wedge 0 \, d\theta$$

for all $\theta \in \mathscr{A}$. Lemma 2.1 yields in particular that the left side of this estimate is finite and it thus follows that $\nabla C(\theta^*)^-$ is $\mathbb{P} \otimes d\theta$-integrable for all $\theta \in \mathscr{A}$. In addition, the above estimate entails that θ^* satisfies the first order condition (4), and so we can use Lemma 2.4 to conclude that θ^* is an optimal policy for problem (1). □

REMARK 2.6. *For problems without fuel constraint ($\overline{\vartheta} \equiv +\infty$), the integrability assertion in Lemma* 2.4 *implies in particular that for an optimal policy $\theta^* \in \mathscr{A}$ the gradient $\nabla C(\theta^*)$ has to be nonnegative, i.e.,*

$$k_S \geq -\mathbb{E}\left[\int_S^{+\infty} c'(t, \theta_t) \, \mu(dt) \,\middle|\, \mathscr{F}_S\right] .$$

Moreover, condition (i) *in Theorem* 2.2 *implies that equality must hold true in the above relation whenever S is a time of intervention. This is in accordance with the first order characterizations obtained for such problems in Bertola* (1998) *and Bank and Riedel* (2001).

3. Construction of an optimal policy. In this section, we shall show how to use the first order characterization of the optimal policy provided by Theorem 2.2 in order to construct the solution to the finite fuel problem (1). The construction will be given in terms of a progressively measurable random process κ specifying a lower bound which the optimal control should respect granted enough fuel is left to do so. This lower bound is characterized as the optional solution κ to the representation problem

$$k_S = -\mathbb{E}\left[\int_S^\infty c'(t, \sup_{s \in [S,t)} \kappa_s) \, \mu(dt) \,\middle|\, \mathscr{F}_S\right] \quad \text{for any} \quad S \in \mathscr{T} \, . \tag{7}$$

Assumptions 1 and 2 ensure existence of a solution to this problem; see Theorem 3 in Bank and El Karoui (2004).

THEOREM 3.1. *Under Assumptions* 1 *and* 2*, the unique minimizer for problem* (1) *is given by*

$$\theta_t^* \triangleq \sup_{s\in[0,t)} \{\kappa_s \wedge \overline{\vartheta}_s\} \vee \underline{\vartheta} \quad (t \geq 0)\,,$$

where κ *is the optional process solving the representation problem* (7).

REMARK 3.2. *Note that the process* κ *does not depend on the bounds* $\underline{\vartheta}, \overline{\vartheta}$ *describing the set of admissible policies. It thus can be viewed as a universal signal process which yields optimal policies for a whole class of finite fuel problems.*

Proof. Let us verify that the policy $\theta^* \in \mathscr{A}$ satisfies the first order conditions derived in Theorem 2.2.

We first compute the lower Snell envelope $\mathbb{S}(\theta^*)$ of $\nabla C(\theta^*)$. To this end, consider $S, T \in \mathscr{T}$ with $S \leq T$ and note that by definition of θ^* and (7) we have

$$\begin{aligned}
&\mathbb{E}\left[\nabla C(\theta^*)_T \,|\, \mathscr{F}_S\right] \\
&\quad = \mathbb{E}\left[\int_T^\infty \left\{c'\left(t,\ \sup_{s\in[0,t)} \{\kappa_s \wedge \overline{\vartheta}_s\} \vee \underline{\vartheta}\right) - c'\left(t,\ \sup_{s\in[T,t)} \kappa_s\right)\right\} \mu(dt) \,\middle|\, \mathscr{F}_S\right] \\
&\quad \geq \mathbb{E}\left[\int_T^\infty \left\{c'\left(t,\ \sup_{s\in[0,t)} \{\kappa_s \wedge \overline{\vartheta}_s\}\right) - c'\left(t,\ \sup_{s\in[T,t)} \kappa_s\right)\right\} \mu(dt) \,\middle|\, \mathscr{F}_S\right] \\
&\quad \geq \mathbb{E}\left[\int_T^\infty \left\{c'\left(t,\ \sup_{s\in[0,t)} \{\kappa_s \wedge \overline{\vartheta}_s\}\right) - c'\left(t,\ \sup_{s\in[S,t)} \kappa_s\right)\right\} \mu(dt) \,\middle|\, \mathscr{F}_S\right] \\
&\quad \geq \mathbb{E}\left[\int_S^\infty \left\{c'\left(t,\ \sup_{s\in[0,t)} \{\kappa_s \wedge \overline{\vartheta}_s\}\right) - c'\left(t,\ \sup_{s\in[S,t)} \kappa_s\right)\right\} \wedge 0\, \mu(dt) \,\middle|\, \mathscr{F}_S\right].
\end{aligned}$$

Note that the last expression no longer depends on $T \geq S$, thus providing a lower bound for the Snell envelope $\mathbb{S}(\theta^*)$. In fact, it coincides with this envelope since we have equality in any of the above estimates for $T = T_S \triangleq \inf\{t \geq S \mid \kappa_t > \overline{\vartheta}_{t+}\}$. This is easy to see for the first of these estimates since, by definition of T_S, $\sup_{s\in[0,T_S]}\{\kappa_s \wedge \overline{\vartheta}_s\} \geq \overline{\vartheta}_{T_S+} \geq \underline{\vartheta}$, whence $\sup_{s\in[0,t)}\{\kappa_s \wedge \overline{\vartheta}_s\} \geq \underline{\vartheta}$ for $t > T_S$. Equality for the second estimate can be deduced from the observation that T_S is a point of increase for $\sup_{s\in[S,t)} \kappa_s$ which yields $\sup_{s\in[S,t)} \kappa_s = \sup_{s\in[T_S,t)} \kappa_s$ for $t \in (T_S, \infty]$. Finally, equality for the third estimate holds true since

$$c'\left(t,\ \sup_{s\in[0,t)} \{\kappa_s \wedge \overline{\vartheta}_s\}\right) - c'\left(t,\ \sup_{s\in[S,t)} \kappa_s\right) \quad \text{is} \quad \begin{cases} \geq 0 & \text{for } t \in (S, T_S]\,, \\ \leq 0 & \text{for } t \in (T_S, +\infty]\,, \end{cases}$$

again by definition of T_S. For later use, let us also note here that the stopping time T_S is actually the largest stopping time which attains

$$\mathbb{S}(\theta^*)_S = \operatorname*{ess\,inf}_{T\in\mathscr{T}([S,\infty])} \mathbb{E}\left[\nabla C(\theta^*)_T \,|\, \mathscr{F}_S\right],$$

since the above difference is always nonpositive and actually strictly negative on some nontrivial time interval starting at T_S whenever $T_S < +\infty$.

Let us now verify the flat off conditions characterizing optimal plans as described in Theorem 2.2. If S is a point of increase for θ^*, we have $\kappa_S \geq \theta^*_{S+} = \sup_{s\in[0,T_S]}\{\kappa_s \wedge$

$\overline{\vartheta}_s\} \geq \underline{\vartheta}$, and so $\sup_{s\in[S,t)} \kappa_s \geq \sup_{s\in[0,t)}\{\kappa_s \wedge \overline{\vartheta}_s\} \geq \underline{\vartheta}$ for $t \in (S,\infty]$. This allows us to conclude that

$$\begin{aligned}\mathbb{S}(\theta^*)_S &= \mathbb{E}\left[\int_S^\infty \left\{c'\left(t,\ \sup_{s\in[0,t)}\{\kappa_s\wedge\overline{\vartheta}_s\}\right) - c'\left(t,\ \sup_{s\in[S,t)}\kappa_s\right)\right\}\wedge 0\,\mu(dt)\,\middle|\,\mathscr{F}_S\right]\\ &= \mathbb{E}\left[\int_S^\infty \left\{c'\left(t,\ \sup_{s\in[0,t)}\{\kappa_s\wedge\overline{\vartheta}_s\}\right) - c'\left(t,\ \sup_{s\in[S,t)}\kappa_s\right)\right\}\mu(dt)\,\middle|\,\mathscr{F}_S\right]\\ &= \mathbb{E}\left[\int_S^\infty c'\left(t,\ \sup_{s\in[0,t)}\{\kappa_s\wedge\overline{\vartheta}_s\}\vee\underline{\vartheta}\right)\mu(dt)\,\middle|\,\mathscr{F}_S\right] + k_S = \nabla C(\theta^*)_S\,.\end{aligned}$$

So, θ^* is indeed flat off $\{\mathbb{S}(\theta^*) = \nabla C(\theta^*)\}$.

If, on the other hand, S is a point of increase for the predictable compensator $A(\theta^*)$ of $\mathbb{S}(\theta^*)$, then, by classical results on optimal stopping, the only stopping time in $\mathscr{T}([S,+\infty])$ attaining $\mathbb{S}(\theta^*)_S$ is S itself. In particular, the maximal stopping time T_S determined above coincides with S: $T_S = S$, i.e., $S = \inf\{t \geq S \mid \kappa_t > \overline{\vartheta}_{t+}\}$ almost surely. This implies $\kappa_S \geq \overline{\vartheta}_{S+}$ and so $\theta^*_{S+} \geq \overline{\vartheta}_{S+}$ $\mathbb{P}$-a.s. We deduce that almost surely $\theta^*_t = \overline{\vartheta}_t$ for any joint point of continuity t for both θ^* and $\overline{\vartheta}$, i.e., for all but at most countably many points t.

Hence, in order to deduce the desired flat off condition $\theta^*_t = \overline{\vartheta}_t$ for $dA(\theta^*)$-a.e. t, it now suffices to note that $A(\theta^*)$ has continuous sample paths. This, however, holds true because $\nabla C(\theta^*)^-$ and thus also $\mathbb{S}(\theta^*)$ are continuous in expectation. □

4. Applications. As an immediate consequence of Theorem 3.1 we obtain an extension of a result in Karatzas (1985) from the Brownian case to our larger class of finite fuel problems.

COROLLARY 4.1. *The optimal control policy in problem* (1) *with finite fuel (*$\overline{\vartheta} \equiv$ const*) is just the optimal control policy with infinite fuel (*$\overline{\vartheta} \equiv +\infty$*) until all fuel has been exhausted.* □

To obtain a more general result, let us note the following version of the dynamic programming principle.

COROLLARY 4.2. *For each stopping time* $S \in \mathscr{T}$*, the process*

$$\theta^S_t \triangleq \sup_{s\in[S,t)}\{\kappa_s\wedge\overline{\vartheta}_s\}\vee\underline{\vartheta}$$

attains

$$\operatorname*{ess\,inf}_{\theta\in\mathscr{A},\,\theta_S=\underline{\vartheta}} \mathbb{E}\left[\int_S^\infty c(t,\theta_t)\,\mu(dt) + \int_S^\infty k_t\,d\theta_t\,\middle|\,\mathscr{F}_S\right]. \tag{8}$$

In particular, $\theta^S_{S+} = \underline{\vartheta}\vee\limsup_{t\searrow S}\kappa_t\wedge\overline{\vartheta}_{S+}$ *describes the initial policy decision one has to take when starting to minimize costs at time* S*.*

Proof. Define $\overline{\vartheta}^S_t \triangleq \underline{\vartheta}$ for $t \leq S$ and $\overline{\vartheta}^S_t \triangleq \overline{\vartheta}_t$ for $t > S$, and note that by Theorem 3.1 θ^S is the optimal policy in the set of admissible policies $\mathscr{A}^S$ corresponding to $\overline{\vartheta}^S$ instead of $\overline{\vartheta}$. It thus attains $\inf_{\theta\in\mathscr{A}^S} C(\theta)$ and also

$$\inf_{\theta\in\mathscr{A},\,\theta_S=\underline{\vartheta}} \mathbb{E}\left[\int_S^\infty c(t,\theta_t)\,\mu(dt) + \int_S^\infty k_t\,d\theta_t\right] = \mathbb{E}\left[\int_S^\infty c(t,\theta^S_t)\,\mu(dt) + \int_S^\infty k_t\,d\theta^S_t\right].$$

It is easy to see that this infimum is actually the expectation of ess inf in (8). This, however, allows us to conclude our assertion, since this essential infimum is always less than $\mathbb{E}\left[\int_S^\infty c(t,\theta^S_t)\,\mu(dt) + \int_S^\infty k_t\,d\theta^S_t \,\middle|\, \mathscr{F}_S\right]$ almost surely. □

The preceding corollary can be used in two ways. On the one hand, it shows that the process κ of (7) can be used to describe optimal solutions not only when starting at time 0, but actually for any arbitrary initial time $S \in \mathscr{T}$. On the other hand, it allows us to deduce κ (at least partially) from the policies attaining (8). This observation yields the following corollary.

COROLLARY 4.3. *Let K be an optional process such that, for $S \in \mathscr{T}$, $K_S = \theta^S_{S+}$ is the initial value of the optimal policy for problem* (8) *when working under the fuel constraint $\theta^S_t \in [\underline{\vartheta}, \overline{\vartheta}_t]$ $(t \geq S)$. Then the optimal policy for problem* (1) *with fuel constraint $\overline{\vartheta}' \leq \overline{\vartheta}$ is given by*

$$\theta'_t \triangleq \sup_{s\in[0,t)} \{K_s \wedge \overline{\vartheta}'_s\} \vee \underline{\vartheta}\,.$$

In particular, the solutions to the problem without fuel constraint ($\overline{\vartheta} \equiv +\infty$) suffice to determine the optimal policies for the problem with an arbitrary dynamic fuel constraint.

4.1. Monotone follower problems. Let us now come back to the special case of a monotone follower problem studied by Karatzas and Shreve (1984) which we used to motivate the formulation of our general finite fuel problem (1) in section 1. We wish to determine a control policy $\theta \in \mathscr{A}$ which minimizes

$$C(\theta) \triangleq \mathbb{E} \int_0^\infty \delta e^{-\delta t} \frac{1}{2} (W_t - \theta_t)^2 \, dt\,,$$

where W is a standard Brownian motion. It follows from Theorem 3.1 that we can solve this problem explicitly for an arbitrary $\overline{\vartheta}$ by providing a solution to the representation problem (7). In our present setting, this amounts to finding a progressively measurable κ such that

$$\mathbb{E}\left[\int_S^\infty \delta e^{-\delta t} \sup_{s\in[S,t)} \kappa_s \, dt \,\middle|\, \mathscr{F}_S \right] = e^{-\delta S} W_S \quad \text{for all} \quad S \in \mathscr{T}. \tag{9}$$

It is intuitively clear (and has been established formally in Karatzas (1985)) that the optimal policy consists in reflecting the controlled Brownian motion at a certain threshold c. This suggests that we consider the ansatz $\kappa_s \triangleq W_s - c$ for some constant $c \in \mathbb{R}$. Plugging this into (9) and using the independence and time-homogeneity of the increments of W, it is easy to see that (9) will be satisfied if we choose

$$c \triangleq \mathbb{E} \int_0^\infty \delta e^{-\delta t} \sup_{s\in[0,t)} W_s \, dt\,.$$

In fact, looking back, we see that the above reasoning will apply not only to Brownian motion but actually to any Lévy process satisfying suitable integrability properties.

COROLLARY 4.4. *Let X be a Lévy process such that $\mathbb{E}\int_0^\infty \delta e^{-\delta t} X_t^2 \, dt < +\infty$. Then the optimal policy for the monotone follower problem*

$$\textit{Minimize } C(\theta) \triangleq \mathbb{E} \int_0^\infty \delta e^{-\delta t} \frac{1}{2} (X_t - \theta_t)^2 \, dt \textit{ over } \theta \in \mathscr{A}$$

is given by

$$\theta^*_t = \sup_{s\in[0,t)} \{(X_s - c) \wedge \overline{\vartheta}_s\} \vee \underline{\vartheta},$$

where $c \triangleq \mathbb{E}\int_0^\infty \delta e^{-\delta t} \sup_{s\in[0,t)} X_s \, dt < +\infty$.

Proof. We merely have to note that $c < +\infty$ follows from the square-integrability condition on X and Doob's maximal inequality for the martingale $(X_t - t\mathbb{E}X_1)_{t\geq 0}$. □

REMARK 4.5. *While Beneš, Shepp, and Witsenhausen* (1980/81) *study a practically identical cost functional, they allow for downward* and *upward displacement of the particle: controls merely have to be of bounded variation. As they show, in this situation the amount of fuel left becomes crucial for the optimal control decision so that there is no longer a universal process like* κ *describing the optimal policy. Indeed, one will accept larger distances of the controlled processes from the origin with little fuel left than with a lot.*

More generally, the above approach will allow us to explicitly describe optimal control policies whenever the representation problem (9) can be solved explicitly for a given process W, not necessarily Brownian motion. This is indeed possible for a large class of diffusions, as shown in Bank and Föllmer (2003).

If, on the other hand, one wishes to consider a nonquadratic cost functional under a dynamic finite fuel constraint, we can use Corollary 4.2 to reduce the construction of an optimal policy to the unconstraint case with infinite fuel and make use of the results of Chow, Menaldi, and Robin (1985) or Karatzas (1985).

4.2. Irreversible investments. Let us finally illustrate how Corollary 4.4 can be used to extend the closed form solutions obtained for certain irreversible investment problems in Kobila (1993) and Scheinkman and Zariphopoulou (2001) to incorporate a dynamic finite fuel constraint.

Kobila (1993) studies the problem of maximizing a reward functional of (essentially) the type

$$R(\theta) \triangleq \mathbb{E}\int_0^\infty e^{-\delta t}\Pi(X_t, \theta_t)\, dt,$$

where X is a geometric Brownian motion and $\Pi = \Pi(x, \vartheta) : (0, +\infty)\times\mathbb{R} \to \mathbb{R}$ describes the reward function. Apart from a number of technical conditions, Π is assumed to be strictly concave in ϑ; see Condition (5.1) in Kobila (1993). All increasing, left-continuous processes θ with $\theta \geq \underline{\vartheta} = \theta_0$ are considered admissible controls, i.e., $\overline{\vartheta} \equiv +\infty$.

Taking a dynamic programming approach, the authors set up and explicitly solve the Hamilton–Jacobi–Bellman equation for this problem. It turns out that the optimal policy consists in keeping the problem's state process (X, θ) away from a "forbidden" region $\mathscr{R}$ of the form

$$\mathscr{R} = \{(x, \vartheta) \in (0, +\infty) \times \mathbb{R} \mid \phi(x) > \vartheta\}$$

for an explicitly given continuous function ϕ. In particular, when starting at time $S \in \mathscr{T}$ in $\theta_S^S = \underline{\vartheta} \in \mathbb{R}$ the optimal policy requires an initial jump to $K_S = \underline{\vartheta} \vee \phi(X_S)$, i.e., to the minimal $\vartheta \geq \underline{\vartheta}$ such that $(X_S, \vartheta) \notin \mathscr{R}$.

It now follows from Corollary 4.4 that the region $\mathscr{R}$ computed in Kobila (1993) can actually be used to solve the same problem with an arbitrary dynamic fuel constraint $\overline{\vartheta} \not\equiv +\infty$: the optimal policy still consists in keeping the state process away from the region $\mathscr{R}$, at least as long as enough fuel is left to do so. If this is not the case, one has to wait until further supply of fuel becomes available (i.e., until $\overline{\vartheta}$ increases) and then use this fuel to move the state process as close as possible to the complement of $\mathscr{R}$.

The problem studied in Scheinkman and Zariphopoulou (2001) can be viewed as the problem in Kobila (1993) with an additional finite fuel constraint: $\underline{\vartheta} = 0$ and $\overline{\vartheta} = 1$. Given our previous observation, this gives a probabilistic explanation for the similarity of the explicit solution computed in Scheinkman and Zariphopoulou (2001) to the results of Kobila (1993).

Acknowledgments. The author is greatly indebted to Thaleia Zariphopoulou for suggesting this problem. He also wishes to express his gratitude to two anonymous referees for their extremely helpful comments, suggestions, and corrections.

REFERENCES

F. M. Baldursson and I. Karatzas (1997), *Irreversible investment and industry equilibrium*, Finance Stoch., 1, pp. 69–89.

P. Bank and N. El Karoui (2004), *A stochastic representation theorem with applications to optimization and obstacle problems*, Ann. Probab., 32, pp. 1030–1067.

P. Bank and H. Föllmer (2003), *American options, multi-armed bandits, and optimal consumption plans: A unifying view*, in Paris-Princeton Lectures on Mathematical Finance, 2002, Lecture Notes in Math. 1814, Springer-Verlag, Berlin, pp. 1–42.

P. Bank and F. Riedel (2001), *Optimal consumption choice with intertemporal substitution*, Ann. Appl. Probab., 11, pp. 750–788.

V. E. Beneš, L. A. Shepp, and H. S. Witsenhausen (1980/81), *Some solvable stochastic control problems*, Stochastics, 4, pp. 39–83.

G. Bertola (1998), *Irreversible investment*, Res. Econom., 52, pp. 3–37.

M. B. Chiarolla (1997), *Singular stochastic control of a singular diffusion process*, Stochastics Stochastics Rep., 62, pp. 31–63.

M. B. Chiarolla and U. G. Haussmann (1994), *The optimal control of the cheap monotone follower*, Stoch. Stoch. Rep., 49, pp. 99–128.

P.-L. Chow, J.-L. Menaldi, and M. Robin (1985), *Additive control of stochastic linear systems with finite horizon*, SIAM J. Control Optim., 23, pp. 858–899.

N. El Karoui (1981), *Les aspects probabilistes du contrôle stochastique*, in Ninth Saint Flour Probability Summer School—1979 (Saint Flour, 1979), Lecture Notes in Math. 876, Springer-Verlag, Berlin, pp. 73–238.

N. El Karoui and I. Karatzas (1988), *Probabilistic aspects of finite-fuel, reflected follower problems*, Acta Appl. Math., 11, pp. 223–258.

N. El Karoui and I. Karatzas (1991), *A new approach to the Skorohod problem, and its applications*, Stochastics Stochastics Rep., 34, pp. 57–82.

W. H. Fleming and H. M. Soner (1993), *Controlled Markov Processes and Viscosity Solutions*, Appl. Math. (N.Y.) 25, Springer-Verlag, New York.

S. D. Jacka (1999), *Keeping a satellite aloft: Two finite fuel stochastic control models*, J. Appl. Probab., 36, pp. 1–20.

S. D. Jacka (2002), *Avoiding the origin: A finite-fuel stochastic control problem*, Ann. Appl. Probab., 12, pp. 1378–1389.

I. Karatzas (1985), *Probabilistic aspects of finite-fuel stochastic control*, Proc. Natl. Acad. Sci. USA, 82, pp. 5579–5581.

I. Karatzas and S. E. Shreve (1984), *Connections between optimal stopping and singular stochastic control* I. *Monotone follower problems*, SIAM J. Control Optim., 22, pp. 856–877.

I. Karatzas and S. E. Shreve (1985), *Connections between optimal stopping and singular stochastic control* II. *Reflected follower problems*, SIAM J. Control Optim., 23, pp. 433–451.

T. Ø. Kobila (1993), *A class of solvable stochastic investment problems involving singular controls*, Stochastics Stochastics Rep., 43, pp. 29–63.

J. A. Scheinkman and T. Zariphopoulou (2001), *Optimal environmental management in the presence of irreversibilities. Intertemporal equilibrium theory: Indeterminacy, bifurcations, and stability*, J. Econom. Theory, 96, pp. 180–207.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 5, pp. 1543–1574

COORDINATION AND GEOMETRIC OPTIMIZATION VIA DISTRIBUTED DYNAMICAL SYSTEMS*

JORGE CORTÉS† AND FRANCESCO BULLO‡

Abstract. This paper discusses dynamical systems for disk-covering and sphere-packing problems. We present facility location functions from geometric optimization and characterize their differentiable properties. We design and analyze a collection of distributed control laws that are related to nonsmooth gradient systems. The resulting dynamical systems promise to be of use in coordination problems for networked robots; in this setting the distributed control laws correspond to local interactions between the robots. The technical approach relies on concepts from computational geometry, nonsmooth analysis, and the dynamical system approach to algorithms.

Key words. distributed dynamical systems, coordination and cooperative control, geometric optimization, disk-covering problem, sphere-packing problem, nonsmooth analysis, Voronoi partitions

AMS subject classifications. 37N35, 68W15, 93D20, 49J52, 05B40

DOI. 10.1137/S0363012903428652

1. Introduction.

Consider n sites $(p_1, \ldots, p_n)$ evolving within a convex polygon Q according to one of the following interaction laws: (i) each site moves away from the closest other site or polygon boundary, (ii) each site moves toward the furthest vertex of its own Voronoi polygon, or (iii) each site moves toward a geometric center (circumcenter, incenter, centroid, etc.) of its own Voronoi polygon. Recall that the Voronoi polygon of the ith site is the closed set of points $q \in Q$ closer to p_i than to any other p_j.

These and related interaction laws give rise to strikingly simple dynamical systems whose behavior remains largely unknown. What are the critical points of such dynamical systems? What is their asymptotic behavior? Are these systems optimizing any aggregate function? In what way do these local interactions give rise to distributed systems? Does any biological ensemble evolve according to these behaviors and are they of any engineering use in coordination problems? These are the questions that motivate this paper.

Coordination in robotics, control, and biology. Coordination problems are becoming increasingly important in numerous engineering disciplines. The deployment of large groups of autonomous vehicles is rapidly becoming possible because of technological advances in computing, networking, and miniaturization of electromechanical systems. These future multiple-vehicle networks will coordinate their actions to perform challenging spatially distributed tasks (e.g., search and recovery

*Received by the editors May 27, 2003; accepted for publication (in revised form) January 20, 2005; published electronically November 14, 2005. This work was supported in part by DARPA/AFOSR MURI award F49620-02-1-0325 and ONR YIP award N00014-03-1-0512. A preliminary version of this paper was presented as *From geometric optimization and nonsmooth analysis to distributed coordination algorithms*, in Proceedings of the IEEE Conference on Decision and Control, Maui, HI, 2003, pp. 3274–3280.

http://www.siam.org/journals/sicon/44-5/42865.html

†Department of Applied Mathematics and Statistics, University of California at Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, (jcortes@ucsc.edu, http://www.ams.ucsc.edu/∼jcortes).

‡Department of Mechanical and Environmental Engineering, University of California at Santa Barbara, 2338 Engineering Bldg II, Santa Barbara, CA 93106 (bullo@engineering.ucsb.edu, http://www.me.ucsb.edu/bullo).

operations, exploration, surveillance, and environmental monitoring for pollution detection and estimation). This future scenario motivates the study of algorithms for autonomy, adaptation, and coordination of multiple-vehicle networks. It is also important to take into careful consideration all constraints on the behavior of the multiple-vehicle network. Coordination algorithms need to be adaptive and distributed in order for the resulting closed-loop network to be scalable, to comply with bandwidth limitations, to tolerate failures, and to adapt to changing environments, topologies, and sensing tasks. The interaction laws introduced above have these properties and, remarkably, they optimize network-wide performance measures for meaningful spatially distributed tasks.

Coordinated group motions are also a widespread phenomenon in biological systems. Some species of fish spend their lives in schools as a defense mechanism against predators. Others travel as swarms in order to protect an area that they have claimed as their own. Flocks of birds are able to travel in large groups and act as one unit. Other animals exhibit remarkable collective behaviors when foraging and selecting food. Certain foraging behaviors include individual animals partitioning their environment in nonoverlapping individual zones whereas other species develop overlapping team areas. These biological network systems possess extraordinary dynamic capabilities without apparently following a group leader. Yet these complex coordinated behaviors emerge while each individual has no global knowledge of the network state and can only plan its motion according to the observation of its closest neighbors.

Facility location, nonsmooth stability analysis, and cooperative control. To analyze the interaction laws introduced above we rely on concepts and methods from various disciplines. Facility location problems play a prominent role in the field of geometric optimization [1, 5]. Facility location pervades a broad spectrum of scientific and technological areas, including resource allocation (where to place mailboxes in a city or cache servers on the internet), quantization and information theory, mesh and grid optimization methods, clustering analysis, data compression, and statistical pattern recognition. Smooth multicenter functions for so-called centroidal Voronoi configurations and smooth distributed dynamical systems are presented in [11, 14]. Multicenter functions are studied in resource allocation problems [13, 29] and in quantization theory [16, 20]. The role of Voronoi tessellations and computational geometry in facility location is discussed in [23, 26].

The notion and computational properties of the generalized gradient are thoroughly studied in nonsmooth analysis [9]. In particular, tools for establishing stability and convergence properties of nonsmooth dynamical systems are presented in [3, 15, 27]. Finally, we refer to [17] for guidelines on how to design dynamical systems for optimization purposes, and to [4] for gradient descent flows in distributed computation in settings with fixed-communication topologies.

Recent years have witnessed a large research effort focused on motion planning and formation control problems for multiple-vehicle systems [18, 22, 19, 24, 30, 31]. Within the literature on behavior-based robotics, heuristic approaches to the design of interaction rules and emerging behaviors have been investigated (see [2] and references therein). Along this specific line of research, no formal results guaranteeing the correctness of the proposed algorithms or their optimality with respect to an aggregate objective are currently available. The aim of this work is to design distributed coordination algorithms for dynamic networks as well as to provide formal verifications of their asymptotic correctness. A key aspect of our treatment is the inherent complexity of studying networks whose communication topology changes along the

system evolution, as opposed to networks with fixed communication topologies. This key aspect is present in the analysis of distributed control laws in [18, 30, 31] and of agreement protocols in [24].

Statement of contributions. We consider two facility location functions from geometric optimization that characterize coverage performance criteria. A collection of sites provides optimal service to a domain of interest if (i) it minimizes the largest distance from any point in the domain to one of the sites, or (ii) it maximizes the minimum distance between any two sites. In other words, if $P = (p_1, \ldots, p_n)$ are n sites evolving within a convex polygon Q, we extremize the *multicenter functions*

$$\max_{q \in Q} \left\{ \min_{i \in \{1,\ldots,n\}} d(q, p_i) \right\}, \qquad \min_{i \neq j \in \{1,\ldots,n\}} \left\{ \tfrac{1}{2} d(p_i, p_j), d(p_i, \partial Q) \right\},$$

where $d(p, q)$ and $d(p, \partial Q)$ are the distances between p and q, and between p and the boundary of Q, respectively. (The role of the $\frac{1}{2}$ factor will become clear later.) We study the differentiable properties of these functions via nonsmooth analysis. We show the functions are globally Lipschitz and regular, we compute their generalized gradients, and we characterize their critical points. Under certain technical conditions, we show that the local minima of the first multicenter function are so-called circumcenter Voronoi configurations, and that these critical points correspond to the solutions of disk-covering problems. Similarly, under analogous technical conditions, we show that the local maxima of the second multicenter function are so-called incenter Voronoi configurations, and that these critical points correspond to the solutions of sphere-packing problems.

Next, we aim to design distributed algorithms that extremize the multicenter functions. Roughly speaking, by distributed we mean that the evolution of each site depends at most on the location of its own Voronoi neighbors. We study the generalized gradient flows induced by the multicenter functions using nonsmooth stability analysis. Although these dynamical systems possess some convergence properties, they are not amenable to distributed implementations. Next, drawing connections with quantization theory, we consider two dynamical systems associated to each multicenter function. First, we consider a novel strategy based on the generalized gradient of the 1-center functions of each site, and second, we consider a geometric centering strategy similar to the well-known Lloyd algorithm [16, 20].

Remarkably, these strategies arising from the nonsmooth gradient information have natural geometric interpretations and are indeed the local interaction rule described earlier. For the first (respectively, second) multicenter function, the first strategy corresponds to the interaction law "move toward the furthest vertex of own Voronoi polygon" (respectively, "move away from the closest other site or polygon boundary"), and the second strategy corresponds to the interaction law "move toward circumcenter of own Voronoi polygon" (respectively, "move toward incenter of own Voronoi polygon"). We prove the uniqueness of the solutions of the resulting distributed dynamical systems and we analyze their asymptotic behavior using nonsmooth stability analysis, showing that the active sites will approach the corresponding centers of their own Voronoi cells.

Two of our results are related to well-known conjectures in the locational optimization literature [13, 29]: (i) that the first multicenter problem is equivalent to a disk-covering problem (how to cover a region with possibly overlapping disks of equal minimum radius), and (ii) that the generalized Lloyd strategy "move toward

circumcenter of own Voronoi polygon" converges to the set of circumcenter Voronoi configurations.

Organization. The paper is organized as follows. Section 2 provides the preliminary concepts on Voronoi partitions, nonsmooth analysis, stability analysis, and gradient flows, and introduces the multicenter problems. Section 3 presents a complete treatment on the functions analysis and algorithm design for the 1-center problems. Section 4 discusses the differentiable properties and the critical points of the multicenter functions. Section 5 introduces a number of dynamical systems (smooth and nonsmooth, distributed and nondistributed) and analyzes their asymptotic correctness.

2. Preliminaries and problem setup. Let $N \in \mathbb{N}$. We denote by $\|\cdot\|$ the Euclidean distance function on $\mathbb{R}^N$ and by $v \cdot w$ the scalar product of the vectors $v, w \in \mathbb{R}^N$. Let $\mathrm{vrs}(v)$ denote the unit vector in the direction of $0 \neq v \in \mathbb{R}^N$, i.e., $\mathrm{vrs}(v) = v/\|v\|$. Given a set S in $\mathbb{R}^N$, we denote its convex hull by $\mathrm{co}(S)$ and its interior set by $\mathrm{int}(S)$. If S is a convex set in $\mathbb{R}^N$, let $\mathrm{proj}_S : \mathbb{R}^N \to S$ denote the orthogonal projection onto S and let $\mathrm{D}_S : \mathbb{R}^N \to \mathbb{R}$ denote the distance function to S. For $R > 0$, let $\overline{B}_N(p, R) = \{q \in \mathbb{R}^N \mid \|p - q\| \leq R\}$ and $B_N(p, R) = \mathrm{int}(\overline{B}_N(p, R))$. A set $\{v_1, \ldots, v_M\}$ of vectors in $\mathbb{R}^N$ *positively spans* $\mathbb{R}^N$ if any $w \in \mathbb{R}^N$ can be written as $w = \sum_{l=1}^{M} a_l v_l$, with $a_l \geq 0$, $l \in \{1, \ldots, M\}$. The following simple lemma, e.g., see [8], characterizes this situation.

LEMMA 2.1. *Given a set $\{v_1, \ldots, v_M\}$ of $M > N$ arbitrary vectors in $\mathbb{R}^N$, the following statements are equivalent:*

(i) *$\{v_1, \ldots, v_M\}$ positively spans $\mathbb{R}^N$;*
(ii) *$0 \in \mathrm{int}(\mathrm{co}\{v_1, \ldots, v_M\})$;*
(iii) *for each $w \in \mathbb{R}^N$, there exists v_i such that $w \cdot v_i > 0$.*

Let Q be a convex simple polygon in $\mathbb{R}^2$. We denote by $\mathrm{Ed}(Q) = \{e_1, \ldots, e_L\}$ and $\mathrm{Ve}(Q) = \{v_1, \ldots, v_L\}$ the set of edges and vertexes of Q, respectively. Let $P = (p_1, \ldots, p_n) \in Q^n \subset (\mathbb{R}^2)^n$ denote the location of n points (which we will call generators) in the space Q. Let $\pi_i : Q^n \to Q$ be the canonical projection onto the ith factor, $\pi_i(p_1, \ldots, p_n) = p_i$. Note that this mapping is surjective, continuous, and open (the latter meaning that open sets of Q^n are mapped onto open sets of Q).

2.1. Voronoi partitions. We present here some relevant concepts on Voronoi diagrams and refer the reader to [12, 23] for comprehensive treatments. A *partition* of Q is a collection of n polygons $\mathcal{W} = \{W_1, \ldots, W_n\}$ with disjoint interiors whose union is Q. Of course, more general types of partitions could be considered (as, for instance, continuous deformations of the previous ones), but these will be sufficient for our purposes. The *Voronoi partition* $\mathcal{V}(P) = (V_1(P), \ldots, V_n(P))$ of Q generated by the points $(p_1, \ldots, p_n)$ is defined by

$$V_i(P) = \{q \in Q \mid \|q - p_i\| \leq \|q - p_j\| \ \forall j \neq i\}.$$

For simplicity, we shall refer to $V_i(P)$ as V_i. Since Q is a convex polygon, the boundary of each V_i is the union of a finite number of segments. If V_i and V_j share an edge, i.e., $V_i \cap V_j$ is neither empty nor a singleton, then p_i is called a *(Voronoi) neighbor* of p_j (and vice versa). All Voronoi neighboring relations are encoded in the mapping $\mathcal{N} : Q^n \times \{1, \ldots, n\} \to 2^{\{1, \ldots, n\}}$, where $\mathcal{N}(P, i)$ is the set of indexes of the Voronoi neighbors of p_i. Of course, $j \in \mathcal{N}(P, i)$ if and only if $i \in \mathcal{N}(P, j)$. We will often omit P and instead write $\mathcal{N}(i)$.

For $P \in Q^n$, the vertexes of the Voronoi partition $\mathcal{V}(P)$ are classified as follows: the vertex v is

- *of type* (a) if it is the center of the circle passing through three generators (say, p_i, p_j, and p_k),
- *of type* (b) if it is the intersection between an edge of Q and the bisector determined by two generators (say, e, p_i, and p_j), and
- *of type* (c) if it is a vertex of Q, i.e., it is determined by two edges of Q and by the generator of a cell containing it (say, e, f, and p_i).

Correspondingly, we shall write $v(i,j,k)$, $v(e,i,j)$, and $v(e,f,i)$, respectively, whenever we are interested in making explicit the elements defining the vertex v. The vertex $v \in \mathrm{Ve}(V_i(P))$ is said to be *nondegenerate* if it is determined by exactly three elements (e.g., as described above, three generators, or an edge and two generators, or two edges and one generator), otherwise it is said to be *degenerate*. Further, the configuration P is said to be *nondegenerate at the ith generator* if all vertexes $v \in \mathrm{Ve}(V_i(P))$ are nondegenerate, otherwise P is *degenerate at the ith generator*. Finally, a configuration P is said to be *nondegenerate* if all its vertexes are nondegenerate, otherwise it is said to be *degenerate*. These concepts are illustrated in Figure 2.1.

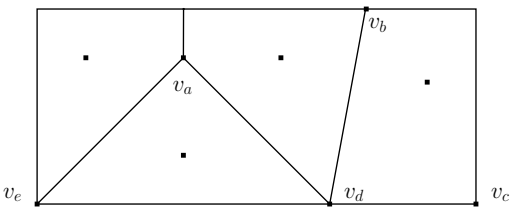


FIG. 2.1. *A Voronoi partition with degenerate and nondegenerate vertexes. Vertexes v_a, v_b, and v_c are nondegenerate vertexes of type* (a), (b), *and* (c), *respectively. Vertexes v_d and v_e are degenerate.*

For $P \in Q^n$, the edges of the Voronoi partition $\mathcal{V}(P)$ are classified as follows: the edge e is

- *of type* (a) if it is a segment of the bisector determined by two generators (say, p_i, p_j), and
- *of type* (b) if it is contained in the boundary of Q, i.e., it is a subset of an edge of Q and it belongs to a single cell (say, the cell of the generator p_i).

Correspondingly, we shall write $e(i,j)$ and $e(i)$, respectively, whenever we are interested in making explicit the elements defining the edge e. Further, when considering an edge of type (a), we let $n_{e(i,j)}$ denote the unit normal to $e(i,j)$ pointing toward $\mathrm{int}(V_i(P))$. When considering an edge of type (b), we let $n_{e(i)}$ denote the unit normal to $e(i)$ pointing toward $\mathrm{int}(Q)$.

2.2. The disk-covering and the sphere-packing problems. We are interested in the following locational optimization problems:

$$\min_{p_1,\dots,p_n} \left\{ \max_{q \in Q} \left\{ \min_{i \in \{1,\dots,n\}} \|q - p_i\| \right\} \right\}, \tag{2.1}$$

$$\max_{p_1,\dots,p_n} \left\{ \min_{\substack{i,j \in \{1,\dots,n\} \\ i \neq j,\ e \in \mathrm{Ed}(Q)}} \left\{ \tfrac{1}{2}\|p_i - p_j\|, \mathrm{D}_e(p_i) \right\} \right\}. \tag{2.2}$$

The optimization problem (2.1) is referred to as the p-center problem in [13, 29]. Throughout the paper, we will refer to it as the multicircumcenter problem. In the context of coverage control of mobile sensor networks [11], the multicircumcenter problem corresponds to considering the worst case scenario, in which no information is available on the distribution of the events taking place in the environment Q. The network therefore tries to minimize the largest possible distance of any point in Q to one of the generators' locations given by $p_1, \dots, p_n$, i.e., to minimize the function

$$\mathcal{H}_{\mathrm{DC}}(P) = \max_{q \in Q} \left\{ \min_{i \in \{1,\dots,n\}} \|q - p_i\| \right\} = \max_{i \in \{1,\dots,n\}} \left\{ \max_{q \in V_i} \|q - p_i\| \right\}.$$

It is conjectured in [29] that this problem can be restated as a disk-covering problem: how to cover a region with (possibly overlapping) disks of minimum radius. The disk-covering problem then reads

$$\min\{R \mid \cup_{i \in \{1,\dots,n\}} \overline{B}_2(p_i, R) \supseteq Q\}.$$

We shall present a proof of this statement in Theorem 4.7 below. Given a polytope W in $\mathbb{R}^N$, its circumcenter, denoted by $\mathrm{CC}(W)$, is the center of the minimum-radius sphere that contains W. The circumradius of W, denoted by $\mathrm{CR}(W)$, is the radius of this sphere. We will say that P is a *circumcenter Voronoi configuration* if $p_i = \mathrm{CC}(V_i(P))$, for all $i \in \{1, \dots, n\}$. We denote by $\mathrm{Ve}_{\mathrm{DC}}(\mathcal{V}(P))$ the set of vertexes of the Voronoi partition where the value $\mathcal{H}_{\mathrm{DC}}(P)$ is attained, i.e., $v \in \mathrm{Ve}_{\mathrm{DC}}(\mathcal{V}(P))$ if there exists i such that $v \in V_i(P)$ and $\|v - p_i\| = \mathcal{H}_{\mathrm{DC}}(P)$. In such cases, we will often refer to both the vertex v and the generator p_i as *active*.

We will refer to the optimization problem (2.2) as the multi-incenter problem. In the context of applications, this problem corresponds to the situation where we are interested in maximizing the coverage of the area Q in such a way that the sensing radius of the generators do not overlap (in order not to interfere with each other) or leave the environment. We therefore consider the maximization of the function

$$\mathcal{H}_{\mathrm{SP}}(P) = \min_{\substack{i,j \in \{1,\dots,n\} \\ i \neq j,\, e \in \mathrm{Ed}(Q)}} \left\{ \tfrac{1}{2}\|p_i - p_j\|, \mathrm{D}_e(p_i) \right\} = \min_{i \in \{1,\dots,n\}} \left\{ \min_{q \notin \mathrm{int}(V_i)} \|q - p_i\| \right\}.$$

A similar conjecture to the one presented above is that the multi-incenter problem can be restated as a sphere-packing problem: how to maximize the coverage of a region with nonoverlapping disks (contained in the region) of maximum radius. The problem reads

$$\max\{R \mid \cup_{i \in \{1,\dots,n\}} \overline{B}_2(p_i, R) \subseteq Q\,,\ B_2(p_i, R) \cap B_2(p_j, R) = \emptyset\}.$$

In Theorem 4.8 we provide a positive answer to this question. Given a polytope W in $\mathbb{R}^N$, its incenter set (or Chebyshev center set; see [6]), denoted by $\mathrm{IC}(W)$, is the set of the centers of maximum-radius spheres contained in W. The inradius of W, denoted by $\mathrm{IR}(W)$, is the common radius of these spheres. We will say that $P \in Q^n$ is an *incenter Voronoi configuration* if $p_i \in \mathrm{IC}(V_i(P))$, for all $i \in \{1, \dots, n\}$. If P is an incenter Voronoi configuration and each Voronoi region $V_i(P)$ has a unique incenter, $\mathrm{IC}(V_i(P)) = \{p_i\}$, then we will say that P is a *generic incenter Voronoi configuration*. We denote by $\mathrm{Ed}_{\mathrm{SP}}(\mathcal{V}(P))$ the set of edges of the Voronoi partition where the value $\mathcal{H}_{\mathrm{SP}}(P)$ is attained; i.e., $e \in \mathrm{Ed}_{\mathrm{SP}}(\mathcal{V}(P))$ if there exists i such that $e \in \mathrm{Ed}(V_i(P))$ and $\mathrm{D}_e(p_i) = \mathcal{H}_{\mathrm{SP}}(P)$. In such cases, we will often refer to both the edge e and the generator p_i as *active*.

2.3. Nonsmooth analysis. The following facts on nonsmooth analysis [9] will be most helpful in analyzing the properties of the locational optimization functions for the disk-covering and the sphere-packing problems, as well as the convergence of the distributed algorithms we will propose to extremize them.

We begin by recalling some basic notions. A function $f : \mathbb{R}^N \to \mathbb{R}$ is said to be *locally Lipschitz at* $x \in \mathbb{R}^N$ if there exist positive constants L_x and ϵ such that $|f(y) - f(y')| \leq L_x \|y - y'\|$ for all $y, y' \in B_N(x, \epsilon)$. The function f is said to be *locally Lipschitz on* $S \subset \mathbb{R}^N$ *if it is locally Lipschitz at* x*, for all* $x \in S$. Note that continuously differentiable functions at x are locally Lipschitz at x. On the other hand, a function $f : \mathbb{R}^N \to \mathbb{R}$ is said to be *regular at* $x \in \mathbb{R}^N$ if for all $v \in \mathbb{R}^N$ the right directional derivative of f at x in the direction of v, denoted by $f'(x; v)$, exists and coincides with the generalized directional derivative of f at x in the direction of v, denoted by $f^o(x; v)$. The interested reader is referred to [9] for the precise definition of these directional derivatives. Again, a continuously differentiable function at x is regular at x. Also, a locally Lipschitz function at x which is convex is regular (cf. Proposition 2.3.6 in [9]).

From Rademacher's theorem [9], we know that locally Lipschitz functions are differentiable almost everywhere (in the sense of Lebesgue measure). If Ω_f denotes the set of points in $\mathbb{R}^N$ at which f fails to be differentiable and S denotes any other set of measure zero, the *generalized gradient* of f is defined by

$$\partial f(x) = \operatorname{co}\left\{ \lim_{i \to +\infty} df(x_i) \mid x_i \to x\,,\ x_i \notin S \cup \Omega_f \right\}.$$

Note that this definition coincides with $df(x)$ if f is continuously differentiable at x. A point $x \in \mathbb{R}^N$ which verifies that $0 \in \partial f(x)$ is called a *critical point of* f. The following result corresponds to Proposition 2.3.12 in [9].

PROPOSITION 2.2. *Let* $f_k : \mathbb{R}^N \to \mathbb{R}$, $k \in \{1, \ldots, m\}$ *be locally Lipschitz functions at* $x \in \mathbb{R}^N$ *and let* $f(x') = \max\{f_k(x') \mid k \in \{1, \ldots, m\}\}$. *Then,*

(i) *f is locally Lipschitz at x,*

(ii) *if $I(x')$ denotes the set of indexes k for which $f_k(x') = f(x')$, we have*

$$\partial f(x) \subset \operatorname{co}\{\partial f_i(x) \mid i \in I(x)\}\,, \tag{2.3}$$

and if f_i, $i \in I(x)$, is regular at x, then equality holds and f is regular at x.

The extrema of Lipschitz functions are characterized by the following result.

PROPOSITION 2.3. *Let f be a locally Lipschitz function at $x \in \mathbb{R}^N$. If f attains a local minimum or maximum at x, then $0 \in \partial f(x)$, i.e., x is a critical point.*

Let $\operatorname{Ln} : 2^{\mathbb{R}^N} \to 2^{\mathbb{R}^N}$ be the set-valued mapping that associates to each subset S of $\mathbb{R}^N$ the set of its least-norm elements $\operatorname{Ln}(S)$. If the set S is convex, then the set $\operatorname{Ln}(S)$ reduces to a singleton and we note the equivalence $\operatorname{Ln}(S) = \operatorname{proj}_S(0)$. In this paper, we shall only apply this function to convex sets. For a locally Lipschitz function f, we consider the *generalized gradient vector field* $\operatorname{Ln}(\partial f) : \mathbb{R}^N \to \mathbb{R}^N$ given by $x \mapsto \operatorname{Ln}(\partial f)(x) = \operatorname{Ln}(\partial f(x))$. The following theorem (cf. [9]) establishes an important feature of this vector field.

THEOREM 2.4. *Let f be a locally Lipschitz function at x. Assume $0 \notin \partial f(x)$. Then, there exists $T > 0$ such that*

$$f(x - t \operatorname{Ln}(\partial f)(x)) \leq f(x) - \frac{t}{2} \|\operatorname{Ln}(\partial f)(x)\|^2\,, \quad 0 < t < T\,.$$

The vector $-\operatorname{Ln}(\partial f)(x)$ is called a direction of descent.

2.4. Stability analysis via nonsmooth Lyapunov functions. Throughout the paper, we will define the solutions of differential equations with discontinuous right-hand sides in terms of differential inclusions [15]. Let $F : \mathbb{R}^N \to 2^{\mathbb{R}^N}$ be a set-valued map. Consider the differential inclusion

$$\dot{x} \in F(x)\,. \tag{2.4}$$

A solution to this equation on an interval $[t_0, t_1] \subset \mathbb{R}$ is defined as an absolutely continuous function $x : [t_0, t_1] \to \mathbb{R}^N$ such that $\dot{x}(t) \in F(x(t))$ for almost all $t \in [t_0, t_1]$. Given $x_0 \in \mathbb{R}^N$, the existence of at least a solution with initial condition x_0 is guaranteed by the following lemma.

LEMMA 2.5. *Let the mapping F be upper semicontinuous with nonempty, compact, and convex values. Then, given $x_0 \in \mathbb{R}^N$, there exists a local solution of* (2.4) *with initial condition x_0.*

Now, consider the differential equation

$$\dot{x}(t) = X(x(t))\,, \tag{2.5}$$

where $X : \mathbb{R}^N \to \mathbb{R}^N$ is measurable and essentially locally bounded. There are various notions of solutions to discontinuous differential equations (see [7, Chapter 1] for a comparative discussion between them). Here, we will understand the solution of this equation in the Filippov sense, which we define in the following. For each $x \in \mathbb{R}^N$, consider the set

$$K[X](x) = \bigcap_{\delta>0} \bigcap_{\mu(S)=0} \mathrm{co}\{X(B_N(x,\delta) \setminus S)\}\,,$$

where μ denotes the usual Lebesgue measure in $\mathbb{R}^N$. Alternatively, one can show [25] that there exists a set S_X of measure zero such that

$$K[X](x) = \mathrm{co}\left\{ \lim_{i\to+\infty} X(x_i) \mid x_i \to x\,,\ x_i \notin S \cup S_X \right\}\,,$$

where S is any set of measure zero. A Filippov solution of (2.5) on an interval $[t_0, t_1] \subset \mathbb{R}$ is defined as a solution of the differential inclusion $\dot{x} \in K[X](x)$. Since the multivalued mapping $K[X] : \mathbb{R}^N \to 2^{\mathbb{R}^N}$ is upper semicontinuous with nonempty, compact, convex values and locally bounded (cf. [15]), the existence of Filippov solutions of (2.5) is guaranteed by Lemma 2.5.

A set M is *weakly invariant* (respectively, *strongly invariant*) for (2.5) if for each $x_0 \in M$, M contains a maximal solution (respectively, all maximal solutions) of (2.5). Given a locally Lipschitz function $f : \mathbb{R}^N \to \mathbb{R}$, the *set-valued Lie derivative of f with respect to X* at x is defined as

$$\widetilde{\mathcal{L}}_X f(x) = \{a \in \mathbb{R} \mid \exists v \in K[X](x) \text{ such that } \zeta \cdot v = a \ \ \forall \zeta \in \partial f(x)\}\,.$$

For each $x \in \mathbb{R}^N$, $\widetilde{\mathcal{L}}_X f(x)$ is a closed and bounded interval in $\mathbb{R}$, possibly empty. If f is continuously differentiable at x, then $\widetilde{\mathcal{L}}_X f(x) = \{df \cdot v \mid v \in K[X](x)\}$. If, in addition, X is continuous at x, then $\widetilde{\mathcal{L}}_X f(x)$ corresponds to the singleton $\{\mathcal{L}_X f(x)\}$, the usual Lie derivative of f in the direction of X at x. The importance of the set-valued Lie derivative stems from the next result [3].

THEOREM 2.6. *Let $x : [t_0, t_1] \to \mathbb{R}^N$ be a Filippov solution of (2.5). Let f be a locally Lipschitz and regular function. Then $\frac{d}{dt}(f(x(t)))$ exists a.e. and $\frac{d}{dt}(f(x(t))) \in \widetilde{\mathcal{L}}_X f(x(t))$ a.e.*

The following result is a generalization of the LaSalle principle for differential equations of the form (2.5) with nonsmooth Lyapunov functions. The formulation is taken from [3] and slightly generalizes the one presented in [27].

THEOREM 2.7 (LaSalle principle). *Let $f : \mathbb{R}^N \to \mathbb{R}$ be a locally Lipschitz and regular function. Let $x_0 \in \mathbb{R}^N$ and let $f^{-1}(\leq f(x_0), x_0)$ be the connected component of $\{x \in \mathbb{R}^N \mid f(x) \leq f(x_0)\}$ containing x_0. Assume the set $f^{-1}(\leq f(x_0), x_0)$ is bounded and assume either $\max \widetilde{\mathcal{L}}_X f(x) \leq 0$ or $\widetilde{\mathcal{L}}_X f(x) = \emptyset$ for all $x \in f^{-1}(\leq f(x_0), x_0)$. Then $f^{-1}(\leq f(x_0), x_0)$ is strongly invariant for (2.5). Let*

$$Z_{X,f} = \{x \in \mathbb{R}^N \mid 0 \in \widetilde{\mathcal{L}}_X f(x)\}.$$

Then, any solution $x : [t_0, +\infty) \to \mathbb{R}^N$ of (2.5) starting from x_0 converges to the largest weakly invariant set M contained in $\overline{Z}_{X,f} \cap f^{-1}(\leq f(x_0), x_0)$. Furthermore, if the set M is a finite collection of points, then the limit of all solutions starting at x_0 exists and equals one of those points.

The proof of the last fact in the theorem statement is the same as in the smooth case, since it only relies on the continuity of the trajectory. The next statement is based on Theorem 2 of [25].

PROPOSITION 2.8. *Under the same assumptions of Theorem 2.7, if $\max \widetilde{\mathcal{L}}_X f(x) < -\epsilon < 0$ a.e. on $\mathbb{R}^N \setminus Z_{X,f}$, then $Z_{X,f}$ is attained in finite time.*

Proof. Let $x : [t_0, +\infty) \to \mathbb{R}^N$ be a Filippov solution starting from x_0. We argue that there must exist T such that $x(T) \in Z_{X,f}$. Otherwise, we have

$$f(x(t)) = f(x(t_0)) + \int_{t_0}^{t} \frac{d}{ds} f(x(s)) ds < f(x(t_0)) - \epsilon(t - t_0) \overset{t \to +\infty}{\longrightarrow} -\infty,$$

contradicting the fact that $f^{-1}(\leq f(x_0), x_0)$ is strongly invariant and bounded. ∎

2.5. Nonsmooth gradient flows. Finally, we are in a position to present the nonsmooth analogue of well-known results on gradient flows. Given a locally Lipschitz and regular function f, consider the following generalized gradient flow:

$$\dot{x}(t) = -\operatorname{Ln}(\partial f)(x(t)). \tag{2.6}$$

Theorem 2.4 guarantees that unless the flow is at a critical point, $-\operatorname{Ln}(\partial f)(x)$ is always a direction of descent at x. In general, the vector field $\operatorname{Ln}(\partial f)$ in (2.6) is discontinuous. We understand its solution in the Filippov sense. Note that since f is locally Lipschitz, $\operatorname{Ln}(\partial f) = df$ almost everywhere. An important observation in this setting is that $K[df](x) = \partial f(x)$ (cf. [25]). The following result, which is a generalization of the discussion in [3], guarantees the convergence of this flow to the set of critical points of f.

PROPOSITION 2.9. *Let $x_0 \in \mathbb{R}^N$ and assume $f^{-1}(\leq f(x_0), x_0)$ is bounded. Then, any solution $x : [t_0, +\infty) \to \mathbb{R}^N$ of (2.6) starting from x_0 converges asymptotically to the set of critical points of f contained in $f^{-1}(\leq f(x_0), x_0)$.*

Proof. Let $a \in \widetilde{\mathcal{L}}_{-\operatorname{Ln}(\partial f)} f(x)$. By definition, there exists $w \in K[-\operatorname{Ln}(\partial f)](x) = -\partial f(x)$ such that $a = w \cdot \zeta$ for all $\zeta \in \partial f(x)$. In particular, for $\zeta = -w \in \partial f(x)$, we have $a = -\|w\|^2 \leq 0$. Therefore, $\max \widetilde{\mathcal{L}}_{-\operatorname{Ln}(\partial f)} f(x) \leq 0$ or $\widetilde{\mathcal{L}}_{-\operatorname{Ln}(\partial f)} f(x) = \emptyset$. Now, resorting to the LaSalle principle (Theorem 2.7), we deduce that any solution

$x : [t_0, +\infty) \to \mathbb{R}^N$ starting from x_0 converges to the largest weakly invariant set contained in $\overline{Z}_{-\operatorname{Ln}(\partial f),f} \cap f^{-1}(\leq f(x_0), x_0)$. Let us see that $Z_{-\operatorname{Ln}(\partial f),f}$ is equal to $L_0 = \{x \in Q^n \mid 0 \in \partial f(x)\}$. Obviously, $L_0 \subset Z_{-\operatorname{Ln}(\partial f),f}$. Conversely, assume $x \in Z_{-\operatorname{Ln}(\partial f),f}$. Then, $0 \in \widetilde{\mathcal{L}}_{-\operatorname{Ln}(\partial f)} f(x)$; i.e., there exists $v \in -\partial f(x)$ such that $\zeta \cdot v = 0$ for all $\zeta \in \partial f(x)$. In particular, for $\zeta = -v$, we get $\|v\|^2 = 0$, that is, $v = 0 \in \partial f(x)$, as desired. Note that $Z_{-\operatorname{Ln}(\partial f),f} = L_0$ is the equilibrium set of (2.6) and therefore is weakly invariant. Finally, we prove that it is also closed. Let $x \in \overline{Z}_{-\operatorname{Ln}(\partial f),f}$ and consider a sequence $\{x_k \in \mathbb{R}^N \mid k \in \mathbb{N}\} \subset Z_{-\operatorname{Ln}(\partial f),f}$ such that $x_k \to x$. Then, using the fact that the multivalued mapping $K[-v]$ is upper semicontinuous, for any $\epsilon > 0$ there exists k_0 such that for $k \geq k_0$, $\partial f(x_k) \subset \partial f(x) + B_N(0, \epsilon)$. Since $x_k \in Z_{-\operatorname{Ln}(\partial f),f}$, then $0 \in \partial f(x) + B_N(0, \epsilon)$ for all $\epsilon > 0$, and this implies that $0 \in \partial f(x)$, i.e., $x \in Z_{-\operatorname{Ln}(\partial f),f}$. Hence the largest weakly invariant set contained in $\overline{Z}_{-\operatorname{Ln}(\partial f),f} \cap f^{-1}(\leq f(x_0), x_0)$ is $Z_{-\operatorname{Ln}(\partial f),f} \cap f^{-1}(\leq f(x_0), x_0) = \{x \in f^{-1}(\leq f(x_0), x_0) \mid 0 \in \partial f(x)\}$. □

3. The 1-center problems. In this section we consider the disk-covering and the sphere-packing problems with a single generator, i.e., $n = 1$. This treatment will give us the necessary insight to tackle later the more involved multicenter version of both problems. When $n = 1$, the minimization of $\mathcal{H}_{\mathrm{DC}}$ simply consists of finding the center of the minimum-radius sphere enclosing the polygon Q. On the other hand, the maximization of $\mathcal{H}_{\mathrm{SP}}$ corresponds to determining the center of the maximum-radius sphere contained in Q. Let us therefore define the functions

$$\begin{aligned} \operatorname{lg}_Q(p) &= \max\{\|q - p\| \mid q \in Q\} = \max\{\|v - p\| \mid v \in \operatorname{Ve}(Q)\}, \\ \operatorname{sm}_Q(p) &= \min\{\|q - p\| \mid q \notin \operatorname{int}(Q)\} = \min\{\operatorname{D}_e(p) \mid e \in \operatorname{Ed}(Q)\}. \end{aligned} \tag{3.1}$$

When $n = 1$, we then have that $\mathcal{H}_{\mathrm{DC}} = \operatorname{lg}_Q : Q \to \mathbb{R}$ and $\mathcal{H}_{\mathrm{SP}} = \operatorname{sm}_Q : Q \to \mathbb{R}$.

3.1. Smoothness and critical points. We here discuss the smoothness properties and the critical points of the 1-center functions. Since the function lg_Q is the maximum of a (finite) set of convex functions in p, it is also a convex function [6]. Therefore, any local minimum of lg_Q is also global.

LEMMA 3.1. *The function* lg_Q *has a unique global minimum, which is the circumcenter of the polygon* Q.

Proof. Let $F : \mathbb{R} \to \mathbb{R}$ be any continuous nondecreasing function. Then

$$F(\operatorname{lg}_Q(p)) = \max\{F(\|v - p\|) \mid v \in \operatorname{Ve}(Q)\}.$$

If we take $F(x) = x^2$, each function $\|v - p\|^2$ is strictly convex, and hence $F(\operatorname{lg}_Q(p))$ is also strictly convex. Therefore, this latter function has a single minimum on Q. Since any global minimum of lg_Q is also a global minimum of $F(\operatorname{lg}_Q(p))$, we conclude the result. □

The function sm_Q is the minimum of a (finite) set of affine (hence, concave) functions defined on the half-planes determined by the edges of Q, and hence it is also a concave function [6] on the intersection of their domains, which is precisely Q. Therefore, any local maximum of sm_Q is also global. However, this maximum is not unique in general.

LEMMA 3.2. *The incenter set of the polygon* Q *is the set of maxima of the function* sm_Q *and it is a segment.*

Proof. It is clear that the set of maxima of sm_Q is $\operatorname{IC}(Q)$. As a consequence of the concavity of sm_Q over the convex domain Q, one deduces that $\operatorname{IC}(Q)$ is a convex

set. Now, assume there are three points p_1, p_2, p_3 in $\mathrm{IC}(Q)$ which are not aligned. Since $B_2(q, \mathrm{IR}(Q)) \subset Q$ for all $q \in \mathrm{co}(p_1, p_2, p_3) \subset \mathrm{IC}(Q)$, and $\mathrm{co}(p_1, p_2, p_3)$ has a nonempty interior, there exist $q_0 \in Q$ and $r > \mathrm{IR}(Q)$ such that $B_2(q_0, r) \subset Q$, which is a contradiction. □

Note that the circumcenter of a polygon can be computed via the finite-step algorithm described in [28]. The incenter set of a polygon can be computed via the following linear program in q and r: maximize the radius r of the sphere centered at q subject to the constraints that the distance between q and each of the polygon edges is greater than or equal to r. Formally, the problem can be expressed as follows. For each $e \in \mathrm{Ed}(Q)$, select a point $q_e \in Q$ belonging to e. Then, we set

$$\begin{aligned} &\text{maximize } r\,, \\ &\text{subject to } (q - q_e) \cdot n_e \geq r \quad \forall\, e \in \mathrm{Ed}(Q)\,. \end{aligned}$$

In what follows, let us examine dynamical systems that compute these geometric centers.

PROPOSITION 3.3. *The functions* $\mathrm{lg}_Q(p)$, $-\mathrm{sm}_Q(p)$ *are locally Lipschitz and regular, and their generalized gradients are given by*

$$\partial \mathrm{lg}_Q(p) = \mathrm{co}\{\mathrm{vrs}(p - v) \mid v \in \mathrm{Ve}(Q)\,,\ \mathrm{lg}_Q(p) = \|p - v\|\}\,, \tag{3.2}$$

$$\partial \mathrm{sm}_Q(p) = \mathrm{co}\{n_e \mid e \in \mathrm{Ed}(Q)\,,\ \mathrm{sm}_Q(p) = \mathrm{D}_e(p)\}\,. \tag{3.3}$$

Moreover,

$$0 \in \partial \mathrm{lg}_Q(p) \Longleftrightarrow p = \mathrm{CC}(Q)\,, \quad 0 \in \partial \mathrm{sm}_Q(p) \Longleftrightarrow p \in \mathrm{IC}(Q)\,, \tag{3.4}$$

and, if $0 \in \mathrm{int}(\partial \mathrm{sm}_Q(p))$, *then* $\mathrm{IC}(Q) = \{p\}$.

Proof. Given the expressions in (3.1) and Proposition 2.2, we deduce that lg_Q and $-\mathrm{sm}_Q$ are locally Lipschitz and regular, and that their generalized gradients are given by (3.2) and (3.3), respectively. Concerning (3.4), the implications from right to left in (3.4) readily follow from Proposition 2.3. As for the other ones, note that it is sufficient to prove that p is a local minimum (respectively, that p is a local maximum). We prove the result for the function lg_Q. The proof for sm_Q is analogous. Assume that $0 \in \partial \mathrm{lg}_Q(p)$. Then there exist vertexes $v_{i_1}, \ldots, v_{i_K}$ of Q with $\mathrm{lg}_Q(p) = \|v_{i_l} - p\|$, $l \in \{1, \ldots, K\}$ such that $0 = \sum_{l \in \{1,\ldots,K\}} \lambda_l \,\mathrm{vrs}(p - v_{i_l})$, where $\sum_{l \in \{1,\ldots,K\}} \lambda_l = 1$, $\lambda_l \geq 0$, $l \in \{1, \ldots, K\}$. Let U be a neighborhood of p and take $q \in U$. One can show that there must exist l^* such that $(p - v_{i_{l^*}}) \cdot (q - p) \geq 0$, since otherwise $0 = 0 \cdot (q - p) = (\sum_{l \in \{1,\ldots,K\}} \lambda_l \,\mathrm{vrs}(p - v_{i_l})) \cdot (q - p) < 0$, which is a contradiction. Then

$$\|q - v_{i_{l^*}}\|^2 = \|q - p\|^2 + \|p - v_{i_{l^*}}\|^2 - 2(q - p) \cdot (v_{i_{l^*}} - p) \geq \|p - v_{i_{l^*}}\|^2\,.$$

Therefore, $\mathrm{lg}_Q(q) \geq \|p - v_{i_{l^*}}\| = \mathrm{lg}_Q(p)$, which shows that p is a local minimum. Finally, if $0 \in \mathrm{int}(\partial \mathrm{sm}_Q(p))$, then one can see that p is a strict local maximum. Furthermore, there cannot be any other local (hence global) maximum of sm_Q, as we now show. Assume $\bar{p} \in \mathrm{IC}(Q)$. By hypothesis, the sphere $B_2(\bar{p}, \mathrm{sm}_Q(p))$ centered at $\bar{p}$ of radius $\mathrm{sm}_Q(p)$ is contained in Q. Consider the vector $\bar{p} - p$. By Lemma 2.1, there exists $e \in \mathrm{Ed}(Q)$ with $\mathrm{D}_e(p) = \mathrm{sm}_Q(p)$ such that $(\bar{p} - p) \cdot n_e > 0$. Therefore, there are points of $B_2(\bar{p}, \mathrm{sm}_Q(p))$ which necessarily belong to the half-plane defined by e where Q is not contained, which is a contradiction. □

3.2. Convergence properties for nonsmooth gradient flows. Here we study the generalized gradient flows arising from the two 1-center functions. An immediate consequence of Propositions 2.9 and 3.3 is the following result.

COROLLARY 3.4. *The gradient flows of the functions* $\lg_Q$ *and* sm_Q*,*

$$\dot{x}(t) = -\operatorname{Ln}(\partial \lg_Q)(x(t))\,, \tag{3.5}$$

$$\dot{x}(t) = \operatorname{Ln}(\partial \mathrm{sm}_Q)(x(t))\,, \tag{3.6}$$

converge asymptotically to the circumcenter $\mathrm{CC}(Q)$ *and the incenter set* $\mathrm{IC}(Q)$*, respectively.*

The following two propositions discuss the convergence properties of the gradient descents.

PROPOSITION 3.5. *If* $0 \in \operatorname{int}(\partial \lg_Q(\mathrm{CC}(Q)))$*, then the flow* (3.5) *reaches* $\mathrm{CC}(Q)$ *in finite time.*

Proof. Let us prove that there exists $\epsilon > 0$ such that $\max \widetilde{\mathcal{L}}_{-\operatorname{Ln}[\lg_Q]} \lg_Q < -\epsilon$ a.e. on $Q \setminus \{\mathrm{CC}(Q)\}$. Take $p \neq \mathrm{CC}(Q)$. We know that each element $a \in \widetilde{\mathcal{L}}_{-\operatorname{Ln}[\lg_Q]} \lg_Q(p)$ can be expressed as $a = -\|w\|^2$, with $-w \in \partial \lg_Q(p)$. Therefore, we have

$$\max \widetilde{\mathcal{L}}_{-\operatorname{Ln}[\lg_Q]} \lg_Q(p) = -\|\operatorname{Ln}[\lg_Q](p)\|^2\,.$$

If there is a single vertex of Q involved in $\partial \lg_Q(p)$, then moving along the direction $-\operatorname{Ln}[\lg_Q](p)$ obviously decreases the distance to that vertex while maintaining constant the norm of the least-norm element, which is 1. If there are two or more vertexes involved, then the generalized gradient at p (cf. (3.2)) can be alternatively described as a polygonal region of the form

$$\{x \in \mathbb{R}^N \mid g_1(x) \leq 0, \ldots, g_s(x) \leq 0\},$$

where each g_r is a linear function whose annihilation corresponds to a set of the form $\mathrm{co}\{\mathrm{vrs}(p - v_{r,1}), \mathrm{vrs}(p - v_{r,2})\}$ for certain vertexes $v_{r,1}$, $v_{r,2}$ of Q. Now, the computation of the least-norm element in $\partial \lg_Q(p)$ can be formulated as the convex problem

$$\begin{aligned} &\text{minimize } \|x\|^2 \\ &\text{subject to } g_1(x) \leq 0, \ldots, g_s(x) \leq 0\,. \end{aligned}$$

Let $x^* = \operatorname{Ln}[\lg_Q](p)$. Let R denote the set of indexes r for which $g_r(x^*) = 0$. Then x^* is a regular point [21], meaning that $dg_r(x^*)$, $r \in R$ are linearly independent vectors. This is because the cardinality of R is at most 2 (since the intersection of two lines already determines a point), and the gradients of the functions g_r are independent when considered pairwise. We apply then the Kuhn–Tucker first-order necessary conditions for optimality [21] to conclude that there must exist $r^* \in R$ such that $g_{r^*}(x^*) = 0$. It is easy to see that r^* must be unique, since otherwise x^* does not have minimum norm. Therefore, we have that $\operatorname{Ln}[\lg_Q](p)$ is determined as the least-norm element in $\mathrm{co}\{\mathrm{vrs}(p - v_{r^*,1}), \mathrm{vrs}(p - v_{r^*,2})\}$. As a consequence, moving along the direction $-\operatorname{Ln}[\lg_Q](p)$ decreases the distance to the vertexes $v_{r^*,1}$, $v_{r^*,2}$, and hence the norm of the least-norm element decreases. If, along the flow (3.5), a new vertex of Q enters in the computation of $\partial \lg_Q(p(t))$, then there can be a jump in the norm of $\operatorname{Ln}[\lg_Q](p(t))$, which by definition will always be decreasing. Finally, note that if $v_{r^*,1}$, $v_{r^*,2}$ are active at the circumcenter, then they cannot be opposite with respect to $\mathrm{CC}(Q)$. If this was the case, then the assumption that 0 lies in $\operatorname{int}(\partial \lg_Q(\mathrm{CC}(Q)))$

would imply that there exists another vertex v of Q, which is active at the circumcenter and lies in the half-plane defined by the line from $v_{r^*,1}$ to $v_{r^*,2}$ which does not contain the point $p(t)$. Therefore, the vertex v would be further away from $p(t)$ than $v_{r^*,1}$ and $v_{r^*,2}$, which is a contradiction. Consequently, we conclude

$$\|\operatorname{Ln}[\lg_Q](p)\| \geq \epsilon = \min\big\{1, \{\|\operatorname{Ln}(\operatorname{co}\{\operatorname{vrs}(\operatorname{CC}(Q)-v), \operatorname{vrs}(\operatorname{CC}(Q)-w)\})\| \mid v, w \in I(\operatorname{CC}(Q)), \operatorname{CC}(Q)-v \neq -(\operatorname{CC}(Q)-w)\}\big\} > 0 \quad \forall p \neq \operatorname{CC}(Q).$$

Resorting now to Proposition 2.8, we deduce that the circumcenter $\operatorname{CC}(Q)$ is attained in finite time. ∎

Remark 3.6. Note that if $0 \in \partial \lg_Q(\operatorname{CC}(Q)) \setminus \operatorname{int}(\partial \lg_Q(\operatorname{CC}(Q)))$, then generically convergence is achieved over an infinite time horizon.

PROPOSITION 3.7. *The flow* (3.6) *reaches the set* $\operatorname{IC}(Q)$ *in finite time.*

Proof. Let $p \notin \operatorname{IC}(Q)$. We know $\min \widetilde{\mathcal{L}}_{\operatorname{Ln}[\operatorname{sm}_Q]} \operatorname{sm}_Q(p) = \|\operatorname{Ln}[\operatorname{sm}_Q](p)\|^2$. Moreover, for all $p \notin \operatorname{IC}(Q)$, we have

$$\|\operatorname{Ln}[\operatorname{sm}_Q](p)\| \geq \epsilon = \min\big\{1, \{\|\operatorname{Ln}(\operatorname{co}\{n_e, n_f\})\| \mid e, f \in \operatorname{Ed}(Q), n_e \neq -n_f\}\big\} > 0.$$

Resorting to Proposition 2.8, we deduce the desired result. ∎

Figure 3.1 shows an example of the implementation of the gradient descent (3.5) and (3.6). Note that if the circumcenter $\operatorname{CC}(Q)$ (respectively, the incenter set $\operatorname{IC}(Q)$) is first computed offline, then the strategy of directly going toward it would converge in a less "erratic" way. Note also that the move-toward-the-center strategy is exponentially fast.

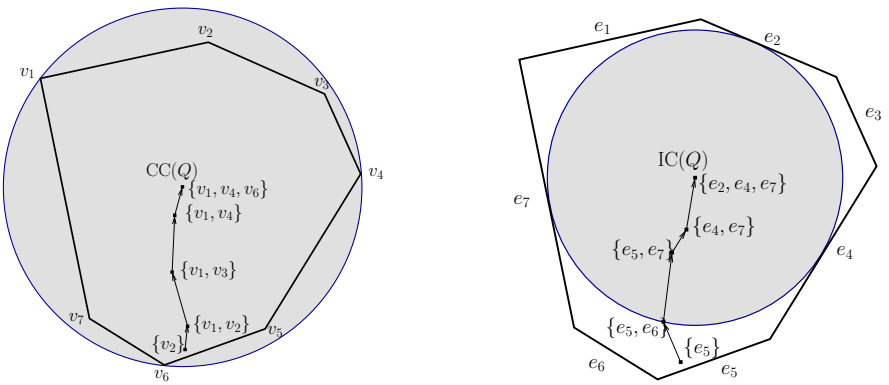


FIG. 3.1. *Illustration of the gradient descent of* $\lg_Q$ *and* sm_Q. *The points where the curve* $t \mapsto p(t)$ *fails to be differentiable correspond to points where there is a new vertex* v *of* Q *such that* $\|p(t) - v\| = \lg_Q(p(t))$ *(respectively, a new edge* e *of* Q *such that* $\operatorname{D}_e(p(t)) = \operatorname{sm}_Q(p(t))$*). The circumcenter and the incenter are attained in finite time according to Propositions* 3.5 *and* 3.7.

Finally, we conclude this section with four results useful for later developments.

LEMMA 3.8. *Let* $q \in Q$, *let* $v(q)$ *be one of the vertexes of* Q *which is furthest away from* q, *and let* $e(q)$ *be one of the edges of* Q *which is nearest to* q. *Then*

(i) $\operatorname{Ln}[\lg_Q](q) \cdot (q - v(q)) \geq 0$, *and the inequality is strict if* $q \neq \operatorname{CC}(Q)$,
(ii) $(q - \operatorname{CC}(Q)) \cdot (q - v(q)) \geq \|q - \operatorname{CC}(Q)\|^2/2$,
(iii) $\operatorname{Ln}[\operatorname{sm}_Q](q) \cdot n_e \geq 0$, *and the inequality is strict if* $q \notin \operatorname{IC}(Q)$, *and*
(iv) $(x - q) \cdot n_e \geq \operatorname{IR}(Q) - \operatorname{D}_e(q) \geq 0$ *for any* $x \in \operatorname{IC}(Q)$, *and the second inequality is strict if* $q \notin \operatorname{IC}(Q)$.

Proof. Let q be a point in Q. If $q = \mathrm{CC}(Q)$, claims (i) and (ii) are obviously satisfied since $\mathrm{Ln}[\mathrm{lg}_Q](q) = 0$. Assume then that $q \neq \mathrm{CC}(Q)$.

Let us prove (i) first. By Proposition 3.3, $0 \notin \partial\, \mathrm{lg}_Q(q)$, and hence $\mathrm{Ln}[\mathrm{lg}_Q](q) \neq 0$. Let us prove $\mathrm{Ln}[\mathrm{lg}_Q](q) \cdot (q - v(q)) > 0$ reasoning by contradiction. If $\mathrm{Ln}[\mathrm{lg}_Q](q) \cdot (q - v(q)) \leq 0$, then $d/dt \left(\| q - t\, \mathrm{Ln}[\mathrm{lg}_Q](q) - v \| \right)_{t=0} = \mathrm{vrs}(q - v) \cdot (-\mathrm{Ln}[\mathrm{lg}_Q](q)) \geq 0$, which implies that $\| q - t\, \mathrm{Ln}[\mathrm{lg}_Q](q) - v \| \geq \| q - v \| = \mathrm{lg}_Q(q)$ for $t > 0$ small enough. On the other hand, invoking Theorem 2.4 we have that $\mathrm{lg}_Q(q) - t \| \mathrm{Ln}[\mathrm{lg}_Q](q) \|^2 / 2 \geq \mathrm{lg}_Q(q - t\, \mathrm{Ln}[\mathrm{lg}_Q](q)) \geq \| q - t\, \mathrm{Ln}[\mathrm{lg}_Q](q) - v \|$. Gathering both facts, we conclude $-t \| \mathrm{Ln}[\mathrm{lg}_Q](q) \|^2 / 2 \geq 0$, which is a contradiction.

Let us now prove (ii). Since $q \neq \mathrm{CC}(Q)$, we have $\| q - v(q) \| > \mathrm{CR}(Q)$. Consider then a ball $\overline{B}_2(v(q), \| q - v(q) \|)$ centered at the vertex $v(q)$ with radius $\| q - v(q) \|$. By definition of the circumcenter, $\mathrm{CC}(Q)$ must lie in the interior of $\overline{B}_2(v(q), \| q - v(q) \|)$. Consequently, $\| \mathrm{CC}(Q) - v(q) \| < \| q - v(q) \|$. Then, from $\| \mathrm{CC}(Q) - v(q) \|^2 = \| \mathrm{CC}(Q) - q \|^2 + \| q - v(q) \|^2 - 2(q - \mathrm{CC}(Q)) \cdot (q - v(q))$, we deduce

$$2(q - \mathrm{CC}(Q)) \cdot (q - v(q)) - \| \mathrm{CC}(Q) - q \|^2 = \| q - v(q) \|^2 - \| \mathrm{CC}(Q) - v(q) \|^2 > 0\,,$$

which implies the desired result.

Let us now prove (iii). If $q \in \mathrm{IC}(Q)$, the claim is obviously satisfied since $\mathrm{Ln}[\mathrm{sm}_Q](q) = 0$. Assume then that $q \notin \mathrm{IC}(Q)$. By Proposition 3.3, $0 \notin \partial\, \mathrm{sm}_Q(q)$, and hence $\mathrm{Ln}[\mathrm{sm}_Q](q) \neq 0$. Let us prove $\mathrm{Ln}[\mathrm{sm}_Q](q) \cdot n_e > 0$ reasoning by contradiction. If $\mathrm{Ln}[\mathrm{sm}_Q](q) \cdot n_e \leq 0$, then $d/dt \left(\mathrm{D}_e(q + t\, \mathrm{Ln}[\mathrm{sm}_Q](q)) \right)_{t=0} = n_e \cdot \mathrm{Ln}[\mathrm{sm}_Q](q) \leq 0$, which implies that $\mathrm{D}_e(q + t\, \mathrm{Ln}[\mathrm{sm}_Q](q)) \leq \mathrm{D}_e(q) = \mathrm{sm}_Q(q)$ for $t > 0$ small enough. On the other hand, invoking Theorem 2.4 for the function $-\mathrm{sm}_Q$, we have that $\mathrm{sm}_Q(q) + t \| \mathrm{Ln}[\mathrm{sm}_Q](q) \|^2 / 2 \leq \mathrm{sm}_Q(q + t\, \mathrm{Ln}[\mathrm{sm}_Q](q)) \leq \mathrm{D}_e(q + t\, \mathrm{Ln}[\mathrm{sm}_Q](q))$. Gathering both facts, we conclude $t \| \mathrm{Ln}[\mathrm{sm}_Q](q) \|^2 / 2 \leq 0$, which is a contradiction.

Let us now prove (iv). By definition, $\mathrm{D}_e(q) \leq \mathrm{IR}(Q)$. This inequality is strict if $q \notin \mathrm{IC}(Q)$. Let $x \in \mathrm{IC}(Q)$. If we take a point O in the edge e, then the function D_e can be expressed as $\mathrm{D}_e(p) = (p - O) \cdot n_e$. Then, we have

$$\mathrm{D}_e(x) = (x - O) \cdot n_e = (x - q) \cdot n_e + (q - O) \cdot n_e = (x - q) \cdot n_e + \mathrm{D}_e(q)\,.$$

Since $\mathrm{D}_e(x) \geq \mathrm{sm}_Q(x) = \mathrm{IR}(Q)$, we conclude that $(x - q) \cdot n_e \geq \mathrm{IR}(Q) - \mathrm{D}_e(q) \geq 0$, and that the second inequality is strict if $q \notin \mathrm{IC}(Q)$. ∎

4. Analysis of the multicenter functions. Here we study the locational optimization functions $\mathcal{H}_{\mathrm{DC}}$ and $\mathcal{H}_{\mathrm{SP}}$ for the disk-covering and sphere-packing problems. We characterize their smoothness properties, generalized gradients, and critical points for arbitrary numbers of generators.

4.1. Smoothness and generalized gradients. We start by providing some alternative expressions and useful quantities. We write

$$\mathcal{H}_{\mathrm{DC}}(P) = \max_{i \in \{1,\dots,n\}} G_i(P)\,, \quad \mathcal{H}_{\mathrm{SP}}(P) = \min_{i \in \{1,\dots,n\}} F_i(P)\,,$$

where

$$G_i(P) = \max_{q \in V_i(P)} \| q - p_i \|\,, \quad F_i(P) = \min_{q \notin \mathrm{int}(V_i(P))} \| q - p_i \|\,.$$

Note that $G_i(P) = \mathrm{lg}_{V_i(P)}(p_i)$ and $F_i(P) = \mathrm{sm}_{V_i(P)}(p_i)$, where, for $i \in \{1, \dots, n\}$,

$$\mathrm{lg}_{V_i} : V_i \to \mathbb{R}\,, \quad \mathrm{sm}_{V_i} : V_i \to \mathbb{R}\,.$$

Proposition 3.3 provides an explicit expression for the generalized gradients of lg_{V_i} and sm_{V_i} when the Voronoi cell V_i is held fixed. Despite the slight abuse of notation, it is convenient to let $\partial \lg_{V_i(P)}(p_i)$ denote $\partial \lg_V(p_i)_{|V=V_i(P)}$ and let $\partial \mathrm{sm}_{V_i(P)}(p_i)$ denote $\partial \mathrm{sm}_V(p_i)_{|V=V_i(P)}$.

In contrast to this analysis at fixed Voronoi partition, the properties of the functions G_i and F_i are strongly affected by the dependence on the Voronoi partition $\mathcal{V}(P)$. We endeavor to characterize these properties in order to study $\mathcal{H}_{\mathrm{DC}}$ and $\mathcal{H}_{\mathrm{SP}}$.

PROPOSITION 4.1. *The functions $G_i, -F_i : Q^n \to \mathbb{R}$ are locally Lipschitz and regular. As a consequence, the locational optimization functions $\mathcal{H}_{DC}, -\mathcal{H}_{SP} : Q^n \to \mathbb{R}$ are locally Lipschitz and regular.*

Proof. (a) *G_i is locally Lipschitz and regular.* The definition of the function G_i admits the following alternative expression:

$$G_i(P) = \max_{v \in \mathrm{Ve}(V_i)} \|p_i - v\| . \tag{4.1}$$

Let P_0 be nondegenerate at the ith generator. Then there exists a neighborhood U of P_0 where the set $\mathcal{N}(i)$ does not change. Let $\{v_1, \ldots, v_{M_1}\}$, $\{w_1, \ldots, w_{M_2}\}$, $\{z_1, \ldots, z_{M_3}\}$ be the vertexes of V_i of types (a), (b), and (c), respectively. Then G_i can be locally written as

$$G_i(P) = \max \left\{ \max_{\ell \in \{1,\ldots,M_1\}} \|v_\ell - p_i\|, \max_{\ell \in \{1,\ldots,M_2\}} \|w_\ell - p_i\|, \max_{\ell \in \{1,\ldots,M_3\}} \|z_\ell - p_i\| \right\}$$

for all $P \in U$. Therefore, G_i restricted to U coincides with the function $\mathcal{G}_{\mathcal{N}(i)} : Q^n \to \mathbb{R}$ defined by

$$\mathcal{G}_{\mathcal{N}(i)}(P) = \max \left\{ \max_{\ell \in \{1,\ldots,M_1\}} \|v_\ell - p_i\|, \max_{\ell \in \{1,\ldots,M_2\}} \|w_\ell - p_i\|, \max_{\ell \in \{1,\ldots,M_3\}} \|z_\ell - p_i\| \right\}. \tag{4.2}$$

The function $\mathcal{G}_{\mathcal{N}(i)}$ is the maximum of a fixed finite set of locally Lipschitz and regular functions and, consequently, locally Lipschitz and regular by Proposition 2.2. We conclude that G_i is both locally Lipschitz and regular at P_0.

Let P_0 be degenerate at the ith generator. Then in any neighborhood U of P_0 there are different sets of neighbors of the ith generator. Indeed, because the number of generators, edges of the boundary Q, and vertexes of Q is finite, there is only a finite number of different sets of neighbors of the ith generator over U, say $\mathcal{N}^1(i), \ldots, \mathcal{N}^L(i)$. This implies that G_i admits the alternative expression $G_i(P) = \min\left\{\mathcal{G}_{\mathcal{N}^1(i)}(P), \ldots, \mathcal{G}_{\mathcal{N}^L(i)}(P)\right\}$ over U. From this expression, one can conclude that G_i is both locally Lipschitz and regular at P_0.

(b) *$-F_i$ is locally Lipschitz and regular.* From the definition of F_i, it is clear that its value at a configuration P is attained at the boundary of the Voronoi region V_i. Therefore, one only minimizes among the edges associated with the Voronoi neighbors $\mathcal{N}(i)$ and the edges of Q with nonempty intersection with V_i. Moreover, one can also see that the minimum must be attained at a point of the form $\mathrm{proj}_e(p_i)$, for some edge e of V_i. Now, consider the function $\mathcal{F}_i : Q^n \to \mathbb{R}$ defined by

$$\mathcal{F}_i(P) = \min \left\{ \min_{j \in \{1,\ldots,n\}} \|p_i - \frac{p_i + p_j}{2}\| , \min_{e \in \mathrm{Ed}(Q)} \mathrm{D}_e(p_i) \right\}. \tag{4.3}$$

We shall prove that $\mathcal{F}_i$ coincides with F_i. Clearly, $\mathcal{F}_i(P) \leq F_i(P)$. If $k \notin \mathcal{N}(i)$, then $(p_i + p_k)/2 \notin V_i$. Since $Q \setminus V_i$ is open, there exists a neighborhood of $(p_i + p_k)/2$ such

that $U \subset Q/V_i$. Therefore,

$$\|p_i - \frac{p_i + p_k}{2}\| > \min_{q \in U} \|p_i - q\| \geq \min_{q \notin \text{int}(V_i)} \|p_i - q\| = F_i(P)\,.$$

If an edge e of Q does not intersect V_i, then $\text{proj}_e(p_i) \notin V_i$. Using again the fact that $Q \backslash V_i$ is open, there exists a neighborhood U of $\text{proj}_e(p_i)$ in $\mathbb{R}^2$ such that $U \cap Q \subset Q \backslash V_i$. Then

$$\text{D}_e(p_i) = \|p_i - \text{proj}_e(p_i)\| > \min_{q \in U \cap Q} \|p_i - q\| \geq F_i(P)\,.$$

As a consequence of the previous inequalities, F_i equals $\mathcal{F}_{\mathcal{N}(i)}$. Resorting now to Proposition 2.2, we conclude that $-F_i$ is locally Lipschitz and regular. □

Next, one can actually prove the following stronger result.

PROPOSITION 4.2. *The locational optimization functions $\mathcal{H}_{DC}, \mathcal{H}_{SP} : Q^n \to \mathbb{R}$ are globally Lipschitz, with Lipschitz constant equal to* 1.

Proof. (a) *$\mathcal{H}_{DC}$ is globally Lipschitz.* Let P, P' be two configurations of the n generators. Without loss of generality, assume that $\mathcal{H}_{\text{DC}}(P) \leq \mathcal{H}_{\text{DC}}(P')$. Let i, j and q_0, $q_0' \in Q$ be such that $\mathcal{H}_{\text{DC}}(P) = G_i(P) = \|q_0 - p_i\|$ and $\mathcal{H}_{\text{DC}}(P') = G_j(P') = \|q_0' - p_j'\|$. Now consider the set $B_2(q_0', G_i(P))$. Then there exists a k such that $p_k \in \overline{B}_2(q_0', G_i(P))$ (otherwise, $\|q_0' - p_l\| > G_i(P)$, which contradicts the definition of the function $\mathcal{H}_{\text{DC}}$). On the other hand, we necessarily have that $p_k' \notin B_2(q_0', G_j(P'))$, since otherwise $\|q_0' - p_k'\| < \|q_0' - p_j'\|$, which implies that $q_0' \notin V_j'$, a contradiction. Finally, we apply the triangle inequality to obtain $\|q_0' - p_k'\| \leq \|q_0' - p_k\| + \|p_k - p_k'\|$. Gathering the previous facts, we have

$$\begin{aligned} |\mathcal{H}_{\text{DC}}(P') - \mathcal{H}_{\text{DC}}(P)| &= G_j(P') - G_i(P) \\ &\leq \|q_0' - p_k'\| - \|q_0' - p_k\| \leq \|p_k - p_k'\| \leq \|P - P'\|\,. \end{aligned}$$

(b) *$\mathcal{H}_{SP}$ is globally Lipschitz.* Let P, P' be two configurations of the n generators. Without loss of generality, assume that $\mathcal{H}_{\text{SP}}(P) \leq \mathcal{H}_{\text{SP}}(P')$. Let i, j and q_0, $q_0' \in Q$ be such that $\mathcal{H}_{\text{SP}}(P) = F_i(P) = \|q_0 - p_i\|$ and $\mathcal{H}_{\text{SP}}(P') = F_j(P') = \|q_0' - p_j'\|$. We treat separately the following two cases: (i) q_0 does not belong to the boundary of Q, and (ii) q_0 belongs to the boundary of Q. In case (i), it necessarily exists $k \in \mathcal{N}(i)$ such that $\|q_0 - p_i\| = \|q_0 - p_k\|$. If $\|q_0 - p_i'\| \geq F_j(P')$, then

$$\begin{aligned} |\mathcal{H}_{\text{SP}}(P') - \mathcal{H}_{\text{SP}}(P)| = F_j(P') - F_i(P) &\leq \|q_0 - p_i'\| - \|q_0 - p_i\| \\ &\leq \|p_i - p_i'\| \leq \|P - P'\|\,. \end{aligned} \tag{4.4}$$

If, on the contrary, $\|q_0 - p_i'\| < F_j(P')$, then $q_0 \in \text{int}(V_i')$. Therefore, $\|q_0 - p_k'\| \geq F_k(P') \geq F_j(P')$. Now we perform the same computation as in (4.4) to conclude $|\mathcal{H}_{\text{SP}}(P') - \mathcal{H}_{\text{SP}}(P)| \leq \|P - P'\|$.

In case (ii), we prove that $\|q_0 - p_i'\| \geq F_j(P')$. Suppose this is not true, i.e., $\|q_0 - p_i'\| < F_j(P')$. Let $m = q_0 + \epsilon(q_0 - p_i')$, with sufficiently small $\epsilon > 0$ such that $\|m - p_i'\| < F_j(P')$. Clearly $m \notin Q$. On the other hand, by definition $B_2(p_i', F_i(P')) \subset V_i'$. Now we have

$$B_2(p_i', F_j(P')) \subset B_2(p_i', F_i(P')) \subset V_i' \subset Q\,.$$

But since $\|m - p_i'\| < F_j(P')$, then $m \in B_2(p_i', F_j(P')) \subset Q$, which is a contradiction. Therefore, $\|q_0 - p_i'\| \geq F_j(P')$, and now the same argument as in (4.4) guarantees that $|\mathcal{H}_{\text{SP}}(P') - \mathcal{H}_{\text{SP}}(P)| \leq \|P - P'\|$. □

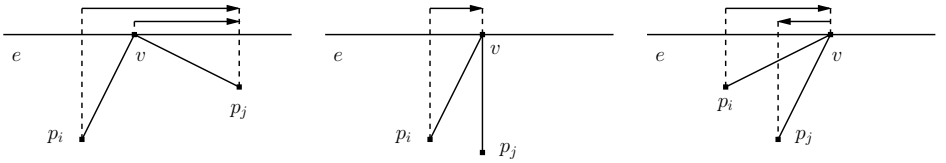


FIG. 4.1. *To illustrate* (4.5) *we draw the vectors* $\mathrm{proj}_e(p_j - v(e,i,j))$ *and* $\mathrm{proj}_e(p_j - p_i)$ *for various locations of* p_i, p_j, *and* e. *The left, center, and right figures correspond to* $\lambda(e,i,j) > 0$, $\lambda(e,i,j) = 0$, $\lambda(e,i,j) < 0$, *respectively.*

We now introduce some quantities that are useful in characterizing the generalized gradient of the functions G_i. Given a vertex of type (b), $v = v(e,i,j)$, determined by the edge e and two generators p_i and p_j, we consider the scalar function $\lambda(e,i,j)$ defined by

$$\mathrm{proj}_e(p_j - v(e,i,j)) = \lambda(e,i,j)\ \mathrm{proj}_e(p_j - p_i)\,, \tag{4.5}$$

where recall that proj_e denotes the orthogonal projection onto the edge e; see Figure 4.1. One can see that $\lambda(e,i,j) + \lambda(e,j,i) = 1$. If e is a segment in the line $ax + by + c = 0$, $(\Delta x_{ij}, \Delta y_{ij}) = p_j - p_i$, $(x_m, y_m) = (p_i + p_j)/2$, then one can show

$$\lambda(e,i,j) = \frac{1}{2} - \frac{(a\Delta x_{ij} + b\Delta y_{ij})(ax_m + by_m + c)}{(a\Delta y_{ij} - b\Delta x_{ij})^2}\,. \tag{4.6}$$

Given a vertex of type (a), $v = v(i,j,k)$, determined by the three generators p_i, p_j, and p_k, we consider the scalar function $\mu(i,j,k)$ defined by

$$\mathrm{proj}_{e_{jk}}(p_\ell - v(i,j,k)) = \mu(i,j,k)\ \mathrm{proj}_{e_{jk}}(p_\ell - p_i)\,, \tag{4.7}$$

where e_{jk} is the bisector of p_j and p_k and where $p_\ell = p_j$ if p_j belongs to the half-plane defined by e_{jk} containing p_i, and $p_\ell = p_k$ otherwise. One can see that $\mu(i,j,k) = \mu(i,k,j)$ and that $\mu(i,j,k) + \mu(j,k,i) + \mu(k,i,j) = 1$. From the expression for λ, one can obtain

$$\mu(i,j,k) = \frac{1}{2} + \frac{(\Delta x_{ij}\Delta x_{jk} + \Delta y_{ij}\Delta y_{jk})(\Delta x_{ik}\Delta x_{jk} + \Delta y_{ik}\Delta y_{jk})}{2(x_k\Delta y_{ij} - x_j\Delta y_{ik} + x_i\Delta y_{jk})^2}. \tag{4.8}$$

Note that, in general, λ and μ are not positive functions. Now we are ready to describe in detail the structure of the generalized gradient of the functions G_i, F_i.

PROPOSITION 4.3. *The generalized gradient of* $G_i : Q^n \to \mathbb{R}$ *at* $P \in Q^n$ *is*

$$\partial G_i(P) = \mathrm{co}\{\partial_v G_i(P) \in (\mathbb{R}^2)^n \mid v \in \mathrm{Ve}(V_i(P)) \text{ such that } G_i(P) = \|p_i - v\|\},$$

where we consider separately the following cases. If $v = v(i,j,k)$ *is a nondegenerate vertex of type* (a), *then*

$$\begin{aligned}&\partial_{v(i,j,k)} G_i(P) = \partial_{v(k,i,j)} G_k(P) = \partial_{v(j,k,i)} G_j(P)\\ &= (0,\dots,\underbrace{\mu(i,j,k)\,\mathrm{vrs}(p_i - v)}_{i\text{th place}},\dots,\underbrace{\mu(j,k,i)\,\mathrm{vrs}(p_j - v)}_{j\text{th place}},\dots,\underbrace{\mu(k,i,j)\,\mathrm{vrs}(p_k - v)}_{k\text{th place}},\dots,0),\end{aligned}$$

where, without loss of generality, we let $i < j < k$. If $v = v(e,i,j)$ is a nondegenerate vertex of type (b)*, then*

$$\partial_{v(e,i,j)} G_i(P) = \partial_{v(e,j,i)} G_j(P) \\ = (0, \dots, \underbrace{\lambda(e,i,j)\operatorname{vrs}(p_i - v)}_{ith\ place}, \dots, \underbrace{\lambda(e,j,i)\operatorname{vrs}(p_j - v)}_{jth\ place}, \dots, 0),$$

where, without loss of generality, we let $i < j$. If $v = v(e,f,i)$ is a nondegenerate vertex of type (c), then

$$\partial_{v(e,f,i)} G_i(P) = (0, \dots, 0, \underbrace{\operatorname{vrs}(p_i - v)}_{ith\ place}, 0, \dots, 0).$$

Finally, if the vertex v is degenerate, i.e., if v is determined by $d > 3$ elements (generators or edges), then there are $\binom{d-1}{2}$ pairs of elements which determine the vertex v together with the generator p_i. In this case, $\partial_v G_i(P)$ is the convex hull of $\partial_{v(\alpha,\beta,\gamma)} G_i(P)$ for all $\binom{d-1}{2}$ such triplets (α,β,γ).

Note that, at all nondegenerate configurations P, the quantity $\partial_v G_i(P)$ is the generalized gradient of the function $(p_1, \dots, p_n) \mapsto \|p_i - v(i,j,k)\|$; however, this interpretation cannot be given when P is degenerate.

Proof. We present the proof for the expression for $\partial G_i(P)$. Let us consider first the case when P is a nondegenerate configuration for the ith generator. According to the proof of Proposition 4.1, G_i coincides with the function $\mathcal{G}_{\mathcal{N}(i)}$ over a neighborhood U of P. Hence, $\partial G_i(P) = \partial \mathcal{G}_{\mathcal{N}(i)}(P)$ which, according to (4.2) and Proposition 2.2, takes the form

$$\operatorname{co}\left\{ \frac{\partial}{\partial P} \|v - p_i\| \mid v \in \operatorname{Ve}(V_i(P)) \text{ such that } \|v - p_i\| = G_i(P) \right\}.$$

If $v = v(i,j,k)$ is a nondegenerate vertex of type (a), then one can compute

$$\frac{\partial}{\partial p_i} \|p_i - v(i,j,k)\| = \operatorname{vrs}(p_i - v)\left(I_2 - \frac{\partial v}{\partial p_i}\right) = \mu(i,j,k)\operatorname{vrs}(p_i - v),$$
$$\frac{\partial}{\partial p_j} \|p_i - v(i,j,k)\| = -\operatorname{vrs}(p_i - v)\left(\frac{\partial v}{\partial p_j}\right) = \mu(j,k,i)\operatorname{vrs}(p_j - v),$$
$$\frac{\partial}{\partial p_\ell} \|p_i - v(i,j,k)\| = 0, \qquad \ell \neq i,j,k,$$

where in the first and second chain of equalities we have used the expression of μ given in (4.8). If $v = v(e,i,j)$ is a nondegenerate vertex of type (b), then one can compute

$$\frac{\partial}{\partial p_i} \|p_i - v(e,i,j)\| = \operatorname{vrs}(p_i - v)\left(I_2 - \frac{\partial v}{\partial p_i}\right) = \lambda(e,i,j)\operatorname{vrs}(p_i - v),$$
$$\frac{\partial}{\partial p_j} \|p_i - v(e,i,j)\| = -\operatorname{vrs}(p_i - v)\left(\frac{\partial v}{\partial p_j}\right) = \lambda(e,j,i)\operatorname{vrs}(p_j - v),$$
$$\frac{\partial}{\partial p_\ell} \|p_i - v(e,i,j)\| = 0, \qquad \ell \neq i,j,$$

where in the first and second chain of equalities we have used the expression of λ given in (4.6). If $v = v(e, f, i)$ is a nondegenerate vertex of type (c), then

$$\frac{\partial}{\partial p_i}\|p_i - v(e,f,i)\| = \mathrm{vrs}(p_i - v)\,,$$
$$\frac{\partial}{\partial p_\ell}\|p_i - v(e,f,i)\| = 0\,, \qquad \ell \neq i\,.$$

If P is a degenerate configuration at the ith generator, then, following the proof of Proposition 4.1, the generalized gradient of G_i can be expressed as the convex hull of the generalized gradients of each of the functions $\mathcal{G}_{\mathcal{N}^1(i)}, \ldots, \mathcal{G}_{\mathcal{N}^L(i)}$. The claim now follows by reproducing the previous discussion for the generalized gradients of each of the functions $\mathcal{G}_{\mathcal{N}^\ell(i)}$, $\ell \in \{1, \ldots, L\}$. □

The expression for $\partial F_i(P)$ can be deduced in an analogous (and simpler) way, since according to the proof of Proposition 4.1, it is not necessary to establish any distinction between the degenerate and the nondegenerate configurations. Accordingly, we state the following result without proof.

PROPOSITION 4.4. *The generalized gradient of $F_i : Q^n \to \mathbb{R}$ at $P \in Q^n$ is*

$$\partial F_i(P) = \mathrm{co}\{\partial_e F_i(P) \in (\mathbb{R}^2)^n \mid e \in \mathrm{Ed}(V_i(P)) \text{ such that } F_i(P) = \mathrm{D}_e(p_i)\}$$

where, if $e = e(i,j)$ is an edge of type (a)*, then*

$$\partial_{e(i,j)} F_i(P) = \partial_{e(j,i)} F_j(P) = \frac{1}{2}(0, \ldots, \underbrace{n_{e(i,j)}}_{i\text{th place}}, \ldots, \underbrace{-n_{e(i,j)}}_{j\text{th place}}, \ldots, 0),$$

and if $e = e(i)$ is an edge of type (b)*, then*

$$\partial_{e(i)} F_i(P) = (0, \ldots, \underbrace{n_{e(i)}}_{i\text{th place}}, \ldots, 0).$$

Next, we give conditions under which the functions λ and μ take positive values.

LEMMA 4.5. *Let $P \in Q^n$ and let $v \in \mathrm{Ve}_{\mathrm{DC}}(\mathcal{V}(P))$. Then*

(i) *if v belongs to an edge e of Q, then there exist generators p_i and p_j such that $\lambda(e,i,j)$ and $\lambda(e,j,i)$ are positive, and*

(ii) *if v belongs to $\mathrm{int}(Q)$, then there exist generators p_i, p_j, and p_k such that $\mu(i,j,k)$, $\mu(j,k,i)$, and $\mu(k,i,j)$ are positive.*

Proof. Consider first the case when v is nondegenerate. If v is in the edge e of Q (i.e., v is of type (b)), let p_i and p_j be the two generators determining it. From the definition of λ, one sees that the values $\lambda(e,i,j) = 0$ and $\lambda(e,j,i) = 0$ correspond to, respectively, p_j and p_i lying on the orthogonal line to e passing through $v(e,i,j)$. If $\lambda(e,i,j) \leq 0$, then there exists $w \in e \cap V_j$ such that $\|p_j - w\| > \|p_j - v\| = \mathcal{H}_{\mathrm{DC}}(P)$, which is a contradiction. Therefore, $\lambda(e,i,j) > 0$. The same argument guarantees $\lambda(e,j,i) > 0$. If v is of type (a) and p_i, p_j, and p_k are the elements determining it, a similar argument leads to the conclusion that $\mu(i,j,k)$, $\mu(j,k,i)$, and $\mu(k,i,j)$ are positive.

Consider the case when v is degenerate. Let $\{i_1, \ldots, i_m\}$ be such that $v \in V_{i_\ell}$, $\ell \in \{1, \ldots, m\}$. Assume v is in an edge e of Q. Let l denote the orthogonal line to the edge e passing through v. We claim that there must exist generators i, j in $\{i_1, \ldots, i_m\}$ on both sides of l (and, therefore, the values of the corresponding

$\lambda(e,i,j)$ and $\lambda(e,j,i)$ are positive; cf. Figure 4.1). Assume this is not the case, i.e., $\{p_{i_1},\dots,p_{i_m}\}$ are contained in one of the closed half-planes defined by l, say l_-. Take $w \in l_+ \cap e$ arbitrarily close to v. Since $\{p_{i_1},\dots,p_{i_m}\} \subset l_-$, we have $\|p_{i_\ell} - w\| > \|p_{i_\ell} - v\|$ for all $\ell \in \{1,\dots,m\}$. On the other hand, since no generator outside the set $\{p_{i_1},\dots,p_{i_m}\}$ is involved in the definition of v, there must exist ℓ^* such that $w \in V_{i_{\ell^*}}$. Therefore, $G_{i_{\ell^*}}(P) \geq \|p_{i_{\ell^*}} - w\| > \|p_{i_{\ell^*}} - v\| = \mathcal{H}_{\mathrm{DC}}(P)$, which is a contradiction. Assume now that $v \in \mathrm{int}(Q)$. Our claim is that for any line l passing through v, there must exist generators on both sides of l (by (4.7), this would imply (ii)). If this is not the case, i.e., $\{p_{i_1},\dots,p_{i_m}\} \subset l_-$, then take $w \in B_2(v,\epsilon) \cap l_+ \cap o$, where o denotes the orthogonal line to l passing through v. As before, $w \in V_{i_{\ell^*}}$ for some ℓ^* and $\|p_{i_{\ell^*}} - w\| > \|p_{i_{\ell^*}} - v\|$, which yields a contradiction. □

This completes our analysis of the generalized gradients of G_i and F_i and, with these results, we return to studying the generalized gradients of $\mathcal{H}_{\mathrm{DC}}$ and $\mathcal{H}_{\mathrm{SP}}$. An immediate consequence of Propositions 2.2 and 4.1 is that

$$\begin{aligned} \partial\mathcal{H}_{\mathrm{DC}}(P) &= \mathrm{co}\{\partial G_i(P) \mid i \in I(P)\}, \\ \partial\mathcal{H}_{\mathrm{SP}}(P) &= \mathrm{co}\{\partial F_i(P) \mid i \in I(P)\}. \end{aligned} \tag{4.9}$$

Furthermore, we can provide the following more detailed characterization.

PROPOSITION 4.6. *Let $P \in Q^n$. For each $i \in \{1,\dots,n\}$, the image by π_i of the generalized gradients of $\mathcal{H}_{DC}$ and $\mathcal{H}_{SP}$ at P is given by*

$$\pi_i(\partial\mathcal{H}_{DC}(P)) = \begin{cases} \pi_i(\partial G_i(P)) & \text{if } i \in I(P),\ \mathrm{Ve}_{\mathrm{DC}}(\mathcal{V}(P)) \subset \mathrm{Ve}(V_i(P)), \\ \mathrm{co}\{\pi_i(\partial G_i(P)), 0\} & \text{if } i \in I(P),\ \mathrm{Ve}_{\mathrm{DC}}(\mathcal{V}(P)) \not\subset \mathrm{Ve}(V_i(P)), \\ 0 & \text{if } i \notin I(P); \end{cases}$$

$$\pi_i(\partial\mathcal{H}_{SP}(P)) = \begin{cases} \pi_i(\partial F_i(P)) & \text{if } i \in I(P),\ \mathrm{Ed}_{\mathrm{SP}}(\mathcal{V}(P)) \subset \mathrm{Ed}(V_i(P)), \\ \mathrm{co}\{\pi_i(\partial F_i(P)), 0\} & \text{if } i \in I(P),\ \mathrm{Ed}_{\mathrm{SP}}(\mathcal{V}(P)) \not\subset \mathrm{Ed}(V_i(P)), \\ 0 & \text{if } i \notin I(P). \end{cases}$$

Proof. From (4.9), if $i \notin I(P)$, then $\pi_i(\partial\mathcal{H}_{\mathrm{DC}}(P)) = 0$, $\pi_i(\partial\mathcal{H}_{\mathrm{SP}}(P)) = 0$. If $i \in I(P)$, then using Proposition 4.3 we deduce that the generators p_j such that ∂G_j has a nonzero entry in the ith place (and hence contributes to the projection by π_i of $\partial\mathcal{H}_{\mathrm{DC}}$) must share a vertex with the ith generator. Analogously, if $i \in I(P)$, then using Proposition 4.4 we deduce that the generators p_j such that ∂F_j has a nonzero entry in the ith place (and hence contributes to the projection by π_i of $\partial\mathcal{H}_{\mathrm{SP}}$) must satisfy $j \in \mathcal{N}(i)$. For the disk-covering function, if v is a common vertex of V_i and V_j, determined by i, j, and a third element α, then $\partial_{v(\alpha,j,i)}G_j = \partial_{v(\alpha,i,j)}G_i$, and the expression for $\pi_i(\partial\mathcal{H}_{\mathrm{DC}}(P))$ then follows. The argument for the expression of $\pi_i(\partial\mathcal{H}_{\mathrm{SP}}(P))$ is analogous. □

4.2. Critical points. Having characterized the generalized gradients of $\mathcal{H}_{\mathrm{DC}}$ and $\mathcal{H}_{\mathrm{SP}}$, we now turn to studying their critical points.

THEOREM 4.7 (Minima of $\mathcal{H}_{\mathrm{DC}}$). *Let $P \in Q^n$ be a nondegenerate configuration and $0 \in \mathrm{int}(\partial\mathcal{H}_{DC}(P))$. Then P is a strict local minimum of $\mathcal{H}_{DC}$, all generators are active, and P is a circumcenter Voronoi configuration.*

Proof. Since P is nondegenerate, note from Proposition 4.3 that $\partial_v G_i$ is a singleton for each $v \in \mathrm{Ve}(V_i(P))$, $i \in \{1,\dots,n\}$. Let $w \in (\mathbb{R}^2)^n$. We claim that moving the configuration of the generators from P in the direction w can only increase the cost. The hypothesis $0 \in \mathrm{int}(\partial\mathcal{H}_{\mathrm{DC}}(P))$ implies by Lemma 2.1 that there exists i

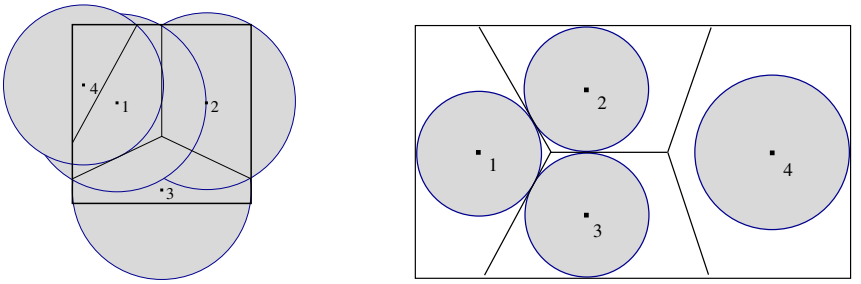


FIG. 4.2. *Local extrema of the disk-covering and the sphere-packing functions in a convex polygonal environment. The configuration on the left corresponds to a local minimum of* $\mathcal{H}_{DC}$ *with* $0 \in \partial\mathcal{H}_{DC}(P)$ *and* $\operatorname{int}(\partial\mathcal{H}_{DC}(P)) = \emptyset$. *The configuration on the right corresponds to a local maximum of* $\mathcal{H}_{SP}$ *with* $0 \in \partial\mathcal{H}_{SP}(P)$ *and* $\operatorname{int}(\partial\mathcal{H}_{SP}(P)) = \emptyset$. *In both configurations, the 4th generator is inactive and noncentered.*

and $v \in \operatorname{Ve}(V_i(P)) \cap \operatorname{Ve}_{\mathrm{DC}}(\mathcal{V}(P))$ such that $w \cdot \partial_v G_i(P) > 0$. Since P is nondegenerate, v will still belong to $V_i(P + \epsilon w)$ for sufficiently small $\epsilon > 0$, and consequently $\mathcal{H}_{\mathrm{DC}}(P + \epsilon w) \geq G_i(P + \epsilon w) > G_i(P) = \mathcal{H}_{\mathrm{DC}}(P)$. Therefore, P is a strict local minimum.

Since π_i is an open map, the set $\pi_i(\operatorname{int}(\partial\mathcal{H}_{\mathrm{DC}}(P)))$ is open for each $i \in \{1, \ldots, n\}$. Therefore, $\pi_i(\operatorname{int}(\partial\mathcal{H}_{\mathrm{DC}}(P))) \neq 0$, and hence all generators are active, i.e., $I(P) = \{1, \ldots, n\}$. Let us see that all generators must also be centered. Assume P is nondegenerate and consider the ith generator. Take $w \in \mathbb{R}^2$ and let $\overline{w} \in (\mathbb{R}^2)^n$ be the vector which has w in the ith place and 0 otherwise. By Lemma 2.1, there exist j and $v \in \operatorname{Ve}(V_j(P)) \cap \operatorname{Ve}_{\mathrm{DC}}(\mathcal{V}(P))$ such that $\overline{w} \cdot \partial_v G_j > 0$. Since $\overline{w} \cdot \partial_v G_j = w \cdot \pi_i(\partial_v G_j) > 0$, then necessarily $\pi_i(\partial_v G_j) \neq 0$, and therefore $v \in V_i(P)$ and $\pi_i(\partial_v G_j) = \pi_i(\partial_v G_i)$. The vertex v is determined by p_i, p_j and a third element, say α. Depending on whether α corresponds to an edge or to another generator, we have that $\pi_i(\partial_v G_i)$ is equal to $\lambda(\alpha, i, j) \operatorname{vrs}(p_i - v)$ or $\mu(\alpha, i, j) \operatorname{vrs}(p_i - v)$. In any case, from Lemma 4.5, we deduce that $\lambda(\alpha, i, j)$ (respectively, $\mu(\alpha, i, j)$) belongs to the interval $(0, 1)$. Therefore, $w \cdot \pi_i(\partial_v G_i) > 0$ implies $w \cdot \operatorname{vrs}(p_i - v) > 0$. Since $\operatorname{vrs}(p_i - v) \in \partial \lg_{V_i(P)}(p_i) = \partial \lg_V(p_i)_{|V=V_i(P)}$ (cf. (3.2)), we conclude from Lemma 2.1 that $0 \in \operatorname{int}(\partial \lg_{V_i(P)}(p_i))$. By Proposition 3.3, this implies that $p_i = \mathrm{CC}(V_i)$. Hence, P is a circumcenter Voronoi configuration. □

THEOREM 4.8 (Maxima of $\mathcal{H}_{\mathrm{SP}}$). *Let* $P \in Q^n$ *and* $0 \in \operatorname{int}(\partial\mathcal{H}_{SP}(P))$. *Then* P *is a strict local maximum of* $\mathcal{H}_{SP}$, *all generators are active, and* P *is a generic incenter Voronoi configuration.*

Proof. The proof of this result is analogous to the proof of Theorem 4.7. Note that $0 \in \operatorname{int}(\partial \operatorname{sm}_{V_i(P)}(p_i))$ implies, by Proposition 3.3, that $\mathrm{IC}(V_i(P)) = \{p_i\}$, and hence P is a generic incenter Voronoi configuration. □

Remark 4.9. Theorems 4.7 and 4.8 precisely provide the interpretation of the multicenter problems that we gave in section 2.2: since all generators are active, they share the same radius. If one drops the hypothesis that 0 belongs to the generalized gradient of the locational optimization function, then one can think of simple examples where P is a local minimum of $\mathcal{H}_{\mathrm{DC}}$ (respectively, local maximum of $\mathcal{H}_{\mathrm{SP}}$), and there are generators which are inactive and noncentered; see Figure 4.2.

5. Dynamical systems for the multicenter problems. In this section, we describe three algorithms that (locally) extremize the multicenter functions for the disk-covering and the sphere-packing problems. We first examine the gradient flow

descent associated with the locational optimization functions $\mathcal{H}_{\mathrm{DC}}$ and $\mathcal{H}_{\mathrm{SP}}$. This flow is guaranteed to find a local critical point, but it has the drawback of being centralized, as we describe later. Then we propose two decentralized flows for each problem. One roughly consists of a distributed implementation of the gradient descent. As we show, it is very much in the spirit of behavior-based robotics. The other one follows the logical strategy given the results in Theorems 4.7 and 4.8: each generator moves toward the circumcenter (alternatively, incenter set) of its own Voronoi polygon. We call them Lloyd flows, since they resemble the original Lloyd algorithm for vector quantization problems, where each quantizer moves toward the centroid or center of mass of its own Voronoi region, see [14, 16, 20]. We present continuous-time versions of the algorithms and discuss their convergence properties. In our setting, the generators' location obeys a first-order dynamical behavior described by

$$\dot{p}_i = u_i(p_1, \dots, p_n)\,, \quad i \in \{1, \dots, n\}\,. \tag{5.1}$$

The dynamical system (5.1) is said to be (strongly) *centralized* if there exists at least an $i \in \{1, \dots, n\}$ such that $u_i(p_1, \dots, p_n)$ cannot be written as a function of the form $u_i(p_i, p_{i_1}, \dots, p_{i_m})$, with $m < n-1$. The dynamical system (5.1) is said to be *Voronoi-distributed* if each $u_i(p_1, \dots, p_n)$ can be written as a function of the form $u_i(p_i, p_{i_1}, \dots, p_{i_m})$, with $i_k \in \mathcal{N}(P, i)$, $k \in \{1, \dots, m\}$. Finally, the dynamical system (5.1) is said to be *nearest-neighbor-distributed* if each $u_i(p_1, \dots, p_n)$ can be written as a function of the form $u_i(p_i, p_{i_1}, \dots, p_{i_m})$, with $\|p_i - p_{i_k}\| \leq \|p_i - p_j\|$ for all $j \in \{1, \dots, n\}$ and $k \in \{1, \dots, m\}$. A nearest-neighbor-distributed dynamical system is also Voronoi-distributed.

It is well known that there are at most $3n-6$ neighborhood relationships in a planar Voronoi diagram [23, section 2.3]. Therefore, the number of Voronoi neighbors of each site is on average less than or equal to 6. (Recall that sites are Voronoi-neighbors if they share an edge, not just a vertex.) We refer to [11] for more details on the distributed character of Voronoi neighborhood relationships.

Note that the set of indexes $\{i_1, \dots, i_m\}$ for a specific generator p_i of a Voronoi-distributed or a nearest-neighbor-distributed dynamical system is not the same for all possible configurations P. In other words, the identity of both the Voronoi neighbors and the nearest neighbors might change along the evolution; i.e., the topology of the dynamical system is *dynamic*.

5.1. Nonsmooth gradient dynamical systems. Consider the (signed) generalized gradient descent flow (2.6) for the locational optimization functions $\mathcal{H}_{\mathrm{DC}}$ and $\mathcal{H}_{\mathrm{SP}}$,

$$\dot{P} = -\operatorname{Ln}(\partial \mathcal{H}_{\mathrm{DC}})(P)\,, \quad \dot{P} = \operatorname{Ln}(\partial \mathcal{H}_{\mathrm{SP}})(P)\,.$$

Alternatively, we may write the following for each $i \in \{1, \dots, n\}$:

$$\dot{p}_i = -\pi_i(\operatorname{Ln}(\partial \mathcal{H}_{\mathrm{DC}})(p_1, \dots, p_n))\,, \tag{5.2}$$

$$\dot{p}_i = \pi_i(\operatorname{Ln}(\partial \mathcal{H}_{\mathrm{SP}})(p_1, \dots, p_n))\,. \tag{5.3}$$

As noted in section 2.4, these vector fields are discontinuous, and therefore we understand their solution in the Filippov sense. Equation (4.9) and Propositions 4.3 and 4.4 provide an expression of the generalized gradients at P, $\partial \mathcal{H}_{\mathrm{DC}}(P)$ and $\partial \mathcal{H}_{\mathrm{SP}}(P)$. One needs to first compute the generalized gradient, then compute the least-norm element, and finally project it to each of the n components; therefore, the expressions

in Proposition 4.6 are not helpful. Note that the least-norm element of convex sets can be computed efficiently, see [6], however closed-form expressions are not available in general.

One can see that the compact set Q^n is strongly invariant for both vector fields $-\operatorname{Ln}(\partial\mathcal{H}_{\mathrm{DC}})$ and $\operatorname{Ln}(\partial\mathcal{H}_{\mathrm{SP}})$. Indeed, the components for each generator of both vector fields point always toward Q. Regarding $-\operatorname{Ln}(\partial\mathcal{H}_{\mathrm{DC}})$, this is a consequence of Proposition 4.3 and of Lemma 4.5. Regarding $\operatorname{Ln}(\partial\mathcal{H}_{\mathrm{SP}})$, this is a consequence of Proposition 4.4.

PROPOSITION 5.1. *For the dynamical system* (5.2) *(respectively,* (5.3)*), the generators' location* $P = (p_1, \ldots, p_n)$ *converges asymptotically to the set of critical points of* $\mathcal{H}_{DC}$ *(respectively, of* $\mathcal{H}_{SP}$*).*

Proof. From Propositions 4.1 and 4.2, $\mathcal{H}_{\mathrm{DC}}$ and $-\mathcal{H}_{\mathrm{SP}}$ are globally Lipschitz and regular over Q^n. The result follows from Proposition 2.9 considering the dynamical system restricted to the strongly invariant and compact domain Q^n. □

Remark 5.2. The gradient dynamical systems enjoy convergence guarantees, but their implementation is centralized for two reasons. First, all functions $G_i(P)$ (respectively, $F_i(P)$) need to be compared in order to determine which generator is active. Second, the least-norm element of the generalized gradients depends on the relative position of the active generators with respect to each other and to the environment.

Remark 5.3. As illustrated in Figure 5.1 the evolution of the gradient dynamical systems may not leave fixed the generators that are already centers (circumcenters or incenters).

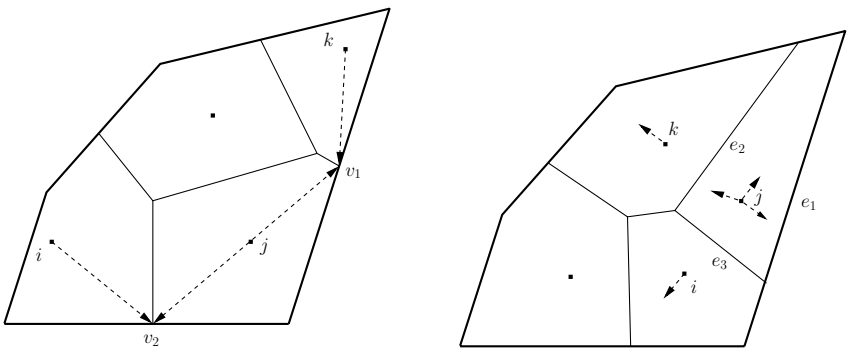


FIG. 5.1. *Illustration of the gradient descent. In the left figure, the only active vertexes at the given configuration are* v_1 *and* v_2*. Although the* j*th generator is in the circumcenter of its own Voronoi region, the control law* (5.2) *will drive it toward the vertex* v*. In the right figure, the only active edges at the given configuration are* e_1*,* e_2*, and* e_3*. Although the* j*th generator is in the incenter of its own Voronoi region, the control law* (5.3) *will drive it away from the edge* e_1*.*

5.2. Nonsmooth dynamical systems based on distributed gradients. In this section, we propose a distributed implementation of the previous gradient dynamical systems and explore its relation with behavior-based rules in multiple-vehicle coordination. Consider the following modifications of the gradient dynamical systems (5.2)–(5.3):

$$\dot{p}_i = -\operatorname{Ln}(\partial \lg_{V_i(P)})(P)\,, \tag{5.4}$$

$$\dot{p}_i = \operatorname{Ln}(\partial \operatorname{sm}_{V_i(P)})(P)\,, \tag{5.5}$$

for $i \in \{1,\dots,n\}$. Note that the system (5.4) is Voronoi-distributed, since $\mathrm{Ln}(\partial \lg_{V_i(P)})(P)$ is determined only by the position of p_i and of its Voronoi neighbors $\mathcal{N}(P,i)$. On the other hand, the system (5.5) is nearest-neighbor-distributed, since $\mathrm{Ln}(\partial \mathrm{sm}_{V_i(P)})(P)$ is determined only by the position of p_i and its nearest neighbors.

For future reference, let $\mathrm{Ln}(\partial \lg_{\mathcal{V}})(P) = (\mathrm{Ln}(\partial \lg_{V_1(P)})(P),\dots,\mathrm{Ln}(\partial \lg_{V_n(P)})(P))$, $\mathrm{Ln}(\partial \mathrm{sm}_{\mathcal{V}})(P) = (\mathrm{Ln}(\partial \mathrm{sm}_{V_1(P)})(P),\dots,\mathrm{Ln}(\partial \mathrm{sm}_{V_n(P)})(P))$, and write

$$\dot{P} = -\mathrm{Ln}(\partial \lg_{\mathcal{V}})(P)\,, \quad \dot{P} = \mathrm{Ln}(\partial \mathrm{sm}_{\mathcal{V}})(P)\,.$$

As for the previous dynamical systems, note that these vector fields are discontinuous, and therefore we understand their solutions in the Filippov sense. One can see that the compact set Q^n is strongly invariant for both vector fields $-\mathrm{Ln}(\partial \lg_{\mathcal{V}})$ and $\mathrm{Ln}(\partial \mathrm{sm}_{\mathcal{V}})$. This fact is a consequence of the expressions for the generalized gradients of lg and sm in Proposition 3.3. Note that in the 1-center case, (5.2) (respectively, (5.3)) coincides with (5.4) (respectively, with (5.5)).

PROPOSITION 5.4. *Let $P \in Q^n$. Then the solutions of the dynamical systems* (5.4) *and* (5.5) *starting at P are unique.*

Proof. (a) *Uniqueness of solution for* (5.4). Let $D_{\lg}$ be the set of $P \in Q^n$ such that P is nondegenerate and $\lg_{V_i(P)}(p_i)$ is attained at a single vertex for all i. Note that $Q^n \setminus D_{\lg}$ has measure zero, and that the vector field $-\mathrm{Ln}(\partial \lg_{\mathcal{V}})$ is differentiable (and hence locally Lipschitz) when restricted to any connected component of $D_{\lg}$. Let P, P' belong to different connected components of $D_{\lg}$, and let $\|P - P'\| \le \epsilon$. Consider all the indexes i at which the values of $\lg_{V_i(P)}(p_i)$ and $\lg_{V_i(P')}(p_i')$ are attained at different vertexes. For these indexes,

$$-\mathrm{Ln}(\partial \lg_{V_i(P)})(p_i) + \mathrm{Ln}(\partial \lg_{V_i(P')})(p_i') = \mathrm{vrs}(v - p_i) - \mathrm{vrs}(w' - p_i')$$

for certain vertexes v and w'. Note that for ϵ small enough, the vertex w' in the Voronoi configuration P' corresponds to a vertex w in the Voronoi configuration P. By construction, p_i and p_i' belong to an $O(\epsilon)$ neighborhood of the bisector b_{vw} determined by v and w, and $n_{vw} \cdot (p_i - p_i') < 0$. In addition, the component of $\mathrm{vrs}(v - p_i) - \mathrm{vrs}(w' - p_i')$ along b_{vw} is $O(\epsilon)$ whereas $n_{vw} \cdot \mathrm{vrs}(v - p_i) > 0$ and $n_{vw} \cdot \mathrm{vrs}(w' - p_i') = n_{vw} \cdot \mathrm{vrs}(w - p_i) + O(\epsilon)$, with $n_{vw} \cdot \mathrm{vrs}(w - p_i) < 0$. Then

$$\begin{aligned}
&\mathrm{vrs}(v - p_i) - \mathrm{vrs}(w' - p_i') \\
&\quad = \mathrm{proj}_{n_{vw}}(\mathrm{vrs}(v - p_i) - \mathrm{vrs}(w' - p_i')) + \mathrm{proj}_{b_{vw}}(\mathrm{vrs}(v - p_i) - \mathrm{vrs}(w' - p_i')) \\
&\qquad\qquad = \mathrm{proj}_{n_{vw}}(\mathrm{vrs}(v - p_i) - \mathrm{vrs}(w' - p_i')) + O(\epsilon)\,,
\end{aligned}$$

and, in turn, for sufficiently small ϵ,

$$\begin{aligned}
&(p_i - p_i') \cdot (\mathrm{vrs}(v - p_i) - \mathrm{vrs}(w' - p_i')) \\
&\qquad = (n_{vw} \cdot (p_i - p_i'))(n_{vw} \cdot (\mathrm{vrs}(v - p_i) - \mathrm{vrs}(w' - p_i'))) + O(\epsilon^2) < 0\,.
\end{aligned}$$

The result now follows from Theorem 1 on page 106 in [15].

(b) *Uniqueness of solution for* (5.5). Let D_{sm} be the set of $P \in Q^n$ such that $\mathrm{sm}_{V_i(P)}(p_i)$ is attained at a single edge for all i. Note that $Q^n \setminus D_{\mathrm{sm}}$ has measure zero, and that the vector field $\mathrm{Ln}(\partial \mathrm{sm}_{\mathcal{V}})$ is differentiable(and hence locally Lipschitz) when restricted to any connected component of D_{sm}. Let P, P' belong to different connected components of D_{sm}, and let $\|P - P'\| \le \epsilon$. Consider all the indexes i at which the values of $\mathrm{sm}_{V_i(P)}(p_i)$ and $\mathrm{sm}_{V_i(P')}(p_i')$ are attained at different edges.

Assume these edges are of type (a) (the type (b) case can be treated analogously). For these indexes,

$$\mathrm{Ln}(\partial \,\mathrm{sm}_{V_i(P)})(p_i) - \mathrm{Ln}(\partial \,\mathrm{sm}_{V_i(P')})(p_i') = \mathrm{vrs}(p_i - p_j) - \mathrm{vrs}(p_i' - p_k')\,,$$

for some uniquely determined p_j and p_k', with $j \neq k$. By construction, p_i and p_i' belong to an $O(\epsilon)$ neighborhood of the bisector b_{jk} determined by p_j and p_k, and $n_{kj} \cdot (p_i - p_i') < 0$. In addition, the component of $\mathrm{vrs}(p_i - p_j) - \mathrm{vrs}(p_i' - p_k')$ along b_{jk} is $O(\epsilon)$ whereas $n_{kj} \cdot \mathrm{vrs}(p_i - p_j) > 0$ and $n_{kj} \cdot \mathrm{vrs}(p_i' - p_k') = n_{kj} \cdot \mathrm{vrs}(p_i - p_k) + O(\epsilon)$, with $n_{kj} \cdot \mathrm{vrs}(p_i - p_k) < 0$. Then

$$\begin{aligned}
&\mathrm{vrs}(p_i - p_j) - \mathrm{vrs}(p_i' - p_k') \\
&\quad = \mathrm{proj}_{n_{kj}}(\mathrm{vrs}(p_i - p_j) - \mathrm{vrs}(p_i' - p_k')) + \mathrm{proj}_{b_{jk}}(\mathrm{vrs}(p_i - p_j) - \mathrm{vrs}(p_i' - p_k')) \\
&\qquad\qquad = \mathrm{proj}_{n_{kj}}(\mathrm{vrs}(p_i - p_j) - \mathrm{vrs}(p_i - p_k)) + O(\epsilon)\,,
\end{aligned}$$

and, in turn, for sufficiently small ϵ,

$$\begin{aligned}
&(p_i - p_i') \cdot (\mathrm{vrs}(p_i - p_j) - \mathrm{vrs}(p_i' - p_k')) \\
&\qquad = (n_{kj} \cdot (p_i - p_i'))(n_{kj} \cdot (\mathrm{vrs}(p_i - p_j) - \mathrm{vrs}(p_i - p_k))) + O(\epsilon^2) < 0\,.
\end{aligned}$$

The result now follows from Theorem 1 on page 106 in [15]. □

Remark 5.5 (relation with behavior-based robotics: move toward the furthest-away vertex). The distributed gradient control law in the disk-covering setting (5.4) has an interesting interpretation in the context of behavior-based robotics. Consider the ith generator. If the maximum of $\mathrm{lg}_{V_i(P)}$ is attained at a single vertex v of its Voronoi cell V_i, then $\mathrm{lg}_{V_i(P)}$ is differentiable at that configuration and its derivative corresponds to $\mathrm{vrs}(p_i - v)$. Therefore, the control law (5.4) corresponds to the behavior "move toward the furthest vertex in own Voronoi cell." If there are two or more vertexes of V_i where the value $\mathrm{lg}_{V_i(P)}(p_i)$ is attained, then (5.4) provides an average behavior by computing the least-norm element in the convex hull of all $\mathrm{vrs}(p_i - v)$ such that $\|p_i - v\| = \mathrm{lg}_{V_i(P)}(p_i)$.

Remark 5.6 (relation with behavior-based robotics: move away from the nearest neighbor). The distributed gradient control law in the sphere-packing setting (5.5) also has an interesting interpretation. For the ith generator, if the minimum of $\mathrm{sm}_{V_i(P)}$ is attained at a single edge e, then $\mathrm{sm}_{V_i(P)}$ is differentiable at that configuration, and its derivative is n_e. The control law (5.5) corresponds to the behavior "move away from the nearest neighbor" (where a neighbor can also be the boundary of the environment). If there are two or more edges where the value $\mathrm{sm}_{V_i(P)}(p_i)$ is attained, then (5.5) provides an average behavior in an analogous manner as before.

PROPOSITION 5.7. *For the dynamical system* (5.4)*, the generators' location* $P = (p_1, \ldots, p_n)$ *converges asymptotically to the largest weakly invariant set contained in the closure of* $A_{\mathrm{DC}}(Q) = \{P \in Q^n \mid i \in I(P) \implies p_i = \mathrm{CC}(V_i)\}$.

Proof. Let $a \in \widetilde{\mathcal{L}}_{-\mathrm{Ln}(\partial \mathrm{lg}_{\mathcal{V}})}\mathcal{H}_{\mathrm{DC}}(P)$. By definition, $a = -\mathrm{Ln}(\partial \,\mathrm{lg}_{\mathcal{V}})(P) \cdot \zeta$, for all $\zeta \in \partial\mathcal{H}_{\mathrm{DC}}(P)$. Let $v \in \mathrm{Ve}_{\mathrm{DC}}(\mathcal{V}(P))$. From Proposition 4.3 and Lemma 4.5, we know that, independently of the degenerate/nondegenerate character of the Voronoi partition at v, there always exist either an edge e of Q and generators p_i and p_j, or generators p_i, p_j, and p_k, such that $\lambda(e, i, j)$, $\lambda(e, j, i) > 0$ (respectively, $\mu(i, j, k)$,

$\mu(j,k,i)$, $\mu(k,i,j) > 0$). If v is a vertex of type (b), then

$$
\begin{aligned}
a &= -\operatorname{Ln}(\partial \lg_{\mathcal{V}})(P) \cdot \partial_v G_i \\
&= -\operatorname{Ln}(\partial \lg_{V_i(P)})(P) \cdot \lambda(e,i,j) \operatorname{vrs}(p_i - v) - \operatorname{Ln}(\partial \lg_{V_j(P)})(P) \cdot \lambda(e,j,i) \operatorname{vrs}(p_j - v).
\end{aligned} \tag{5.6}
$$

From Lemma 3.8(i) we conclude that $a \leq 0$, and the inequality is strict if either $p_i \neq \mathrm{CC}(V_i)$ or $p_j \neq \mathrm{CC}(V_j)$. The same conclusion can be derived if v is a vertex of type (a). Therefore, $\max \widetilde{\mathcal{L}}_{-\operatorname{Ln}(\partial \lg_{\mathcal{V}})} \mathcal{H}_{\mathrm{DC}}(P) \leq 0$ or $\widetilde{\mathcal{L}}_{-\operatorname{Ln}(\partial \lg_{\mathcal{V}})} \mathcal{H}_{\mathrm{DC}}(P) = \emptyset$. Now, resorting to the LaSalle principle (Theorem 2.7), we deduce that the solution $P : [0, +\infty) \to Q^n$ starting from P_0 converges to the largest weakly invariant set contained in $\overline{Z}_{-\operatorname{Ln}(\partial \lg_{\mathcal{V}}), \mathcal{H}_{\mathrm{DC}}} \cap \mathcal{H}_{\mathrm{DC}}^{-1}(\leq \mathcal{H}_{\mathrm{DC}}(P_0), P_0) \cap Q^n$.

Let us see that $Z_{-\operatorname{Ln}(\partial \lg_{\mathcal{V}}), \mathcal{H}_{\mathrm{DC}}} \cap Q^n$ is equal to $A_{\mathrm{DC}}(Q)$. Take a configuration $P \in A_{\mathrm{DC}}(Q)$. Then $\operatorname{Ln}(\partial \lg_{V_i(P)})(P) = 0$ if $i \in I(P)$, and $\pi_i(\zeta) = 0$ if $i \notin I(P)$, for any $\zeta \in \partial \mathcal{H}_{\mathrm{DC}}(P)$ (cf. Proposition 4.6). Consequently, $0 = -\operatorname{Ln}(\partial \lg_{\mathcal{V}})(P) \cdot \zeta$, for all $\zeta \in \partial \mathcal{H}_{\mathrm{DC}}(P)$, and so $0 \in \widetilde{\mathcal{L}}_{-\operatorname{Ln}(\partial \lg_{\mathcal{V}})} \mathcal{H}_{\mathrm{DC}}(P)$. Therefore, $A_{\mathrm{DC}}(Q) \subset Z_{-\operatorname{Ln}(\partial \lg_{\mathcal{V}}), \mathcal{H}_{\mathrm{DC}}}$. Now, consider $P \in Z_{-\operatorname{Ln}(\partial \lg_{\mathcal{V}}), \mathcal{H}_{\mathrm{DC}}}$. Then $0 \in \widetilde{\mathcal{L}}_{-\operatorname{Ln}(\partial \lg_{\mathcal{V}})} \mathcal{H}_{\mathrm{DC}}(P)$, that is, $0 = -\operatorname{Ln}(\partial \lg_{\mathcal{V}})(P) \cdot \zeta$, for all $\zeta \in \partial \mathcal{H}_{\mathrm{DC}}(P)$. If P is nondegenerate, we deduce from (5.6) and Lemma 3.8 that all the active generators are centered, i.e., $P \in A_{\mathrm{DC}}(Q)$. If P is degenerate, consider a degenerate vertex v where the value of $\mathcal{H}_{\mathrm{DC}}(P)$ is attained. For simplicity, we deal with the case where v is contained in an edge e of Q (the case $v \in \operatorname{int}(Q)$ is treated analogously). From Lemma 4.5 we know that there exist generators p_i, p_j determining v on opposite sides of l, the orthogonal line to the edge e passing through v. From (5.6) and Lemma 3.8 we deduce that both p_i and p_j are centered. Now, for each generator p_k with $v \in V_k$ in the same side of l as p_i (respectively, p_j), we consider the triplet (e,j,k) (respectively, (e,i,k)). Again resorting to (5.6) and Lemma 3.8, we conclude that p_k is also centered. Finally, if a generator p_k with $v \in V_k$ is such that $p_k \in l$, any of the triplets (e,j,k) or (e,i,k) can be invoked in a similar argument to ensure that p_k is centered. Therefore, $P \in A_{\mathrm{DC}}(Q)$, and hence $(Z_{-\operatorname{Ln}(\partial \lg_{\mathcal{V}}), \mathcal{H}_{\mathrm{DC}}} \cap Q^n) \subset A_{\mathrm{DC}}(Q)$. □

PROPOSITION 5.8. *For the dynamical system* (5.5)*, the generators' location* $P = (p_1, \ldots, p_n)$ *converges asymptotically to the largest weakly invariant set contained in the closure of* $A_{\mathrm{SP}}(Q) = \{P \in Q^n \mid i \in I(P) \implies p_i \in \mathrm{IC}(V_i)\}$.

Proof. Let $a \in \widetilde{\mathcal{L}}_{\operatorname{Ln}(\partial \operatorname{sm}_{\mathcal{V}})} \mathcal{H}_{\mathrm{SP}}(P)$. By definition, $a = \operatorname{Ln}[\operatorname{sm}_{\mathcal{V}}](P) \cdot \zeta$, for all $\zeta \in \partial \mathcal{H}_{\mathrm{SP}}(P)$. Let $e \in \mathrm{Ed}_{\mathrm{SP}}(\mathcal{V}(P))$. If e is an edge of type (a), i.e., a segment of the bisector determined by p_i and p_j, we compute (cf. Proposition 4.4)

$$
\begin{aligned}
a &= \operatorname{Ln}(\partial \operatorname{sm}_{\mathcal{V}})(P) \cdot \partial_e F_i \\
&= \operatorname{Ln}(\partial \operatorname{sm}_{V_i(P)})(P) \cdot \pi_i(\partial_e F_i) + \operatorname{Ln}(\partial \operatorname{sm}_{V_j(P)})(P) \cdot \pi_j(\partial_e F_i)\,.
\end{aligned} \tag{5.7}
$$

From Lemma 3.8(iii) we conclude that $a \geq 0$, and the inequality is strict if either $p_i \notin \mathrm{IC}(V_i)$ or $p_j \notin \mathrm{IC}(V_j)$. The same conclusion can be derived if e is a vertex of type (b). Therefore, $\min \widetilde{\mathcal{L}}_{\operatorname{Ln}(\partial \operatorname{sm}_{\mathcal{V}})} \mathcal{H}_{\mathrm{SP}}(P) \geq 0$ or $\widetilde{\mathcal{L}}_{\operatorname{Ln}(\partial \operatorname{sm}_{\mathcal{V}})} \mathcal{H}_{\mathrm{SP}}(P) = \emptyset$. Now, applying the LaSalle principle (Theorem 2.7) with the function $-\mathcal{H}_{\mathrm{SP}}$, we deduce that the solution $P : [0, +\infty) \to Q^n$ starting from P_0 converges to the largest weakly invariant set contained in $\overline{Z}_{\operatorname{Ln}(\partial \operatorname{sm}_{\mathcal{V}}), \mathcal{H}_{\mathrm{SP}}} \cap \mathcal{H}_{\mathrm{SP}}^{-1}(\leq \mathcal{H}_{\mathrm{SP}}(P_0), P_0) \cap Q^n$. From (5.7), and resorting to Proposition 4.6 and Lemma 3.8, one can also show that $Z_{\operatorname{Ln}(\partial \operatorname{sm}_{\mathcal{V}}), \mathcal{H}_{\mathrm{SP}}} \cap Q^n$ is equal to $A_{\mathrm{SP}}(Q)$. □

Remark 5.9. The sets $A_{\mathrm{DC}}(Q)$ and $A_{\mathrm{SP}}(Q)$ are not closed in general. If $\dim Q = 1$, then it can be seen that they indeed are. In higher dimensions one can find sequences

$\{P_k \in Q^n \mid k \in \mathbb{N}\}$ in these sets which converge to configurations P where not all active generators are centered.

5.3. Distributed dynamical systems based on geometric centering. Here, we propose alternative distributed dynamical systems for the multicenter functions. Our design is directly inspired by the results in Theorems 4.7 and 4.8 on the critical points of the multicenter functions $\mathcal{H}_{\mathrm{DC}}$ and $\mathcal{H}_{\mathrm{SP}}$. For $i \in \{1, \dots, n\}$, consider the dynamical systems

$$\dot{p}_i = \mathrm{CC}(V_i) - p_i\,, \tag{5.8}$$

$$\dot{p}_i \in \mathrm{IC}(V_i) - p_i\,. \tag{5.9}$$

Alternatively, we may write $\dot{P} = \mathrm{CC}(\mathcal{V}(P)) - P$ and $\dot{P} \in \mathrm{IC}(\mathcal{V}(P)) - P$. Note that both systems are Voronoi-distributed. Also, note that the vector field (5.8) is continuous, since the circumcenter of a polygon depends continuously on the location of its vertexes, and the location of the vertexes of the Voronoi partition depends continuously on the location of the generators; see [23]. However, (5.9) is a differential inclusion, since the incenter sets may not be singletons. By Lemma 2.5, the existence of solutions to (5.9) is guaranteed by the following result.

PROPOSITION 5.10. *Consider the set-valued map* $\mathrm{IC}(\mathcal{V}) - \mathrm{Id} : Q^n \to 2^{(\mathbb{R}^2)^n}$ *given by* $P \mapsto \mathrm{IC}(\mathcal{V}(P)) - P$. *Then* $\mathrm{IC}(\mathcal{V}) - \mathrm{Id}$ *is upper semicontinuous with nonempty, compact, and convex values.*

Proof. Clearly, the map $\mathrm{IC}(\mathcal{V}) - \mathrm{Id}$ takes nonempty and compact values. From Lemma 3.2, we also know that it takes convex values. Furthermore, since the identity map is continuous, it suffices to check that $P \mapsto \mathrm{IC}(\mathcal{V}(P))$ is upper semicontinuous. We then have to verify that, given $P_0 \in Q^n$, for each $\epsilon > 0$ there exists $\delta > 0$ such that

$$\mathrm{IC}(\mathcal{V}(P)) \subset \mathrm{IC}(\mathcal{V}(P_0)) + B_{2n}(0, \epsilon) \quad \text{if } \|P - P_0\| \leq \delta\,. \tag{5.10}$$

Now, for each i, if $\mathrm{IC}(V_i(P_0))$ is not a singleton, then it is a segment (cf. Lemma 3.2) whose extremal points $q_{i1}(P_0)$, $q_{i2}(P_0)$ are the intersection points of some bisectors of the edges of the Voronoi cell. It is clear that $q_{i\alpha}(P) \to q_{i\alpha}(P_0)$ when $P \to P_0$ for $\alpha = 1, 2$. Therefore, given $\epsilon > 0$, one can choose $\delta_i > 0$ such that if $\|P - P_0\| \leq \delta_i$, then $\|q_{i\alpha}(P) - q_{i\alpha}(P_0)\| \leq \epsilon/n$. Since $\mathrm{IC}(V_i(P))$ is contained in the segment joining $q_{i1}(P)$ and $q_{i2}(P)$, we deduce $\mathrm{IC}(V_i(P)) \subset \mathrm{IC}(V_i(P_0)) + B_2(0, \epsilon/n)$. On the other hand, if $\mathrm{IC}(V_i(P_0))$ is a singleton, then it coincides with the intersection points $q_{i1}(P_0), \dots, q_{im}(P_0)$ of some bisectors of the edges of the Voronoi cell. The above reasoning also guarantees that there exits $\delta_i > 0$ such that $q_{i\alpha}(P) \in \mathrm{IC}(V_i(P_0)) + B_2(0, \epsilon/n)$, $\alpha = 1, \dots, m$, if $\|P - P_0\| \leq \delta_i$. Since $\mathrm{IC}(V_i(P))$ is contained in one of the segments joining the points $q_{i1}(P), \dots, q_{im}(P)$, we again deduce $\mathrm{IC}(V_i(P)) \subset \mathrm{IC}(V_i(P_0)) + B_2(0, \epsilon/n)$. The statement in (5.10) follows by taking the minimum of $\delta_1, \dots, \delta_n$. □

Having established the existence of solutions, one can also see that the compact set Q^n is strongly invariant for the vector field $\mathrm{CC}(\mathcal{V}) - \mathrm{Id}$ and for the differential inclusion $\mathrm{IC}(\mathcal{V}) - \mathrm{Id}$. Next, we characterize the asymptotic convergence of the dynamical systems under study.

PROPOSITION 5.11. *For the dynamical system* (5.8) *(respectively,* (5.9)*), the generators' location* $P = (p_1, \dots, p_n)$ *converges asymptotically to the largest weakly invariant set contained in the closure of* $A_{\mathrm{DC}}(Q)$ *(respectively, in the closure of* $A_{\mathrm{SP}}(Q)$*).*

Proof. The proof of this result is parallel to the proofs of Propositions 5.7 and 5.8. The sequence of steps is the same as before, though now one resorts to Lemma 3.8(ii) and Lemma 3.8(iv). The only additional observation is that when computing the set-valued Lie derivative for (5.9), one has that $a \in \widetilde{\mathcal{L}}_{\mathrm{IC}(\mathcal{V})-\mathrm{Id}}\mathcal{H}_{\mathrm{SP}}(P)$ if and only if there exists $x \in \mathrm{IC}(\mathcal{V}(P))$ such that $a = (x - P) \cdot \zeta$, for any $\zeta \in \partial\mathcal{H}_{\mathrm{SP}}(P)$. The application of Lemma 3.8 guarantees that $a \geq 0$ and that the inequality is strict if any of the active generators is not in its corresponding incenter set. ∎

5.4. Simulations. To illustrate the performance of the distributed coordination algorithms, we include some simulation results. The algorithms are implemented in `Mathematica` as a single centralized program. We compute the bounded Voronoi diagram of a collection of points using the `Mathematica` package `ComputationalGeometry`. We compute the circumcenter of a polygon via the algorithm in [28] and the incenter set via the `LinearProgramming` solver in `Mathematica`. Measuring displacements in meters, we consider the domain determined by the vertexes

$$\{(0,0),(2.5,0),(3.45,1.5),(3.5,1.6),(3.45,1.7),(2.7,2.1),(1.,2.4),(.2,1.2)\}.$$

In Figures 5.2 and 5.3, we illustrate the performance of the dynamical systems (5.4) and (5.8), respectively, minimizing the multicircumcenter function $\mathcal{H}_{\mathrm{DC}}$. In Figures 5.4 and 5.5, we illustrate the performance of the dynamical systems (5.5) and (5.9), respectively, maximizing the multi-incenter function $\mathcal{H}_{\mathrm{SP}}$. Observing the final configurations in the four figures, one can verify, visually and numerically, that the active generators are asymptotically centered as forecast by our analysis.

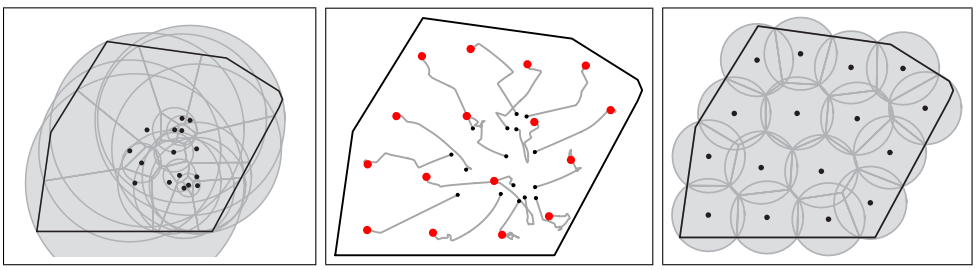


FIG. 5.2. *"Toward the furthest" algorithm for* 16 *generators in a convex polygonal environment. The left (respectively, right) figure illustrates the initial (respectively, final) locations and Voronoi partition. The central figure illustrates the network evolution. After* 2 *seconds, the multicenter function is approximately* .39504 *meters.*

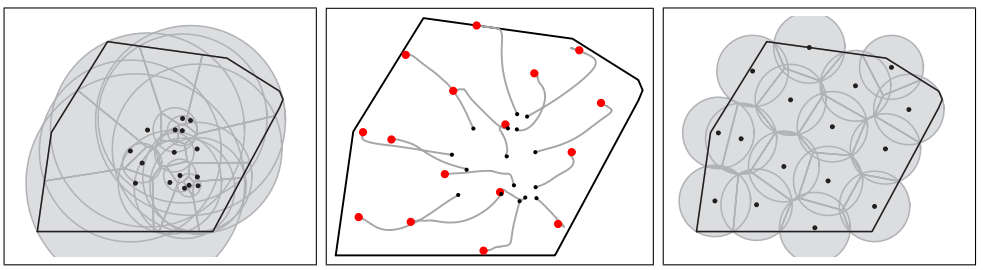


FIG. 5.3. *"Move-toward-the-circumcenter" algorithm for* 16 *generators in a convex polygonal environment. The left (respectively, right) figure illustrates the initial (respectively, final) locations and Voronoi partition. The central figure illustrates the network evolution. After* 20 *seconds, the multicenter function is approximately* 0.43273 *meters.*

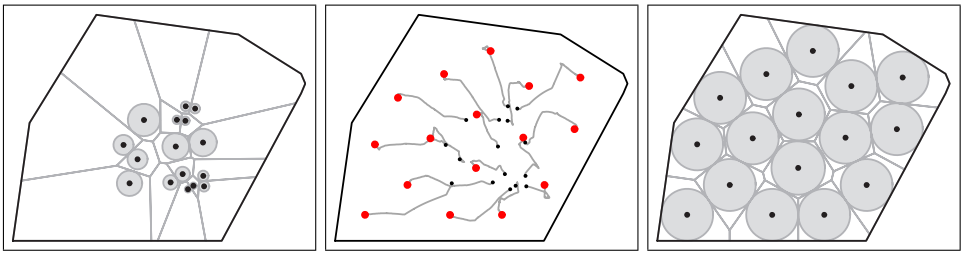


FIG. 5.4. *"Away-from-closest" algorithm for* 16 *generators in a convex polygonal environment. The left (respectively, right) figure illustrates the initial (respectively, final) locations and Voronoi partition. The central figure illustrates the network evolution. After* 2 *seconds, the multicenter function is approximately* .26347 *meters.*

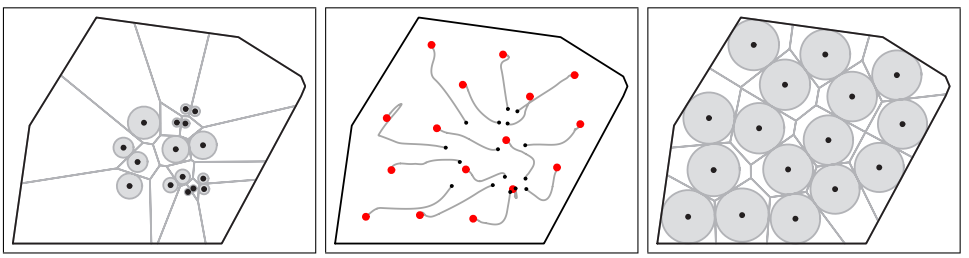


FIG. 5.5. *"Move-toward-the-incenter" algorithm for* 16 *generators in a convex polygonal environment. The left (respectively, right) figure illustrates the initial (respectively, final) locations and Voronoi partition. The central figure illustrates the network evolution. After* 20 *seconds, the multicenter function is approximately* .2498 *meters.*

6. Conclusions. We have introduced two multicenter functions that provide quality-of-service measures for mobile networks. We have shown that both functions are globally Lipschitz, and we have computed their generalized gradients. Furthermore, under certain technical conditions, we have characterized via nonsmooth analysis their critical points as center Voronoi configurations and as solutions of disk-covering and sphere-packing problems. We have also considered various algorithms that extremize the multicenter functions. First, we considered the nonsmooth gradient flows induced by their respective generalized gradients. Second, we devised a novel strategy based on the generalized gradients of the 1-center functions of each generator. Third, we introduced and characterized a geometric centering strategy with resemblances to the classical Lloyd algorithm. We have unveiled the remarkable geometric interpretations of these algorithms, discussed their distributed character and analyzed their asymptotic behavior using nonsmooth stability analysis.

Future directions of research include the following: (i) sharpening the asymptotic convergence results for the proposed dynamical systems (e.g., proving that all generators will asymptotically be centered), (ii) considering the setting of convex polytopes in $\mathbb{R}^N$, for $N > 2$, (iii) understanding in what sense the proposed multicircumcenter and the multi-incenter problems can be shown to be dual, and (iv) analyzing other meaningful geometric optimization problems and their relations with cooperative behaviors.

Symbol	: **Description and page(s) when applicable**
$A_{\mathrm{DC}}(Q)$	: Set of configurations $P \in Q^n$ where all active generators are in the circumcenter of its own Voronoi region, 1567
$A_{\mathrm{SP}}(Q)$	: Set of configurations $P \in Q^n$ where all active generators are in the incenter set of its own Voronoi region, 1568
$\mathrm{CC}(Q)$	: Circumcenter of polytope Q, 1548
$\mathrm{CR}(Q)$	: Circumradius of polytope Q, 1548
D_S	: Distance function to the convex set S, 1546
$\mathrm{Ed}(Q)$	: Edges of polygon Q, 1546
$\mathrm{Ed}_{\mathrm{SP}}(\mathcal{V}(P))$	: Edges where the value of $\mathcal{H}_{\mathrm{SP}}(P)$ is attained, 1546
$e(i)$	: Edge of $\mathcal{V}(P)$ belonging to V_i and to the boundary of Q, 1547
$e(i,j)$	: Edge of $\mathcal{V}(P)$ determined by p_i and p_j, 1547
$F_i(P)$	: Smallest distance from p_i to the boundary of $V_i(P)$, 1556
$G_i(P)$	: Largest distance from p_i to the boundary of $V_i(P)$, 1556
∂f	: Generalized gradient of the locally Lipschitz function f, 1549
$\mathcal{H}_{\mathrm{DC}}$	: Multicircumcenter function, 1548
$\mathcal{H}_{\mathrm{SP}}$	: Multi-incenter function, 1548
$\mathrm{IC}(Q)$	: Incenter set of polytope Q, 1548
$\mathrm{IR}(Q)$	: Inradius of polytope Q, 1548
$K[X]$	: Filippov mapping associated with a measurable and essentially locally bounded mapping $X : \mathbb{R}^N \to \mathbb{R}^N$, 1550
$\lambda(e,i,j)$	: Scalar function associated with the vertex $v(e,i,j)$, 1559
$\mathrm{Ln}(S)$	: Least-norm element of the convex set S, 1549
$\mathrm{lg}_Q(p)$	: Largest distance from p to the boundary of Q, 1552
$\mu(i,j,k)$	: Scalar function associated with the vertex $v(i,j,k)$, 1559
$\mathcal{N}(P,i)$, $\mathcal{N}(i)$	: Set of neighbors of the ith generator at configuration P, 1546
$n_{e(i,j)}$	: Unit normal to $e(i,j)$ pointing toward $\mathrm{int}(V_i(P))$, 1547
$n_{e(i)}$	: Unit normal to $e(i)$ pointing toward $\mathrm{int}(Q)$, 1547
proj_S	: Orthogonal projection onto the convex set S, 1546
π_i	: Canonical projection from Q^n onto the ith factor, 1546
$\widetilde{\mathcal{L}}_X f$	: Set-valued Lie derivative of f with respect to X, 1550
$\mathrm{sm}_Q(p)$	: Smallest distance from p to the boundary of Q, 1552
$v(i,j,k)$	: Vertex of $\mathcal{V}(P)$ determined by p_i, p_j, and p_k, 1547
$v(e,i,j)$	: Vertex of $\mathcal{V}(P)$ determined by $e \in \mathrm{Ed}(Q)$ and p_i, p_j, 1547
$v(e,f,i)$	: Vertex of $\mathcal{V}(P)$ determined by $e, f \in \mathrm{Ed}(Q)$, and p_i, 1547
$\mathrm{Ve}_{\mathrm{DC}}(\mathcal{V}(P))$	: Vertexes of $\mathcal{V}(P)$ where the value of $\mathcal{H}_{\mathrm{DC}}(P)$ is attained, 1548
$\mathrm{vrs}(v)$	: Unit vector in the direction of $0 \neq v \in \mathbb{R}^N$, 1546
$\mathrm{Ve}(Q)$	: Vertexes of polygon Q, 1546
$\mathcal{V}(P)$	: Voronoi partition of Q generated by $P = (p_1, \ldots, p_n)$, 1546
$Z_{X,f}$	: Set formed by points $x \in \mathbb{R}^N$ such that 0 belongs to $\widetilde{\mathcal{L}}_X f(x)$, 1551

REFERENCES

[1] P. K. Agarwal and M. Sharir, *Efficient algorithms for geometric optimization*, ACM Comput. Surveys, 30 (1998), pp. 412–458.

[2] R. C. Arkin, *Behavior-Based Robotics*, MIT Press, Cambridge, MA, 1998.

[3] A. Bacciotti and F. Ceragioli, *Stability and stabilization of discontinuous systems and nonsmooth Lyapunov functions*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 361–376.

[4] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Athena Scientific, Belmont, MA, 1997.

[5] V. Boltyanski, H. Martini, and V. Soltan, *Geometric Methods and Optimization Problems*, Comb. Optim. 4, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999.

[6] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

[7] F. Ceragioli, *Discontinuous Ordinary Differential Equations and Stabilization*, Ph.D. thesis, Università di Firenze, 2000.

[8] H. Choset, *Nonsmooth analysis, convex analysis, and their applications to motion planning*, Internat. J. Comput. Geom. Appl., 9 (1999), pp. 447–469.

[9] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Canadian Mathematical Society Series of Monographs and Advanced Texts, John Wiley, New York, 1983.

[10] J. Cortés and F. Bullo, *From geometric optimization and nonsmooth analysis to distributed coordination algorithms*, in Proceedings of the IEEE Conference on Decision and Control, Maui, HI, 2003, pp. 3274–3280.

[11] J. Cortés, S. Martínez, T. Karatas, and F. Bullo, *Coverage control for mobile sensing networks*, IEEE Trans. Robotics and Automation, 20 (2004), pp. 243–255.

[12] M. de Berg, M. van Kreveld, and M. Overmars, *Computational Geometry: Algorithms and Applications*, Springer-Verlag, New York, 1997.

[13] Z. Drezner, ed., *Facility Location: A Survey of Applications and Methods*, Springer Series in Operations Research, Springer-Verlag, New York, 1995.

[14] Q. Du, V. Faber, and M. Gunzburger, *Centroidal Voronoi tessellations: Applications and algorithms*, SIAM Rev., 41 (1999), pp. 637–676.

[15] A. F. Filippov, *Differential Equations with Discontinuous Righthand Sides*, Math. Appl. 18, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1988.

[16] R. M. Gray and D. L. Neuhoff, *Quantization*, IEEE Trans. Inform. Theory, 44 (1998), pp. 2325–2383.

[17] U. Helmke and J. Moore, *Optimization and Dynamical Systems*, Springer-Verlag, New York, 1994.

[18] A. Jadbabaie, J. Lin, and A. S. Morse, *Coordination of groups of mobile autonomous agents using nearest neighbor rules*, IEEE Trans. Automat. Control, 48 (2003), pp. 988–1001.

[19] Y. Liu, K. M. Passino, and M. M. Polycarpou, *Stability analysis of m-dimensional asynchronous swarms with a fixed communication topology*, IEEE Trans. Automat. Control, 48 (2003), pp. 76–95.

[20] S. P. Lloyd, *Least squares quantization in PCM*, IEEE Trans. Inform. Theory, 28 (1982), pp. 129–137.

[21] D. G. Luenberger, *Linear and Nonlinear Programming*, 2nd ed., Addison–Wesley, Reading, MA, 1984.

[22] P. Ögren, E. Fiorelli, and N. E. Leonard, *Cooperative control of mobile sensor networks: Adaptive gradient climbing in a distributed environment*, IEEE Trans. Automat. Control, 49 (2004), pp. 1292–1302.

[23] A. Okabe, B. Boots, K. Sugihara, and S. N. Chiu, *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, 2nd ed., Wiley Series in Probability and Statistics, John Wiley, New York, 2000.

[24] R. Olfati-Saber and R. M. Murray, *Consensus problems in networks of agents with switching topology and time-delays*, IEEE Trans. Automat. Control, 49 (2004), pp. 1520–1533.

[25] B. Paden and S. Sastry, *A calculus for computing Filippov's differential inclusion with application to the variable structure control of robot manipulators*, IEEE Trans. Circuits Systems, 34 (1987), pp. 73–82.

[26] J.-M. Robert and G. T. Toussaint, *Computational geometry and facility location*, in Proceedings of the International Conference on Operations Research and Management Science, Vol. B, Manila, The Philippines, 1990, pp. 1–19.

[27] D. Shevitz and B. Paden, *Lyapunov stability theory of nonsmooth systems*, IEEE Trans. Automat. Control, 39 (1994), pp. 1910–1914.

[28] S. Skyum, *A simple algorithm for computing the smallest enclosing circle*, Inform. Process.

Lett., 37 (1991), pp. 121–125.

[29] A. Suzuki and Z. Drezner, *The p-center location problem in an area*, Location Science, 4 (1996), pp. 69–82.

[30] I. Suzuki and M. Yamashita, *Distributed anonymous mobile robots: Formation of geometric patterns*, SIAM J. Comput., 28 (1999), pp. 1347–1363.

[31] H. Tanner, A. Jadbabaie, and G. J. Pappas, *Stable flocking of mobile agents, Part* II: *Dynamic topology*, in Proceedings of the IEEE Conference on Decision and Control, Maui, HI, 2003, pp. 2016–2021.

CORRECTION TO "COORDINATION AND GEOMETRIC OPTIMIZATION VIA DISTRIBUTED DYNAMICAL SYSTEMS"

Because of a production error, the page numbers in the list of symbols on page 1572 are incorrect. The list should read as follows.

Symbol	: **Description and page(s) when applicable**
$A_{\mathrm{DC}}(Q)$	: Set of configurations $P \in Q^n$ where all active generators are in the circumcenter of its own Voronoi region, 1567
$A_{\mathrm{SP}}(Q)$	: Set of configurations $P \in Q^n$ where all active generators are in the incenter set of its own Voronoi region, 1568
$\mathrm{CC}(Q)$	: Circumcenter of polytope Q, 1548
$\mathrm{CR}(Q)$	: Circumradius of polytope Q, 1548
D_S	: Distance function to the convex set S, 1546
$\mathrm{Ed}(Q)$	: Edges of polygon Q, 1546
$\mathrm{Ed}_{\mathrm{SP}}(\mathcal{V}(P))$	: Edges where the value of $\mathcal{H}_{\mathrm{SP}}(P)$ is attained, 1548
$e(i)$	: Edge of $\mathcal{V}(P)$ belonging to V_i and to the boundary of Q, 1547
$e(i,j)$	: Edge of $\mathcal{V}(P)$ determined by p_i and p_j, 1547
$F_i(P)$	: Smallest distance from p_i to the boundary of $V_i(P)$, 1556
$G_i(P)$	: Largest distance from p_i to the boundary of $V_i(P)$, 1556
∂f	: Generalized gradient of the locally Lipschitz function f, 1549
$\mathcal{H}_{\mathrm{DC}}$	: Multicircumcenter function, 1548
$\mathcal{H}_{\mathrm{SP}}$	: Multi-incenter function, 1548
$\mathrm{IC}(Q)$	: Incenter set of polytope Q, 1548
$\mathrm{IR}(Q)$	: Inradius of polytope Q, 1548
$K[X]$	: Filippov mapping associated with a measurable and essentially locally bounded mapping $X : \mathbb{R}^N \to \mathbb{R}^N$, 1550
$\lambda(e,i,j)$	: Scalar function associated with the vertex $v(e,i,j)$, 1559
$\mathrm{Ln}(S)$	: Least-norm element of the convex set S, 1549
$\mathrm{lg}_Q(p)$	: Largest distance from p to the boundary of Q, 1552
$\mu(i,j,k)$	: Scalar function associated with the vertex $v(i,j,k)$, 1559
$\mathcal{N}(P,i)$, $\mathcal{N}(i)$	: Set of neighbors of the ith generator at configuration P, 1546
$n_{e(i,j)}$	: Unit normal to $e(i,j)$ pointing toward $\mathrm{int}(V_i(P))$, 1547
$n_{e(i)}$	: Unit normal to $e(i)$ pointing toward $\mathrm{int}(Q)$, 1547
proj_S	: Orthogonal projection onto the convex set S, 1546
π_i	: Canonical projection from Q^n onto the ith factor, 1546
$\widetilde{\mathcal{L}}_X f$	: Set-valued Lie derivative of f with respect to X, 1550
$\mathrm{sm}_Q(p)$	: Smallest distance from p to the boundary of Q, 1552
$v(i,j,k)$	: Vertex of $\mathcal{V}(P)$ determined by p_i, p_j, and p_k, 1547
$v(e,i,j)$	: Vertex of $\mathcal{V}(P)$ determined by $e \in \mathrm{Ed}(Q)$ and p_i, p_j, 1547
$v(e,f,i)$	: Vertex of $\mathcal{V}(P)$ determined by $e, f \in \mathrm{Ed}(Q)$, and p_i, 1547
$\mathrm{Ve}_{\mathrm{DC}}(\mathcal{V}(P))$	: Vertexes of $\mathcal{V}(P)$ where the value of $\mathcal{H}_{\mathrm{DC}}(P)$ is attained, 1548
$\mathrm{vrs}(v)$	: Unit vector in the direction of $0 \neq v \in \mathbb{R}^N$, 1546
$\mathrm{Ve}(Q)$	: Vertexes of polygon Q, 1546

$\mathcal{V}(P)$: Voronoi partition of Q generated by $P = (p_1, \ldots, p_n)$, 1546

$Z_{X,f}$: Set formed by points $x \in \mathbb{R}^N$ such that 0 belongs to $\widetilde{\mathcal{L}}_X f(x)$, 1551

SIAM J. CONTROL OPTIM.
Vol. 44, No. 5, pp. 1575–1597

EXPONENTIAL STABILIZATION OF LAMINATED BEAMS WITH STRUCTURAL DAMPING AND BOUNDARY FEEDBACK CONTROLS*

JUN-MIN WANG†, GEN-QI XU‡, AND SIU-PANG YUNG§

Abstract. We study the boundary stabilization of laminated beams with structural damping which describes the slip occurring at the interface of two-layered objects. By using an invertible matrix function with an eigenvalue parameter and an asymptotic technique for the first order matrix differential equation, we find out an explicit asymptotic formula for the matrix fundamental solutions and then carry out the asymptotic analyses for the eigenpairs. Furthermore, we prove that there is a sequence of generalized eigenfunctions that forms a Riesz basis in the state Hilbert space, and hence the spectrum determined growth condition holds. Furthermore, exponential stability of the closed-loop system can be deduced from the eigenvalue expressions. In particular, the semigroup generated by the system operator is a C_0-group due to the fact that the three asymptotes of the spectrum are parallel to the imaginary axis.

Key words. Riesz basis, laminated beams, exponential stability

AMS subject classifications. 93C20, 93D15, 35B35, 35P10

DOI. 10.1137/040610003

1. Introduction.

The vibration suppression of the laminated beams due to the demand for advanced performance has been one of the main research topics in smart materials and structures. These composite laminates usually have superior structural properties such as adaptability, and the design of their piezoelectric materials can be used as both actuators and sensors. The detailed physical background can be found in [10] and the references therein. In [4], Hansen and Spies derived three mathematical models for two-layered beams with structural damping due to the interfacial slip. Our interest in this paper is to study the first model in [4] which is closely related to the Timoshenko beam theory. The equations for this beam model are

$$\begin{cases} mw_{tt} + (G(\psi - w_x))_x = 0, & 0 < x < 1,\ t \geq 0, \\ I_m(3s_{tt} - \psi_{tt}) - G(\psi - w_x) - (D(3s_x - \psi_x))_x = 0, & 0 < x < 1,\ t \geq 0, \\ I_m s_{tt} + G(\psi - w_x) + \frac{4}{3}\gamma s + \frac{4}{3}\beta I_m s_t - (Ds_x)_x = 0, & 0 < x < 1,\ t \geq 0, \end{cases} \tag{1.1}$$

where $w(x,t)$ denotes the transverse displacement, $\psi(x,t)$ represents the rotation angle and $s(x,t)$ is proportional to the amount of slip along the interface at time t and longitudinal spatial variable x, respectively, and $m > 0$ is the density of the beams, $G, I_m, D, \gamma > 0$ are the shear stiffness, mass moment of inertia, flexural rigidity, and adhesive stiffness of the beams together with $\beta > 0$ as the adhesive damping parameter. Moreover, $\sqrt{G/m}$ and $\sqrt{D/I_m}$ are two wave speeds and we always assume

*Received by the editors June 15, 2004; accepted for publication (in revised form) April 26, 2005; published electronically November 14, 2005.
http://www.siam.org/journals/sicon/44-5/61000.html

†School of Computational and Applied Mathematics, University of the Witwatersrand, Private 3, Wits 2050, Johannesburg, South Africa (wangjc@graduate.hku.hk).

‡Department of Mathematics, Tianjin University, Tianjin 300072, People's Republic of China (xugq_2001@yahoo.com). The research of this author was supported by the National Natural Science Foundation of China.

§Department of Mathematics, The University of Hong Kong, Hong Kong, People's Republic of China (spyung@hku.hk). The research of this author was supported by an HKRGC Earmarked Research Grant.

that they are different in the present paper (see [7]). We refer to [4] for the detailed derivation of the mathematical model and its physical parameters. It is easy to find that if the slip s is assumed to be identically zero, then the first two equations of system (1.1) can be reduced exactly to the Timoshenko beam system. The third equation in (1.1) describes the dynamics of the slip. For convenience, if we introduce another variable ξ of the effective rotation angle by

$$\xi = 3s - \psi, \tag{1.2}$$

then (1.1) changes to

$$\begin{cases} mw_{tt} + (G(3s - \xi - w_x))_x = 0, & 0 < x < 1,\ t \geq 0, \\ I_m\xi_{tt} - G(3s - \xi - w_x) - (D\xi_x)_x = 0, & 0 < x < 1,\ t \geq 0, \\ I_m s_{tt} + G(3s - \xi - w_x) + \frac{4}{3}\gamma s + \frac{4}{3}\beta I_m s_t - (Ds_x)_x = 0, & 0 < x < 1,\ t \geq 0. \end{cases} \tag{1.3}$$

For system (1.3), we impose the cantilever boundary conditions, which can be easily obtained from the principle of virtual work (see [4]),

$$\begin{cases} w(0,t) = 0, \quad \xi(0,t) = 0, \quad s(0,t) = 0, \\ \xi_x(1,t) = u_2(t), \quad s_x(1,t) = 0, \quad 3s(1,t) - \xi(1,t) - w_x(1,t) = u_1(t), \end{cases} \tag{1.4}$$

where $u_1(t)$ and $u_2(t)$ are boundary control forces, and the initial conditions (for $0 < x < 1$)

$$(w, \xi, s)\Big|_{t=0} = (w_0, \xi_0, s_0) \quad \text{and} \quad (w_t, \xi_t, s_t)\Big|_{t=0} = (w_1, \xi_1, s_1). \tag{1.5}$$

We point out that due to the action of the slip s, the uncontrolled system (1.3) with boundary conditions (1.4) ($u_1 = u_2 \equiv 0$) in [4] can achieve the asymptotic stability but it does not reach the exponential stability (see Corollary 2.3 and Note 2.1).

In this paper, the following boundary feedback controls are proposed to exponentially stabilize systems (1.3) and (1.4):

$$u_2(t) = -k_2\xi_t(1,t), \quad u_1(t) = k_1 w_t(1,t), \tag{1.6}$$

where k_1 and k_2 are positive constant feedback gains. Then the boundary conditions become

$$\begin{cases} w(0,t) = 0, \quad \xi(0,t) = 0, \quad s(0,t) = 0, \\ \xi_x(1,t) = -k_2\xi_t(1,t), \quad s_x(1,t) = 0, \quad 3s(1,t) - \xi(1,t) - w_x(1,t) = k_1 w_t(1,t), \end{cases} \tag{1.7}$$

and the closed-loop system has both internal damping and boundary controls.

Our goal is to show that the closed-loop system (1.3) with (1.7) is exponentially stable in the state Hilbert space. This will follow from proving the following three aspects: (i) the closed-loop system is dissipative in the state space and the system operator has compact resolvents; (ii) there exist three asymptotes of frequencies for the system which are parallel to the imaginary axis from the left side; (iii) the generalized eigenfunctions of the system form a Riesz basis in the state space and hence the spectrum determined growth condition, and the exponential stability holds for the system. Among these, (i) is easy to verify while (ii) and (iii) are very difficult to solve. Our interests in this paper are mainly concentrated on the asymptotically spectral analysis and the proof of Riesz basis for the system.

Now let us briefly outline the contents of this paper. In the next section, the well-poseness of the system will be established. Asymptotic estimates of the eigenvalues for the system will be given in section 3. This is the foundation that we shall use to investigate the exponential stability and basis property for the system. Section 4 is devoted to the asymptotic expansion of the corresponding eigenfunctions. Finally, in the last section, we obtain a more profound result, namely, the existence of a sequence of the generalized eigenfunctions of the system that forms a Riesz basis in the state Hilbert space. Consequently, the spectrum determined growth condition and the exponential stability are concluded. Furthermore, the semigroup generated by the system operator is actually a C_0-group based on the spectrum distribution of the system.

2. Well-posedness of the system. We start our investigation by formulating the problem on the state Hilbert space. Let

$$\mathcal{H} := \big(H_E^1(0,1) \times L^2(0,1)\big)^3 \tag{2.1}$$

with

$$H_E^i(0,1) := \{f \in H^i(0,1) \mid f(0) = 0\} \quad \text{for } i = 1,2, \tag{2.2}$$

where $H^i(0,1)$ $(i = 1,2)$ denote the usual Sobolev spaces. The inner product in $\mathcal{H}$ is defined by

$$\begin{aligned}\langle Y_1, Y_2\rangle_{\mathcal{H}} :&= m\langle z_1, z_2\rangle_{L^2} + G\langle 3s_1 - \xi_1 - w_1', 3s_2 - \xi_2 - w_2'\rangle_{L^2} + I_m\langle \varphi_1, \varphi_2\rangle_{L^2} \\ &\quad + D\langle \xi_1', \xi_2'\rangle_{L^2} + 3I_m\langle h_1, h_2\rangle_{L^2} + 3D\langle s_1', s_2'\rangle_{L^2} + 4\gamma\langle s_1, s_2\rangle_{L^2},\end{aligned} \tag{2.3}$$

where $Y_i := [w_i, z_i, \xi_i, \varphi_i, s_i, h_i]^\top \in \mathcal{H}$ with $i = 1,2$, in which the superscript $\top$ denotes the transpose of a vector or a matrix, $\langle\cdot,\cdot\rangle_{L^2}$ is the inner product on $L^2(0,1)$, and the prime represents the differentiation with respect to x. In view of system (1.3) and (1.7), we define a linear operator $\mathcal{A} : \mathcal{D}(\mathcal{A}) \subset \mathcal{H} \to \mathcal{H}$ in Hilbert space $\mathcal{H}$ by

$$\mathcal{A}\begin{bmatrix} w \\ z \\ \xi \\ \varphi \\ s \\ h \end{bmatrix} := \begin{bmatrix} z \\ \frac{G}{m}(\xi' + w'' - 3s') \\ \varphi \\ \frac{G}{I_m}(3s - \xi - w') + \frac{D}{I_m}\xi'' \\ h \\ \frac{G}{I_m}(\xi + w' - 3s) - \frac{4}{3}\frac{\gamma}{I_m}s - \frac{4}{3}\beta h + \frac{D}{I_m}s'' \end{bmatrix} \tag{2.4}$$

with

$$\mathcal{D}(\mathcal{A}) := \left\{ [w,z,\xi,\varphi,s,h]^\top \in \mathcal{H} \;\middle|\; \begin{array}{l} w \in H_E^2(0,1),\ \xi \in H_E^2(0,1),\ s \in H_E^2(0,1), \\ z \in H_E^1(0,1),\ \varphi \in H_E^1(0,1),\ h \in H_E^1(0,1), \\ \xi'(1) = -k_2\varphi(1),\ s'(1) = 0, \\ 3s(1) - \xi(1) - w'(1) = k_1 z(1) \end{array} \right\}. \tag{2.5}$$

If we set $Y := [w, w_t, \xi, \xi_t, s, s_t]^\top$, then the closed-loop system (1.3), (1.5), and (1.7) can be formulated into an abstract evolution equation in $\mathcal{H}$:

$$\begin{cases} \dfrac{d}{dt}Y(t) = \mathcal{A}Y(t), \quad t > 0, \\ Y(0) := [w_0,\ w_1,\ \xi_0,\ \xi_1,\ s_0,\ s_1]^\top. \end{cases} \tag{2.6}$$

THEOREM 2.1. *Let $\mathcal{A}$ be defined by* (2.4) *and* (2.5). *Then $\mathcal{A}$ is dissipative in $\mathcal{H}$. In addition, $\mathcal{A}^{-1}$ exists and is compact on $\mathcal{H}$. Therefore, $\mathcal{A}$ generates a C_0-semigroup $e^{\mathcal{A}t}$ of contractions on $\mathcal{H}$ and the spectrum $\sigma(\mathcal{A})$ consists of isolated eigenvalues only.*

Proof. Since for any $[w, z, \xi, \varphi, s, h]^\top \in \mathcal{D}(\mathcal{A})$,

$$\begin{aligned}
&\big\langle \mathcal{A}[w,z,\xi,\varphi,s,h]^\top,\ [w,z,\xi,\varphi,s,h]^\top \big\rangle_{\mathcal{H}} \\
&= \Big\langle \Big[z,\ \frac{G}{m}(\xi' + w'' - 3s'),\ \varphi,\ \frac{G}{I_m}(3s - \xi - w') + \frac{D}{I_m}\xi'',\ h, \\
&\frac{G}{I_m}(\xi + w' - 3s) - \frac{4}{3}\frac{\gamma}{I_m}s - \frac{4}{3}\beta h + \frac{D}{I_m}s''\Big]^\top,\ [w,z,\xi,\varphi,s,h]^\top \Big\rangle_{\mathcal{H}} \\
&= G\langle \xi' + w'' - 3s',\ z\rangle_{L^2} + G\langle 3h - \varphi - z',\ 3s - \xi - w'\rangle_{L^2} \\
&+ \langle G(3s - \xi - w') + D\xi'',\ \varphi\rangle_{L^2} + D\langle \varphi', \xi'\rangle_{L^2} + 3D\langle h', s'\rangle_{L^2} + 4\gamma\langle h, s\rangle_{L^2} \\
&+ \langle G(\xi + w' - 3s) - \frac{4}{3}\gamma s - \frac{4}{3}I_m\beta h + Ds'',\ 3h\rangle_{L^2} \\
&= G\big[\xi(x) + w'(x) - 3s(x)\big]\overline{z(x)}\Big|_0^1 + D\xi'(x)\overline{\varphi(x)}\Big|_0^1 + 3Ds'(x)\overline{h(x)}\Big|_0^1 \\
&- G\langle 3s - \xi - w',\ 3h - \varphi - z'\rangle_{L^2} + G\langle 3h - \varphi - z',\ 3s - \xi - w'\rangle_{L^2} \\
&- D\langle \xi',\ \varphi'\rangle_{L^2} + D\langle \varphi', \xi'\rangle_{L^2} + 3D\langle h', s'\rangle_{L^2} + 4\gamma\langle h, s\rangle_{L^2} \\
&- 3D\langle s',\ h'\rangle_{L^2} - 4\gamma\langle s, h\rangle_{L^2} - 4\beta I_m\langle h, h\rangle_{L^2} \\
&= -k_1 G|z(1)|^2 - k_2 D|\varphi(1)|^2 - G\langle 3s - \xi - w',\ 3h - \varphi - z'\rangle_{L^2} \\
&+ G\langle 3h - \varphi - z',\ 3s - \xi - w'\rangle_{L^2} - D\langle \xi',\ \varphi'\rangle_{L^2} + D\langle \varphi', \xi'\rangle_{L^2} + 3D\langle h', s'\rangle_{L^2} \\
&+ 4\gamma\langle h, s\rangle_{L^2} - 3D\langle s',\ h'\rangle_{L^2} - 4\gamma\langle s, h\rangle_{L^2} - 4\beta I_m\langle h, h\rangle_{L^2},
\end{aligned}$$

it follows that

$$\operatorname{Re}\big\langle \mathcal{A}[w,z,\xi,\varphi,s,h]^\top, [w,z,\xi,\varphi,s,h]^\top \big\rangle_{\mathcal{H}} = -k_1 G|z(1)|^2 - k_2 D|\varphi(1)|^2 - 4\beta I_m\|h\|_{L^2}^2 \le 0.$$

Hence, $\mathcal{A}$ is dissipative in $\mathcal{H}$. We accomplish the proof by showing that $0 \in \rho(\mathcal{A})$ because from Theorem 4.6 of [6], if $\mathcal{A}^{-1}$ exists, $\mathcal{A}$ must be densely defined in $\mathcal{H}$. Therefore, the Lumer–Phillips theorem can be applied to conclude that $\mathcal{A}$ generates a C_0-semigroup $e^{\mathcal{A}t}$ of contractions on $\mathcal{H}$.

To do so, for each $F := [u_1, u_2, \eta_1, \eta_2, v_1, v_2]^\top \in \mathcal{H}$, we seek $Y := [w, z, \xi, \varphi, s, h]^\top \in \mathcal{D}(\mathcal{A})$ such that

$$\mathcal{A}Y = F$$

which yields

$$\begin{cases} z = u_1, \quad G(\xi' + w'' - 3s') = mu_2, \\ \varphi = \eta_1, \quad G(3s - \xi - w') + D\xi'' = I_m\eta_2, \\ h = v_1, \quad 3G(\xi + w' - 3s) - 4\gamma s - 4\beta I_m h + 3Ds'' = 3I_m v_2, \\ \xi'(1) = -k_2\varphi(1), \quad s'(1) = 0, \\ 3s(1) - \xi(1) - w'(1) = k_1 z(1), \quad w(0) = \xi(0) = s(0) = 0. \end{cases} \tag{2.7}$$

From the first equation of (2.7), we have

$$G\big(\xi(x) + w'(x) - 3s(x)\big) = Gw'(0) + m\int_0^x u_2(r)dr. \tag{2.8}$$

By eliminating the term $G\big(\xi(x) + w'(x) - 3s(x)\big)$ from the second and the third equations of (2.7), it follows that

$$D\xi''(x) = I_m\eta_2(x) + Gw'(0) + m\int_0^x u_2(r)dr \tag{2.9}$$

and

$$3Ds''(x) - 4\gamma s(x) = 3I_m v_2(x) + 4\beta I_m v_1(x) - 3\left[Gw'(0) + m\int_0^x u_2(r)dr\right]. \tag{2.10}$$

A simple computation of (2.9), yields

$$\xi(x) = -k_2\eta_1(1)x - \frac{G}{D}w'(0)\left(x - \frac{x^2}{2}\right) - \widehat{\xi}(x), \tag{2.11}$$

where

$$\widehat{\xi}(x) := \frac{I_m}{D}\int_0^1 K_1(x,r)\eta_2(r)dr + \frac{m}{D}\int_0^1 K_2(x,r)u_2(r)dr \tag{2.12}$$

and

$$K_1(x,r) := \begin{cases} r, & 0 \le r < x, \\ x, & x \le r \le 1, \end{cases} \qquad K_2(x,r) := \begin{cases} x - \dfrac{x^2}{2} - \dfrac{r^2}{2}, & 0 \le r \le x, \\ x(1-r), & x \le r \le 1. \end{cases}$$

Similarly, it follows from (2.10) that

$$s(x) = a\sinh(bx) + \frac{G}{D}w'(0)\frac{1-\cosh(bx)}{b^2} + \frac{4\beta I_m}{3Db}\int_0^x \sinh(b(x-r))v_1(r)dr + \widehat{s}(x), \tag{2.13}$$

where a will be given later in (2.18), and

$$b := \sqrt{\frac{4\gamma}{3D}}, \quad \widehat{s}(x) := \frac{1}{b}\int_0^x \sinh(b(x-r))\left[\frac{I_m}{D}v_2(r) - \frac{m}{D}\int_0^r u_2(t)dt\right]dr. \tag{2.14}$$

Substitute (2.11) and (2.13) into (2.8), and integrate from 0 to x respect to x, to obtain

$$\begin{aligned} w(x) = {} & 3a\int_0^x \sinh(br)dr + w'(0)\left[\frac{3G}{Db^2}\int_0^x (1-\cosh(br))dr + \frac{G}{D}\left(\frac{x^2}{2} - \frac{x^3}{6}\right) + x\right] \\ & + \frac{4\beta I_m}{bD}\int_0^x (x-r)\sinh(b(x-r))v_1(r)dr - \frac{1}{2}\xi'(1)x^2 + \widehat{w}(x), \end{aligned} \tag{2.15}$$

where $\widehat{w}(x)$ is given by

$$\widehat{w}(x) := 3\int_0^x \widehat{s}(r)dr - \int_0^x \widehat{\xi}(r)dr + \frac{m}{G}\int_0^x (x-r)u_2(r)dr. \tag{2.16}$$

Using the boundary conditions $s'(1)=0$ in (2.13) and $3s(1)-\xi(1)-w'(1)=k_1u_1(1)$ in (2.8), respectively, we obtain that

$$\begin{cases} ab\cosh b - \dfrac{G}{D}w'(0)\dfrac{\sinh b}{b} + \dfrac{4\beta I_m}{3D}\displaystyle\int_0^1 \cosh(b(1-r))v_1(r)dr + \widehat{s}'(1) = 0, \\ Gw'(0) + m\displaystyle\int_0^1 u_2(r)dr = -k_1Gu_1(1). \end{cases} \tag{2.17}$$

Thus, a and $w'(0)$ in (2.13) and (2.15), respectively, can be obtained as follows:

$$\begin{cases} a = \dfrac{G}{D}w'(0)\dfrac{\sinh b}{b^2\cosh b} - \dfrac{4\beta I_m}{3Db\cosh b}\displaystyle\int_0^1 \cosh(b(1-r))v_1(r)dr - \dfrac{\widehat{s}'(1)}{b\cosh b}, \\ w'(0) = -\dfrac{m}{G}\displaystyle\int_0^1 u_2(r)dr - k_1u_1(1). \end{cases} \tag{2.18}$$

Hence, there is a solution $Y=[w,z,\xi,\varphi,s,h]^\top \in \mathcal{D}(\mathcal{A})$ so that $\mathcal{A}Y=F$, which in turn implies that $\mathcal{A}^{-1}$ exists. Finally, by the Sobolev embedding theorem, we can claim that $\mathcal{A}^{-1}$ is compact on $\mathcal{H}$ and thus the spectrum $\sigma(\mathcal{A})$ consists of isolated eigenvalues only (see [5]). The proof is complete. □

As a consequence of Theorem 2.1, we have the following corollary.

COROLLARY 2.2. *Let $\mathcal{A}$ be defined by* (2.4) *and* (2.5)*, and let $T(t)$ be a C_0-semigroup on $\mathcal{H}$ generated by $\mathcal{A}$. Then $T(t)$ is asymptotically stable in $\mathcal{H}$, i.e.,*

$$\lim_{t\to\infty}\|T(t)Y\| = 0 \quad \forall\, Y\in\mathcal{H}.$$

Proof. Since $T(t)$ is a C_0-semigroup of contractions on $\mathcal{H}$, the proof will be accomplished by showing that there is no eigenvalue on the imaginary axis (see [3, p. 130]). Assume that $\lambda=i\tau,\ \tau\in\mathbb{R}$ is an eigenvalue of $\mathcal{A}$ and $Y:=[w,z,\xi,\varphi,s,h]^\top \in \mathcal{D}(\mathcal{A})$ is an eigenfunction associated with λ. Then we have

$$z = i\tau w, \quad \varphi = i\tau\xi, \quad h = i\tau s,$$

and

$$\mathrm{Re}\langle \mathcal{A}Y,\ Y\rangle_{\mathcal{H}} = -k_1G|z(1)|^2 - k_2D|\varphi(1)|^2 - 4\beta I_m\|h\|_{L^2} \equiv 0.$$

Thus, it follows that

$$h(x) = i\tau s(x) \equiv 0, \quad z(1) = i\tau w(1) \equiv 0, \quad \varphi(1) = i\tau\xi(1) \equiv 0$$

and functions w and ξ satisfy the following equations:

$$\begin{cases} m\tau^2 w(x) + G\big(\xi'(x) + w''(x)\big) = 0, \quad 0<x<1, \\ I_m\tau^2\xi(x) - G\big(\xi(x) + w'(x)\big) + D\xi''(x) = 0, \quad 0<x<1, \\ \xi(x) + w'(x) = 0, \quad 0<x<1, \\ w(0) = \xi(0) = w(1) = \xi(1) = 0, \quad \xi'(1) = -k_2\varphi(1) = 0,\ w'(1) = -\xi(1) - k_1z(1) = 0. \end{cases} \tag{2.19}$$

By a direct computation, we obtain that (2.19) has a unique trivial solution only. Thus $w(x) = \xi(x) \equiv 0$ and hence $Y \equiv 0$, which contradicts that Y is an eigenfunction. Therefore, no eigenvalue exists on the imaginary axis. The proof is complete. □

COROLLARY 2.3. *If $k_1 = k_2 \equiv 0$ in* (1.6)*, that is, there is no control imposed on system* (1.3)*, then there is no eigenvalue on the imaginary axis. So the uncontrolled system* (1.3) *with its boundary conditions* (1.4) *is also asymptotically stable.*

Proof. Similar to the proof of Corollary 2.2, if $k_1 = k_2 \equiv 0$ and assume that $\lambda = i\tau,\ \tau \in \mathbb{R}$ is an eigenvalue with $Y := [w, z, \xi, \varphi, s, h]^\top$ being an eigenfunction, then it follows that $s \equiv 0$ and the functions w and ξ satisfy the following equations:

$$\begin{cases} m\tau^2 w(x) = 0, \quad 0 < x < 1, \\ I_m \tau^2 \xi(x) + D\xi''(x) = 0, \quad 0 < x < 1, \\ w(0) = \xi(0) = \xi'(1) = 0, \quad w'(1) = -\xi(1). \end{cases} \tag{2.20}$$

Therefore, one has $w = \xi \equiv 0$ and hence $Y \equiv 0$. The proof is complete. □

Note 2.1. We should note here that if $k_1 = k_2 \equiv 0$ in (1.6), then system (1.3) with its boundary conditions (1.4) cannot achieve the exponential stability. This is because of the fact that from the asymptotes of the system given later in (3.28), if $k_1 = k_2 \equiv 0$, then the eigenvalues of the first and second branches are very close to the imaginary axis as their moduli go to the infinity.

Let us now formulate the eigenvalue problem for the operator $\mathcal{A}$. If $\lambda \in \sigma(\mathcal{A})$ and $Y_\lambda := [w, z, \xi, \varphi, s, h]^\top \in \mathcal{D}(\mathcal{A})$ is a corresponding eigenfunction, then it is routine to verify that $\mathcal{A}Y_\lambda = \lambda Y_\lambda$ implies that $z = \lambda w,\ \varphi = \lambda\xi,\ h = \lambda s$ with w, ξ as well as s satisfying the following characteristic equations, for $0 < x < 1$:

$$\begin{cases} m\lambda^2 w(x) + G(3s' - \xi' - w'')(x) = 0, \\ I_m\lambda^2 \xi(x) - G(3s - \xi - w')(x) - D\xi''(x) = 0, \\ I_m\lambda^2 s(x) + G(3s - \xi - w')(x) + \frac{4}{3}\gamma s(x) + \frac{4}{3}\beta\lambda I_m s(x) - Ds''(x) = 0, \\ w(0) = 0, \quad \xi(0) = 0, \quad s(0) = 0, \\ \xi'(1) = -\lambda k_2 \xi(1), \quad s'(1) = 0, \quad 3s(1) - \xi(1) - w'(1) = \lambda k_1 w(1). \end{cases} \tag{2.21}$$

For brevity in notation, from now on, we define

$$r_1 := \sqrt{\frac{m}{G}}, \quad r_2 := \sqrt{\frac{I_m}{D}}, \quad d_1 := \frac{G}{D}, \quad d_2 := \frac{\gamma}{D}, \quad d_3 := 3d_1 + \frac{4}{3}d_2. \tag{2.22}$$

(2.21) then becomes

$$\begin{cases} r_1^2\lambda^2 w(x) + 3s'(x) - \xi'(x) - w''(x) = 0, \\ r_2^2\lambda^2 \xi(x) - 3d_1 s(x) + d_1\xi(x) + d_1 w'(x) - \xi''(x) = 0, \\ r_2^2\lambda^2 s(x) + d_3 s(x) - d_1\xi(x) - d_1 w'(x) + \frac{4}{3}\beta r_2^2\lambda s(x) - s''(x) = 0, \\ w(0) = 0, \quad \xi(0) = 0, \quad s(0) = 0, \\ \xi'(1) = -\lambda k_2\xi(1), \quad s'(1) = 0, \quad 3s(1) - \xi(1) - w'(1) = \lambda k_1 w(1). \end{cases} \tag{2.23}$$

Clearly, (2.23) is a coupled system of ordinary differential equations. In order to solve these equations, we shall use the matrix operator pencil method (see [8]). Let

$$w_1 := w, \quad w_2 := w', \quad \xi_1 := \xi, \quad \xi_2 := \xi', \quad s_1 := s, \quad s_2 := s' \tag{2.24}$$

and

$$\Phi := [w_1,\ w_2,\ \xi_1,\ \xi_2,\ s_1,\ s_2]^\top. \tag{2.25}$$

Then (2.23) becomes

$$\begin{cases} T^D(x,\lambda)\Phi(x) := \Phi'(x) + M(\lambda)\Phi(x) = 0, \\ T^R(x,\lambda)\Phi(x) := W^0(\lambda)\Phi(0) + W^1(\lambda)\Phi(1) = 0, \end{cases} \tag{2.26}$$

where

(2.27)

$$W^0(\lambda) := \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ & & O_{3\times 6} & & & \end{bmatrix}, \quad W^1(\lambda) := \begin{bmatrix} & & O_{3\times 6} & & & \\ 0 & 0 & \lambda k_2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ \lambda k_1 & 1 & 1 & 0 & -3 & 0 \end{bmatrix},$$

and

$$M(\lambda) := D_0 - \lambda D_1 - \lambda^2 D_2 \tag{2.28}$$

with D_0, D_1, and D_2 being three matrices defined by

$$D_0 := \begin{bmatrix} 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -3 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & -d_1 & -d_1 & 0 & 3d_1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & d_1 & d_1 & 0 & -d_3 & 0 \end{bmatrix}, \quad D_1 := \begin{bmatrix} O_{4\times 4} & O_{4\times 2} \\ O_{2\times 4} & D_{11} \end{bmatrix}, \tag{2.29}$$

$$D_2 := \begin{bmatrix} r_1^2 D_{21} & O_{2\times 2} & O_{2\times 2} \\ O_{2\times 2} & r_2^2 D_{21} & O_{2\times 2} \\ O_{2\times 2} & O_{2\times 2} & r_2^2 D_{21} \end{bmatrix}, \quad D_{11} := \begin{bmatrix} 0 & 0 \\ \frac{4}{3}\beta r_2^2 & 0 \end{bmatrix}, \quad D_{21} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}. \tag{2.30}$$

THEOREM 2.4. *The characteristic equation* (2.21) *is equivalent to the first order linear system* (2.26). *Also* $\lambda \in \sigma(\mathcal{A})$ *if and only if* (2.26) *has a nontrivial solution.*

3. Asymptotic behavior of eigenfrequencies. In this section, we are looking for the asymptotic expressions for the eigenvalues of $\mathcal{A}$. It will be accomplished by expanding the characteristic determinant $\Delta(\lambda)$ of (2.26) via an asymptotic expression of the fundamental matrix solution, which can be obtained by modifying a standard technique of Birkhoff–Langer (see [1]) and later of Tretter (see [8] or [9]) for tackling the matrix operator pencils. A key step is an invertible matrix transformation which is very powerful and universal in the sense that it can be applied to a lot of other coupled problems.

To begin, we shall diagonalize the leading term $\lambda^2 D_2$ in (2.28). For each $0 \neq \lambda \in \mathbb{C}$, define an invertible matrix in λ by

$$P(\lambda) := \begin{bmatrix} P_1(\lambda) & & \\ & P_2(\lambda) & \\ & & P_2(\lambda) \end{bmatrix}, \quad P_1(\lambda) := \begin{bmatrix} r_1\lambda & r_1\lambda \\ r_1^2\lambda^2 & -r_1^2\lambda^2 \end{bmatrix}, \tag{3.1}$$

$$P_2(\lambda) := \begin{bmatrix} r_2\lambda & r_2\lambda \\ r_2^2\lambda^2 & -r_2^2\lambda^2 \end{bmatrix}.$$

For any $\lambda \neq 0$, a simple computation shows that

$$P^{-1}(\lambda) := \begin{bmatrix} P_1^{-1}(\lambda) & & \\ & P_2^{-1}(\lambda) & \\ & & P_2^{-1}(\lambda) \end{bmatrix} \tag{3.2}$$

with

$$P_1^{-1}(\lambda) := \begin{bmatrix} \frac{1}{2r_1\lambda} & \frac{1}{2r_1^2\lambda^2} \\ \frac{1}{2r_1\lambda} & \frac{-1}{2r_1^2\lambda^2} \end{bmatrix}, \quad P_2^{-1}(\lambda) := \begin{bmatrix} \frac{1}{2r_2\lambda} & \frac{1}{2r_2^2\lambda^2} \\ \frac{1}{2r_2\lambda} & \frac{-1}{2r_2^2\lambda^2} \end{bmatrix}.$$

So matrix $P(\lambda)$ is a polynomial of degree 2 in λ. Define

$$\Psi(x) := P^{-1}(\lambda)\Phi(x), \quad \widehat{T}^D(x,\lambda) := P(\lambda)^{-1}T^D(x,\lambda)P(\lambda). \tag{3.3}$$

Then we have

$$\widehat{T}^D(x,\lambda)\Psi(x) = \Psi'(x) - \widehat{M}(\lambda)\Psi(x) = 0, \tag{3.4}$$

where

$$\begin{aligned}
\widehat{M}(\lambda) &= -P(\lambda)^{-1}M(\lambda)P(\lambda) \\
&= -P(\lambda)^{-1}\begin{bmatrix} 0 & -1 & 0 & 0 & 0 & 0 \\ -r_1^2\lambda^2 & 0 & 0 & 1 & 0 & -3 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & -d_1 & -d_1 - r_2^2\lambda^2 & 0 & 3d_1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & d_1 & d_1 & 0 & -d_3 - r_2^2(\frac{4}{3}\beta\lambda + \lambda^2) & 0 \end{bmatrix} P(\lambda) \\
&= -\begin{bmatrix} -\frac{1}{2} & \frac{-1}{2r_1\lambda} & 0 & \frac{1}{2r_1^2\lambda^2} & 0 & \frac{-3}{2r_1^2\lambda^2} \\ \frac{1}{2} & \frac{-1}{2r_1\lambda} & 0 & \frac{-1}{2r_1^2\lambda^2} & 0 & \frac{3}{2r_1^2\lambda^2} \\ 0 & \frac{-d_1}{2r_2^2\lambda^2} & \frac{-d_1}{2r_2^2\lambda^2} - \frac{1}{2} & \frac{-1}{2r_2\lambda} & \frac{3d_1}{2r_2^2\lambda^2} & 0 \\ 0 & \frac{d_1}{2r_2^2\lambda^2} & \frac{d_1}{2r_2^2\lambda^2} + \frac{1}{2} & \frac{-1}{2r_2\lambda} & \frac{-3d_1}{2r_2^2\lambda_2^2} & 0 \\ 0 & \frac{d_1}{2r_2^2\lambda^2} & \frac{d_1}{2r_2^2\lambda^2} & 0 & -\frac{1}{2} - \frac{2}{3}\frac{\beta}{\lambda} - \frac{d_3}{2r_2^2\lambda^2} & \frac{-1}{2r_2\lambda} \\ 0 & \frac{-d_1}{2r_2^2\lambda^2} & \frac{-d_1}{2r_2^2\lambda^2} & 0 & \frac{1}{2} + \frac{2}{3}\frac{\beta}{\lambda} + \frac{d_3}{2r_2^2\lambda^2} & \frac{-1}{2r_2\lambda} \end{bmatrix} P(\lambda) \\
&= -\begin{bmatrix} -r_1\lambda & 0 & \frac{1}{2}d_4 & -\frac{1}{2}d_4 & -\frac{3}{2}d_4 & \frac{3}{2}d_4 \\ 0 & r_1\lambda & -\frac{1}{2}d_4 & \frac{1}{2}d_4 & \frac{3}{2}d_4 & -\frac{3}{2}d_4 \\ -\frac{1}{2}d_6 & \frac{1}{2}d_6 & -r_2\lambda - \frac{d_7}{2\lambda} & -\frac{d_7}{2\lambda} & \frac{3d_7}{2\lambda} & \frac{3d_7}{2\lambda} \\ \frac{1}{2}d_6 & -\frac{1}{2}d_6 & \frac{d_7}{2\lambda} & r_2\lambda + \frac{d_7}{2\lambda} & -\frac{3d_7}{2\lambda} & -\frac{3d_7}{2\lambda} \\ \frac{1}{2}d_6 & -\frac{1}{2}d_6 & \frac{d_7}{2\lambda} & \frac{d_7}{2\lambda} & -r_2\lambda - \frac{2}{3}\beta r_2 - \frac{d_8}{2\lambda} & -\frac{2}{3}\beta r_2 - \frac{d_8}{2\lambda} \\ -\frac{1}{2}d_6 & \frac{1}{2}d_6 & -\frac{d_7}{2\lambda} & -\frac{d_7}{2\lambda} & \frac{2}{3}\beta r_2 + \frac{d_8}{2\lambda} & r_2\lambda + \frac{2}{3}\beta r_2 + \frac{d_8}{2\lambda} \end{bmatrix}
\end{aligned}$$

with

$$d_4 := \frac{r_2^2}{r_1^2}, \quad d_5 := \frac{r_1^2}{r_2^2}, \quad d_6 := d_1 d_5, \quad d_7 := \frac{d_1}{r_2}, \quad d_8 := \frac{d_3}{r_2}. \tag{3.5}$$

It is seen from the above that $\widehat{M}(\lambda)$ can be written as

$$\widehat{M}(\lambda) := \lambda\widehat{M}_1 + \widehat{M}_0 + \lambda^{-1}\widehat{M}_{-1}, \tag{3.6}$$

where

$$\widehat{M}_1 := \operatorname{diag}[r_1, -r_1, r_2, -r_2, r_2, -r_2] \tag{3.7}$$

and

$$\widehat{M}_0 := \begin{bmatrix} O_{2\times 2} & \frac{1}{2}d_4\widehat{M}_{01} & -\frac{3}{2}d_4\widehat{M}_{01} \\ -\frac{1}{2}d_6\widehat{M}_{01} & O_{2\times 2} & O_{2\times 2} \\ \frac{1}{2}d_6\widehat{M}_{01} & O_{2\times 2} & \frac{2}{3}\beta r_2\widehat{M}_{02} \end{bmatrix}, \quad \widehat{M}_{-1} := \begin{bmatrix} O_{2\times 2} & O_{2\times 4} \\ O_{4\times 2} & \widehat{M}_{-11} \end{bmatrix} \tag{3.8}$$

with

$$\widehat{M}_{01} := \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}, \quad \widehat{M}_{02} := \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}, \quad \widehat{M}_{-11} := \begin{bmatrix} \frac{1}{2}d_7\widehat{M}_{02} & -\frac{3}{2}d_7\widehat{M}_{02} \\ -\frac{1}{2}d_7\widehat{M}_{02} & \frac{1}{2}d_8\widehat{M}_{02} \end{bmatrix}.$$

On the basis of these transformations, we are now in a position to find an asymptotic expression for the fundamental matrix solution of system (3.4).

THEOREM 3.1. *Let $0 \neq \lambda \in \mathbb{C}$, and let $\widehat{M}(\lambda)$ be given by* (3.6) *and assume that $r_1 \neq r_2$. For $x \in [0,1]$, set*

$$E(x,\lambda) := \operatorname{diag}\left[e^{r_1\lambda x},\ e^{-r_1\lambda x},\ e^{r_2\lambda x},\ e^{-r_2\lambda x},\ e^{r_2\lambda x},\ e^{-r_2\lambda x}\right]. \tag{3.9}$$

Then there exists a fundamental matrix solution $\widehat{\Psi}(x,\lambda)$ for system (3.4)*, which satisfies*

$$\Psi'(x) = \widehat{M}(\lambda)\Psi(x) \tag{3.10}$$

such that for large enough $|\lambda|$,

$$\widehat{\Psi}(x,\lambda) = \left(\widehat{\Psi}_0(x) + \frac{\widetilde{\Theta}(x,\lambda)}{\lambda}\right) E(x,\lambda), \tag{3.11}$$

where

$$\widehat{\Psi}_0(x) := \operatorname{diag}[1,\ 1,\ 1,\ 1,\ e_1(x),\ e_2(x)] \tag{3.12}$$

and

$$\widetilde{\Theta}(x,\lambda) := \widehat{\Psi}_1(x) + \lambda^{-1}\widehat{\Psi}_2(x) + \cdots \tag{3.13}$$

with all entries uniformly bounded in $[0,1]$. Here,

$$e_1(x) := e^{\frac{2}{3}\beta r_2 x} \quad \text{and} \quad e_2(x) := e^{-\frac{2}{3}\beta r_2 x}. \tag{3.14}$$

Proof. Since $\widehat{M}_1$ given by (3.7) is a diagonal matrix, it follows that $E(x,\lambda)$ given by (3.9) is a fundamental matrix solution to (3.10) which involves only the leading order terms, that is, to say

$$E'(x,\lambda) = \lambda\widehat{M}_1 E(x,\lambda).$$

Now we look for a fundamental matrix solution of (3.10) in the form of

$$\widehat{\Psi}(x,\lambda)=\left(\widehat{\Psi}_0(x)+\lambda^{-1}\widehat{\Psi}_1(x)+\cdots+\lambda^{-n}\widehat{\Psi}_n(x)+\cdots\right)E(x,\lambda).$$

The left-hand side of (3.10) is

$$\begin{aligned}\widehat{\Psi}'(x,\lambda)=&\left(\widehat{\Psi}_0'(x)+\lambda^{-1}\widehat{\Psi}_1'(x)+\cdots+\lambda^{-n}\widehat{\Psi}_n'(x)+\cdots\right)E(x,\lambda)\\&+\lambda\left(\widehat{\Psi}_0(x)+\lambda^{-1}\widehat{\Psi}_1(x)+\cdots+\lambda^{-n}\widehat{\Psi}_n(x)+\cdots\right)\widehat{M}_1E(x,\lambda).\end{aligned}$$

Compare it with the right-hand side of (3.10),

$$\big(\lambda\widehat{M}_1+\widehat{M}_0+\lambda^{-1}\widehat{M}_{-1}\big)\big(\widehat{\Psi}_0(x)+\lambda^{-1}\widehat{\Psi}_1(x)+\cdots+\lambda^{-n}\widehat{\Psi}_n(x)+\cdots\big)E(x,\lambda),$$

to give, according to the coefficients of $\lambda^1,\lambda^0,\lambda^{-1},\ldots,\lambda^{-n},\ldots$, that

$$\begin{gathered}\widehat{\Psi}_0(x)\widehat{M}_1-\widehat{M}_1\widehat{\Psi}_0(x)=0,\\ \widehat{\Psi}_0'(x)-\widehat{M}_0\widehat{\Psi}_0(x)+\widehat{\Psi}_1(x)\widehat{M}_1-\widehat{M}_1\widehat{\Psi}_1(x)=0,\\ \widehat{\Psi}_1'(x)-\widehat{M}_0\widehat{\Psi}_1(x)-\widehat{M}_{-1}\widehat{\Psi}_0(x)+\widehat{\Psi}_2(x)\widehat{M}_1-\widehat{M}_1\widehat{\Psi}_2(x)=0,\\ \vdots\\ \widehat{\Psi}_n'(x)-\widehat{M}_0\widehat{\Psi}_n(x)-\widehat{M}_{-1}\widehat{\Psi}_{n-1}(x)+\widehat{\Psi}_{n+1}(x)\widehat{M}_1-\widehat{M}_1\widehat{\Psi}_{n+1}(x)=0,\\ \vdots\end{gathered}$$

Using the arguments in [8, p. 135] (or [1]), we conclude that there is an asymptotic fundamental matrix solution $\widehat{\Psi}(x,\lambda)$ for system (3.10). It remains to show that the leading order term $\widehat{\Psi}_0(x)$ is given by (3.12). Indeed, since $\widehat{\Psi}_0(x)$ can be determined by the matrix equations

$$\widehat{\Psi}_0(x)\widehat{M}_1-\widehat{M}_1\widehat{\Psi}_0(x)=0 \tag{3.15}$$

and

$$\widehat{\Psi}_0'(x)-\widehat{M}_0\widehat{\Psi}_0(x)+\widehat{\Psi}_1(x)\widehat{M}_1-\widehat{M}_1\widehat{\Psi}_1(x)=0, \tag{3.16}$$

where $\widehat{M}_1$ and $\widehat{M}_0$ are given in (3.7), (3.8), respectively, it follows that if $\widehat{\Psi}_0$ is known, then one can deduce the leading order term $\widehat{\Psi}_1$ of $\widetilde{\Theta}(x,\rho)$ in (3.13) from (3.16) and

$$\widehat{\Psi}_1'(x)-\widehat{M}_0\widehat{\Psi}_1(x)-\widehat{M}_{-1}\widehat{\Psi}_0(x)+\widehat{\Psi}_2(x)\widehat{M}_1-\widehat{M}_1\widehat{\Psi}_2(x)=0$$

with $\widehat{M}_{-1}$ being given in (3.8). Similarly, we obtain all the terms $\widehat{\Psi}_1,\widehat{\Psi}_2,\ldots,\widehat{\Psi}_n,\ldots$ of $\widetilde{\Theta}(x,\lambda)$ in (3.13). So, the proof will be accomplished if we would find the leading order term $\widehat{\Psi}_0$ in (3.11).

Let us denote by $c_{ij}(x)$ the (i,j)-entry of the matrix $\widehat{\Psi}_0(x)$ with $i,j=1,2,\ldots,6$. Since $\widehat{M}_1$ is diagonal, it follows from (3.15) and $r_1\neq r_2$ that the entries $c_{ij}(x)$ of $\widehat{\Psi}_0$ satisfy

$$\begin{cases}c_{ij}(x)=0 & \text{if}\quad 1\le i\le 2,\ 1\le j\le 6,\ i\neq j,\\ c_{ij}(x)=0 & \text{if}\quad 3\le i\le 4,\ 1\le j\le 6,\ i\neq j,\ j\neq i+2,\\ c_{ij}(x)=0 & \text{if}\quad 5\le i\le 6,\ 1\le j\le 6,\ i\neq j,\ j\neq i-2,\end{cases}$$

and the entries $c_{ii}(x)$ $(i = 1, 2, \ldots, 6)$, $c_{35}(x)$, $c_{53}(x)$, $c_{46}(x)$, and $c_{64}(x)$ can be found by substituting them into (3.16) to obtain

$$\begin{cases} c'_{ii}(x) = 0 \quad \text{for } \ i = 1, 2, 3, 4, \\ c'_{55}(x) = \frac{2}{3}\beta r_2 c_{55}(x), \quad c'_{66}(x) = -\frac{2}{3}\beta r_2 c_{66}(x), \\ c'_{35}(x) = 0, \quad c'_{53}(x) = \frac{2}{3}\beta r_2 c_{53}(x), \\ c'_{46}(x) = 0, \quad c'_{64}(x) = -\frac{2}{3}\beta r_2 c_{64}(x). \end{cases} \tag{3.17}$$

(3.12) then follows from $\widehat{\Psi}_0(0) = I$. The proof is complete. □

By virtue of transformation for $\widehat{\Psi}(x, \lambda)$ in (3.3), we have immediately the following corollary, which shows the relationship between (2.26) and (3.4).

COROLLARY 3.2. *Let $0 \neq \lambda \in \mathbb{C}$, let $r_1 \neq r_2$, and let $\widehat{\Psi}(x, \lambda)$ given by* (3.11) *be a fundamental matrix solution of system* (3.4). *Then*

$$\widehat{\Phi}(x, \lambda) := P(\lambda)\widehat{\Psi}(x, \lambda) \tag{3.18}$$

is a fundamental matrix solution for the first order linear system (2.26).

We are now ready to estimate the asymptotic eigenfrequencies of the system. Note that the eigenvalues of the first order linear system in (2.26) are given by the zeros of the characteristic determinant

$$\Delta(\lambda) := \det\left(T^R \widehat{\Phi}(x, \lambda)\right), \quad \lambda \in \mathbb{C}, \tag{3.19}$$

where operator T^R is given in (2.26) and $\widehat{\Phi}(x, \lambda)$ is any fundamental matrix of $T^D(x, \lambda)\Phi(x) = 0$ (see [8]). We shall derive the asymptotic expansion of eigenfrequencies by substituting (3.11) and (3.18) into (3.19), together with the boundary conditions in (2.26). In fact, since

$$T^R \widehat{\Phi}(x, \lambda) = W^0(\lambda)P(\lambda)\widehat{\Psi}(0, \lambda) + W^1(\lambda)P(\lambda)\widehat{\Psi}(1, \lambda), \tag{3.20}$$

using (2.27) and (3.1), a simple computation gives

$$W^0(\lambda)P(\lambda) = \begin{bmatrix} r_1\lambda & r_1\lambda & 0 & 0 & 0 & 0 \\ 0 & 0 & r_2\lambda & r_2\lambda & 0 & 0 \\ 0 & 0 & 0 & 0 & r_2\lambda & r_2\lambda \\ & & & O_{3\times 6} & & \end{bmatrix}$$

and

$$W^1(\lambda)P(\lambda) = \begin{bmatrix} & & & O_{3\times 6} & & \\ 0 & 0 & r_2 r_3\lambda^2 & r_2 r_4\lambda^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & r_2^2\lambda^2 & -r_2^2\lambda^2 \\ r_1 r_5\lambda^2 & r_1 r_6\lambda^2 & r_2\lambda & r_2\lambda & -3r_2\lambda & -3r_2\lambda \end{bmatrix},$$

where

$$r_3 := k_2 + r_2, \quad r_4 := k_2 - r_2, \quad r_5 := k_1 + r_1, \quad r_6 := k_1 - r_1. \tag{3.21}$$

Once again for notational simplicity, set

$$[a]_1 := a + \mathcal{O}(\lambda^{-1}).$$

Since $\widehat{\Psi}_0(0) = I$ and $E(0,\lambda) = I$, a direct computation yields

$$W^0(\lambda)P(\lambda)\widehat{\Psi}(0,\lambda) = \begin{bmatrix} \lambda[r_1]_1 & \lambda[r_1]_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda[r_2]_1 & \lambda[r_2]_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda[r_2]_1 & \lambda[r_2]_1 \\ & & O_{3\times 6} & & & \end{bmatrix}$$

and

$$W^1(\lambda)P(\lambda)\widehat{\Psi}(1,\lambda) = \begin{bmatrix} & & O_{3\times 6} & & & \\ 0 & 0 & \lambda^2 E_3[r_2r_3]_1 & \lambda^2 E_4[r_2r_4]_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & [r_2^2]_1\lambda^2 E_3E_5 & -[r_2^2]_1\lambda^2 E_4E_6 \\ \lambda^2 E_1[r_1r_5]_1 & \lambda^2 E_2[r_1r_6]_1 & \lambda E_3[r_2]_1 & \lambda E_4[r_2]_1 & -3\lambda E_3E_5[r_2]_1 & -3\lambda E_4E_6[r_2]_1 \end{bmatrix},$$

where

$$\begin{cases} E_1 := e^{r_1\lambda}, \quad E_2 := e^{-r_1\lambda}, \quad E_3 := e^{r_2\lambda}, \quad E_4 := e^{-r_2\lambda}, \\ E_5 := e_1(1) = e^{\frac{2}{3}\beta r_2}, \quad E_6 := e_2(1) = e^{-\frac{2}{3}\beta r_2}. \end{cases} \tag{3.22}$$

Hence,

$$T^R\widehat{\Phi}(x,\lambda) = \begin{bmatrix} \lambda[r_1]_1 & \lambda[r_1]_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda[r_2]_1 & \lambda[r_2]_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda[r_2]_1 & \lambda[r_2]_1 \\ 0 & 0 & \lambda^2 E_3[r_2r_3]_1 & \lambda^2 E_4[r_2r_4]_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & [r_2^2]_1\lambda^2 E_3E_5 & -[r_2^2]_1\lambda^2 E_4E_6 \\ \lambda^2 E_1[r_1r_5]_1 & \lambda^2 E_2[r_1r_6]_1 & \lambda E_3[r_2]_1 & \lambda E_4[r_2]_1 & -3\lambda E_3E_5[r_2]_1 & -3\lambda E_4E_6[r_2]_1 \end{bmatrix}.$$

Therefore,

$$\begin{aligned} \Delta(\lambda) = \det(T^R\widehat{\Phi}(x,\lambda)) &= \lambda^9 r_1^2 r_2^5 \times \det\begin{bmatrix} [1]_1 & [1]_1 \\ [r_5]_1E_1 & [r_6]_1E_2 \end{bmatrix} \\ &\quad \times \det\begin{bmatrix} [1]_1 & [1]_1 \\ [r_3]_1E_3 & [r_4]_1E_4 \end{bmatrix} \times \det\begin{bmatrix} [1]_1 & [1]_1 \\ -E_3E_5[1]_1 & E_4E_6[1]_1 \end{bmatrix} \\ &= r_1^2 r_2^5 \lambda^9 \Delta_1(\lambda)\Delta_2(\lambda)\Delta_3(\lambda), \end{aligned}$$

where

$$\begin{cases} \Delta_1(\lambda) := r_6E_2 - r_5E_1 + \mathcal{O}(\lambda^{-1}), \\ \Delta_2(\lambda) := r_4E_4 - r_3E_3 + \mathcal{O}(\lambda^{-1}), \\ \Delta_2(\lambda) := E_4E_6 + E_3E_5 + \mathcal{O}(\lambda^{-1}) \end{cases} \tag{3.23}$$

with r_i $(i = 3,4,5,6)$ being given in (3.21) and E_i $(i = 1,2,\dots,6)$ in (3.22), respectively. With all these preparations, we come to the proof of the asymptotic behavior of the eigenvalues.

THEOREM 3.3. *Let $r_1 \neq r_2$ and let $\Delta(\lambda)$ be the characteristic determinant of system (2.26). Then the following asymptotic expansion for $\Delta(\lambda)$ holds:*

$$\Delta(\lambda) = r_1^2 r_2^5 \lambda^9 \Delta_1(\lambda)\Delta_2(\lambda)\Delta_3(\lambda) \tag{3.24}$$

with $\Delta_i(\lambda)$ being given in (3.23). If $k_i \neq r_i$ $(i = 1, 2)$, then there are three branches of asymptotic eigenvalues given by (as $|n| \to \infty$ and $n \in \mathbb{Z}$)

$$\begin{cases} \lambda_{jn} = \mu_j + r_j^{-1} n\pi i + \mathcal{O}(n^{-1}) & \text{for } j = 1, 2, \\ \lambda_{3n} = \mu_3 + r_2^{-1}(n + \frac{1}{2})\pi i + \mathcal{O}(n^{-1}), \end{cases} \tag{3.25}$$

where

$$\mu_j := \begin{cases} \frac{1}{2r_j} \ln \frac{k_j - r_j}{k_j + r_j}, & k_j > r_j \\ \frac{1}{2r_j} \left(\ln \frac{r_j - k_j}{k_j + r_j} + \pi i \right), & k_j < r_j \end{cases} \qquad \text{for } j = 1, 2 \tag{3.26}$$

and

$$\mu_3 := -\frac{2}{3}\beta. \tag{3.27}$$

Moreover, we have, as $|n| \to \infty$,

$$\mathrm{Re}\lambda_{jn} \to \frac{1}{2r_j} \ln \left| \frac{k_j - r_j}{k_j + r_j} \right| < 0 \ \text{ for } j = 1, 2 \quad \text{and} \quad \mathrm{Re}\lambda_{3n} \to \mu_3 < 0. \tag{3.28}$$

Furthermore, if k_1 and k_2 satisfy the conditions

$$k_1 \neq \begin{cases} \dfrac{\alpha_1 + 1}{1 - \alpha_1} r_1 & \text{for } k_1 > r_1, \\ \dfrac{1 - \alpha_1}{\alpha_1 + 1} r_1 & \text{for } k_1 < r_1, \end{cases} \qquad \alpha_1 := \left| \frac{k_2 - r_2}{k_2 + r_2} \right|^{r_1/r_2}, \quad 0 < \alpha_1 < 1, \tag{3.29}$$

and

$$k_1 \neq \begin{cases} \dfrac{\alpha_2 + 1}{1 - \alpha_2} r_1 & \text{for } k_1 > r_1, \\ \dfrac{1 - \alpha_2}{\alpha_2 + 1} r_1 & \text{for } k_1 < r_1, \end{cases} \qquad \alpha_2 := e^{-\frac{4}{3}\beta r_1}, \quad 0 < \alpha_2 < 1, \tag{3.30}$$

then the zeros of $\Delta(\lambda)$ are simple when their moduli are sufficiently large.

Proof. By $\Delta(\lambda) = 0$ and (3.24), it follows that

$$\Delta_1(\lambda)\Delta_2(\lambda)\Delta_3(\lambda) = 0 \tag{3.31}$$

and

$$\Delta_i(\lambda) = 0 \quad \text{for } i = 1, 2, 3.$$

Let $\Delta_1(\lambda) = 0$. Then we obtain

$$r_6 E_2 - r_5 E_1 + \mathcal{O}(\lambda^{-1}) = 0, \tag{3.32}$$

which is equivalent to (from (3.21) and (3.23))

$$(k_1 - r_1)e^{-r_1\lambda} - (k_1 + r_1)e^{r_1\lambda} + \mathcal{O}(\lambda^{-1}) = 0. \tag{3.33}$$

Since the solutions of the equation

$$(k_1 - r_1)e^{-r_1\lambda} - (k_1 + r_1)e^{r_1\lambda} = 0$$

are given by

$$\tilde{\lambda}_{1n} = \mu_1 + r_1^{-1} n\pi i, \quad n \in \mathbb{Z},$$

it follows from Rouché's theorem that the solutions to (3.33) are in the form

$$\lambda_{1n} = \tilde{\lambda}_{1n} + \mathcal{O}(n^{-1}) = \mu_1 + r_1^{-1} n\pi i + \mathcal{O}(n^{-1}), \quad n \in \mathbb{Z} \text{ and } |n| \to \infty. \tag{3.34}$$

Similarly, let $\Delta_2(\lambda) = 0$. Then the equation

$$(k_2 - r_2)e^{-r_2\lambda} - (k_2 + r_2)e^{r_2\lambda} + \mathcal{O}(\lambda^{-1}) = 0 \tag{3.35}$$

has the solutions

$$\lambda_{2n} = \mu_2 + r_2^{-1} n\pi i + \mathcal{O}(n^{-1}), \quad n \in \mathbb{Z} \text{ and } |n| \to \infty. \tag{3.36}$$

Also, let $\Delta_3(\lambda) = 0$. The equation

$$e_2(1)e^{-r_2\lambda} + e_1(1)e^{r_2\lambda} + \mathcal{O}(\lambda^{-1}) = 0 \tag{3.37}$$

has the solutions

$$\lambda_{3n} = \mu_3 + r_2^{-1}\Big(n + \frac{1}{2}\Big)\pi i + \mathcal{O}(n^{-1}), \quad n \in \mathbb{Z} \text{ and } |n| \to \infty. \tag{3.38}$$

Finally, by a direct computation, it follows from (3.29) and (3.30) that k_1 and k_2 satisfy the following conditions:

$$\frac{1}{r_1}\ln\left|\frac{k_1 - r_1}{k_1 + r_1}\right| \neq \frac{1}{r_2}\ln\left|\frac{k_2 - r_2}{k_2 + r_2}\right|, \quad \frac{1}{2r_1}\ln\left|\frac{k_1 - r_1}{k_1 + r_1}\right| \neq -\frac{2}{3}\beta.$$

Thus $\mu_1 \neq \mu_2$ and $\mu_1 \neq \mu_3$. The last assertion is then concluded. The proof is complete. □

THEOREM 3.4. *Suppose $r_1 \neq r_2$ and $k_i \neq r_i$ $(i = 1, 2)$. Let $\mathcal{A}$ be defined by* (2.4) *and* (2.5). *Then all eigenvalues of $\mathcal{A}$ have the asymptotic expressions given by* (3.25). *Moreover, if k_1 and k_2 satisfy conditions* (3.29) *and* (3.30), *then all eigenvalues of the system with sufficiently large moduli are simple.*

4. Asymptotic behavior of eigenfunctions. In this section, we shall consider the asymptotic behavior for eigenfunctions of $\mathcal{A}$. It will be used in the proof of the Riesz basis in the last section.

THEOREM 4.1. *Suppose $r_1 \neq r_2$ and $k_i \neq r_i$ $(i = 1, 2)$. Let $\sigma(\mathcal{A}) := \{\lambda_{1n}, \lambda_{2n}, \lambda_{3n}, n \in \mathbb{Z}\}$ be the eigenvalues of $\mathcal{A}$ with λ_{jn} $(j = 1, 2, 3)$ being given in* (3.25). *Then the corresponding eigenfunctions*

$$\Big\{[w_{jn}, \lambda_{jn}w_{jn}, \xi_{jn}, \lambda_{jn}\xi_{jn}, s_{jn}, \lambda_{jn}s_{jn}]^\top, \quad j = 1, 2, 3,\ n \in \mathbb{Z}\Big\}$$

have the following asymptotic expressions for $|n| \to \infty,\ n \in \mathbb{Z}$*:*

(4.1)
$$\begin{cases} w'_{1n}(x) = \frac{1}{2}(e^{-r_1\lambda_{1n}x} + e^{r_1\lambda_{1n}x}) + \mathcal{O}(n^{-1}), \quad w'_{jn}(x) = \mathcal{O}(n^{-1}) \ \text{for}\ j = 2,3, \\ \lambda_{1n}w_{1n}(x) = \frac{1}{2}r_1^{-1}(e^{r_1\lambda_{1n}x} - e^{-r_1\lambda_{1n}x}) + \mathcal{O}(n^{-1}), \ \lambda_{jn}w_{jn}(x) = \mathcal{O}(n^{-1}) \ \text{for}\ j = 2,3, \\ \xi'_{2n}(x) = \frac{1}{2}(e^{-r_2\lambda_{2n}x} + e^{r_2\lambda_{2n}x}) + \mathcal{O}(n^{-1}), \quad \xi'_{jn}(x) = \mathcal{O}(n^{-1}) \ \text{for}\ j = 1,3, \\ \lambda_{2n}\xi_{2n}(x) = \frac{1}{2}r_2^{-1}(e^{r_2\lambda_{2n}x} - e^{-r_2\lambda_{2n}x}) + \mathcal{O}(n^{-1}), \quad \lambda_{jn}\xi_{jn}(x) = \mathcal{O}(n^{-1}) \ \text{for}\ j = 1,3, \\ s'_{3n}(x) = \frac{1}{2}(e_2(x)e^{-r_2\lambda_{3n}x} + e_1(x)e^{r_2\lambda_{3n}x}) + \mathcal{O}(n^{-1}), \ s'_{jn}(x) = \mathcal{O}(n^{-1}) \ \text{for}\ j = 1,2, \\ \lambda_{3n}s_{3n}(x) = \frac{1}{2}r_2^{-1}(e_1(x)e^{r_2\lambda_{3n}x} - e_2(x)e^{-r_2\lambda_{3n}x}) + \mathcal{O}(n^{-1}), \\ \lambda_{jn}s_{jn}(x) = \mathcal{O}(n^{-1}) \quad \text{for}\ j = 1,2, \end{cases}$$

where r_1, r_2 *are given in* (2.22) *and* $e_1(x), e_2(x)$ *are given in* (3.14)*, respectively. Moreover,* $\{[w_{jn}, \lambda_{jn}w_{jn}, \xi_{jn}, \lambda_{jn}\xi_{jn}, s_{jn}, \lambda_{jn}s_{jn}]^\top (j = 1,2,3,\ n \in \mathbb{Z})\}$ *are approximately normalized in* $\mathcal{H}$ *in the sense that there exist positive constants* c_1 *and* c_2 *independent of* n *such that* $(j = 1,2,3)$

$$c_1 \le \|w'_{jn}\|_{L^2}, \|\lambda_{jn}w_{jn}\|_{L^2}, \|\xi'_{jn}\|_{L^2}, \|\lambda_{jn}\xi_{jn}\|_{L^2}, \|s'_{jn}\|_{L^2}, \|\lambda_{jn}s_{jn}\|_{L^2} \le c_2 \tag{4.2}$$

for all integers n*.*

Proof. Note that the jth component of $\Phi(x) = [w_1(x), w_2(x), \xi_1(x), \xi_2(x), s_1(x), s_2(x)]^\top$ in (2.25) with respect to the eigenvalue λ can be obtained by taking the determinant of the matrices which are replaced one of the rows of $T^R\widehat{\Phi}$ in (3.20) by $e_j^\top(\widehat{\Phi}(x,\lambda))$ so that their determinants are not zero, where e_j is the jth column of the identity matrix. Indeed, we have from (3.18) that $\widehat{\Phi}(x,\lambda) = P(\lambda)\widehat{\Psi}(x,\lambda)$ and hence

$$\widehat{\Phi}(x,\lambda) = \begin{bmatrix} \widehat{\Phi}_{11}(x,\lambda) & O_{2\times 2} & O_{2\times 2} \\ O_{2\times 2} & \widehat{\Phi}_{22}(x,\lambda) & O_{2\times 2} \\ O_{2\times 2} & O_{2\times 2} & \widehat{\Phi}_{33}(x,\lambda) \end{bmatrix}, \tag{4.3}$$

where

$$\widehat{\Phi}_{ii}(x,\lambda) := \begin{bmatrix} r_i\lambda e^{r_i\lambda x}[1 + \mathcal{O}(\lambda^{-1})] & r_i\lambda e^{-r_i\lambda x}[1 + \mathcal{O}(\lambda^{-1})] \\ r_i^2\lambda^2 e^{r_i\lambda x}[1 + \mathcal{O}(\lambda^{-1})] & -r_i^2\lambda^2 e^{-r_i\lambda x}[1 + \mathcal{O}(\lambda^{-1})] \end{bmatrix} \quad \text{for}\ \ i = 1,2 \tag{4.4}$$

and

$$\widehat{\Phi}_{33}(x,\lambda) := \begin{bmatrix} r_2\lambda e^{r_2\lambda x}e_1(x)[1 + \mathcal{O}(\lambda^{-1})] & r_2\lambda e^{-r_2\lambda x}e_2(x)[1 + \mathcal{O}(\lambda^{-1})] \\ r_2^2\lambda^2 e^{r_2\lambda x}e_1(x)[1 + \mathcal{O}(\lambda^{-1})] & -r_2^2\lambda^2 e^{-r_2\lambda x}e_2(x)[1 + \mathcal{O}(\lambda^{-1})] \end{bmatrix} \tag{4.5}$$

with $e_i(x)$ $(i = 1,2)$ being given in (3.14).

Thus, the first component of $\Phi(x)$ is given by

$$\begin{aligned} w_1(x,\lambda) &= r_1^{-2} r_2^{-5} \lambda^{-8} \det \begin{bmatrix} \lambda[r_1]_1 & \lambda[r_1]_1 \\ \lambda e^{r_1\lambda x}[r_1]_1 & \lambda e^{-r_1\lambda x}[r_1]_1 \end{bmatrix} \\ &\quad \times \det \begin{bmatrix} \lambda[r_2]_1 & \lambda[r_2]_1 \\ \lambda^2 E_3[r_2 r_3]_1 & \lambda^2 E_4[r_2 r_4]_1 \end{bmatrix} \times \det \begin{bmatrix} \lambda[r_2]_1 & \lambda[r_2]_1 \\ -\lambda^2 E_3 E_5[r_2^2]_1 & \lambda^2 E_4 E_6[r_2^2]_1 \end{bmatrix} \\ &= \left(e^{-r_1\lambda x} - e^{r_1\lambda x} + \mathcal{O}(\lambda^{-1})\right) \left(r_4 e^{-r_2\lambda} - r_3 e^{r_2\lambda} + \mathcal{O}(\lambda^{-1})\right) \\ &\quad \times \left(e_2(1) e^{-r_2\lambda} + e_1(1) e^{r_2\lambda} + \mathcal{O}(\lambda^{-1})\right). \end{aligned}$$

By (3.23), (3.25), and (3.31), we conclude that

$$w_1(x,\lambda) = \begin{cases} r_7(\lambda)\left(e^{-r_1\lambda x} - e^{r_1\lambda x} + \mathcal{O}(\lambda^{-1})\right) & \text{if } \lambda = \lambda_{1n}, \\ \mathcal{O}(\lambda^{-1}) & \text{if } \lambda = \lambda_{2n} \text{ or } \lambda_{3n}, \end{cases} \tag{4.6}$$

where $r_7(\lambda)$ is bounded in λ and has the form

$$r_7(\lambda) := \left(r_4 e^{-r_2\lambda} - r_3 e^{r_2\lambda}\right)\left(e_2(1) e^{-r_2\lambda} + e_1(1) e^{r_2\lambda}\right). \tag{4.7}$$

Similarly, we have

$$\begin{aligned} w_2(x,\lambda) &= r_1^{-2} r_2^{-5} \lambda^{-8} \det \begin{bmatrix} \lambda[r_1]_1 & \lambda[r_1]_1 \\ \lambda^2 e^{r_1\lambda x}[r_1^2]_1 & -\lambda^2 e^{-r_1\lambda x}[r_1^2]_1 \end{bmatrix} \\ &\quad \times \det \begin{bmatrix} \lambda[r_2]_1 & \lambda[r_2]_1 \\ \lambda^2 E_3[r_2 r_3]_1 & \lambda^2 E_4[r_2 r_4]_1 \end{bmatrix} \times \det \begin{bmatrix} \lambda[r_2]_1 & \lambda[r_2]_1 \\ -\lambda^2 E_3 E_5[r_2^2]_1 & \lambda^2 E_4 E_6[r_2^2]_1 \end{bmatrix} \\ &= -r_1\lambda \left(e^{-r_1\lambda x} + e^{r_1\lambda x} + \mathcal{O}(\lambda^{-1})\right) \left(r_4 e^{-r_2\lambda} - r_3 e^{r_2\lambda} + \mathcal{O}(\lambda^{-1})\right) \\ &\quad \times \left(e_2(1) e^{-r_2\lambda} + e_1(1) e^{r_2\lambda} + \mathcal{O}(\lambda^{-1})\right) \end{aligned}$$

and

$$w_2(x,\lambda) = \begin{cases} -\lambda r_1 r_7(\lambda)\left(e^{-r_1\lambda x} + e^{r_1\lambda x} + \mathcal{O}(\lambda^{-1})\right) & \text{if } \lambda = \lambda_{1n}, \\ r_1\lambda[\mathcal{O}(\lambda^{-1})] & \text{if } \lambda = \lambda_{2n} \text{ or } \lambda_{3n}. \end{cases} \tag{4.8}$$

Also, along the same line,

$$\begin{aligned} w_3(x,\lambda) &= r_1^{-2} r_2^{-5} \lambda^{-8} \det \begin{bmatrix} \lambda[r_1]_1 & \lambda[r_1]_1 \\ \lambda^2 E_1[r_1 r_5]_1 & \lambda^2 E_2[r_1 r_6]_1 \end{bmatrix} \\ &\quad \times \det \begin{bmatrix} \lambda[r_2]_1 & \lambda[r_2]_1 \\ \lambda e^{r_2\lambda x}[r_2]_1 & \lambda e^{-r_2\lambda x}[r_2]_1 \end{bmatrix} \times \det \begin{bmatrix} \lambda[r_2]_1 & \lambda[r_2]_1 \\ -\lambda^2 E_3 E_5[r_2^2]_1 & \lambda^2 E_4 E_6[r_2^2]_1 \end{bmatrix} \\ &= \left(r_6 e^{-r_1\lambda} - r_5 e^{r_1\lambda} + \mathcal{O}(\lambda^{-1})\right) \left(e^{-r_2\lambda x} - e^{r_2\lambda x} + \mathcal{O}(\lambda^{-1})\right) \\ &\quad \times \left(e_2(1) e^{-r_2\lambda} + e_1(1) e^{r_2\lambda} + \mathcal{O}(\lambda^{-1})\right) \end{aligned}$$

and from (3.23), (3.25), and (3.31), we obtain that

$$w_3(x,\lambda) = \begin{cases} r_8(\lambda)\left(e^{-r_2\lambda x} - e^{r_2\lambda x} + \mathcal{O}(\lambda^{-1})\right) & \text{if } \lambda = \lambda_{2n}, \\ \mathcal{O}(\lambda^{-1}) & \text{if } \lambda = \lambda_{1n} \text{ or } \lambda_{3n} \end{cases} \tag{4.9}$$

with

$$r_8(\lambda) := \left(r_6 e^{-r_1\lambda} - r_5 e^{r_1\lambda}\right)\left(e_2(1)e^{-r_2\lambda} + e_1(1)e^{r_2\lambda}\right). \tag{4.10}$$

Furthermore,

$$\begin{aligned} w_4(x,\lambda) &= r_1^{-2} r_2^{-5}\lambda^{-8} \det\begin{bmatrix} \lambda[r_1]_1 & \lambda[r_1]_1 \\ \lambda^2 E_1[r_1 r_5]_1 & \lambda^2 E_2[r_1 r_6]_1 \end{bmatrix} \\ &\quad \times \det\begin{bmatrix} \lambda[r_2]_1 & \lambda[r_2]_1 \\ \lambda^2 e^{r_2\lambda x}[r_2^2]_1 & -\lambda^2 e^{-r_2\lambda x}[r_2^2]_1 \end{bmatrix} \times \det\begin{bmatrix} \lambda[r_2]_1 & \lambda[r_2]_1 \\ -\lambda^2 E_3 E_5[r_2^2]_1 & \lambda^2 E_4 E_6[r_2^2]_1 \end{bmatrix} \\ &= -r_2\lambda\left(r_6 e^{-r_1\lambda} - r_5 e^{r_1\lambda} + \mathcal{O}(\lambda^{-1})\right)\left(e^{-r_2\lambda x} + e^{r_2\lambda x} + \mathcal{O}(\lambda^{-1})\right) \\ &\quad \times \left(e_2(1)e^{-r_2\lambda} + e_1(1)e^{r_2\lambda} + \mathcal{O}(\lambda^{-1})\right) \end{aligned}$$

and

$$w_4(x,\lambda) = \begin{cases} -\lambda r_2 r_8(\lambda)\left(e^{-r_2\lambda x} + e^{r_2\lambda x} + \mathcal{O}(\lambda^{-1})\right) & \text{if } \lambda = \lambda_{2n}, \\ r_2\lambda[\mathcal{O}(\lambda^{-1})] & \text{if } \lambda = \lambda_{1n} \text{ or } \lambda_{3n}. \end{cases} \tag{4.11}$$

Also, the fifth component of $\Phi(x)$ can be given by

$$\begin{aligned} w_5(x,\lambda) &= -r_1^{-2} r_2^{-4}\lambda^{-8} \det\begin{bmatrix} \lambda[r_1]_1 & \lambda[r_1]_1 \\ \lambda^2 E_1[r_1 r_5]_1 & \lambda^2 E_2[r_1 r_6]_1 \end{bmatrix} \\ &\quad \times \det\begin{bmatrix} \lambda[r_2]_1 & \lambda[r_2]_1 \\ \lambda^2 E_3[r_2 r_3]_1 & \lambda^2 E_4[r_2 r_4]_1 \end{bmatrix} \times \det\begin{bmatrix} \lambda[r_2]_1 & \lambda[r_2]_1 \\ \lambda e^{r_2\lambda x} e_1(x)[r_2]_1 & \lambda e^{-r_2\lambda x} e_2(x)[r_2]_1 \end{bmatrix} \\ &= -\left(r_6 e^{-r_1\lambda} - r_5 e^{r_1\lambda} + \mathcal{O}(\lambda^{-1})\right)\left(r_4 e^{-r_2\lambda} - r_3 e^{r_2\lambda} + \mathcal{O}(\lambda^{-1})\right) \\ &\quad \times \left(e_2(x)e^{-r_2\lambda x} - e_1(x)e^{r_2\lambda x} + \mathcal{O}(\lambda^{-1})\right), \end{aligned}$$

and we conclude from (3.23), (3.25), and (3.31) that

$$w_5(x,\lambda) = \begin{cases} r_9(\lambda)\left(e_2(x)e^{-r_2\lambda x} - e_1(x)e^{r_2\lambda x} + \mathcal{O}(\lambda^{-1})\right) & \text{if } \lambda = \lambda_{3n}, \\ \mathcal{O}(\lambda^{-1}) & \text{if } \lambda = \lambda_{1n} \text{ or } \lambda_{2n} \end{cases} \tag{4.12}$$

with

$$r_9(\lambda) := -\left(r_6 e^{-r_1\lambda} - r_5 e^{r_1\lambda} + \mathcal{O}(\lambda^{-1})\right)\left(r_4 e^{-r_2\lambda} - r_3 e^{r_2\lambda} + \mathcal{O}(\lambda^{-1})\right). \tag{4.13}$$

For the last component of $\Phi(x)$, one has

$$\begin{aligned} w_6(x,\lambda) &= r_1^{-2} r_2^{-4}\lambda^{-8} \det\begin{bmatrix} \lambda[r_2]_1 & \lambda[r_2]_1 \\ -\lambda^2 e^{r_2\lambda x} e_1(x)[r_2^2]_1 & \lambda^2 e^{-r_2\lambda x} e_2(x)[r_2^2]_1 \end{bmatrix} \\ &\quad \times \det\begin{bmatrix} \lambda[r_2]_1 & \lambda[r_2]_1 \\ \lambda^2 E_3[r_2 r_3]_1 & \lambda^2 E_4[r_2 r_4]_1 \end{bmatrix} \times \det\begin{bmatrix} \lambda[r_1]_1 & \lambda[r_1]_1 \\ \lambda^2 E_1[r_1 r_5]_1 & \lambda^2 E_2[r_1 r_6]_1 \end{bmatrix} \\ &= r_2\lambda\left(r_6 e^{-r_1\lambda} - r_5 e^{r_1\lambda} + \mathcal{O}(\lambda^{-1})\right)\left(r_4 e^{-r_2\lambda} - r_3 e^{r_2\lambda} + \mathcal{O}(\lambda^{-1})\right) \\ &\quad \times \left(e_2(x)e^{-r_2\lambda x} + e_1(x)e^{r_2\lambda x} + \mathcal{O}(\lambda^{-1})\right) \end{aligned}$$

and

(4.14)
$$w_6(x,\lambda)=\begin{cases}-\lambda r_2 r_9(\lambda)\left(e_2(x)e^{-r_2\lambda x}+e_1(x)e^{r_2\lambda x}+\mathcal{O}(\lambda^{-1})\right) & \text{if } \lambda=\lambda_{3n},\\ r_2\lambda[\mathcal{O}(\lambda^{-1})] & \text{if } \lambda=\lambda_{1n} \text{ or } \lambda_{2n}.\end{cases}$$

On the basis of above computations, (4.1) can then be deduced from (4.6)–(4.14) by setting

$$w_n(x)=-\frac{w_1(x,\lambda)}{2r_1\lambda r_7(\lambda)},\quad \xi_n(x)=-\frac{w_3(x,\lambda)}{2r_2\lambda r_8(\lambda)},\quad s_n(x)=-\frac{w_5(x,\lambda)}{2r_2\lambda r_9(\lambda)} \tag{4.15}$$

in (4.6)–(4.14), respectively. Finally, it follows from (3.25) that

$$\begin{cases}\|e^{-r_j\lambda_{jn}x}\|_{L^2}=\dfrac{1-e^{-2r_j\mu_j}}{2r_j\mu_j}+\mathcal{O}(n^{-1}) \quad \text{for } j=1,2,\\ \|e^{r_j\lambda_{jn}x}\|_{L^2}=\dfrac{e^{2r_j\mu_j}-1}{2r_j\mu_j}+\mathcal{O}(n^{-1}) \quad \text{for } j=1,2,\\ \|e^{-r_2\lambda_{3n}x}\|_{L^2}=\dfrac{1-e^{-2r_2\mu_3}}{2r_2\mu_3}+\mathcal{O}(n^{-1}),\\ \|e^{r_2\lambda_{3n}x}\|_{L^2}=\dfrac{e^{2r_2\mu_3}-1}{2r_2\mu_3}+\mathcal{O}(n^{-1}),\end{cases} \tag{4.16}$$

where μ_j $(j=1,2,3)$ are given in (3.26) and (3.27). These together with (4.1) yield (4.2). The proof is complete. □

5. The Riesz basis property and exponential stability of the system. In the previous sections, we have obtained the asymptotic expressions of eigenpairs of $\mathcal{A}$ and concluded that there are three asymptotes for the spectrum $\sigma(\mathcal{A})$ with their asymptotic expressions in (3.25). In this section, we shall prove that there exists a sequence of generalized eigenfunctions of $\mathcal{A}$ which forms a Riesz basis for $\mathcal{H}$. Furthermore, the exponential stability of the system can be determined by its spectrum distribution.

For these purposes, we introduce another equivalent inner product on $\mathcal{H}$. Let $Y_j:=[w_j,z_j,\xi_j,\varphi_j,s_j,h_j]^\top\in\mathcal{H}$ $(j=1,2)$ define a new inner product in $\mathcal{H}$ by

(5.1)
$$[Y_1,\ Y_2]_H:=\langle w_1',w_2'\rangle_{L^2}+\langle z_1,z_2\rangle_{L^2}+\langle\xi_1',\xi_2'\rangle_{L^2}+\langle\varphi_1,\varphi_2\rangle_{L^2}+\langle s_1',s_2'\rangle_{L^2}+\langle h_1,h_2\rangle_{L^2},$$

and write its induced norm of (5.1) by $\|\cdot\|_H$. One can easily check that $\mathcal{H}$ is a Hilbert space under this new inner product. From now on, we shall consider our problem in $\mathcal{H}$ associated with this new inner product of (5.1). For convenience, define another Hilbert space

$$\mathcal{L}:=\left(L^2(0,1)\right)^6 \tag{5.2}$$

with an inner product (for any $X_j:=[w_j,z_j,\xi_j,\varphi_j,s_j,h_j]^\top\in\mathcal{L},\ j=1,2$)

$$\langle X_1,\ X_2\rangle_{\mathcal{L}}:=\langle w_1,w_2\rangle_{L^2}+\langle z_1,z_2\rangle_{L^2}+\langle\xi_1,\xi_2\rangle_{L^2}+\langle\varphi_1,\varphi_2\rangle_{L^2}+\langle s_1,s_2\rangle_{L^2}+\langle h_1,h_2\rangle_{L^2}$$

and define the subspaces of $\mathcal{H}$ and $\mathcal{L}$, respectively, by

(5.3)
$$\begin{cases} \mathcal{H}_1 := \{Y \in \mathcal{H} \mid Y = [w, z, 0, 0, 0, 0]^\top\}, \\ \mathcal{H}_2 := \{Y \in \mathcal{H} \mid Y = [0, 0, \xi, \varphi, 0, 0]^\top\}, \\ \mathcal{H}_3 := \{Y \in \mathcal{H} \mid Y = [0, 0, 0, 0, s, h]^\top\}, \end{cases} \quad \begin{cases} \mathcal{L}_1 := \{X \in \mathcal{L} \mid X = [w, z, 0, 0, 0, 0]^\top\}, \\ \mathcal{L}_2 := \{X \in \mathcal{L} \mid X = [0, 0, \xi, \varphi, 0, 0]^\top\}, \\ \mathcal{L}_3 := \{X \in \mathcal{L} \mid X = [0, 0, 0, 0, s, h]^\top\}. \end{cases}$$

Obviously, we have

$$\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2 \oplus \mathcal{H}_3 \quad \text{and} \quad \mathcal{L} = \mathcal{L}_1 \oplus \mathcal{L}_2 \oplus \mathcal{L}_3, \tag{5.4}$$

where the sign $\oplus$ denotes the direct sum in the sense of orthogonality with respect to the inner products $[\cdot, \cdot]_H$ and $\langle \cdot, \cdot \rangle_{\mathcal{L}}$ in $\mathcal{H}$ and $\mathcal{L}$, respectively.

Before continuing, let us recall some forms of notation. For a closed operator $\mathbf{A}$ in a Hilbert space $\mathbf{H}$, a nonzero element $\phi \in \mathbf{H}$ is called a generalized eigenvector of $\mathbf{A}$, corresponding to an eigenvalue λ of $\mathbf{A}$, if there is an integer $\nu \geq 1$ such that $(\lambda I - \mathbf{A})^\nu \phi = 0$. If $\nu = 1$, then ϕ is an eigenvector. A sequence $\{\phi_n\}_{n=1}^\infty$ in $\mathbf{H}$ is called a Riesz basis for $\mathbf{H}$ if there exists an orthonormal basis $\{e_n\}_{n=1}^\infty$ in $\mathbf{H}$ and a linear bounded invertible operator T such that

$$T\phi_n = e_n, \quad n = 1, 2, \dots .$$

Let $\{\lambda_n\}_{n=1}^\infty = \sigma(\mathbf{A})$ be the spectrum of $\mathbf{A}$. Suppose each λ_n has finite algebraic multiplicity m_n, and let $\{\psi_{n_i}\}_1^{m_n}$ be the set of generalized eigenvectors of $\mathbf{A}$ corresponding to λ_n. If $\{\psi_{n_i} \mid 1 \leq i \leq m_n,\ n = 1, 2, \dots\}$ form a Riesz basis for $\mathbf{H}$, then the C_0-semigroup generated by $\mathbf{A}$ can be represented as

$$e^{\mathbf{A}t} x = \sum_{n=1}^{\infty} e^{\lambda_n t} \sum_{j=1}^{m_n} a_{nj} f_{nj}(t) \psi_{nj} \quad \forall\, x = \sum_{n=1}^{\infty} \sum_{j=1}^{m_n} a_{nj} \psi_{nj} \in \mathbf{H}, \tag{5.5}$$

where $f_{nj}(t)$ are the polynomials of t with order not greater than m_n. In particular, if m_n has a uniform upper bound and $\{\psi_{n_i}\}_1^{m_n}$ is the eigenvector (not generalized eigenvector) set of $\mathcal{A}$ with respect to λ_n for all sufficiently large n, then the spectrum determined growth condition holds, i.e., $\omega(\mathbf{A}) = s(\mathbf{A})$, where $\omega(\mathbf{A})$ is the growth bound of $e^{\mathbf{A}t}$, and $s(\mathbf{A})$ is the spectral bound of $\mathbf{A}$ (see [2]).

To establish the Riesz basis property for the root space of the operator $\mathcal{A}$, we recall a result of Bari's theorem in [11].

THEOREM 5.1. *Let $\mathbf{H}$ be a separable Hilbert space and let $\{e_n;\ n \in \mathbb{Z}\}$ be an orthonormal basis for $\mathbf{H}$. If $\{f_n;\ n \in \mathbb{Z}\}$ is an ω-independent sequence that is quadratically close to $\{e_n;\ n \in \mathbb{Z}\}$, then $\{f_n;\ n \in \mathbb{Z}\}$ is a Riesz basis for $\mathbf{H}$.*

LEMMA 5.2. *Let $\{\phi_n(x);\ n \in \mathbb{N}\}$ and $\{1, \psi_n(x);\ n \in \mathbb{N}\}$ be two subsets in $L^2(0,1)$ defined by, respectively,*

$$\phi_n(x) := \sin n\pi x \quad \text{and} \quad \psi_n(x) := \cos n\pi x \quad \forall\, x \in (0,1),\ n \in \mathbb{N}.$$

Then $\{\phi_n(x);\ n \in \mathbb{N}\}$ and $\{1, \psi_n(x);\ n \in \mathbb{N}\}$ are two orthogonal bases in $L^2(0,1)$. Moreover, for any scalars $\alpha, \beta \neq 0 \in \mathbb{C}$ the vector family $\{\Psi_n := [\cosh(\alpha + in\pi)x, \beta \sinh(\alpha + in\pi)x]^\top, n \in \mathbb{Z}\}$ forms a Riesz basis on the Hilbert space $L^2(0,1) \times L^2(0,1)$.

Proof. The first assertion is a direct result in [11] and it is easily verified that

$$\left\{ \begin{bmatrix} 0 \\ \sin n\pi x \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} \cos n\pi x \\ 0 \end{bmatrix} \right\}_{n \in \mathbb{N}}$$

constitutes a Riesz basis on $L^2(0,1)\times L^2(0,1)$. So the sequence

$$\left\{\begin{bmatrix}\cos n\pi x\\ \sin n\pi x\end{bmatrix},\quad \begin{bmatrix}1\\0\end{bmatrix},\quad \begin{bmatrix}\cos n\pi x\\ -\sin n\pi\end{bmatrix}\right\}_{n\in\mathbb{N}}$$

also forms a Riesz basis on $L^2(0,1)\times L^2(0,1)$. Let T be an invertible matrix function in $(0,1)$ given by

$$T:=\begin{bmatrix}\cosh\alpha x & i\sinh\alpha x\\ \beta\sinh\alpha x & i\beta\cosh\alpha x\end{bmatrix}\quad \text{with } |T|=i\beta \text{ for each } x\in(0,1).$$

Then we obtain, for $n\in\mathbb{N}$,

$$\begin{bmatrix}\cosh(\alpha+in\pi)x\\ \beta\sinh(\alpha+in\pi)x\end{bmatrix}=T\begin{bmatrix}\cos n\pi x\\ \sin n\pi x\end{bmatrix},\quad \begin{bmatrix}\cosh(\alpha-in\pi)x\\ \beta\sinh(\alpha-in\pi)x\end{bmatrix}=T\begin{bmatrix}\cos n\pi x\\ -\sin n\pi x\end{bmatrix},$$

and

$$\begin{bmatrix}\cosh\alpha x\\ \beta\sinh\alpha x\end{bmatrix}=T\begin{bmatrix}1\\0\end{bmatrix}.$$

Thus, the sequence $\{\Psi_n:=[\cosh(\alpha+in\pi)x,\ \beta\sinh(\alpha+in\pi)x]^\top, n\in\mathbb{Z}\}$ also forms a Riesz basis on the Hilbert space $L^2(0,1)\times L^2(0,1)$. The second assertion is concluded. □

THEOREM 5.3. *Suppose $r_1\neq r_2$ and $k_i\neq r_i$ $(i=1,2)$. Let $\mathcal{A}$ be defined by* (2.4) *and* (2.5)*, and let*

$$\Psi_{jn}:=[w'_{jn},\ \lambda_{jn}w_{jn},\ \xi'_{jn},\ \lambda_{jn}\xi_{jn},\ s'_{jn},\ \lambda_{jn}s_{jn}]^\top,\quad (j=1,2,3,\ n\in\mathbb{Z}), \tag{5.6}$$

where the entries are given as (4.1) *corresponding to the eigenvalues λ_{jn}. Then $\{\Psi_{1n},\Psi_{2n},\Psi_{3n};\ n\in\mathbb{Z}\}$ forms a Riesz basis in Hilbert space $\mathcal{L}$ provided that $\{\Psi_{jn},\ j=1,2,3,\ n\in\mathbb{Z}\}$ is ω-linearly independent in $\mathcal{L}$.*

Proof. Let three vector families be given by

$$\begin{aligned}
\Phi_{1n}&:=\big[\cosh(r_1\mu_1+in\pi)x,\ r_1^{-1}\sinh(r_1\mu_1+in\pi)x,0,0,0,0\big]^\top,\\
\Phi_{2n}&:=\big[0,0,\cosh(r_2\mu_2+in\pi)x,\ r_2^{-1}\sinh(r_2\mu_2+in\pi)x,0,0\big]^\top,\\
\Phi_{3n}&:=\Big[0,0,0,0,\cosh i\Big(n+\frac{1}{2}\Big)\pi x,\ r_3^{-1}\sinh i\Big(n+\frac{1}{2}\Big)\pi x\Big]^\top.
\end{aligned}$$

Then one concludes from Lemma 5.2 that the families $\{\Phi_{jn},\ n\in\mathbb{Z}\}$ $(j=1,2,3)$ are the Riesz bases for $\mathcal{L}_j$, respectively. Also, by using the asymptotic expressions of both eigenvalues (3.25) and their eigenfunctions (4.1), it follows that

$$\begin{cases}
\|w'_{1n}-\cosh(r_1\mu_1+in\pi)x\|_{L^2}=\mathcal{O}(n^{-1}),\\
\|r_1\lambda_{1n}w_{1n}-\sinh(r_1\mu_1+in\pi)x\|_{L^2}=\mathcal{O}(n^{-1}),\\
\|\xi'_{2n}-\cosh(r_2\mu_2+in\pi)x\|_{L^2}=\mathcal{O}(n^{-1}),\\
\|r_2\lambda_{2n}\xi_{1n}-\sinh(r_2\mu_2+in\pi)x\|_{L^2}=\mathcal{O}(n^{-1}),\\
\|s'_{3n}-\cosh i(n+\frac{1}{2})\pi x\|_{L^2}=\mathcal{O}(n^{-1}),\\
\|r_3\lambda_{3n}s_{3n}-\sinh i(n+\frac{1}{2})\pi x\|_{L^2}=\mathcal{O}(n^{-1}).
\end{cases} \tag{5.7}$$

Hence, we obtain

$$\begin{aligned} &\|\Psi_{1n} - \Phi_{1n}\|_{\mathcal{L}_1} = \mathcal{O}(n^{-1}),\\ &\|\Psi_{2n} - \Phi_{2n}\|_{\mathcal{L}_2} = \mathcal{O}(n^{-1}),\\ &\|\Psi_{3n} - \Phi_{3n}\|_{\mathcal{L}_3} = \mathcal{O}(n^{-1}). \end{aligned} \tag{5.8}$$

Therefore, by Theorem 5.1 (Bari's theorem), $\{\Psi_{1n}, \Psi_{2n}, \Psi_{3n}; n \in \mathbb{Z}\}$ forms a Riesz basis for $\mathcal{L}$ provided that $\{\Psi_{jn};\ j = 1, 2, 3, n \in \mathbb{Z}\}$ is ω-linearly independent in $\mathcal{L}$. □

As a consequence of this theorem, we obtain the main results of the paper.

THEOREM 5.4. *Suppose $r_1 \neq r_2$ and $k_i \neq r_i$ $(i = 1, 2)$. Let $\mathcal{A}$ be defined by* (2.4) *and* (2.5). *Then there exists a sequence of generalized eigenfunctions of $\mathcal{A}$ which forms a Riesz basis for $\mathcal{H}$.*

Proof. In Theorem 4.1, we have obtained the asymptotic expressions of the eigenfunctions of $\mathcal{A}$ corresponding to the eigenvalues with large moduli. Without loss of generality, we may assume that

$$Y_{jn} := [w_{jn}, \lambda_{jn} w_{jn}, \xi_{jn}, \lambda_{jn}\xi_{jn}, s_{jn}, \lambda_{jn} s_{jn}]^\top \quad \text{for } j = 1, 2, 3 \text{ and } n \in \mathbb{Z}$$

is an eigenfunction corresponding to the eigenvalue λ_{jn} in which the entries have the asymptotic expansions given in (4.1). Then $\{Y_{jn};\ j = 1, 2, 3, n \in \mathbb{Z}\}$ is ω-linearly independent in $\mathcal{H}$. The proof will be completed via an isomorphic mapping between two Hilbert spaces $\mathcal{H}$ and $\mathcal{L}$ that maps Y_{jn} to Ψ_{jn}, where $\{\Psi_{jn};\ j = 1, 2, 3, n \in \mathbb{Z}\}$ is a sequence given by (5.6).

To do this, for any $F := [f_1, f_2, g_1, g_2, u_1, u_2]^\top \in \mathcal{H}$, we define a linear bounded operator $\mathcal{T} : \mathcal{H} \to \mathcal{L}$ by

$$\mathcal{T}F := [f_1', f_2, g_1', g_2, u_1', u_2]^\top := \widehat{F}.$$

Since $[F,\ Y_{jn}]_H = \langle \widehat{F},\ \Psi_{jn} \rangle_{\mathcal{L}}$, it is easy to prove that $\mathcal{T}$ is isomorphic and satisfies

$$\|\mathcal{T}F\|_{\mathcal{L}} = \|\widehat{F}\|_{\mathcal{L}} = \|F\|_H. \tag{5.9}$$

In particular, for $j = 1, 2, 3$ and $n \in \mathbb{Z}$,

$$\mathcal{T}Y_{jn} = \Psi_{jn} \quad \text{and} \quad \|Y_{jn}\|_H = \|\Psi_{jn}\|_{\mathcal{L}}.$$

Moreover, $\{Y_{jn};\ j = 1, 2, 3, n \in \mathbb{Z}\}$ is ω-linearly independent in $\mathcal{H}$, so is $\{\Psi_{jn};\ j = 1, 2, 3, n \in \mathbb{Z}\}$ in $\mathcal{L}$. Therefore, by Theorem 5.3, $\{\Psi_{jn};\ j = 1, 2, 3, n \in \mathbb{Z}\}$ forms a Riesz basis in $\mathcal{L}$. Hence, $\{Y_{jn}; j = 1, 2, 3,\ n \in \mathbb{Z}\}$ forms a Riesz basis for $\mathcal{H}$. The proof is complete. □

THEOREM 5.5. *Suppose $r_1 \neq r_2$ and $k_i \neq r_i$ $(i = 1, 2)$. Let $\mathcal{A}$ be defined by* (2.4) *and* (2.5), *and let $T(t)$ be a C_0-semigroup generated by $\mathcal{A}$ in $\mathcal{H}$. Then $T(t)$ is exponentially stable, and in fact it is a C_0-group in $\mathcal{H}$.*

Proof. As a direct consequence of Theorems 3.4 and 5.4, the spectrum determined growth condition $\omega(\mathbf{A}) = s(\mathbf{A})$ for $T(t)$ holds. Furthermore, Corollary 2.2 implies that there is no eigenvalue on the imaginary axis. This, together with (3.25) and the spectrum determined growth condition, shows that $T(t)$ is an exponentially stable semigroup on $\mathcal{H}$. Moreover, $T(t)$ is also a C_0-group in $\mathcal{H}$. This is because of the fact that the spectrum of $\mathcal{A}$ distributes in a vertical strip due to Theorems 3.3 and 3.4. □

Acknowledgment. The authors would like to thank the referees for their very valuable suggestions and comments.

REFERENCES

[1] G. D. Birkhoff and R. E. Langer, *The boundary problems and developments associated with a system of ordinary linear differential equations of the first order,* Proc. Amer. Acad. Arts Sci., 58 (1923), pp. 49–128.

[2] B. Z. Guo, *Riesz basis approach to the stabilization of a flexible beam with a tip mass,* SIAM J. Control Optim., 39 (2001), pp. 1736–1747.

[3] Z. H. Luo, B. Z. Guo, and O. Morgül, *Stability and Stabilization of Infinite Dimensional Systems with Applications,* Spring-Verlag, London, 1999.

[4] S. W. Hansen and R. Spies, *Structural damping in a laminated beams due to interfacial slip,* J. Sound Vibration, 204 (1997), pp. 183–202.

[5] T. Kato, *Perturbation Theory of Linear Operators,* Springer-Verlag, New York, 1976.

[6] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations,* Springer-Verlag, New York, 1983.

[7] D. L. Russell, *Mathematical models for the elastic beam and their control-theoretic implications,* in Semigroups, Theory and Applications, Vol. II (Trieste, 1984), Pitman Res. Notes Math. Ser. 152, Longman Scientific and Technical, Harlow, UK, 1986, pp. 177–216.

[8] C. Tretter, *Spectral problems for systems of differential equations $y' + A_0 y = \lambda A_1 y$ with λ-polynomial boundary conditions,* Math. Nachr., 214 (2000), pp. 129–172.

[9] C. Tretter, *Boundary eigenvalue problems for differential equations $N\eta = \lambda P\eta$ with λ-polynomial boundary conditions,* J. Differential Equations, 170 (2001), pp. 408–471.

[10] M. L. Zhu, S. W. R. Lee, H. L. Li, T. Y. Zhang, and P. Tong, *Modeling and analysis of torsional vibration induced by extension-twisting coupling of anisotropic composite laminates with piezoelectric actuators,* Smart Mater. Struc., 11 (2002), pp. 55–62.

[11] R. Young, *An introduction to Nonharmonic Fourier Series,* Revised 1st ed., Academic Press, New York, 2001.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 5, pp. 1598–1613

THE REGULARITY OF THE WAVE EQUATION WITH PARTIAL DIRICHLET CONTROL AND COLOCATED OBSERVATION*

BAO-ZHU GUO† AND XU ZHANG‡

Abstract. In this paper we analyze a multidimensional controlled wave equation on a bounded domain, subject to partial Dirichlet control and colocated observation. By means of a partial Fourier transform, it is shown that the system is well-posed and regular in the sense of D. Salamon and G. Weiss. The corresponding feedthrough operator is found to be the identity operator on the input space.

Key words. wave equation, transfer function, well-posed and regular system, partial Dirichlet control and colocated observation, partial Fourier transform

AMS subject classifications. 35J05, 93C20, 93C25

DOI. 10.1137/040610702

1. Introduction.

A very general class of linear infinite-dimensional systems for which there is a well established theory parallel to that for finite-dimensional systems is the class of *well-posed and regular linear systems* (see [5]). This generic framework covers many systems governed by partial differential equations with actuators and sensors supported on isolated points, on a subdomain, or on a part of the boundary of the spatial region. There are many papers in this field (e.g., [7], [13], [14], [15], [16], [20], [21], [24], [25], [26], [27], [34], [35], [36], [38], and the references therein). Recently, the regular linear system theory has been generalized to the time-varying case in [22]. We refer to [5] for a nice earlier summary of well-posed system theory.

Well-posedness and regularity are two new crucial concepts introduced in linear infinite-dimensional systems theory under the above-mentioned framework. It is notable that these two concepts are completely different from those one usually uses in partial differential equations. For the reader's convenience, we shall recall their definitions and other related notions in section 2. As remarked in [4], very little is known about the well-posedness or the regularity of controlled infinite-dimensional systems. In [2], the well-posedness of the wave equation with Dirichlet input and colocated output in a two-dimensional (2-D) disk was proved by a direct method. The well-posedness of the same equation on a bounded open domain of $\mathbb{R}^n (n \geq 2)$ with a smooth boundary was proved in [1] using microlocal analysis. The well-posedness and regularity of the multidimensional heat equation with both Dirichlet- and Neumann-type boundary control has been established in [3]. To the best of our knowledge, [3] is the first article dealing with the regularity of a multidimensional partial differential equation system, although well-posedness and regularity have been well-established

*Received by the editors June 28, 2004; accepted for publication (in revised form) May 28, 2005; published electronically November 14, 2005.
http://www.siam.org/journals/sicon/44-5/61070.html

†Corresponding author. Academy of Mathematics and System Sciences, Academia Sinica, Beijing 100080, China, and School of Computational and Applied Mathematics, University of the Witwatersrand, Johannesburg, South Africa (bzguo@iss.ac.cn). This author was supported by the National Natural Science Foundation of China and the National Research Foundation of South Africa.

‡School of Mathematics, Sichuan University, Chengdu 610064, China, and Departamento de Matemáticas, Facultad de Ciencias, Universidad Autónoma de Madrid, 28049 Madrid, Spain (xu.zhang@uam.es). This author was supported by the FANEDD of China No. 200119, the NSFC under grant 10371084, the Program for New Century Excellent Talents in University of China, and grant BFM2002-03345 from the Spanish MCYT.

for many one-dimensional systems (see [11]). The regularity of the wave equation in a 2-D disk with Dirichlet control and colocated observation was first obtained in [12]. However, the same problem for a general bounded domain in $\mathbb{R}^n$ has remained open.

The aim of this paper is to give a positive solution to the above-mentioned problem. More precisely, we consider the following multidimensional wave equation with partial Dirichlet control and colocated observation:

$$\begin{cases} w_{tt}(x,t) - \Delta w(x,t) = 0, & x \in \Omega,\ t > 0, \\ w(x,t) = 0, & x \in \Gamma_1,\ t > 0, \\ w(x,t) = u(x,t), & x \in \Gamma_0,\ t > 0, \\ y(x,t) = -\dfrac{\partial(-\Delta)^{-1} w_t(x,t)}{\partial \nu}, & x \in \Gamma_0,\ t > 0. \end{cases} \tag{1.1}$$

Here, $\Omega \subset \mathbb{R}^n$ $(n \geq 2)$ is a bounded domain with the smooth boundary $\partial\Omega = \overline{\Gamma_0} \cup \overline{\Gamma_1}$, both Γ_0 and Γ_1 are disjoint parts of the boundary relatively open in $\partial\Omega$ with $\mathrm{int}(\Gamma_0) \neq \emptyset$, and ν is the unit normal vector of Γ_0 pointing towards the exterior of Ω. In system (1.1), u is the *input function* (or *control*) and y is the *output function* (or *output*). Put $\mathcal{H} = L^2(\Omega) \times H^{-1}(\Omega)$ and $U = L^2(\Gamma_0)$. The following result comes from Proposition 2.2 of [1] and Theorem 4.2 of [19, p. 46] (see also [17]).

THEOREM 1.1. *Let $T > 0$, $(w_0, w_1) \in \mathcal{H}$, and $u \in L^2(0,T;U)$. Then there exists a unique solution $(w, w_t) \in C([0,T];\mathcal{H})$ to (1.1) satisfying $w(\cdot,0) = w_0$ and $w_t(\cdot,0) = w_1$. Moreover, there exists a constant $C > 0$, independent of (w_0, w_1, u), such that*

$$\|(w(\cdot,T), w_t(\cdot,T))\|_{\mathcal{H}}^2 + \|y\|_{L^2(0,T;U)}^2 \leq C \left[\|(w_0,w_1)\|_{\mathcal{H}}^2 + \|u\|_{L^2(0,T;U)}^2 \right].$$

Theorem 1.1 implies that the system described by (1.1) is well-posed with state space $\mathcal{H}$, input space U, and output space U (the precise definition of these concepts will be given in the next section). We mention that Proposition 2.2 of [1] says that there exists a $C^* > 0$ independent of u such that

$$\|y\|_{L^2(0,T;U)}^2 \leq C^* \|u\|_{L^2(0,T;U)}^2 \ \text{ when } (w_0,w_1) = 0.$$

However, as was indicated in [2] and [37], Theorem 1.1 can be derived from here with relative ease.

The main goal of this paper is to show that the system described by (1.1) is regular as well. Our result reads as follows.

THEOREM 1.2. *System (1.1) is regular. More precisely, if $w(\cdot,0) = w_t(\cdot,0) = 0$ and $u(x,t) \equiv u(x)$ is a step input with some $u \in U$, then the corresponding output y satisfies*

$$\lim_{\sigma \to 0} \int_{\Gamma_0} \left| \frac{1}{\sigma} \int_0^\sigma y(x,t)dt - u(x) \right|^2 dx = 0.$$

This result allows us to study dynamic stabilization, optimal control, or other problems for system (1.1) using a theory that is parallel in many ways to the finite-dimensional theory; see, e.g., [6]. Also, as we shall explain in section 2, Theorem 1.2 states that system (1.1) has feedthrough operator $\mathbb{D} = I$, where I is the identity operator on U.

This paper is organized as follows: In the next section, we introduce the background and the necessary preliminaries about well-posed and regular systems. The proof of Theorem 1.2 is given in section 3.

2. Preliminaries. In this section, we shall briefly recall some background about infinite-dimensional well-posed and regular systems (see [5], [27], [30], [31], [32], [33], [34]).

Let X, U, and Y be three Hilbert spaces. Denote by $\|\cdot\|$ the norm of X (induced by its inner product). In what follows, we choose X, U, and Y to be the state, input, and output spaces, respectively, of an infinite-dimensional linear system. This system is described by the equations

$$\begin{cases} \dot{x}(t) = \mathbb{A}x(t) + \mathbb{B}u(t), \quad x(0) = x_0 \in X, \\ y(t) = \mathbb{C}_e x(t) + \mathbb{D}_e u(t), \end{cases} \tag{2.1}$$

where the (usually unbounded) operator $\mathbb{A}$ generates a C_0-semigroup $\mathbb{T}(\cdot)$ on X, $\mathbb{B}$ is a control operator from U to X, $\mathbb{C}_e$ is an observation operator from X to Y, and $\mathbb{D}_e$ is a bounded operator from U to Y. In (2.1), $u(t) \in U$, $x(t) \in X$, and $y(t) \in Y$ are called the input, the state, and the output, respectively. The input function $u(\cdot)$ is assumed to be in the space $L^2_{loc}(0, \infty; U)$, but the representation (2.1) is valid only if $u \in H^1_{loc}(0, \infty; U)$ and $\mathbb{A}x(0) + \mathbb{B}u(0) \in X$ (see [28] for details). For the case that both $\mathbb{B}$ and $\mathbb{C}_e$ are bounded, a nice theory for system (2.1) has been summarized in the book [9]. The framework of well-posed system theory is, however, mainly concerned with the case where neither $\mathbb{B}$ nor $\mathbb{C}_e$ is bounded.

Let us recall some basic notation. The Hilbert space X_{-1} is defined as the completion of X with respect to the norm

$$\|x\|_{-1} = \|(\beta - \mathbb{A})^{-1}x\| \qquad \forall\, x \in X,$$

and the space X_1 is the space $D(\mathbb{A})$ with the norm

$$\|x\|_1 = \|(\beta - \mathbb{A})x\| \qquad \forall\, x \in D(\mathbb{A}),$$

where $\beta \in \rho(\mathbb{A})$, the resolvent set of $\mathbb{A}$. It is easy to verify that both X_{-1} and X_1 are independent of the choice of β. It was shown in [30] that $X_{-1} = D(\mathbb{A}^*)'$, the dual space of $D(\mathbb{A}^*)$ with respect to the pivot X. Identifying X with its dual space, we have the following continuous, dense inclusions:

$$X_1 \hookrightarrow X \hookrightarrow X_{-1}.$$

DEFINITION 2.1. *System* (2.1) *is said to be well-posed if the following hold:*

(a) $\mathbb{A}$ *generates a* C_0*-semigroup* $\mathbb{T}(\cdot)$ *on* X.

(b) $\mathbb{B} \in \mathcal{L}(U, X_{-1})$ *is an admissible control operator for* $\mathbb{T}(\cdot)$*, i.e., for some (and hence for any)* $t > 0$ *there exists* $C_t > 0$ *such that*

$$\left\| \int_0^t \mathbb{T}(t-\tau)\mathbb{B}u(\tau)d\tau \right\|^2 \le C_t \int_0^t \|u(t)\|_U^2 dt \qquad \forall\, u \in L^2(0, t; U).$$

(c) *The domain* $D(\mathbb{C}_e) \supset D(\mathbb{A})$*. If we denote by* $\mathbb{C}$ *the restriction of* $\mathbb{C}_e$ *to* $D(\mathbb{A})$*, then* $\mathbb{C} \in \mathcal{L}(X_1, Y)$ *is an admissible observation operator for* $\mathbb{T}(\cdot)$*, which means that for some (and hence for any)* $t > 0$*, there exists* $C_t' > 0$ *such that*

$$\int_0^t \|\mathbb{C}\mathbb{T}(\cdot)x\|_Y^2 dt \le C_t' \|x\|^2 \qquad \forall\, x \in D(\mathbb{A}).$$

(d) *The input-output map is bounded; i.e., for some (and hence for any)* $t > 0$, *there exists* $C_t'' > 0$ *such that*

$$\int_0^t \|y(t)\|_Y^2 dt \le C_t'' \int_0^t \|u(t)\|_U^2 dt \qquad \forall\, u \in L^2(0,t;U) \text{ when } x_0 = 0.$$

It should be noted that the definition above is not the standard one given by [5] or [8], but it is equivalent to Weiss's definition (see [16], [23], [27]). From [31], $\mathbb{B}$ is admissible for $\mathbb{T}(\cdot)$ if and only if the adjoint operator $\mathbb{B}^*$ is admissible for $\mathbb{T}^*(\cdot)$, the adjoint C_0-semigroup of $\mathbb{T}(t)$.

Roughly speaking, a well-posed system is a system for which both the state and output depend continuously on the initial state and input function of the system.

If system (2.1) is well-posed, then the weak solution of (2.1) can be represented as (see [5], [28])

$$\begin{cases} x(t) = \mathbb{T}(t)x_0 + \displaystyle\int_0^t \mathbb{T}(t-\tau)\mathbb{B}u(\tau)d\tau \in C([0,\infty);X) \\ \qquad \forall\, x_0 \in X, u \in L^2_{loc}(0,\infty;U), \\ y(t) = \mathbb{C}_\Lambda\left[x(t) - (\lambda - \mathbb{A})^{-1}\mathbb{B}u(t)\right] + \mathbb{H}(\lambda)u(t) \in L^2_{loc}(0,\infty;Y) \\ \qquad \forall\, u \in L^2_{loc}(0,\infty;U), \end{cases} \tag{2.2}$$

where $\mathbb{C}_\Lambda x = \lim_{\lambda\to+\infty} \mathbb{C}\lambda(\lambda - \mathbb{A})^{-1}x$ for all $x \in D(\mathbb{C}_\Lambda)$ is by definition the Λ-extension of $\mathbb{C}$, where $D(\mathbb{C}_\Lambda)$ is the subspace of X for which the associated limit exists (see [5]). $\mathbb{H}(\lambda)$ is called the transfer function which is defined in some right-half planes and is an analytic $\mathcal{L}(U,Y)$-valued function. It can be shown that if $\hat{u}(\lambda)$ exists, then

$$\hat{y}(\lambda) = \mathbb{H}(\lambda)\hat{u}(\lambda) \quad \text{ when } x_0 = 0, \tag{2.3}$$

where $\hat{}$ denotes the Laplace transform. In terms of the operators from (2.1), we have (see [28])

$$\mathbb{H}(\lambda) = \mathbb{C}_e(\lambda - \mathbb{A})^{-1}\mathbb{B} + \mathbb{D}_e.$$

The transfer function $\mathbb{H}(\lambda)$ can be determined by the triple of operators $(\mathbb{A}, \mathbb{B}, \mathbb{C})$ up to an additive constant bounded operator in the following way (see [8]):

$$\frac{\mathbb{H}(\lambda) - \mathbb{H}(\beta)}{\lambda - \beta} = -\mathbb{C}(\lambda - \mathbb{A})^{-1}(\beta - \mathbb{A})^{-1}\mathbb{B} \qquad \forall\, \lambda, \beta \in \mathcal{C}_\rho^+,\ \lambda \ne \beta, \tag{2.4}$$

where $\mathcal{C}_\rho^+ = \{\lambda \in \mathcal{C} |\ \mathrm{Re}\lambda > \rho\}$ for some $\rho > 0$ and $\mathcal{C}$ stands for the complex plane. Using the transfer function, the boundedness of the input-output map described in condition (d) of Definition 2.1 can be expressed as the boundedness of the transfer function on an open right complex half plane (see [8], [11], [16])

$$\sup_{\mathrm{Re}\lambda \ge \alpha > \rho} \|\mathbb{H}(\lambda)\|_{\mathcal{L}(U,Y)} < \infty \tag{2.5}$$

for some $\alpha \in \mathbb{R}$.

The paper [32] introduced an important subclass of well-posed systems, the so-called *regular systems*, for which the representation (2.2) becomes much simpler.

DEFINITION 2.2. *System* (2.1) *is said to be regular if it is well-posed and there exists an operator* $\mathbb{D} \in \mathcal{L}(U,Y)$ *such that, for* $x_0 = 0$ *and* $u(t) \equiv u \in U$*, the output* y *of* (2.1) *satisfies*

$$\lim_{t\to 0} \frac{1}{t} \int_0^t y(\tau)d\tau = \mathbb{D}u \tag{2.6}$$

in the strong topology of Y*. The above* $\mathbb{D}$ *and property* (2.6) *are called the feedthrough operator and the* regularity *of system* (2.1)*, respectively.*

It was shown in [34] that, in the frequency domain, (2.6) is equivalent to

$$\lim_{\lambda\in\mathbb{R},\, \lambda\to+\infty} \mathbb{H}(\lambda)u = \mathbb{D}u \qquad \forall\, u \in U. \tag{2.7}$$

If a well-posed system is regular, then (2.2) can be written as

$$\begin{cases} x(t) = \mathbb{T}(t)x_0 + \displaystyle\int_0^t \mathbb{T}(t-\tau)\mathbb{B}u(\tau)d\tau \in C([0,\infty);X), \\ \qquad x_0 \in X,\ u \in L^2_{loc}(0,\infty;U), \\ y(t) = \mathbb{C}_\Lambda x(t) + \mathbb{D}u(t) \in L^2_{loc}(0,\infty;Y),\ u \in L^2_{loc}(0,\infty;U). \end{cases} \tag{2.8}$$

In this case, the transfer function is uniquely determined by the quadruple of operators $(\mathbb{A},\mathbb{B},\mathbb{C},\mathbb{D})$ and can be represented as

$$\mathbb{H}(\lambda) = \mathbb{D} + \mathbb{C}_\Lambda(\lambda - \mathbb{A})^{-1}\mathbb{B}. \tag{2.9}$$

It is seen that the representations (2.8) and (2.9) resemble that for finite-dimensional systems.

Roughly speaking, a *well-posed regular* system is like a linear finite-dimensional system among the infinite-dimensional systems but with the feature of allowing both control and observation operators to be unbounded in some sense. Unlike stability, controllability, observability, etc., which have finite-dimensional counterparts, regularity is an important but new concept in linear infinite-dimensional systems under the elegant framework of well-posed linear systems theory.

Now let us introduce a special class of well-posed systems: the colocated second-order linear systems. It is well known that "passivity," which was introduced in connection with circuit theory in the 1950s (see [10]), is a very important concept in control system design. It means that the increase of energy stored in the system does not exceed the energy that enters from the external world. For such a system, the transfer function is positive real, and negative output feedback produces a dissipative system, which is stable in the sense of Lyapunov. For a long time, it has been known by engineers that a partial differential equation describing a mechanical system, like a flexible structure in which the power flow into the system is the scalar product $\langle u, y\rangle$ (e.g., when u is force and y is velocity), leads to a positive-real system (2.1) in which $U = Y$ and $\mathbb{A}^* + \mathbb{A} \le 0, \mathbb{C} = \mathbb{B}^*$ if actuators and sensors are designed in a "colocated" fashion. The particular case $\mathbb{A} + \mathbb{A}^* = 0$ corresponds to energy preserving systems. This means that the measurement and control action are made dual in some sense. In [11] and [35], an abstract setting of a second-order passive system of the following type was studied. The state space is $X = D(\mathbb{A}_0^{1/2}) \times H$, and the input and output spaces are the same $U = Y$ (see also [2], [37]):

$$\begin{cases} \ddot{x}(t) + \mathbb{A}_0 x(t) = \mathbb{B}_0 u(t), \\ y(t) = \mathbb{B}_0^* \dot{x}(t), \end{cases} \tag{2.10}$$

where

(i) $\mathbb{A}_0 : D(\mathbb{A}_0)(\subset H) \to H$ is an unbounded positive self-adjoint operator in the Hilbert space H;

(ii) $\mathbb{B}_0 \in \mathcal{L}(U, (D(\mathbb{A}_0^{1/2}))')$;

(iii) $\mathbb{B}_0^* \in \mathcal{L}(D(\mathbb{A}_0^{1/2}), U)$ is defined as

$$(\mathbb{B}_0^* x, u)_U = \langle x, \mathbb{B}_0 u \rangle_{D(\mathbb{A}_0^{1/2}) \times (D(\mathbb{A}_0^{1/2}))'} \qquad \forall\, x \in D(\mathbb{A}_0^{1/2});$$

(iv) an extension $\tilde{\mathbb{A}}_0 \in \mathcal{L}(D(\mathbb{A}_0^{1/2}), (D(\mathbb{A}_0^{1/2}))')$ of $\mathbb{A}_0$ is defined by

$$\langle \tilde{\mathbb{A}}_0 x, z \rangle_{(D(\mathbb{A}_0^{1/2}))' \times D(\mathbb{A}_0^{1/2})} = (\mathbb{A}_0^{1/2} x, \mathbb{A}_0^{1/2} z)_H \qquad \forall\, x, z \in D(\mathbb{A}_0^{1/2}).$$

It was found in [11] that if system (2.10) is well-posed, its transfer function is uniquely determined by the pair $(\mathbb{A}_0, \mathbb{B}_0)$:

$$H(\lambda) = \lambda \mathbb{B}_0^* (\lambda^2 + \tilde{\mathbb{A}}_0)^{-1} \mathbb{B}_0. \tag{2.11}$$

Actually, it was indicated in [2] and [37] that, for this system, the boundedness of the transfer function on some open right half complex plane implies automatically the admissibility of $\left[\begin{smallmatrix} 0 \\ \mathbb{B}_0 \end{smallmatrix}\right]$ for the associated semigroup generated by $\mathbb{A} = \left[\begin{smallmatrix} 0 & I \\ -\mathbb{A}_0 & 0 \end{smallmatrix}\right]$. This system is closely related (via feedback) to the example in [29].

To end this section, we return to our wave equation (1.1) with control $u \in L^2_{loc}(0, \infty; U), U = L^2(\Gamma_0)$. We formulate our problem in the framework of (2.10), although it is already available in the literature (see, e.g., [1]).

Let $H = H^{-1}(\Omega)$ be the dual space of the usual Sobolev space $H_0^1(\Omega)$ (with respect to the pivot space $L^2(\Omega)$). Let A_0 be the positive self-adjoint operator in H induced by the bilinear form $a(\cdot, \cdot)$ defined by

$$\langle A_0 f, g \rangle_{H^{-1}(\Omega) \times H_0^1(\Omega)} = a(f, g) = \int_\Omega \nabla f(x) \overline{\nabla g(x)} dx \qquad \forall\, f, g \in H_0^1(\Omega). \tag{2.12}$$

By means of the Lax–Milgram theorem, A_0 is a canonical isomorphism from $D(A_0) = H_0^1(\Omega)$ to H. If we introduce the Laplacian $-\Delta : H^2(\Omega) \cap H_0^1(\Omega) \to L^2(\Omega)$, then it is easy to show that $A_0 f = -\Delta f$ for $f \in H^2(\Omega) \cap H_0^1(\Omega)$ and that $A_0^{-1} g = (-\Delta)^{-1} g$ for any $g \in L^2(\Omega)$. Hence, A_0 is an extension of usual Laplacian to the space $H_0^1(\Omega)$.

It is well known that $D(A_0^{1/2}) = L^2(\Omega)$. Define the Dirichlet map

$$\Upsilon \in \mathcal{L}(L^2(\Gamma_0), L^2(\Omega)),$$

i.e., $\Upsilon u = v$ by

$$\begin{cases} \Delta v = 0 & \text{in } \Omega, \\ v|_{\Gamma_1} = 0, \quad v|_{\Gamma_0} = u. \end{cases} \tag{2.13}$$

Using the Dirichlet map, we can rewrite the first three equations in (1.1) as

$$\ddot{w} + A_0(w - \Upsilon u) = 0. \tag{2.14}$$

We identify H with its dual H'. Then the following relations hold:

$$D(A_0^{1/2}) \hookrightarrow H \hookrightarrow (D(A_0^{1/2}))'.$$

An extension $\tilde{A}_0 \in \mathcal{L}(D(A_0^{1/2}), (D(A_0^{1/2}))')$ of A_0 is defined by

$$\langle \tilde{A}_0 f, g\rangle_{(D(A_0^{1/2}))'\times D(A_0^{1/2})} = (A_0^{1/2} f, A_0^{1/2} g)_H \qquad \forall\ f, g \in D(A_0^{1/2}). \tag{2.15}$$

Hence, (2.14) can be rewritten in $\mathcal{H}_{-1}$ as

$$\ddot{w} + \tilde{A}_0 w = B_0 u, \tag{2.16}$$

where $B_0 \subset \mathcal{L}(U, (D(A_0^{1/2}))')$ is given by

$$B_0 u = \tilde{A}_0 \Upsilon u \qquad \forall\ u \in U. \tag{2.17}$$

Define $B_0^* \in \mathcal{L}(D(A_0^{1/2}), U)$ by

$$(B_0^* f, u)_U = \langle f, B_0 u\rangle_{D(A_0^{1/2})\times (D(A_0^{1/2}))'} \qquad \forall\ f \in D(A_0^{1/2}).$$

Then for any $f \in D(A_0^{1/2})$ and $u \in C_0^\infty(\Gamma_0)$, we have

$$\begin{aligned}
&\langle f, B_0 u\rangle_{D(A_0^{1/2})\times (D(A_0^{1/2}))'} = \langle \tilde{A}_0 f, \tilde{A}_0^{-1} B_0 u\rangle_{D(A_0^{1/2})\times (D(A_0^{1/2}))'} \\
&= (A_0^{1/2} f, A_0^{1/2} \tilde{A}_0^{-1} B_0 u)_H = (A_0^{-1} A_0^{1/2} f, A_0^{-1} A_0^{1/2} \tilde{A}_0^{-1} B_0 u)_{H_0^1(\Omega)} \\
&= (A_0^{-1/2} f, A_0^{-1/2} \Upsilon u)_{H_0^1(\Omega)} = (f, \Upsilon u)_{L^2(\Omega)} \\
&= (A_0 A_0^{-1} f, \Upsilon u)_{L^2(\Omega)} = -\left(\frac{\partial (-\Delta)^{-1} f}{\partial \nu}, u\right)_U.
\end{aligned}$$

In the last step, we used the fact that

$$\int_\Omega \nabla v \nabla \phi = 0 \qquad \forall\ \phi \in H_0^1(\Omega)$$

holds for any classical solution v of (2.13). Since $C_0^\infty(\Gamma_0)$ is dense in $L^2(\Gamma_0)$, we obtain

$$B_0^* = -\frac{\partial (-\Delta)^{-1}}{\partial \nu}\bigg|_{\Gamma_0}. \tag{2.18}$$

Now, we have formulated system (1.1) into an abstract form of the second-order system (2.10) in the state space $\mathcal{H}$:

$$\begin{cases} \ddot{w}(t) + \tilde{A}_0 w(t) = B_0 u(t), \\ y(t) = B_0^* \dot{w}, \end{cases} \tag{2.19}$$

where B_0 and B_0^* are defined by (2.17) and (2.18), respectively.

The main contribution of this paper is to show that system (2.19) is regular with feedthrough operator $\mathbb{D} = I$.

3. Proof of Theorem 1.2. From (2.19), we see that system (1.1) is in the framework of form (2.10) discussed in section 2. Since system (1.1) is well-posed, it follows from (2.11) that the transfer function of system (1.1) is

$$H(\lambda) = \lambda B_0^* (\lambda^2 + \tilde{A}_0)^{-1} B_0, \tag{3.1}$$

where $\tilde{A}_0$, B_0, and B_0^* are given by (2.15), (2.17), and (2.18), respectively. Moreover, from the well-posedness and (2.5), it follows that there exists a positive number $\alpha > 0$ such that

$$\sup_{\mathrm{Re}\lambda \geq \alpha} \|H(\lambda)\|_{\mathcal{L}(U)} = M < \infty. \tag{3.2}$$

To begin, we show the following proposition.

PROPOSITION 3.1. *Theorem* 1.2 *is valid if for any* $u \in C_0^\infty(\Gamma_0)$ *the solution* u_ε *to the equation*

$$\begin{cases} u_\varepsilon(x) - \varepsilon^2 \Delta u_\varepsilon(x) = 0, & x \in \Omega, \\ u_\varepsilon(x) = 0, & x \in \Gamma_1, \\ u_\varepsilon(x) = u(x), & x \in \Gamma_0 \end{cases} \tag{3.3}$$

satisfies

$$\lim_{\varepsilon \to 0} \int_{\Gamma_0} \left| \varepsilon \frac{\partial u_\varepsilon(x)}{\partial \nu} - u(x) \right|^2 dx = 0, \tag{3.4}$$

where ε *are real and positive numbers.*

Proof. In light of the equivalence between (2.6) and (2.7), in order to prove Theorem 1.2 we need only to show that

$$\lim_{\lambda \in \mathbb{R},\, \lambda \to +\infty} H(\lambda) u = u \tag{3.5}$$

for any $u \in L^2(\Gamma_0) = U$ in the strong topology of U, where $H(\lambda)$ is given by (3.1). We claim that in order to show (3.5), it suffices to show that (3.5) is satisfied for all $u \in C_0^\infty(\Gamma_0)$. Indeed, for any $u \in U$ and any given $\delta > 0$, since $C_0^\infty(\Gamma_0)$ is dense in $L^2(\Gamma_0)$, if (3.5) is valid for $u \in C_0^\infty(\Gamma_0)$, then one can find $u_0 \in C_0^\infty(\Gamma_0)$ and the real number $\beta > \alpha$ such that

$$\|u_0 - u\|_U < \min\left\{ \frac{\delta}{3M}, \frac{\delta}{3} \right\}, \qquad \sup_{\lambda \in \mathbb{R},\, \lambda > \beta} \|H(\lambda) u_0 - u_0\|_U < \frac{\delta}{3},$$

where M and α are given in (3.2). Therefore,

$$\sup_{\lambda \in \mathbb{R},\, \lambda > \beta} \|H(\lambda) u - u\|_U = \sup_{\lambda \in \mathbb{R},\, \lambda > \beta} \|H(\lambda) u_0 - u_0 + H(\lambda)(u - u_0) - u + u_0\|_U < \delta.$$

This shows that (3.5) is valid for any $u \in U$.

Now assume that $u \in C_0^\infty(\Gamma_0)$, and put

$$w_\lambda(x) = ((\lambda^2 + \tilde{A}_0)^{-1} B_0 u)(x).$$

Then w_λ satisfies

$$\begin{cases} \lambda^2 w_\lambda(x) - \Delta w_\lambda(x) = 0, & x \in \Omega, \\ w_\lambda(x) = 0, & x \in \Gamma_1, \\ w_\lambda(x) = u(x), & x \in \Gamma_0, \end{cases} \tag{3.6}$$

and

$$(H(\lambda) u)(x) = -\lambda \frac{\partial((-\Delta)^{-1} w_\lambda)(x)}{\partial \nu} \qquad \forall\, x \in \Gamma_0. \tag{3.7}$$

Since $u \in C_0^\infty(\Gamma_0)$, there exists a unique classical solution to (3.6). Take a function $v \in H^2(\Omega)$ such that

$$\begin{cases} \Delta v(x) = 0, & x \in \Omega, \\ v(x) = 0, & x \in \Gamma_1, \\ v(x) = u(x), & x \in \Gamma_0. \end{cases} \tag{3.8}$$

Then (3.6) can be written as

$$\begin{cases} \lambda^2 w_\lambda(x) - \Delta(w_\lambda(x) - v(x)) = 0, \quad x \in \Omega, \\ (w_\lambda - v)\big|_{\partial\Omega} = 0, \end{cases} \tag{3.9}$$

or equivalently

$$-\lambda^2((-\Delta)^{-1} w_\lambda)(x) = w_\lambda(x) - v(x).$$

Hence (3.7) becomes

$$(H(\lambda)u)(x) = \frac{1}{\lambda}\frac{\partial w_\lambda(x)}{\partial \nu} - \frac{1}{\lambda}\frac{\partial v(x)}{\partial \nu}. \tag{3.10}$$

Letting $u_\varepsilon(x) = w_\lambda(x)$ with $\varepsilon = \lambda^{-1}$ and noting that $\frac{\partial v(x)}{\partial \nu}$ is independent of λ, we conclude the required result. □

The rest of this section is devoted to proving that the solution u_ε of (3.3) with $u \in C_0^\infty(\Gamma_0)$ satisfies (3.4). We shall go a little bit further. Indeed, we will show that there exists a constant $C > 0$ such that for all $\varepsilon \in (0, 1)$, any solution $u_\varepsilon \in H^2(\Omega)$ of

$$\left(\varepsilon^2 \Delta - 1\right) u_\varepsilon(x) = 0, \quad x \in \Omega,$$

satisfies the following inequality:

$$\left\| \varepsilon \frac{\partial u_\varepsilon}{\partial \nu} - u_\varepsilon \right\|^2_{L^2(\partial\Omega)} \leq C\varepsilon \left\| u_\varepsilon \right\|^2_{H^{3/2}(\partial\Omega)}.$$

This will be performed by estimating the Dirichlet–Neumann map by means of easy Fourier analysis tools after applying a diffeomorphism to reduce locally our geometry to the half-space. Notice that the Dirichlet–Neumann map for the Laplacian in a manifold was more precisely computed in [18] by using symbolic calculus of pseudodifferential operators.

Proof of Theorem 1.2. By Proposition 3.1, we need only to show that the solution u_ε of (3.3) with $u \in C_0^\infty(\Gamma_0)$ satisfies (3.4) . We assume $0 < \varepsilon < 1$ throughout the proof.

For any $x_0 \in \partial\Omega$, suppose without loss of generality that in an open neighborhood $V_{x_0} \subset \mathbb{R}^n$ of x_0,

$$V_{x_0} \cap \Omega = \{(x', x_n) = (x_1, x_2, \dots, x_{n-1}, x_n) \in V_{x_0},\ x_n - \phi(x') > 0\}$$

for some $\phi \in C^3(\mathbb{R}^{n-1})$. Then the unit outward normal vector to $V_{x_0} \cap \partial\Omega$ at $(x', \phi(x'))$ is defined by

$$\nu(x') = \frac{\left(\partial_{x_1}\phi(x'), \dots, \partial_{x_{n-1}}\phi(x'), -1\right)}{\sqrt{1 + |\nabla\phi(x')|^2}}.$$

Let us use the geodesic normal coordinates as follows. Let

$$(h,s)=(h_1,h_2,\dots,h_{n-1},s)\in\mathbb{R}^n.$$

We introduce a diffeomorphism by

$$\Psi(h,s)=(h,\phi(h))-s\nu(h)$$

such that

(i) $\Psi^{-1}(\Omega_{x_0})=B_r=\{(h,s)\in\mathbb{R}^n,|(h,s)|<r\}$;
(ii) $\Psi^{-1}(\Omega_{x_0}\cap\Omega)=B_r^+=\{(h,s)\in B_r,s>0\}$;
(iii) $\Psi^{-1}(\Omega_{x_0}\cap\partial\Omega)=\{(h,s)\in B_r,s=0\}=\{|h|<r\}\times\{0\}$

for some $r>0$ and an open neighborhood $\Omega_{x_0}(\subset V_{x_0})$ of x_0, where $|\cdot|$ denotes the Euclidean norm. Using the diffeomorphism $\Psi:B_r\to\Omega_{x_0}$, the normal derivative on the boundary becomes

$$\frac{\partial}{\partial\nu}=-\partial_s,$$

and the operator in the first equation of (3.3) can be written in the form

$$\Delta-\frac{1}{\varepsilon^2}=\partial_s^2+P\,(h,s,-i\partial_h)+\ell(h,s)\partial_s-\frac{1}{\varepsilon^2},$$

where $\partial_h=(\partial_{h_1},\dots,\partial_{h_{n-1}})$, ℓ is a continuous function, and P is a second-order elliptic differential operator in the h variables only.

The proof is now divided into three steps.

Step 1. *Flattening and localization.* We first flatten the local domain $\Omega_{x_0}\cap\Omega$ with the above diffeomorphism Ψ and set

$$\tilde{u}_\varepsilon(h,s)=u_\varepsilon(\Psi(h,s)),\qquad \tilde{u}(h)=u_\varepsilon(\Psi(h,0)). \tag{3.11}$$

Then $\tilde{u}_\varepsilon$ satisfies

$$\begin{cases}\partial_s^2\tilde{u}_\varepsilon(h,s)+\displaystyle\sum_{i,j=1}^{n-1}a_{ij}(h,s)\partial_{h_i}\partial_{h_j}\tilde{u}_\varepsilon(h,s)+Q\tilde{u}_\varepsilon(h,s)-\frac{1}{\varepsilon^2}\tilde{u}_\varepsilon(h,s)=0,\\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (h,s)\in B_r^+,\\ \tilde{u}_\varepsilon(h,0)=\tilde{u}(h),\qquad |h|<r,\end{cases} \tag{3.12}$$

where Q is a linear differential operator of order 1 with continuous coefficients in B_r and $(a_{ij})_{1\le i,j\le n-1}$ is a strictly positive definite symmetric matrix of continuous functions of (h,s) in B_r. Assume that $\lambda_0>0$ is a constant such that

$$\sum_{i,j=1}^{n-1}a_{ij}(h,s)\xi_i\xi_j\ge\lambda_0|\xi|^2\quad\forall\,\xi=(\xi_1,\xi_2,\dots,\xi_{n-1})\in\mathbb{R}^{n-1},\quad(h,s)\in B_r. \tag{3.13}$$

Let $\mu_0>0$ be such that $\mu_0<\frac{\lambda_0}{(n-1)^2}$. Since a_{ij} is continuous in B_r, one can find a scalar $\rho\in(0,r)$ such that

$$|a_{ij}(h,s)-a_{ij}(0,0)|\le\mu_0\qquad\forall\,i,j=1,2,\dots,n-1,\quad(h,s)\in B_\rho^+. \tag{3.14}$$

Second, we introduce a cutoff function $\varphi = \varphi(h,s) \in C_0^\infty(B_\rho)$ such that $0 \le \varphi \le 1$ and $\varphi = 1$ in $B_{\rho/2}$. Set, for all $(h,s) \in \mathbb{R}^{n-1} \times \mathbb{R}^+$,

$$\chi_\varepsilon(h,s) = \varphi(h,s)\tilde{u}_\varepsilon(h,s), \qquad f(h) = \varphi(h,0)\tilde{u}(h). \tag{3.15}$$

Then one can check that $\chi_\varepsilon \in H^2(\mathbb{R}^{n-1} \times \mathbb{R}^+)$ and $\chi_\varepsilon(h,s) = 0$ in $\mathbb{R}^{n-1} \times \{s \ge \rho\}$. By (3.12), χ_ε satisfies

$$\begin{cases} \partial_s^2 \chi_\varepsilon(h,s) + \displaystyle\sum_{i,j=1}^{n-1} a_{ij}(0,0)\partial_{h_i}\partial_{h_j}\chi_\varepsilon(h,s) - \frac{1}{\varepsilon^2}\chi_\varepsilon(h,s) \\ \qquad = G\chi_\varepsilon(h,s) + L\tilde{u}_\varepsilon(h,s), \qquad (h,s) \in \mathbb{R}^{n-1} \times \mathbb{R}^+, \\ \chi_\varepsilon(h,0) = f(h), \qquad h \in \mathbb{R}^{n-1}, \end{cases} \tag{3.16}$$

where

$$\begin{cases} G\chi_\varepsilon(h,s) &= \displaystyle\sum_{i,j=1}^{n-1} [a_{ij}(0,0) - a_{ij}(h,s)]\partial_{h_i}\partial_{h_j}\chi_\varepsilon(h,s), \\ L\tilde{u}_\varepsilon(h,s) &= -\varphi(h,s)Q\tilde{u}_\varepsilon(h,s) + [\partial_s^2, \varphi]\tilde{u}_\varepsilon(h,s) \\ &\quad + \displaystyle\sum_{i,j=1}^{n-1} a_{ij}(h,s)[\partial_{h_i}\partial_{h_j}, \varphi]\tilde{u}_\varepsilon(h,s) \end{cases} \tag{3.17}$$

with

$$[\partial_s^2, \varphi]\tilde{u}_\varepsilon = 2\partial_s\varphi\partial_s\tilde{u}_\varepsilon + \partial_s^2\varphi\tilde{u}_\varepsilon, \;\; [\partial_{h_i}\partial_{h_j}, \varphi]\tilde{u}_\varepsilon = \partial_{h_i}\varphi\partial_{h_j}\tilde{u}_\varepsilon + \partial_{h_j}\varphi\partial_{h_i}\tilde{u}_\varepsilon + \partial_{h_i}\partial_{h_j}\varphi\tilde{u}_\varepsilon.$$

Clearly, G and L are two linear differential operators of order 2 and order 1, respectively.

Step 2. *Partial Fourier transform.* Fix s, for any $\chi(\cdot,s) \in L^2(\mathbb{R}^{n-1})$. From now on, we denote by $\widehat{\chi}(\xi,s)$ the partial Fourier transform of $\chi(h,s)$ with respect to h, i.e.,

$$\widehat{\chi}(\xi,s) = \int_{\mathbb{R}^{n-1}} \chi(h,s)e^{-i\langle h,\xi\rangle}dh.$$

Applying the above partial Fourier transform to system (3.16), it becomes

$$\begin{cases} \partial_s^2\widehat{\chi_\varepsilon}(\xi,s) - \dfrac{1}{\varepsilon^2}(\varepsilon^2\xi^\top A\xi + 1)\widehat{\chi_\varepsilon}(\xi,s) = \widehat{G\chi_\varepsilon}(\xi,s) + \widehat{L\tilde{u}_\varepsilon}(\xi,s), \\ \qquad (\xi,s) \in \mathbb{R}^{n-1} \times \mathbb{R}^+, \\ \widehat{\chi_\varepsilon}(\xi,0) = \widehat{f}(\xi), \qquad \xi \in \mathbb{R}^{n-1}, \end{cases} \tag{3.18}$$

where $A = \{a_{ij}(0,0)\}_{1\le i,j\le n-1}$ is a positive definite symmetric matrix. Notice that

$$\widehat{\chi_\varepsilon}(\xi,s) = 0 \qquad \forall (\xi,s) \in \mathbb{R}^{n-1} \times [\rho, +\infty)\,. \tag{3.19}$$

To analyze the solution of (3.18) satisfying (3.19), we decompose $\widehat{\chi_\varepsilon}(\xi,s)$ as follows. Let

$$\widehat{\chi_\varepsilon}(\xi,s) = w_\varepsilon(\xi,s) + v_\varepsilon(\xi,s), \qquad (\xi,s) \in \mathbb{R}^{n-1} \times \mathbb{R}^+, \tag{3.20}$$

where w_ε satisfies

$$\begin{cases} \partial_s^2 w_\varepsilon(\xi,s) - \dfrac{1}{\varepsilon^2}(\varepsilon^2 \xi^\top A\xi + 1) w_\varepsilon(\xi,s) = 0, \qquad (\xi,s) \in \mathbb{R}^{n-1} \times \mathbb{R}^+, \\ w_\varepsilon(\xi,0) = \widehat{f}(\xi), \qquad \xi \in \mathbb{R}^{n-1}, \\ \lim\limits_{s\to+\infty} w_\varepsilon(\xi,s) = 0, \qquad \xi \in \mathbb{R}^{n-1}, \end{cases} \tag{3.21}$$

and v_ε satisfies

$$\begin{cases} \partial_s^2 v_\varepsilon(\xi,s) - \dfrac{1}{\varepsilon^2}(\varepsilon^2 \xi^\top A\xi + 1) v_\varepsilon(\xi,s) = \widehat{G\chi_\varepsilon}(\xi,s) + \widehat{L\widetilde{u}_\varepsilon}(\xi,s), \\ \qquad (\xi,s) \in \mathbb{R}^{n-1} \times \mathbb{R}^+, \\ v_\varepsilon(\xi,0) = 0, \qquad \xi \in \mathbb{R}^{n-1}, \\ v_\varepsilon(\xi,s) = -\widehat{f}(\xi) e^{-s\frac{\sqrt{\varepsilon^2\xi^\top A\xi+1}}{\varepsilon}}, \qquad (\xi,s) \in \mathbb{R}^{n-1} \times [\rho,+\infty)\,. \end{cases} \tag{3.22}$$

The validity of the last equality comes from (3.19) and the following explicit expression of the solution of (3.21):

$$w_\varepsilon(\xi,s) = \widehat{f}(\xi) e^{-s\frac{\sqrt{\varepsilon^2\xi^\top A\xi+1}}{\varepsilon}}. \tag{3.23}$$

We claim that there exists a constant $C > 0$ such that for all $\varepsilon \in (0,1)$

$$\int_{\mathbb{R}^{n-1}} |\varepsilon \partial_s w_\varepsilon(\xi,0) + w_\varepsilon(\xi,0)|^2 d\xi \le C\varepsilon^2 \|u_\varepsilon\|^2_{H^1(\partial\Omega)}. \tag{3.24}$$

Indeed, by (3.23), we get

$$\int_{\mathbb{R}^{n-1}} |\varepsilon \partial_s w_\varepsilon(\xi,0) + w_\varepsilon(\xi,0)|^2 d\xi = \int_{\mathbb{R}^{n-1}} \left(\frac{\varepsilon^2 \xi^\top A\xi}{\sqrt{\varepsilon^2\xi^\top A\xi + 1} + 1} \right)^2 |\widehat{f}(\xi)|^2 d\xi$$

$$\le \int_{\mathbb{R}^{n-1}} \varepsilon^2 \xi^\top A\xi |\widehat{f}(\xi)|^2 d\xi,$$

and (3.24) follows easily.

Now we need to bound the quantity $\int_{\mathbb{R}^{n-1}} |\varepsilon \partial_s v_\varepsilon(\xi,0) + v_\varepsilon(\xi,0)|^2 d\xi$ uniformly with respect to ε. This will be done in the next step.

Step 3. *Estimate of* $\varepsilon\partial_s v_\varepsilon(\cdot,0) + v_\varepsilon(\cdot,0)$. We will estimate $\partial_s v_\varepsilon(\cdot,0)$ by means of a classical trace theorem. This requires the computation of $\partial_s^2 v_\varepsilon$ and $\partial_s v_\varepsilon$. To do it, we estimate $\widehat{L\widetilde{u}_\varepsilon}$ and $\widehat{G\chi_\varepsilon}$ first. Throughout the proof, C denotes several positive constants independent of ε.

(a) *Estimate of* $\widehat{L\widetilde{u}_\varepsilon}$ *and* $\widehat{G\chi_\varepsilon}$. Clearly, we have

$$\left\| \widehat{L\widetilde{u}_\varepsilon} \right\|_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)} \le C \left\| u_\varepsilon \right\|_{H^1(\Omega)}. \tag{3.25}$$

By (3.14) and the Plancherel formula, it follows that

$$\begin{aligned} \|\widehat{G\chi_\varepsilon}\|_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)} &= (2\pi)^{\frac{n-1}{2}} \|G\chi_\varepsilon\|_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)} \\ &\le (2\pi)^{\frac{n-1}{2}} \mu_0 \sum_{i,j=1}^{n-1} \|\partial_{h_i}\partial_{h_j}\chi_\varepsilon\|_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)} \\ &\le \mu_0 (n-1)^2 \||\xi|^2 \widehat{\chi_\varepsilon}\|_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)}. \end{aligned} \tag{3.26}$$

From (3.26) and noting (3.20), we find

$$\|\widehat{G\chi_\varepsilon}\|_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)} \le \mu_0(n-1)^2 \left\||\xi|^2 w_\varepsilon\right\|_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)} + \mu_0(n-1)^2 \left\||\xi|^2 v_\varepsilon\right\|_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)}. \tag{3.27}$$

On the other hand, multiplying (3.22) by $-|\xi|^2\overline{v_\varepsilon}$ and then integrating by parts over $\mathbb{R}^{n-1}\times\mathbb{R}^+$, taking (3.13) and the last equality of (3.22) into account, we have

$$\lambda_0 \left\||\xi|^2 v_\varepsilon\right\|_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)} \le \|\widehat{G\chi_\varepsilon}\|_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)} + \|\widehat{L\tilde{u}_\varepsilon}\|_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)}. \tag{3.28}$$

Substituting (3.28) into (3.27), we get

$$\begin{aligned} &\left(1-\tfrac{\mu_0(n-1)^2}{\lambda_0}\right)\left\|\widehat{G\chi_\varepsilon}\right\|_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)} \\ &\le \mu_0(n-1)^2\left\||\xi|^2 w_\varepsilon\right\|_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)} + \tfrac{\mu_0(n-1)^2}{\lambda_0}\left\|\widehat{L\tilde{u}_\varepsilon}\right\|_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)}. \end{aligned} \tag{3.29}$$

Moreover, from (3.23) and (3.13), we have

$$\begin{aligned} \left\||\xi|^2 w_\varepsilon\right\|_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)} &= \left(\int_{\mathbb{R}^{n-1}} \left||\xi|^2\,\widehat{f}(\xi)\right|^2 \left(\int_0^{+\infty} e^{-2s\frac{\sqrt{\varepsilon^2\xi^\top A\xi+1}}{\varepsilon}} ds\right) d\xi\right)^{1/2} \\ &= \left\|\sqrt{\frac{\varepsilon}{2\sqrt{\varepsilon^2\xi^\top A\xi+1}}}\,|\xi|^2\,\widehat{f}\right\|_{L^2(\mathbb{R}^{n-1})} \\ &\le \sqrt{\tfrac{1}{2\sqrt{\lambda_0}}}\left\||\xi|^{3/2}\,\widehat{f}\right\|_{L^2(\mathbb{R}^{n-1})} \le C\,\|u_\varepsilon\|_{H^{3/2}(\partial\Omega)}. \end{aligned} \tag{3.30}$$

Finally, it follows from (3.29), (3.30), and (3.25) that

$$\left\|\widehat{G\chi_\varepsilon}\right\|_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)} \le C\left(\|u_\varepsilon\|_{H^{3/2}(\partial\Omega)} + \|u_\varepsilon\|_{H^1(\Omega)}\right). \tag{3.31}$$

(b) *Estimate of* $\partial_s^2 v_\varepsilon$. Multiplying (3.22) by $\overline{\partial_s^2 v_\varepsilon}$ and then integrating by parts over $\mathbb{R}^{n-1}\times\mathbb{R}^+$, we obtain, noticing the last equality of (3.22),

$$\|\partial_s^2 v_\varepsilon\|^2_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)} \le \left(\|\widehat{G\chi_\varepsilon}\|_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)} + \|\widehat{L\tilde{u}_\varepsilon}\|_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)}\right)\|\partial_s^2 v_\varepsilon\|_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)}.$$

This together with (3.25) and (3.31) gives

$$\|\partial_s^2 v_\varepsilon\|_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)} \le C[\|u_\varepsilon\|_{H^{3/2}(\partial\Omega)} + \|u_\varepsilon\|_{H^1(\Omega)}]. \tag{3.32}$$

(c) *Estimate of* $\partial_s v_\varepsilon$. Noticing the last equality of (3.22), multiplying (3.22) by $-\overline{v_\varepsilon}$, and integrating by parts over $\mathbb{R}^{n-1}\times\mathbb{R}^+$, we also have

$$\begin{aligned} &\|\partial_s v_\varepsilon\|^2_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)} + \frac{1}{\varepsilon^2}\|v_\varepsilon\|^2_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)} \\ &\le \|\varepsilon(\widehat{G\chi_\varepsilon} + \widehat{L\tilde{u}_\varepsilon})\|_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)}\left\|\frac{v_\varepsilon}{\varepsilon}\right\|_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)}. \end{aligned}$$

Thus,

$$\|\partial_s v_\varepsilon\|_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)} \le \varepsilon\left(\|\widehat{G\chi_\varepsilon}\|_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)} + \|\widehat{L\tilde{u}_\varepsilon}\|_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)}\right).$$

This together with (3.25) and (3.31) gives

$$\|\partial_s v_\varepsilon\|_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)} \le C[\|u_\varepsilon\|_{H^{3/2}(\partial\Omega)} + \|u_\varepsilon\|_{H^1(\Omega)}]. \tag{3.33}$$

(d) *Estimate of* $\partial_s v_\varepsilon(\cdot,0)$. We use the following standard inequality:

$$\begin{aligned}\int_{\mathbb{R}^{n-1}} |\partial_s v_\varepsilon(\xi,0)|^2\, d\xi &= -2\int_{\mathbb{R}^{n-1}}\int_0^{+\infty} \mathrm{Re}\left(\partial_s v_\varepsilon(\xi,s)\overline{\partial_s^2 v_\varepsilon}(\xi,s)\right) ds d\xi \\ &\le \|\partial_s v_\varepsilon\|^2_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)} + \left\|\partial_s^2 v_\varepsilon\right\|^2_{L^2(\mathbb{R}^{n-1}\times\mathbb{R}^+)}.\end{aligned} \tag{3.34}$$

This together with (3.32) and (3.33) gives the desired estimate for v_ε:

$$\int_{\mathbb{R}^{n-1}} |\varepsilon\partial_s v_\varepsilon(\xi,0) + v_\varepsilon(\xi,0)|^2 d\xi \le C\varepsilon^2[\|u_\varepsilon\|^2_{H^{3/2}(\partial\Omega)} + \|u_\varepsilon\|^2_{H^1(\Omega)}]. \tag{3.35}$$

Here we used the fact that $v_\varepsilon(\cdot,0)=0$ given by the second equation of (3.22).

Combining (3.20), the estimates (3.24) and (3.35) imply

$$\int_{\mathbb{R}^{n-1}} |\varepsilon\partial_s \widehat{\chi_\varepsilon}(\xi,0) + \widehat{\chi_\varepsilon}(\xi,0)|^2 d\xi \le C\varepsilon^2[\|u_\varepsilon\|^2_{H^{3/2}(\partial\Omega)} + \|u_\varepsilon\|^2_{H^1(\Omega)}], \tag{3.36}$$

and hence by the Parseval formula, we obtain

$$\int_{\mathbb{R}^{n-1}} |\varepsilon\partial_s \chi_\varepsilon(s,0) + \chi_\varepsilon(s,0)|^2 ds \le C\varepsilon^2[\|u_\varepsilon\|^2_{H^{3/2}(\partial\Omega)} + \|u_\varepsilon\|^2_{H^1(\Omega)}]. \tag{3.37}$$

By (3.15), we deduce from (3.37) that

$$\int_{|s|<\rho/2} |\varepsilon\partial_s \tilde{u}_\varepsilon(s,0) + \tilde{u}_\varepsilon(s,0)|^2 ds \le C\varepsilon^2[\|u_\varepsilon\|^2_{H^{3/2}(\partial\Omega)} + \|u_\varepsilon\|^2_{H^1(\Omega)}], \tag{3.38}$$

which implies by the change of coordinates involving Ψ that

$$\int_{\tilde{\Omega}_{x_0}\cap\partial\Omega} \left|\varepsilon\frac{\partial u_\varepsilon(x)}{\partial\nu} - u_\varepsilon(x)\right|^2 dx \le C\varepsilon^2[\|u_\varepsilon\|^2_{H^{3/2}(\partial\Omega)} + \|u_\varepsilon\|^2_{H^1(\Omega)}], \tag{3.39}$$

where $\tilde{\Omega}_{x_0}\subset\Omega_{x_0}$ is an open neighborhood of $x_0\in\partial\Omega$. Since x_0 is arbitrarily chosen, one easily deduces from (3.39) that

$$\left\|\varepsilon\frac{\partial u_\varepsilon}{\partial\nu} - u_\varepsilon\right\|^2_{L^2(\partial\Omega)} \le C\varepsilon^2[\|u_\varepsilon\|^2_{H^{3/2}(\partial\Omega)} + \|u_\varepsilon\|^2_{H^1(\Omega)}]. \tag{3.40}$$

Now, multiplying (3.3) by u_ε and integrating by parts, we find

$$\|\varepsilon\nabla u_\varepsilon\|^2_{L^2(\Omega)} + \|u_\varepsilon\|^2_{L^2(\Omega)} = \varepsilon^2\int_{\partial\Omega}\frac{\partial u_\varepsilon(x)}{\partial\nu}u_\varepsilon(x)dx.$$

Hence, using the Cauchy–Schwarz inequality, we get

$$\begin{aligned}\varepsilon^2\|u_\varepsilon\|^2_{H^1(\Omega)} &\le \varepsilon\left(\left\|\varepsilon\frac{\partial u_\varepsilon}{\partial\nu} - u_\varepsilon\right\|_{L^2(\partial\Omega)} + \|u_\varepsilon\|_{L^2(\partial\Omega)}\right)\|u_\varepsilon\|_{L^2(\partial\Omega)} \\ &\le \frac{\varepsilon}{2C}\left\|\varepsilon\frac{\partial u_\varepsilon}{\partial\nu} - u_\varepsilon\right\|^2_{L^2(\partial\Omega)} + \left(1+\frac{C}{2}\right)\varepsilon\|u_\varepsilon\|^2_{L^2(\partial\Omega)}.\end{aligned}$$

Substituting the above formula into (3.40), we have finally proved that there exists a constant $C > 0$ such that for all $\varepsilon \in (0, 1)$ any solution $u_\varepsilon \in H^2(\Omega)$ of

$$\left(\varepsilon^2 \Delta - 1\right) u_\varepsilon(x) = 0, \qquad x \in \Omega,$$

satisfies

$$\left\| \varepsilon \frac{\partial u_\varepsilon}{\partial \nu} - u_\varepsilon \right\|^2_{L^2(\partial\Omega)} \leq C\varepsilon \|u_\varepsilon\|^2_{H^{3/2}(\partial\Omega)}. \tag{3.41}$$

Therefore,

$$\lim_{\varepsilon \to 0} \left\| \varepsilon \frac{\partial u_\varepsilon}{\partial \nu} - u \right\|_{L^2(\Gamma_0)} = 0.$$

This completes the proof of Theorem 1.2. □

Acknowledgments. The special case of Theorem 1.2 in a 2-D disk was first proved in [12] when Bao-Zhu Guo was visiting INRIA in Metz, France in 2002. The authors would like to thank the anonymous referees for their careful reading, helpful suggestions, and many corrections of the manuscript.

REFERENCES

[1] K. Ammari, *Dirichlet boundary stabilization of the wave equation*, Asymptot. Anal., 30 (2002), pp. 117–130.
[2] K. Ammari and M. Tucsnak, *Stabilization of second order evolution equations by a class of unbounded feedbacks*, ESAIM Control Optim. Calc. Var., 6 (2001), pp. 361–386.
[3] C. I. Byrnes, D. S. Gilliam, V. I. Shubov, and G. Weiss, *Regular linear systems governed by a boundary controlled heat equation*, J. Dynam. Control Systems, 8 (2002), pp. 341–370.
[4] A. Cheng and K. Morris, *Well-posedness of boundary control systems*, SIAM J. Control Optim., 42 (2003), pp. 1244–1265.
[5] R. F. Curtain, *The Salamon–Weiss class of well-posed infinite dimensional linear systems: A survey*, IMA J. Math. Control Inform., 14 (1997), pp. 207–223.
[6] R. F. Curtain, *Linear operator inequalities for strongly stable weakly regular linear systems*, Math. Control Signals Systems, 14 (2001), pp. 299–337.
[7] R. F. Curtain, H. Logemann, and O. J. Staffans, *Absolute-stability results in infinite dimensions*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 460 (2004), pp. 2171–2196.
[8] R. F. Curtain and G. Weiss, *Well posedness of triples of operators (in the sense of linear systems theory)*, in Control and Estimation of Distributed Parameter Systems, Internat. Ser. Numer. Math. 91, F. Kappel, K. Kunisch, and W. Schappacher, eds., Birkhäuser, Basel, 1989, pp. 41–59.
[9] R. F. Curtain and H. Zwart, *An Introduction to Infinite Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1995.
[10] E. A. Guillemin, *Synthesis of Passive Networks*, Wiley, New York, 1957.
[11] B. Z. Guo and Y. H. Luo, *Controllability and stability of a second order hyperbolic system with colocated sensor/actuator*, Systems Control Lett., 46 (2002), pp. 45–65.
[12] B. Z. Guo and C. Z. Xu, *Regularity of the Transfer Function of a Wave Equation in a 2-D Disk with Dirichlet Control and Colocated Observation*, manuscript, 2002.
[13] B. Z. Guo and G. Q. Xu, *Riesz bases and exact controllability of C_0-groups with one-dimensional input operators*, Systems Control Lett., 52 (2004), pp. 221–232.
[14] B. Jacob and H. Zwart, *Equivalent conditions for stabilizability of infinite-dimensional systems with admissible control operators*, SIAM J. Control Optim., 37 (1999), pp. 1419–1455.
[15] B. Jacob and H. Zwart, *Exact observability of diagonal systems with a finite-dimensional output operator*, Systems Control Lett., 43 (2001), pp. 101–109.
[16] B. Jacob and H. Zwart, *Properties of the realization of inner functions*, Math. Control Signals Systems, 15 (2002), pp. 356–379.
[17] I. Lasiecka, J. L. Lions, and R. Triggiani, *Nonhomogeneous boundary value problems for second order hyperbolic operators*, J. Math. Pures Appl. (9), 65 (1986), pp. 149–192.

[18] J. Lee and G. Uhlmann, *Determining anisotropic real-analytic conductives by boundary measurements*, Comm. Pure Appl. Math., 42 (1989), pp. 1097–1112.
[19] J. L. Lions, *Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués,* Tome 1*, Contrôlabilité exacte*, Recherches en Mathématiques Appliquées 8, Masson, Paris, 1988.
[20] R. Rebarber and G. Weiss, *Necessary conditions for exact controllability with a finite-dimensional input space*, Systems Control Lett., 40 (2000), pp. 217–227.
[21] D. L. Russell and G. Weiss, *A general necessary condition for exact observability*, SIAM J. Control Optim., 32 (1994), pp. 1–23.
[22] R. Schnaubelt, *Feedbacks for nonautonomous regular linear systems*, SIAM J. Control Optim., 41 (2002), pp. 1141–1165.
[23] D. Salamon, *Infinite dimensional systems with unbounded control and observation: A functional analytic approach*, Trans. Amer. Math. Soc., 300 (1987), pp. 383–431.
[24] D. Salamon, *Realization theory in Hilbert space*, Math. Systems Theory, 21 (1989), pp. 147–164.
[25] O. J. Staffans, *Quadratic optimal control of well-posed linear systems*, SIAM J. Control Optim., 37 (1998), pp. 131–164.
[26] O. J. Staffans, *Admissible factorizations of Hankel operators induce well-posed linear systems*, Systems Control Lett., 37 (1999), pp. 301–307.
[27] O. J. Staffans, *Passive and conservative continuous-time impedance and scattering systems, part* I: *Well-posed systems*, Math. Control Signals Systems, 15 (2002), pp. 291–315.
[28] O. J. Staffans and G. Weiss, *Transfer functions of regular linear systems part* II: *The system operator and the Lax-Phillips semigroup*, Trans. Amer. Math. Soc., 354 (2002), pp. 3229–3262.
[29] M. Tucsnak and G. Weiss, *How to get a conservative well-posed linear system out of thin air. Part* II. *Controllability and stability*, SIAM J. Control Optim., 42 (2003), pp. 907–935.
[30] G. Weiss, *Admissibility of unbounded control operators*, SIAM J. Control Optim., 27 (1989), pp. 527–545.
[31] G. Weiss, *Admissibility observation operators for linear semigroups*, Israel J. Math., 65 (1989), pp. 17–43.
[32] G. Weiss, *The representation of regular linear systems in Hilbert spaces*, in Control and Estimation of Distributed Parameter Systems, F. Kappel, K. Kunisch, and W. Schappacher, eds., Internat. Ser. Numer. Math. 91, Birkhäuser, Basel, 1989, pp. 401–416.
[33] G. Weiss, *Regular linear systems with feedback*, Math. Control Signals Systems, 7 (1994), pp. 23–57.
[34] G. Weiss, *Transfer functions of regular linear systems* I: *Characterizations of regularity*, Trans. Amer. Math. Soc., 342 (1994), pp. 827–854.
[35] G. Weiss, *Optimal control of systems with a unitary semigroup and with colocated control and observation*, Systems Control Lett., 48 (2003), pp. 329–340.
[36] G. Weiss and R. F. Curtain, *Dynamic stabilization of regular linear systems*, IEEE Trans. Automat. Control, 42 (1997), pp. 4–21.
[37] G. Weiss, O. J. Staffans, and M. Tucsnak, *Well-posed linear systems—a survey with emphasis on conservative systems*, Int. J. Appl. Math. Comput. Sci., 11 (2001), pp. 7–13.
[38] C.-Z. Xu and G. Sallet, *On spectrum and Riesz basis assignment of infinite-dimensional linear systems by bounded linear feedbacks*, SIAM J. Control Optim., 34 (1996), pp. 521–541.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 5, pp. 1614–1635

REGULARITY OF THE FREE BOUNDARY IN AN OPTIMIZATION PROBLEM RELATED TO THE BEST SOBOLEV TRACE CONSTANT*

JULIÁN FERNÁNDEZ BONDER†, JULIO D. ROSSI‡, AND NOEMI WOLANSKI†

Abstract. In this paper we study the regularity properties of a free boundary problem arising in the optimization of the best Sobolev trace constant in the immersion $H^1(\Omega) \hookrightarrow L^q(\partial\Omega)$ for functions that vanish in a subset of Ω. This problem is also related to a minimization problem for Steklov eigenvalues.

Key words. Sobolev trace constant, free boundaries, eigenvalue optimization problems

AMS subject classifications. 35J20, 35P30, 49K20

DOI. 10.1137/040613615

1. Introduction.

The study of Sobolev inequalities and of optimal constants is a subject of interest in the analysis of PDEs and related topics. It has been widely studied in the past by many authors and is still an area of intensive research. See, for instance, the book [1] and, for recent developments in this field, see the articles [6, 11, 9, 17] and the survey [7] among others.

The optimal Sobolev constant and its corresponding extremals (if they exist) are related to eigenvalue problems. In the case of the best Sobolev trace embedding $H^1(\Omega) \to L^q(\partial\Omega)$, where Ω is a bounded smooth domain in $\mathbb{R}^N$, the best constant and the extremal (that exists for $1 \le q < 2_* = 2(N-1)/(N-2)$ since the immersion is compact) give rise to the following elliptic problem with nonlinear boundary conditions:

$$\begin{cases} -\Delta u + u = 0 & \text{in } \Omega, \\ \dfrac{\partial u}{\partial \nu} = \lambda u^{q-1} & \text{on } \partial\Omega. \end{cases}$$

The constant λ depends on the normalization of the extremal u. For instance, if u is chosen so that $\|u\|_{L^q(\partial\Omega)} = 1$, then $\lambda = S$, the best Sobolev trace constant. In the linear case, $q = 2$, this problem becomes an eigenvalue problem that is known as the *Steklov eigenvalue problem* [19].

In this paper we are interested in the best Sobolev trace constant among functions that vanish in a subset of Ω. We try to optimize this best constant when varying the subset in the class of measurable sets with prescribed positive measure α. In a previous article [10], we proved that there exists an optimal set. In this paper we focus our attention on regularity properties of these optimal sets.

*Received by the editors August 18, 2004; accepted for publication (in revised form) February 24, 2005; published electronically November 22, 2005. Supported by ANPCyT PICT 03-05009, 03-13719, and 03-10608, CONICET PIP0660/98 and PEI6388/04, UBA X052 and X066, and Fundación Antorchas 13900-5. J. D. Rossi and N. Wolanski are members of CONICET.

http://www.siam.org/journals/sicon/44-5/61361.html

†Departamento de Matemática, FCEyN, UBA (1428) Buenos Aires, Argentina (jfbonder@dm.uba.ar, http://mate.dm.uba.ar/∼jfbonder; wolanski@dm.uba.ar, http://mate.dm.uba.ar/∼wolanski).

‡Consejo Superior de Investigaciones Científicas (CSIC), Serrano 117, Madrid, Spain. On leave from Departamento de Matemática, FCEyN, UBA (1428) Buenos Aires, Argentina (jrossi@dm.uba.ar, http://mate.dm.uba.ar/∼jrossi).

More precisely, in [10] we studied the following problem. Let

$$\mathcal{J}(v) = \int_\Omega |\nabla v|^2 + v^2\, dx,$$
$$\mathcal{A}_\alpha = \{v \in H^1(\Omega) \ / \ \|v\|_{L^q(\partial\Omega)} = 1 \text{ and } |\{v > 0\}| = \alpha\}.$$

Then the problem is as follows:

$$(P_\alpha) \qquad \text{Find } \phi_0 \in \mathcal{A}_\alpha \text{ such that } S(\alpha) := \inf_{v\in\mathcal{A}_\alpha} \mathcal{J}(v) = \mathcal{J}(\phi_0).$$

In [10] we proved that there exists a solution ϕ_0 to (P_α), but the approach in [10] does not give any regularity properties of ϕ_0 or of the *hole* $\{\phi_0 = 0\}$.

In this paper we consider a different approach. Instead of minimizing $\mathcal{J}(v)$ over $\mathcal{A}_\alpha$ we penalize the functional and minimize without the measure restriction. This approach has been used with great success by many authors starting with the work [2] (see also, for instance, [3, 15, 16, 20]). Thus, let

$$\mathcal{J}_\varepsilon(v) = \int_\Omega |\nabla v|^2 + v^2\, dx + F_\varepsilon(|\{v > 0\}|), \tag{1.1}$$

where

$$F_\varepsilon(s) = \begin{cases} \dfrac{1}{\varepsilon}(s-\alpha) & \text{if } s \geq \alpha, \\ \varepsilon(s-\alpha) & \text{if } s < \alpha. \end{cases}$$

The penalized problem is to minimize $\mathcal{J}_\varepsilon$ over the class

$$\mathcal{K}_1 = \{v \in H^1(\Omega) \,/\, \|v\|_{L^q(\partial\Omega)} = 1\}.$$

For technical reasons, it is better to minimize in the class

$$\mathcal{K} = \{v \in H^1(\Omega) \,/\, \|v\|_{L^q(\Gamma_N)} = 1,\ v = \varphi_0 \text{ on } \Gamma_D\},$$

where $\emptyset \neq \Gamma_N \subset \partial\Omega$, $\Gamma_D = \partial\Omega \setminus \Gamma_N$ is the closure of a relatively open set of the boundary and $\varphi_0 \in H^1(\Omega)$, $\varphi_0 \geq c_0 > 0$ on Γ_D. We will only need to assume that $\Gamma_D \neq \emptyset$ at the end of our arguments. See section 4, Lemma 4.3.

So the penalized problem is as follows:

$$(P_\varepsilon) \qquad \text{Find } u_\varepsilon \in \mathcal{K} \text{ such that } \mathcal{J}_\varepsilon(u_\varepsilon) = \inf_{v\in\mathcal{K}} \mathcal{J}_\varepsilon(v).$$

Observe that minimizing $\mathcal{J}_\varepsilon$ over $\mathcal{K}$ gives a problem with mixed boundary conditions. We believe that this problem has independent interest.

The main idea is to prove that for ε small any minimizer u_ε of $\mathcal{J}_\varepsilon$ in $\mathcal{K}$ satisfies $|\{u_\varepsilon > 0\}| = \alpha$; therefore, the penalization term F_ε vanishes, and hence we have a minimizer of our original problem. This allows us to avoid the passage to the limit (as $\varepsilon \to 0$) where uniform bounds are needed. Proving regularity of the minimizers of $\mathcal{J}_\varepsilon$ and their free boundaries, $\partial\{u_\varepsilon > 0\}$, is easier than the original problem, thanks to the results of [4].

The main theorem in this article is the following.

THEOREM 1.1. *For every $\varepsilon > 0$ there exists a solution $u_\varepsilon \in \mathcal{K}$ to (P_ε). Moreover, any such solution is a locally Lipschitz continuous function and the free boundary*

$\partial\{u_\varepsilon > 0\}$ is locally a $C^{1,\beta}$ surface up to a set of zero $\mathcal{H}^{N-1}$ measure. In the case $N = 2$ the free boundary is locally a $C^{1,\beta}$ surface. Moreover, if $\Gamma_D \neq \emptyset$, for ε small we have $|\{u_\varepsilon > 0\}| = \alpha$.

Outline of the paper. In section 2 we begin our analysis of problem (P_ε) for fixed ε. First we prove the existence of a minimizer, local Lipschitz regularity, and nondegeneracy near the free boundary (Theorem 2.1). Then we prove that a minimizer u_ε of (P_ε) is a weak solution to the free boundary problem

$$\begin{cases} -\Delta u + u = 0 & \text{in } \{u > 0\} \cap \Omega, \\ \dfrac{\partial u}{\partial \nu} = \lambda_\varepsilon & \text{on } \partial\{u > 0\} \cap \Omega, \end{cases}$$

where λ_ε is a positive constant (Theorem 2.6).

In section 3, again for fixed ε, we analyze the regularity of the free boundary and show that, up to a set of zero $\mathcal{H}^{N-1}$ measure, $\partial\{u_\varepsilon > 0\}$ is locally a $C^{1,\beta}$ surface and, in the case $N = 2$, the free boundary has no exceptional points (Theorem 3.1). The proof of this result follows almost exactly the lines in [4], so we note only the significant differences and refer the reader to [4] for further details.

In section 4 we analyze the behavior of the solutions to (P_ε) for small ε. We prove that if $\Gamma_D \neq \emptyset$, the positivity set of the minimizer u_ε has measure α (Theorem 4.1).

Finally, in section 5, we go back to our original problem and show, under some mild assumptions on the solutions ϕ_0 to (P_α), that they are also solutions to (P_ε) for small ε, so they inherit the properties of the solutions to (P_ε) (Theorem 5.1). These extra assumptions are satisfied, for instance, if Ω is a ball (Corollary 5.1). In the general case, without the assumption that $\Gamma_D \neq \emptyset$, we prove that the set of α's for which there is a solution to (P_α) with smooth free boundary is dense in $(0, |\Omega|)$ (Theorem 5.2). Then, we show that the minimizers of (P_ε) converge (up to a subsequence) to a solution to (P_α) (Theorem 5.3). We believe that this last result might be of interest in numerical approximations.

2. The penalized problem. In this section, we consider the penalized problem (P_ε) stated in the introduction and prove the existence of a minimizer and some regularity properties.

THEOREM 2.1. *There exists a solution to the problem (P_ε). Moreover, any such solution u_ε has the following properties:*

(1) *u_ε is locally Lipschitz continuous in Ω.*

(2) *For every $D \subset\subset \Omega$, there exist constants $C, c > 0$ such that for every $x \in D \cap \{u_\varepsilon > 0\}$,*

$$c\,\mathrm{dist}(x, \partial\{u_\varepsilon > 0\}) \leq u_\varepsilon(x) \leq C\,\mathrm{dist}(x, \partial\{u_\varepsilon > 0\}).$$

(3) *For every $D \subset\subset \Omega$, there exists a constant $c > 0$ such that for $x \in \partial\{u > 0\}$ and $B_r(x) \subset D$,*

$$c \leq \frac{|B_r(x) \cap \{u_\varepsilon > 0\}|}{|B_r(x)|} \leq 1 - c.$$

The constants may depend on ε.

The proof will be divided into a series of steps for the reader's convenience.

Proof of existence. Let $(u_n) \subset \mathcal{K}$ be a minimizing sequence for $\mathcal{J}_\varepsilon$. Then $\mathcal{J}_\varepsilon(u_n)$ is bounded and so $\|u_n\|_{H^1(\Omega)} \leq C$. Therefore there exists a subsequence (that we still

call u_n) and a function $u_\varepsilon \in H^1(\Omega)$ such that

$$\begin{aligned} u_n &\rightharpoonup u_\varepsilon \quad \text{weakly in } H^1(\Omega),\\ u_n &\to u_\varepsilon \quad \text{strongly in } L^q(\partial\Omega),\\ u_n &\to u_\varepsilon \quad \text{a.e. } \Omega. \end{aligned}$$

Thus,

$$\begin{aligned} &\|u_\varepsilon\|_{L^q(\Gamma_N)} = 1,\\ &u_\varepsilon = \varphi_0 \quad \text{on} \quad \Gamma_D,\\ &|\{u_\varepsilon > 0\}| \le \liminf_{n\to\infty} |\{u_n > 0\}|, \quad \text{and}\\ &\|u_\varepsilon\|_{H^1(\Omega)} \le \liminf_{n\to\infty} \|u_n\|_{H^1(\Omega)}. \end{aligned}$$

Hence $u_\varepsilon \in \mathcal{K}$ and

$$\mathcal{J}_\varepsilon(u_\varepsilon) \le \liminf_{n\to\infty} \mathcal{J}_\varepsilon(u_n) = \inf_{v\in\mathcal{K}} \mathcal{J}_\varepsilon(v);$$

therefore u_ε is a minimizer of $\mathcal{J}_\varepsilon$ in $\mathcal{K}$. □

Remark 1. Any minimizer u_ε of $\mathcal{J}_\varepsilon$ satisfies the inequality

$$\Delta u - u \ge 0 \quad \text{in } \Omega. \tag{2.1}$$

In fact, this can be seen by performing one-side perturbations. Namely, we let $v = u_\varepsilon - t\varphi$ with $t > 0$ and $\varphi \in C_0^\infty(\Omega)$, $\varphi \ge 0$, to get

$$\int_\Omega \nabla u_\varepsilon \nabla\varphi + u_\varepsilon\varphi \le 0.$$

In the remainder of the section we will remove the subscript ε from the solution of (P_ε).

For the proof of properties (1)–(3), we apply the ideas developed in [4]. To this end, we need a series of lemmas.

LEMMA 2.1. *Let $u \in \mathcal{K}$ be a solution to (P_ε). There exists a constant $C = C(N,\Omega,\varepsilon)$ such that for every ball $B_r \subset\subset \Omega$*

$$\frac{1}{r} \fint_{\partial B_r} u \ge C \quad \textit{implies} \quad u > 0 \quad \textit{in} \quad B_r.$$

Proof. The idea is similar to that of Lemma 3.2 in [4]. Let v be the solution to

$$\begin{cases} v = u & \text{in } \overline{\Omega \setminus B_r},\\ \Delta v = v & \text{in } B_r. \end{cases} \tag{2.2}$$

Then $v \in \mathcal{K}$, $v > 0$ in B_r. We claim that

$$\|u - v\|^2_{H^1(\Omega)} = \|u\|^2_{H^1(\Omega)} - \|v\|^2_{H^1(\Omega)}. \tag{2.3}$$

In fact,

$$\int_{B_r} \nabla v \nabla(v-u) + v(v-u)\,dx = 0$$

since $v - u \in H_0^1(B_r)$. This implies

$$\int_{B_r} \nabla u \nabla v + uv \, dx = \int_{B_r} |\nabla v|^2 + v^2 \, dx. \tag{2.4}$$

This equality implies the claim since $u = v$ in $\Omega \setminus B_r$.

By (2.1), $u \le v$ in B_r. Now, by (2.3), since u is a minimizer and $u = v$ in $\Omega \setminus B_r$, we have

$$\begin{aligned} \int_\Omega |\nabla(u-v)|^2 + (u-v)^2 \, dx &\le -F_\varepsilon(|\{u>0\}|) + F_\varepsilon(|\{v>0\}|) \\ &\le C_\varepsilon \, |\{u=0\} \cap B_r|. \end{aligned} \tag{2.5}$$

Now, as in [4], the idea is to control $|\{u=0\} \cap B_r|$ from above by the left-hand side of (2.5). By replacing $u(x)$ by $u(x_0 + rx)/r$ we can assume that $B_r = B_1(0)$. For $|z| \le \frac{1}{2}$ we consider the change of variables from B_1 into itself such that z becomes the new origin. We call $u_z(x) = u\big((1-|x|)z+x\big)$, $v_z(x) = v\big((1-|x|)z+x\big)$ and define

$$r_\xi = \inf \left\{ r \text{ such that } \frac{1}{8} \le r \le 1 \quad \text{and} \quad u_z(r\xi) = 0 \right\}$$

if this set is nonempty. Observe that this change of variables leaves the boundary fixed.

Now, for almost every $\xi \in \partial B_1$ we have

$$v_z(r_\xi \xi) = \int_{r_\xi}^1 \frac{d}{dr}(u_z - v_z)(r\xi) \, dr \le \sqrt{1 - r_\xi} \left(\int_{r_\xi}^1 |\nabla(u_z - v_z)(r\xi)|^2 \, dr \right)^{1/2}. \tag{2.6}$$

Let us see that

$$v_z(r_\xi \xi) \ge C(N,\Omega)(1 - r_\xi) \fint_{\partial B_1} u. \tag{2.7}$$

In fact, $v_z(r_\xi \xi) = v\big((1-r_\xi)z + r_\xi \xi\big)$, and if $|(1-r_\xi)z + r_\xi \xi| \le \frac{3}{4}$, by the Harnack inequality applied to a solution to $\Delta v - r^2 v = 0$ in B_1 with $r \le 1$,

$$v_z(r_\xi \xi) \ge C_N v(0).$$

Clearly (2.7) follows from

$$v(0) \ge \alpha(N) \fint_{\partial B_1} v = \alpha(N) \fint_{\partial B_1} u. \tag{2.8}$$

But (2.8) is a consequence of the mean value property of solutions to the Schrödinger equation $\Delta v - r^2 v = 0$, namely,

$$v(0) = \frac{1}{J(r)} \fint_{\partial B_1(0)} v,$$

where $J(r) = \Gamma(N/2)(\frac{r}{2})^{1-\frac{N}{2}} I_{\frac{N-2}{2}}(r)$ and $I_{\frac{N-2}{2}}$ is the Bessel function. In particular,

$$J(0) = 1.$$

See Theorem 9.9 in [18] for this result.

Now, if $|(1-r_\xi)z + r_\xi\xi| \geq \frac{3}{4}$, we prove by a comparison argument that inequality (2.7) also holds. In fact, first observe that we can assume that $⨍_{\partial B_1} v = ⨍_{\partial B_1} u = 1$. Then, by (2.8), $v \geq C_N\alpha$ in $B_{3/4}$. Let $w(x) = e^{-\lambda|x|^2} - e^{-\lambda}$. There exists $\lambda = \lambda(N,\alpha)$ such that

$$\begin{cases} \Delta w \geq w & \text{in } B_1 \setminus B_{3/4}, \\ w \leq C_N\alpha & \text{in } \partial B_{3/4}, \\ w = 0 & \text{in } \partial B_1, \end{cases}$$

so that, since $\Delta v \leq v$, there holds that $v \geq w \geq C(1-|x|)$ in $B_1 \setminus B_{3/4}$. Therefore,

$$v_z(r_\xi\xi) \geq C\Big(1 - |(1-r_\xi)z + r_\xi\xi|\Big) ⨍_{\partial B_1} u \geq C(1-r_\xi) ⨍_{\partial B_1} u$$

since $|z| \leq \frac{1}{2}$. Thus (2.7) holds for every $r_\xi \geq \frac{1}{8}$.

By (2.6) and (2.7) we have

$$c\sqrt{1-r_\xi} ⨍_{\partial B_1} u \leq \left(\int_{r_\xi}^1 |\nabla(u_z - v_z)|^2(r\xi)\, dr \right)^{1/2}.$$

Hence

$$\begin{aligned} c^2 \int_{\partial B_1} (1-r_\xi)\, dS_\xi \left(⨍_{\partial B_1} u \right)^2 &\leq \int_{\partial B_1} \int_{r_\xi}^1 |\nabla(u_z - v_z)|^2(r\xi)\, dr\, dS_\xi \\ &\leq C \int_{B_1} |\nabla(u_z - v_z)|^2\, dx. \end{aligned}$$

Since

$$\int_{\partial B_1} (1-r_\xi)\, dS_\xi \geq \int_{B_1 \setminus B_{1/4}(z)} \chi_{\{u=0\}}\, dx,$$

we have

$$\begin{aligned} c^2 |\{x \in B_1 \setminus B_{1/4}(z) \,/\, u(x) = 0\}| \left(⨍_{\partial B_1} u \right)^2 &\leq C \int_{B_1} |\nabla(u_z - v_z)|^2\, dx \\ &\leq K \int_{B_1} |\nabla(u-v)|^2\, dx. \end{aligned}$$

Finally, we integrate over $z \in B_{1/2}(0)$ and use (2.5) to obtain

$$\begin{aligned} |B_1 \cap \{u=0\}| \left(⨍_{\partial B_1} u \right)^2 &\leq K \int_{B_1} |\nabla(u-v)|^2\, dx \\ &\leq K C_\varepsilon |B_1 \cap \{u=0\}|. \end{aligned} \tag{2.9}$$

Therefore we either have $u > 0$ a.e. in B_1 or else $⨍_{\partial B_1} u \leq \sqrt{KC_\varepsilon}$.

Hence we deduce that if

$$⨍_{\partial B_1} u \geq \sqrt{KC_\varepsilon} = C(N, \Omega, \varepsilon),$$

then $|B_1 \cap \{u = 0\}| = 0$. Thus by (2.5) $u = v > 0$ in B_1. □

Now we can prove the Lipschitz continuity of the minimizer u.

Proof of Theorem 2.1(1). The proof follows as in [4, Lemma 3.3]. In fact, let $D \subset\subset D' \subset\subset \Omega$ and $x \in D$. Let $r > 0$ be the largest number such that $B_r(x) \subset \{u > 0\} \cap D'$. As in [4] we prove by using Lemma 2.1 that $\{u > 0\}$ is open and

$$\frac{1}{r} \fint_{\partial B_r(x)} u \leq C$$

with C independent of either u or x. Since $u > 0$ in $B_r(x)$, it is a solution to

$$\Delta u = u \quad \text{in } B_r(x).$$

In fact, let v be the solution to $\Delta v = v$ in $B_r(x)$, $v = u$ on $\Omega \setminus B_r(x)$. Then,

$$0 \leq \|u - v\|^2_{H^1(\Omega)} = \|u\|^2_{H^1(\Omega)} - \|v\|^2_{H^1(\Omega)} = \mathcal{J}_\varepsilon(u) - \mathcal{J}_\varepsilon(v) \leq 0.$$

Thus, $u = v$ in $B_r(x)$.

Hence, there is a universal constant such that

$$|\nabla u(x)| \leq C \left\{ r\|u\|_{L^\infty(B_r(x))} + \frac{1}{r} \fint_{\partial B_r(x)} u \right\}.$$

Now, since u is subharmonic in Ω and $D' \subset\subset \Omega$, there holds that u is bounded in D' by a constant that depends on the H^1 norm of u in Ω, which is bounded by a constant that depends only on Ω and ε. Therefore,

$$|\nabla u(x)| \leq C$$

with C depending only on N, Ω, ε, D, and D'. □

In order to prove the nondegeneracy of u we need the following lemma (see [4, Lemma 3.4]).

LEMMA 2.2. *Let $u \in \mathcal{K}$ be a solution to (P_ε). For $0 < \kappa < 1$ there exists a constant $c = c(\kappa, N, \Omega, \varepsilon)$ such that for every ball $B_r(x_0) \subset\subset \Omega$,*

$$\frac{1}{r} \fint_{\partial B_r} u \leq c \quad \textit{implies that} \quad u = 0 \quad \textit{in} \quad B_{\kappa r}.$$

Proof. As in [4, Lemma 3.4], we consider the function

$$\phi_s^N(x) = \begin{cases} \dfrac{s}{N-2} \left(\left(\dfrac{s}{|x|} \right)^{N-2} - 1 \right) & \text{for} \quad N \geq 3, \\ s \log \dfrac{s}{|x|} & \text{for} \quad N = 2, \\ s - |x| & \text{for} \quad N = 1. \end{cases} \tag{2.10}$$

For simplicity let us take $\bar{u}(x) = \dfrac{1}{r} u(x_0 + rx)$,

$$\bar{F}_\varepsilon(s) = \begin{cases} \dfrac{1}{\varepsilon} \left(s - \dfrac{\alpha}{r^N} \right) & \text{if } s > \dfrac{\alpha}{r^N}, \\ \varepsilon \left(s - \dfrac{\alpha}{r^N} \right) & \text{if } s \leq \dfrac{\alpha}{r^N}, \end{cases}$$

and

$$\bar{\mathcal{J}}_\varepsilon(w) = \int_{\Omega^r} |\nabla w|^2 + r^2 w^2 + \bar{F}_\varepsilon(|\{w > 0\}|),$$

where $\Omega^r = \frac{1}{r}(\Omega - x_0)$. Thus, $\mathcal{J}_\varepsilon(u) = r^N \bar{\mathcal{J}}_\varepsilon(\bar{u})$.

Now, let $v(x) = \frac{\gamma\sqrt{\kappa}}{-\phi_\kappa^N(\sqrt{\kappa})} \max(-\phi_\kappa^N(x), 0)$, where, since $\bar{u}$ is subharmonic,

$$\gamma := \frac{1}{\sqrt{\kappa}} \sup_{B_{\sqrt{\kappa}}} \bar{u} \le C_1(N,\kappa) \fint_{\partial B_1} \bar{u} = C_1(N,\kappa) \frac{1}{r} \fint_{\partial B_r(x_0)} u.$$

Hence, $v \ge \bar{u}$ on $\partial B_{\sqrt{\kappa}}$, and therefore if

$$w = \begin{cases} \min(\bar{u}, v) & \text{in} \quad B_{\sqrt{\kappa}}, \\ \bar{u} & \text{in} \quad \Omega^r \setminus B_{\sqrt{\kappa}}, \end{cases}$$

there holds that

$$\begin{aligned}
&\int_{B_\kappa} |\nabla \bar{u}|^2 + r^2 \bar{u}^2 \, dx + |B_\kappa \cap \{\bar{u} > 0\}| \\
&\quad = \bar{\mathcal{J}}_\varepsilon(\bar{u}) - \int_{\Omega^r \setminus B_\kappa} |\nabla \bar{u}|^2 + r^2 \bar{u}^2 \, dx + |B_\kappa \cap \{\bar{u} > 0\}| - \bar{F}_\varepsilon(|\{\bar{u} > 0\}|) \\
&\quad \le \bar{\mathcal{J}}_\varepsilon(w) - \int_{\Omega^r \setminus B_\kappa} |\nabla \bar{u}|^2 + r^2 \bar{u}^2 \, dx + |B_\kappa \cap \{\bar{u} > 0\}| - \bar{F}_\varepsilon(|\{\bar{u} > 0\}|) \\
&\quad = \int_{B_{\sqrt{\kappa}} \setminus B_\kappa} |\nabla w|^2 + r^2 w^2 \, dx - \int_{B_{\sqrt{\kappa}} \setminus B_\kappa} |\nabla \bar{u}|^2 + r^2 \bar{u}^2 \, dx + |B_\kappa \cap \{\bar{u} > 0\}| \\
&\qquad\qquad + \bar{F}_\varepsilon(|\{w > 0\}|) - \bar{F}_\varepsilon(|\{\bar{u} > 0\}|) \\
&\quad \le \int_{B_{\sqrt{\kappa}} \setminus B_\kappa} |\nabla w|^2 + r^2 w^2 \, dx - \int_{B_{\sqrt{\kappa}} \setminus B_\kappa} |\nabla \bar{u}|^2 + r^2 \bar{u}^2 \, dx + (1 - \varepsilon) \, |B_\kappa \cap \{\bar{u} > 0\}|,
\end{aligned}$$

since $w = 0$ in B_κ, $w = \bar{u}$ in $\Omega^r \setminus B_{\sqrt{\kappa}}$. We have also used that $\bar{F}_\varepsilon(A) - \bar{F}_\varepsilon(B) \ge \varepsilon(A - B)$ if $A \ge B$ and $\{w > 0\} \subset \{\bar{u} > 0\}$. This inclusion follows from the fact that $w \le \bar{u}$. Thus,

$$\begin{aligned}
&\int_{B_\kappa} |\nabla \bar{u}|^2 + r^2 \bar{u}^2 \, dx + \varepsilon |B_\kappa \cap \{\bar{u} > 0\}| \\
&\quad \le \int_{B_{\sqrt{\kappa}} \setminus B_\kappa} |\nabla w|^2 + r^2 w^2 \, dx - \int_{B_{\sqrt{\kappa}} \setminus B_\kappa} |\nabla \bar{u}|^2 + r^2 \bar{u}^2 \\
&\quad = \int_{B_{\sqrt{\kappa}} \setminus B_\kappa} |\nabla \bar{u} - \nabla(\bar{u} - v)^+|^2 - |\nabla \bar{u}|^2 \, dx + r^2 \int_{B_{\sqrt{\kappa}} \setminus B_\kappa} \left(\bar{u} - (\bar{u} - v)^+\right)^2 - \bar{u}^2 \, dx \\
&\quad = -\int_{B_{\sqrt{\kappa}} \setminus B_\kappa} \nabla(\bar{u} - v)^+ \nabla(\bar{u} + v) \, dx - r^2 \int_{B_{\sqrt{\kappa}} \setminus B_\kappa} (\bar{u} - v)^+ (\bar{u} + v) \, dx \\
&\quad = -\int_{B_{\sqrt{\kappa}} \setminus B_\kappa} \nabla(\bar{u} - v)^+ \nabla \bar{u} \, dx - r^2 \int_{B_{\sqrt{\kappa}} \setminus B_\kappa} (\bar{u} - v)^+ \bar{u} \, dx \\
&\qquad\quad - \int_{B_{\sqrt{\kappa}} \setminus B_\kappa} \nabla(\bar{u} - v)^+ \nabla v \, dx - r^2 \int_{B_{\sqrt{\kappa}} \setminus B_\kappa} (\bar{u} - v)^+ v \, dx
\end{aligned}$$

$$\leq -2\int_{B_{\sqrt{\kappa}}\setminus B_\kappa} \nabla(\bar{u}-v)^+\nabla v\,dx - 2r^2\int_{B_{\sqrt{\kappa}}\setminus B_\kappa}(\bar{u}-v)^+ v\,dx$$
$$\leq 2\int_{\partial B_\kappa}\bar{u}\nabla v\,\eta\,dS \leq C_2(N,\kappa)\,\gamma\int_{\partial B_\kappa}\bar{u}.$$

Therefore,

$$\int_{B_\kappa}|\nabla\bar{u}|^2 + r^2\bar{u}^2\,dx + \varepsilon|B_\kappa\cap\{\bar{u}>0\}| \leq C_2(N,\kappa)\,\gamma\int_{\partial B_\kappa}\bar{u}. \tag{2.11}$$

Here we have used that $\min(\bar{u},v) = \bar{u}-(\bar{u}-v)^+$, $\Delta v = 0$ in $B_{\sqrt{\kappa}}\setminus B_\kappa$, $v=0$ on ∂B_κ, and $(\bar{u}-v)^+ = 0$ on $\partial B_{\sqrt{\kappa}}$.

Recall that γ is controlled by $\frac{1}{r}⨍_{\partial B_r(x_0)}u$; thus γ will be small if $\frac{1}{r}⨍_{\partial B_r(x_0)}u$ is small.

On the other hand, by standard estimates,

$$\begin{aligned}\int_{\partial B_\kappa}\bar{u} &\leq C_3(N,\kappa)\int_{B_\kappa}|\nabla\bar{u}|+\bar{u}\,dx\\ &\leq C_3(N,\kappa)\Big\{\frac{1}{2}\int_{B_\kappa}|\nabla\bar{u}|^2\,dx + \frac{1}{2}|B_\kappa\cap\{\bar{u}>0\}| + \gamma|B_\kappa\cap\{\bar{u}>0\}|\Big\}\\ &\leq C_3(N,\kappa)\Big\{\int_{B_\kappa}|\nabla\bar{u}|^2 + r^2\bar{u}^2\,dx + |B_\kappa\cap\{\bar{u}>0\}|\Big\}\end{aligned}$$

if $\gamma \leq 1/2$.

Thus, by (2.11), if γ is small enough ($\gamma \leq 1/2$ and $C_2(N,\kappa)C_3(N,\kappa)\gamma < 1$), we deduce that $|B_\kappa\cap\{\bar{u}>0\}| = 0$. That is, $u=0$ in $B_{r\kappa}(x_0)$ and the lemma is proved. □

We can now prove the nondegeneracy of u.

Proof of Theorem 2.1(2). Let $x\in\{u>0\}$ and $r = \operatorname{dist}(x,\{u=0\})$. As we proved in (2.8), since $\Delta u = u$ in $B_r(x)$, there holds that

$$u(x) \geq \alpha(N)⨍_{\partial B_r(x)}u.$$

Since $u(x)>0$,

$$\frac{1}{r}⨍_{\partial B_r(x)}u \geq c,$$

where c is the constant in Lemma 2.2 for $\kappa = 1/2$. Thus,

$$u(x) \geq c\alpha\, r.$$

The upper bound clearly follows from the Lipschitz continuity of u. Hence (2) is proved. □

Proof of Theorem 2.1(3). In order to prove the uniform positive density of $\{u>0\}$ and $\{u=0\}$ at every free boundary point we proceed as in [4, Lemma 3.7]. The only difference is that the function v that we have to take is the one in (2.2).

This ends the proof of Theorem 2.1. □

COROLLARY 2.1. *Let $u \in \mathcal{K}$ be a solution to (P_ε). Let $D \subset\subset \Omega$. There exist constants $c, C > 0$ depending only on N, Ω, D, and ε such that for $B_r(x) \subset D$ and $x \in \partial\{u > 0\}$,*

$$c \leq \frac{1}{r} \fint_{\partial B_r(x)} u \leq C. \tag{2.12}$$

Proof. The proof follows easily from Lemmas 2.1 and 2.2. ∎

LEMMA 2.3. *Let $u \in \mathcal{K}$ be a solution to (P_ε). Then u satisfies, for every $\varphi \in C_0^\infty(\Omega)$ such that $\operatorname{supp}\varphi \subset \{u > 0\}$,*

$$\int_\Omega \nabla u \nabla \varphi + u\varphi \, dx = 0. \tag{2.13}$$

Moreover, the application

$$\lambda(\varphi) := -\int_\Omega \nabla u \nabla \varphi + u\varphi \, dx$$

from $C_0^\infty(\Omega)$ into $\mathbb{R}$ defines a nonnegative Radon measure with support on $\Omega \cap \partial\{u > 0\}$.

Proof. The proof follows exactly as in [4, Lemma 4.2]. ∎

THEOREM 2.2. *Let $u \in \mathcal{K}$ be a solution to (P_ε). Let $D \subset\subset \Omega$. Then there exist constants $C, c > 0$ such that for $B_r(x) \subset D$ and $x \in \partial\{u > 0\}$,*

$$c\, r^{N-1} \leq \int_{B_r(x)} d\lambda \leq C\, r^{N-1}.$$

Proof. For n large enough, let $u_n = u * \rho_n$, where ρ_n are the standard mollifiers. Then

$$\begin{aligned}\int_{B_r(x)} \lambda * \rho_n \, dx &= \int_{B_r(x)} \Delta u_n - u_n \, dx = \int_{\partial B_r(x)} \nabla u_n \cdot \nu \, dS - \int_{B_r(x)} u_n \\ &\leq \omega_{N-1} \sup_{\partial B_r(x)} |\nabla u_n| \, r^{N-1} \leq C\, r^{N-1}\end{aligned}$$

since $|\nabla u_n| \leq |\nabla u| \leq C$ for a certain constant C depending on D. By taking the limit for $n \to \infty$ we get

$$\int_{B_r(x)} d\lambda \leq C\, r^{N-1}.$$

The other inequality follows as in the proof of Theorem 4.3 in [4] by taking as $G_y(z)$ the (positive) Green function of $-\Delta + Id$ with homogeneous Dirichlet boundary conditions in the ball $B_r(x)$. Then for $0 < \kappa < 1/2$ and $y \in B_{\kappa r}(x)$ one uses the inequality

$$v(y) \geq Cv(x) \geq C\alpha \fint_{\partial B_r(x)} u$$

for v the solution to $\Delta v - v = 0$ in $B_r(x)$, $v = u$ on $\partial B_r(x)$, which follows from the Harnack inequality and (2.8). ∎

THEOREM 2.3 (representation theorem). *Let $u \in \mathcal{K}$ be a solution to (P_ε). Then*

(1) $\mathcal{H}^{N-1}(D \cap \partial\{u > 0\}) < \infty$ *for every* $D \subset\subset \Omega$.
(2) *There exists a Borel function* q_u *such that*

$$\Delta u - u = q_u \, \mathcal{H}^{N-1} \lfloor \partial\{u > 0\}.$$

(3) *For* $D \subset\subset \Omega$ *there are constants* $0 < c \le C < \infty$ *depending on* N, Ω, D, *and the constants in* (2.12) *such that for* $B_r(x) \subset D$ *and* $x \in \partial\{u > 0\}$,

$$c \le q_u(x) \le C, \quad c\, r^{N-1} \le \mathcal{H}^{N-1}(B_r(x) \cap \partial\{u > 0\}) \le C\, r^{N-1}.$$

Proof. The proof follows exactly as in Theorem 4.5 in [4]. □

Remark 2. Let $u \in \mathcal{K}$ be a solution to (P_ε) and let $D \subset\subset \Omega$. Then $D \cap \partial\{u > 0\}$ has finite perimeter. Thus, the reduced boundary $\partial_{\mathrm{red}}\{u > 0\}$ is defined as well as the measure theoretic normal $\nu(x)$ for $x \in \partial_{\mathrm{red}}\{u > 0\}$. See [8].

If the free boundary $\partial\{u > 0\}$ is a regular surface, then $q_u = -\partial_\nu u$. In Theorem 2.4 it is shown that this is true for almost all points in the reduced boundary.

PROPOSITION 2.1. *Let* $u \in \mathcal{K}$ *be a solution to* (P_ε) *and let* $B_{\rho_k}(x_k) \subset \Omega$ *be a sequence of balls with* $\rho_k \to 0$, $x_k \to x_0 \in \Omega$, *and* $u(x_k) = 0$. *Let*

$$u_k(x) := \frac{1}{\rho_k} u(x_k + \rho_k x).$$

We call u_k *a blow-up sequence with respect to* $B_{\rho_k}(x_k)$. *Since* u *is locally Lipschitz continuous, there exists a blow-up limit* $u_0 : \mathbb{R}^N \to \mathbb{R}$ *satisfying* (2.12) *with the same constants, when* $x_k \in \partial\{u > 0\}$ *and such that for a subsequence,*

$$\begin{aligned}
&u_k \to u_0 \quad \text{in} \quad C^\alpha_{\mathrm{loc}}(\mathbb{R}^N) \quad \text{for every} \quad 0 < \alpha < 1,\\
&\nabla u_k \to \nabla u_0 \quad \text{weakly}\star \text{ in} \quad L^\infty_{\mathrm{loc}}(\mathbb{R}^N),\\
&\partial\{u_k > 0\} \to \partial\{u_0 > 0\} \quad \text{locally in Hausdorff distance},\\
&\chi_{\{u_k > 0\}} \to \chi_{\{u_0 > 0\}} \quad \text{in} \quad L^1_{\mathrm{loc}}(\mathbb{R}^N),\\
&\Delta u_0 = 0 \quad \text{in} \quad \{u_0 > 0\}.
\end{aligned}$$

Moreover, if $x_k \in \partial\{u > 0\}$, *then* $0 \in \partial\{u_0 > 0\}$.

Proof. The proof follows as in [4, section 4.7], observing that $\Delta u_k - \rho_k^2 u_k = 0$ in $\{u_k > 0\}$. □

THEOREM 2.4 (identification of q_u). *Let* $u \in \mathcal{K}$ *be a solution to* (P_ε). *Then, for almost every* $x_0 \in \partial_{\mathrm{red}}\{u > 0\}$,

$$u(x_0 + x) = q_u(x_0) \langle x, \nu(x_0) \rangle^- + o(|x|) \quad \text{for} \quad x \to 0$$

with $\nu(x_0)$ *the outward unit normal de* $\partial\{u > 0\}$ *in the measure theoretic sense.*

Proof. The proof follows exactly as in Theorem 4.8 and Remark 4.9 in [4]. □

Remark 3. Observe that by Theorem 2.1(3)

$$\mathcal{H}^{N-1}(\partial\{u > 0\} \setminus \partial_{\mathrm{red}}\{u > 0\}) = 0.$$

See [8].

Now we get a more precise identification of q_u.

THEOREM 2.5. *Let* $u \in \mathcal{K}$ *be a solution to* (P_ε) *and let* q_u *be the function in Theorem* 2.4. *Then there exists a constant* λ_u *such that*

$$\limsup_{\substack{x \to x_0 \\ u(x) > 0}} |\nabla u(x)| = \lambda_u \qquad \text{for every } x_0 \in \Omega \cap \partial\{u > 0\}, \tag{2.14}$$

$$q_u(x_0) = \lambda_u, \qquad \mathcal{H}^{N-1} \text{ a.e. } x_0 \in \Omega \cap \partial\{u > 0\}. \tag{2.15}$$

Moreover, if B is a ball contained in $\{u = 0\}$ touching the boundary $\partial\{u > 0\}$ at x_0, then

$$\limsup_{\substack{x \to x_0 \\ u(x)>0}} \frac{u(x)}{\operatorname{dist}(x, B)} = \lambda_u. \tag{2.16}$$

Proof. We follow the ideas of [2, Theorem 3] and [16, Theorem 5.1 and Lemma 5.2].

Let $x_0, x_1 \in \partial\{u > 0\}$ and $\rho_k \to 0^+$. For $i = 0, 1$, let $x_{i,k} \to x_i$ with $u(x_{i,k}) = 0$ such that $B_{\rho_k}(x_{i,k}) \subset \Omega$ and such that the blow-up sequence

$$u_{i,k}(x) = \frac{1}{\rho_k} u(x_{i,k} + \rho_k x)$$

has a limit $u_i(x) = \lambda_i \langle x, \nu_i \rangle^-$, with $0 < \lambda_i < \infty$ and ν_i a unit vector. We will prove that $\lambda_0 = \lambda_1$. From this, the theorem will follow as in [16].

Assume that $\lambda_1 < \lambda_0$. Then we will perturb the minimizer u near x_0 and x_1 and get an admissible function with less energy, which is a contradiction. We perform a perturbation that increases the measure of the positivity set in a neighborhood of $x_{0,k}$ and decreases its measure in a neighborhood of $x_{1,k}$. We perform this perturbation in such a way that we change the measure of the positivity set in an amount of essentially order $o(\rho_k^N)$.

To this end, we take a nonnegative C_0^∞ symmetric function Φ supported in the unit interval and for $t > 0$ small, we define

$$\tau_k(x) = \begin{cases} x + t\rho_k \Phi\Big(\dfrac{|x - x_{0,k}|}{\rho_k}\Big)\nu_0 & \text{for } x \in B_{\rho_k}(x_{0,k}), \\ x - t\rho_k \Phi\Big(\dfrac{|x - x_{1,k}|}{\rho_k}\Big)\nu_1 & \text{for } x \in B_{\rho_k}(x_{1,k}), \\ x & \text{elsewhere,} \end{cases}$$

which is a diffeomorphism if t is small enough. Now, let

$$v_k(x) = u(\tau_k^{-1}(x)),$$

which are admissible functions. Moreover, since $\|D\tau_k^{-1}\| \le C$ independent of k for t small enough, there holds that

$$\|\nabla v_k\|_{L^\infty} \le C$$

independent of k.

Also, we have

$$F_\varepsilon(|\{v_k > 0\}|) - F_\varepsilon(|\{u > 0\}|) = o(t)\, \rho_k^N + o(\rho_k^N). \tag{2.17}$$

In fact, $v_k = u$ in $\Omega \setminus (B_{\rho_k}(x_{0,k}) \cup B_{\rho_k}(x_{1,k}))$ and

$$|\{v_k > 0\} \cap B_{\rho_k}(x_{i,k})| - |\{u > 0\} \cap B_{\rho_k}(x_{i,k})| = \\ = (-1)^i\, \rho_k^N \Big(t \int_{B_1 \cap \{y_1 = 0\}} \Phi(|y|)\, d\mathcal{H}_y^{N-1} + o_i(t) \Big) + o(\rho_k^N),$$

since $\Phi(|y|)$ is radially symmetric and $\chi_{\{u_{i,k}>0\}} \to \chi_{\{\langle x,\nu_i\rangle<0\}}$ in $L^1_{\rm loc}(\mathbb{R}^N)$.

Similar computations that also involve the development of ∇v_k in terms of ∇u and $D\tau_k$ give

$$\begin{aligned}\int_\Omega |\nabla v_k|^2\,dx - \int_\Omega |\nabla u|^2\,dx \\ = \rho_k^N\Big((\lambda_1^2-\lambda_0^2)\,t\int_{B_1(0)\cap\{y_1=0\}}\Phi(|y|)\,d\mathcal{H}_y^{N-1}+o(t)\Big)+o(\rho_k^N).\end{aligned} \tag{2.18}$$

See [2] or [16] for detailed computations.

It remains to estimate the difference of the L^2 norms. Since $u(x_{i,k})=0$ there holds that

$$u(x)\le C\rho_k^N \quad \text{in } B_{\rho_k}(x_{i,k}).$$

On the other hand,

$$0=u(x_{i,k})=v_k(\tau_k(x_{i,k}))=v_k(x_{i,k}+(-1)^i\,t\rho_k\Phi(0)\nu_i).$$

Thus,

$$v_k(z)\le C\big|z-x_{i,k}-(-1)^i\,t\rho_k\Phi(0)\nu_i\big|\le K\rho_k \quad \text{if } z\in B_{\rho_k}(x_{i,k}).$$

Therefore,

$$\int_\Omega v_k^2\,dx-\int_\Omega u^2\,dx=o(\rho_k^N). \tag{2.19}$$

Thus, we get from (2.17), (2.18), and (2.19), for t small enough and k large enough, that

$$\mathcal{J}_\varepsilon(v_k)<\mathcal{J}_\varepsilon(u),$$

a contradiction. □

Summing up, we have the following theorem,

THEOREM 2.6. *Let $u\in\mathcal{K}$ be a solution to (P_ε). Then u is a weak solution to the free boundary problem*

$$\begin{aligned}-\Delta u+u&=0 && \text{in } \{u>0\}\cap\Omega,\\ \frac{\partial u}{\partial\nu}&=\lambda_u && \text{on } \partial\{u>0\}\cap\Omega,\end{aligned}$$

where λ_u is the constant in Theorem 2.5. More precisely, $\mathcal{H}^{N-1}$ a.e. point $x_0\in\partial\{u>0\}$ belongs to $\partial_{red}\{u>0\}$ and

$$u(x_0+x)=\lambda_u\langle x,\nu(x_0)\rangle^-+o(|x|) \quad \text{for } x\to 0.$$

Finally, we get an estimate of the gradient of u that will be needed in order to get the regularity of the free boundary.

THEOREM 2.7. *Let $u\in\mathcal{K}$ be a solution to (P_ε). Given $D\subset\subset\Omega$, there exist constants $C=C(N,\varepsilon,D)$, $r_0=r_0(N,D)>0$, and $\gamma=\gamma(N,\varepsilon,D)>0$ such that if $x_0\in D\cap\partial\{u>0\}$ and $r<r_0$, then*

$$\sup_{B_r(x_0)}|\nabla u|\le\lambda_u(1+Cr^\gamma).$$

Proof. The proof follows the lines of the proof of Theorem 4.1 in [5].

Let $U_k = \left(|\nabla u| - \lambda_u - \frac{1}{k}\right)^+$ and $U_0 = \left(|\nabla u| - \lambda_u\right)^+$. By (2.14) we know that U_k vanishes in a neighborhood of the free boundary. Also, the support of U_k is contained in $\{u > 0\}$. Therefore U_k satisfies

$$\Delta U_k \geq U_k \quad \text{in } \Omega \cap \{u > 0\}$$

and vanishes in a neighborhood of the free boundary. We extend U_k by zero into $\{u = 0\}$ and set

$$h_k(r) = \sup_{B_r(x_0)} U_k, \qquad h_0(r) = \sup_{B_r(x_0)} U_0$$

for any $r < r_0 = \operatorname{dist}(D, \partial\Omega)$ and $x_0 \in D \cap \partial\{u > 0\}$.

Then, $h_k(r) - U_k$ is a supersolution of $\Delta v = v$ in the ball $B_r(x_0)$ and

$$\begin{aligned} h_k(r) - U_k &\geq 0 && \text{in } B_r(x_0) \\ &= h_k(r) && \text{in } B_r(x_0) \cap \{u = 0\}. \end{aligned}$$

Applying the weak Harnack inequality (see [12, p. 246]) with $1 \leq p < N/(N-2)$, we get

$$\inf_{B_{r/2}(x_0)} \left(h_k(r) - U_k\right) \geq cr^{-N/p} \|h_k(r) - U_k\|_{L^p(B_r(x_0))} \geq c h_k(r),$$

since, by Theorem 2.1(3), $|B_r(x_0) \cap \{u = 0\}| \geq cr^N$. Taking now $k \to \infty$ we obtain

$$\inf_{B_{r/2}(x_0)} \left(h_0(r) - U_0\right) \geq c h_0(r)$$

for some $0 < c < 1$, which is the same as

$$\sup_{B_{r/2}(x_0)} U_0 \leq (1 - c) h_0(r).$$

Therefore

$$h_0\left(\frac{r}{2}\right) \leq (1 - c) h_0(r),$$

from which it follows that $h_0(r) \leq Cr^\gamma$ for some $C > 0$, $0 < \gamma < 1$, and now the conclusion of the theorem follows. □

3. Regularity of the free boundary. At this point we have that our minimizer u_ε meets the conditions of the regularity theory developed in [4], the only difference being the equation satisfied by u_ε in $\{u_\varepsilon > 0\}$.

We recall some definitions and point out the only significant difference with [4]. The rest of the proof of the regularity then follows as sections 7 and 8 of [4] with only minor modifications.

Throughout this section we remove the subscript ε.

DEFINITION 3.1 (flat free boundary points). *Let $0 < \sigma_+, \sigma_- \leq 1$ and $\tau > 0$. We say that u is of class*

$$F(\sigma_+, \sigma_-; \tau) \quad \textit{in} \quad B_\rho = B_\rho(0)$$

if

(1) $0 \in \partial\{u > 0\}$ *and*

$$\begin{aligned} u &= 0 && \text{for} \quad x_N \geq \sigma_+\rho, \\ u(x) &\geq -\lambda(x_N + \sigma_-\rho) && \text{for} \quad x_N \leq -\sigma_-\rho; \end{aligned}$$

(2) $|\nabla u| \leq \lambda(1+\tau)$ *in* B_ρ.

If the origin is replaced by x_0 *and the direction* e_N *by the unit vector* ν, *we say that* u *is of class* $F(\sigma_+, \sigma_-; \tau)$ *in* $B_\rho(x_0)$ *in direction* ν.

Observe that the results in section 2 imply that the minimizer u of $\mathcal{J}_\varepsilon$ is in the class $F(\sigma, 1; \sigma)$ in $B_\rho(x_0)$ in direction $\nu_u(x_0)$ for every $x_0 \in \partial_{red}\{u > 0\}$ with $\sigma = \sigma(\rho) \to 0$ as $\rho \to 0$.

The following lemma (Lemma 7.2 in [4]) is the only one that requires a nonobvious modification.

LEMMA 3.1. *There is a constant* $C = C(N)$ *such that* $u \in F(\sigma, 1; \sigma)$ *in* $B_\rho(x_0)$ *in direction* ν *implies* $u \in F(2\sigma, C\sigma; \sigma)$ *in* $B_{\rho/2}(x_0)$ *in direction* ν.

Proof. Clearly, by a change of variables, we may assume that $x_0 = 0$ and $\nu = e_N$. Let $\bar{u}(x) = u(\rho x)/\lambda\rho$; then $|\nabla \bar{u}| \leq 1 + \sigma$ and $\bar{u} \in F(\sigma, 1; \sigma)$ in B_1. That is, $\bar{u} = 0$ if $x_N > \sigma$. Define

$$\eta(x') = \begin{cases} \exp\left(-\dfrac{9|x'|^2}{1 - 9|x'|^2}\right) & \text{for } |x'| < \frac{1}{3}, \\ 0 & \text{otherwise} \end{cases}$$

and choose $s \geq 0$ maximal with the property that $\bar{u} = 0$ in $x_N > \sigma - s\eta(x')$.

Now, the proof follows as in Lemma 7.2 of [4] with the only difference being that the comparison function v must be the solution to $\Delta v = \rho^2 \bar{u}$ in $D = B_1 \cap \{x_N < \sigma - s\eta(x')\}$ instead of a harmonic function. The estimate

$$\partial_{-\nu} v \leq 1 + C\sigma$$

follows from

$$|\nabla(v + x_N)| \leq C\Big[\sup_D (v + x_N) + \rho^2\Big] \leq C\sigma$$

in $D \cap B_{1/2}$ if $\rho^2 \leq C\sigma$, since

$$\Delta(v + x_N) = \rho^2 \bar{u} \qquad \text{in } D,$$

$v + x_N \leq C\sigma$ in D, and $|\bar{u}(x)| \leq 2$. □

Once this lemma is established the following regularity result follows.

THEOREM 3.1. *Let* $u \in \mathcal{K}$ *be a solution to* (P_ε). *Then* $\partial_{red}\{u > 0\}$ *is a* $C^{1,\beta}$ *surface locally in* Ω, *and the remainder of the free boundary has zero* $\mathcal{H}^{N-1}$ *measure. Moreover, if* $N = 2$, *then the whole free boundary is a* $C^{1,\beta}$ *surface.*

4. Behavior of the minimizer for small ε. To complete the analysis of the problem, we now show that if ε is small enough, then

$$|\{u_\varepsilon > 0\}| = \alpha.$$

To this end, we need to prove that the constant $\lambda_\varepsilon := \lambda_{u_\varepsilon}$ is bounded from above and below by positive constants independent of ε. We perform this task in a series of lemmas.

LEMMA 4.1. *Let $u_\varepsilon \in \mathcal{K}$ be a solution to (P_ε). Then there exist constants $C, c > 0$ independent of ε such that*

$$c \le |\{u_\varepsilon > 0\}| \le \alpha + C\varepsilon. \tag{4.1}$$

Proof. As $\mathcal{J}_\varepsilon(u_\varepsilon)$ is bounded from above uniformly in ε we obtain

$$F_\varepsilon(|\{u_\varepsilon > 0\}|) \le C.$$

Hence

$$|\{u_\varepsilon > 0\}| \le \alpha + C\varepsilon.$$

For the lower bound, we proceed as follows: by the Sobolev trace embedding, for some $1 < p < 2$ such that $p(N-1)/(N-p) > q$,

$$1 \le \|u_\varepsilon\|_{L^q(\partial\Omega)} \le C\|u_\varepsilon\|_{W^{1,p}(\Omega)} \le C\|u_\varepsilon\|_{H^1(\Omega)}|\{u_\varepsilon > 0\}|^\theta$$

for some exponent θ that depends only on p. Since $\|u_\varepsilon\|_{H^1(\Omega)}$ is uniformly bounded, the lower bound follows. ∎

LEMMA 4.2. *Let $u_\varepsilon \in \mathcal{K}$ be a solution to (P_ε). Then there exists a constant $C > 0$ independent of ε such that*

$$\lambda_\varepsilon := \lambda_{u_\varepsilon} \le C.$$

Proof. Let $D \subset\subset \Omega$ smooth, such that $\omega = |D| > \alpha$ and $|\Omega \setminus D| < c$, where c is the constant in Lemma 4.1. Then,

$$|D \cap \{u_\varepsilon > 0\}| \le \alpha + C\varepsilon < \omega$$

for ε small enough. On the other hand,

$$|D \cap \{u_\varepsilon > 0\}| \ge |\{u_\varepsilon > 0\}| - |\Omega \setminus D| \ge c - |\Omega \setminus D| > 0.$$

Therefore by the relative isoperimetric inequality, we have

$$\mathcal{H}^{N-1}(D \cap \partial\{u_\varepsilon > 0\}) \ge c_0 \min\{|D \cap \{u_\varepsilon > 0\}|, |D \cap \{u_\varepsilon = 0\}|\}^{\frac{N-1}{N}} \ge c_1 > 0.$$

Now take $\varphi \in C_0^\infty(\Omega)$ as a test function in Lemma 2.3 such that $0 \le \varphi \le 1$, $\varphi \equiv 1$ in D, and $\|\nabla\varphi\|_\infty \le C = C(\operatorname{dist}(D, \partial\Omega))$ to get, since $\|u_\varepsilon\|_{H^1(\Omega)}$ is bounded independently of ε,

$$C \ge \int_\Omega \nabla u^\varepsilon \nabla\varphi \, dx + \int_\Omega u_\varepsilon \varphi \, dx = \lambda_\varepsilon(\varphi) \ge \lambda_\varepsilon \mathcal{H}^{N-1}(D \cap \partial_{\mathrm{red}}\{u_\varepsilon > 0\}).$$

This completes the proof of the lemma. ∎

The proof of the uniform lower bound follows similarly to Lemma 6 in [2]. We only make a sketch of the proof for the reader's convenience. It is at this point where we need the hypothesis that $\Gamma_D \ne \emptyset$.

LEMMA 4.3. *Let $\Gamma_D \ne \emptyset$ be the closure of a relatively open subset of $\partial\Omega$. Let $\varphi_0 \in H^1(\Omega)$ with $\varphi_0 \ge c_0 > 0$ in Γ_D. Let $u_\varepsilon \in \mathcal{K}$ be a solution to (P_ε). Then*

(1) *u_ε is positive in a neighborhood of Γ_D (depending on ε);*

(2) *there exists a constant* $c > 0$ *independent of* ε *such that*

$$c < \lambda_\varepsilon := \lambda_{u_\varepsilon}.$$

Proof. Let us first prove (1). In fact, arguing as in (2.9), given $y_0 \in \Gamma_D$ there exists a constant $K > 0$ independent of ε such that

$$|\Omega_r \cap \{u = 0\}| \left(\frac{1}{r} \fint_{\partial\Omega_r} u \right)^2 \leq K \int_{\Omega_r} |\nabla(u - v)|^2 \, dx,$$

where $\Omega_r = \Omega \cap B_r(y_0)$ and v is the solution of

$$\begin{cases} \Delta v = v & \text{in } \Omega_r, \\ v = u & \text{on } \partial\Omega_r. \end{cases}$$

Therefore,

$$\left(\frac{c_0}{r} \right)^2 |\Omega_r \cap \{u = 0\}| \leq K(\|u\|^2_{H^1(\Omega_r)} - \|v\|^2_{H^1(\Omega_r)}) \leq \frac{C}{\varepsilon} |\Omega_r \cap \{u = 0\}|.$$

Thus, $u > 0$ in Ω_r for small r depending on ε.

In order to see (2) we proceed as in [2, Lemma 6]. Let $y_0 \in \Gamma_D$ and let D_t with $0 \leq t \leq 1$ be a family of open sets with smooth boundary and uniformly (in ε and t) bounded curvatures such that D_0 is an exterior tangent ball at y_0, D_1 contains a free boundary point, $D_t \cap \partial\Omega \subset \Gamma_D$, and $D_0 \subset\subset D_t$ for $t > 0$.

Let $t \in (0, 1)$ be the first time that D_t touches the free boundary and let $x_0 \in \partial D_t \cap \partial\{u_\varepsilon > 0\} \cap \Omega$. Now, take w as the solution to $\Delta w = w$ in $D_t \setminus \overline{D}_0$ with $w = c_0$ on ∂D_0 and $w = 0$ on ∂D_t. Thus $w \leq u_\varepsilon$ in $D_t \cap \Omega$ and $\partial_{-\nu} w(x_0) \geq c\, c_0$ with c independent of ε; therefore, for r small enough,

$$\frac{1}{r} \fint_{\partial B_r(x_0)} u_\varepsilon \geq \frac{1}{r} \fint_{\partial B_r(x_0)} w \geq \bar{c}\, c_0$$

with $\bar{c}$ independent of ε.

If v_0 is the solution to

$$\begin{cases} \Delta v = v & \text{in } B_r(x_0), \\ v = u & \text{on } \partial B_r(x_0), \end{cases}$$

then, by (2.9), we have

$$\begin{aligned} c|B_r(x_0) \cap \{u_\varepsilon = 0\}| &\leq |B_r(x_0) \cap \{u_\varepsilon = 0\}| \left(\frac{1}{r} \fint_{\partial B_r(x_0)} u_\varepsilon \right)^2 \\ &\leq K \int_{B_r(x_0)} |\nabla(u_\varepsilon - v_0)|^2 \, dx \\ &\leq K(\|u_\varepsilon\|^2_{H^1(B_r(x_0))} - \|v_0\|^2_{H^1(B_r(x_0))}). \end{aligned}$$

Now let $\delta_r = |B_r(x_0) \cap \{u_\varepsilon = 0\}|$ and let $x_1 \in \partial\{u_\varepsilon > 0\}$ be such that the free boundary is smooth in a neighborhood of x_1. We perturb $\{u_\varepsilon > 0\}$ in a neighborhood of x_1 so that the measure of the perturbed set is increased by an amount δ_r (cf. Theorem 2.5).

Let Φ be a smooth nonnegative function supported in $B_\kappa(x_1)$ with $\kappa > 0$ small. For $x \in B_\kappa(x_1)$ we write $x = \sigma + s\nu(\sigma)$ with $\sigma \in \partial\{u_\varepsilon > 0\}$ and $s \in \mathbb{R}$, where $\nu(\sigma)$ is the outer unit normal to the free boundary at σ. We define the change of variables $y = x - \Phi(\sigma)\tau\nu(\sigma)$ with $\tau > 0$ small and the deformed set $\mathcal{D}_{\delta_r}$ such that $\mathcal{D}_{\delta_r} \cap B_\kappa(x_1) = \{y \,/\, x \in \{u_\varepsilon > 0\} \cap B_\kappa(x_1)\}$. Observe that if r is small we can perform this perturbation in such a way that it decreases the measure of $\{u_\varepsilon > 0\}$ by exactly δ_r. Also, observe that $\delta_r \to 0$ as $r \to 0$.

Now let v_r be the solution of

$$\begin{cases} \Delta v = v & \text{in } \mathcal{D}_{\delta_r}, \\ v = 0 & \text{on } \partial\mathcal{D}_{\delta_r} \cap B_\kappa(x_1), \\ v = u_\varepsilon & \text{on } \partial B_\kappa(x_1) \cap \overline{\mathcal{D}}_r. \end{cases} \tag{4.2}$$

Then v_r verifies

$$\frac{\partial v_r}{\partial \nu} = -\lambda_\varepsilon + o(\delta_r).$$

On the other hand,

$$u_\varepsilon = \lambda_\varepsilon \delta_r + o_\varepsilon(\delta_r) \qquad \text{on } \partial\{v_r > 0\} \cap B_\kappa(x_1).$$

Thus

$$\begin{aligned} \int_{B_\kappa(x_1)} |\nabla v_r|^2 + v_r^2 \, dx - \int_{B_\kappa(x_1)} |\nabla u_\varepsilon|^2 + (u_\varepsilon)^2 \, dx \\ &= \int_{B_\kappa(x_1)} |\nabla(u_\varepsilon - v_r)|^2 + (u_\varepsilon - v_r)^2 \, dx \\ &= -\int_{\partial\{v_r > 0\} \cap B_\kappa(x_1)} \frac{\partial v_r}{\partial \nu} u_\varepsilon \, dS \\ &= \lambda_\varepsilon^2 \delta_r + o_\varepsilon(\delta_r). \end{aligned}$$

Now we extend v_r by zero to $B_\kappa(x_1) \setminus \mathcal{D}_{\delta_r}$ and define

$$w_r = \begin{cases} v_r & \text{in } B_\kappa(x_1), \\ v_0 & \text{in } B_r(x_0), \\ u & \text{elsewhere.} \end{cases}$$

Then $|\{w_r > 0\}| = |\{u_\varepsilon > 0\}|$ and $w_r = u_\varepsilon$ on $\partial\Omega$; thus

$$\begin{aligned} 0 \le \mathcal{J}_\varepsilon(w_r) - \mathcal{J}_\varepsilon(u_\varepsilon) &= \int_\Omega |\nabla w_r|^2 + w_r^2 \, dx - \int_\Omega |\nabla u_\varepsilon|^2 + (u_\varepsilon)^2 \, dx \\ &= \int_{B_r(x_0)} |\nabla v_0|^2 + v_0^2 \, dx - \int_{B_\kappa(x_0)} |\nabla u_\varepsilon|^2 + (u_\varepsilon)^2 \, dx \\ &\quad + \int_{B_\kappa(x_1)} |\nabla v_r|^2 + v_r^2 \, dx - \int_{B_\kappa(x_1)} |\nabla u_\varepsilon|^2 + (u_\varepsilon)^2 \, dx \\ &\le -c\delta_r + \lambda_\varepsilon^2 \delta_r + o_\varepsilon(\delta_r) \end{aligned}$$

for every $r > 0$ small. Therefore, $\lambda_\varepsilon^2 \ge c/2$. ∎

Now we are in a position to prove the main result of this section, namely, that for ε small the measure of the positivity set is exactly α.

THEOREM 4.1. *Let $\Gamma_D \neq \emptyset$ be the closure of a relatively open subset of $\partial\Omega$. Let $\varphi_0 \in H^1(\Omega)$ with $\varphi_0 \geq c_0 > 0$ in Γ_D. Let $u_\varepsilon \in \mathcal{K}$ be a solution to (P_ε). Then, for ε small*

$$|\{u_\varepsilon > 0\}| = \alpha. \tag{4.3}$$

Proof. Arguing by contradiction, assume first that $|\{u_\varepsilon > 0\}| > \alpha$. Let $x_1 \in \partial\{u_\varepsilon > 0\} \cap \Omega$ be a regular point. We will proceed as in the proof of the previous lemma. Given $\delta > 0$, we perturb the domain $\{u_\varepsilon > 0\}$ in a neighborhood of x_1, $B_\kappa(x_1)$, decreasing its measure by δ. We choose δ small so that the measure of the perturbed set is still larger than α. Then we let v be the solution to (4.2) extended by zero to the rest of $B_\kappa(x_1)$ and equal to u in the rest of Ω. We have

$$\begin{aligned} 0 \leq \mathcal{J}_\varepsilon(v) - \mathcal{J}_\varepsilon(u_\varepsilon) &= \int_\Omega |\nabla v|^2 + v^2 - \int_\Omega |\nabla u_\varepsilon|^2 + (u_\varepsilon)^2 + F_\varepsilon(|\{v > 0\}|) \\ &\qquad - F_\varepsilon(|\{u_\varepsilon > 0\}|) \\ &\leq \lambda_\varepsilon^2 \delta + o_\varepsilon(\delta) - \frac{1}{\varepsilon}\delta \leq \left(C^2 - \frac{1}{\varepsilon}\right)\delta + o_\varepsilon(\delta) < 0 \end{aligned}$$

if $\varepsilon < \varepsilon_0$, and then $\delta < \delta_0(\varepsilon)$, a contradiction.

Now assume that $|\{u_\varepsilon > 0\}| < \alpha$. We proceed as in the previous case but this time we perturb in a neighborhood of x_1 the set $\{u_\varepsilon > 0\}$, increasing the measure by δ. Then we construct the function v as before, and if δ is small enough, $|\{v > 0\}| < \alpha$. Then

$$\begin{aligned} 0 \leq \mathcal{J}_\varepsilon(v) - \mathcal{J}_\varepsilon(u_\varepsilon) &= \int_\Omega |\nabla v|^2 + v^2 - \int_\Omega |\nabla u_\varepsilon|^2 + (u_\varepsilon)^2 + F_\varepsilon(|\{v > 0\}|) \\ &\qquad - F_\varepsilon(|\{u_\varepsilon > 0\}|) \\ &\leq -\lambda_\varepsilon^2 \delta + o_\varepsilon(\delta) + \varepsilon\delta \leq (-c^2 + \varepsilon)\delta + o_\varepsilon(\delta) < 0 \end{aligned}$$

if $\varepsilon < \varepsilon_1$, and then $\delta < \delta_0(\varepsilon)$. Again, a contradiction that ends the proof. ∎

As a consequence of the previous theorem, we get the following corollary.

COROLLARY 4.1. *Let $\Gamma_D \neq \emptyset$ be the closure of a relatively open subset of $\partial\Omega$. Let $\varphi_0 \in H^1(\Omega)$ with $\varphi_0 \geq c_0 > 0$ in Γ_D. Then there exists a minimizer u of $\mathcal{J}(v)$ in the set*

$$\mathcal{K}_\alpha = \{v \in H^1(\Omega) \,/\, \|v\|_{L^q(\Gamma_N)} = 1,\ v = \varphi_0 \text{ on } \Gamma_D,\ |\{v > 0\}| = \alpha\}.$$

This minimizer can be chosen in such a way that it is locally Lipschitz continuous in Ω and the free boundary $\partial\{u > 0\} \cap \Omega$ is locally a $C^{1,\beta}$ surface up to a set of zero $\mathcal{H}^{N-1}$ measure. In the case $N = 2$ the free boundary is locally a $C^{1,\beta}$ surface.

Proof. From our previous results we have (4.3) for every ε small enough. Therefore we can take $u = u_\varepsilon$ and the desired regularity of u and its free boundary follows from the results of sections 2 and 3. ∎

5. Main results. In this section we go back to our original minimization problem related to the best Sobolev trace constant. Here we prove that any extremal is a locally Lipschitz continuous function and the boundary of the hole $\partial\{u > 0\} \cap \Omega$ is locally $C^{1,\beta}$ up to a set of zero $\mathcal{H}^{N-1}$ measure.

We begin with the following theorem.

THEOREM 5.1. *Let ϕ_0 be a minimizer for (P_α). Assume that there exists a positive constant c such that $\phi_0 > c$ in a ball $B_0' \subset \Omega$ (resp., on $B_0' \cap \partial\Omega$, where B_0' is a ball centered at $\partial\Omega$). Then ϕ_0 is a minimizer of $\mathcal{J}_\varepsilon$ in*

$$\mathcal{K}_2 = \{v \in H^1(\Omega) \,/\, \|v\|_{L^q(\partial\Omega)} = 1,\ v = \phi_0 \text{ in } B_0\}$$

(resp., ϕ_0/k is a minimizer of $\mathcal{J}_\varepsilon$ in $\mathcal{K} = \{v \in H^1(\Omega) \,/\, \|v\|_{L^q(\Gamma_N)} = 1,\ v = \phi_0/k \text{ on } \Gamma_D\}$ with $\Gamma_D = \partial\Omega \cap B_0$, $\Gamma_N = \partial\Omega \setminus \Gamma_D$). Here, B_0 is a ball compactly contained in B_0' and $k = \|\phi_0\|_{L^q(\Gamma_N)}$.

In particular, ϕ_0 is locally Lipschitz continuous in Ω and the free boundary $\partial\{\phi_0 > 0\} \cap \Omega$ is locally a $C^{1,\beta}$ surface up to a set of zero $\mathcal{H}^{N-1}$ measure. In the case $N = 2$ the free boundary is locally a $C^{1,\beta}$ surface.

Proof. We will make the proof for the first case; the second one follows in the same way.

Let ε be small enough so that any minimizer u_ε of $\mathcal{J}_\varepsilon$ in $\mathcal{K}_2$ verifies that $|\{u_\varepsilon > 0\}| = \alpha$. Then it follows that ϕ_0 is one such minimizer and thus the conclusions of the theorem follow. In fact, as ϕ_0 minimizes (P_α) we have

$$\mathcal{J}_\varepsilon(\phi_0) = \int_\Omega |\nabla\phi_0|^2 + |\phi_0|^2\,dx \le \int_\Omega |\nabla v|^2 + |v|^2\,dx \tag{5.1}$$

for any $v \in H^1(\Omega)$ such that $\|v\|_{L^q(\partial\Omega)} = 1$ and $|\{v > 0\}| = \alpha$. In particular (5.1) holds for $v = u_\varepsilon$. Thus

$$\mathcal{J}_\varepsilon(\phi_0) \le \mathcal{J}_\varepsilon(u_\varepsilon) = \inf_{v\in\mathcal{K}_2} \mathcal{J}_\varepsilon(v).$$

This ends the proof. ∎

In particular, by the symmetry results for minimizers of (P_α) in balls of [10] we have the following corollary.

COROLLARY 5.1. *Let $\Omega = B(x_0, r)$ be a ball and let ϕ_0 be a minimizer of (P_α). Then ϕ_0 is locally Lipschitz continuous in $B(x_0, r)$ and the free boundary $\partial\{\phi_0 > 0\} \cap B(x_0, r)$ is locally a $C^{1,\beta}$ surface up to a set of zero $\mathcal{H}^{N-1}$ measure. In the case $N = 2$ the free boundary is locally a $C^{1,\beta}$ surface.*

Proof. In [10] it was proved that any minimizer ϕ_0 of (P_α) in the case that Ω is a ball $B_r(x_0)$ satisfies that, for any $c_0 > 0$, $\{\phi_0 \ge c_0\} \cap \partial B_r(x_0)$ is a spherical cap. Since $\|\phi_0\|_{L^q(\partial\Omega)} = 1$, there exists $c_0 > 0$ such that $\{\phi_0 \ge c_0\} \cap \partial B_r(x_0) \ne \emptyset$. Hence the conditions of Theorem 5.1 are satisfied. ∎

In the general case, for the problem (P_α) we can prove that the set of α's for which there exist minimizers with smooth free boundary is dense in $(0, |\Omega|)$. More precisely, we have the following theorem.

THEOREM 5.2. *For any $0 < \alpha < |\Omega|$ there exists $\alpha_\varepsilon \to \alpha$ as $\varepsilon \to 0$ such that there exists a solution ϕ_ε of (P_{α_ε}) which is locally Lipschitz continuous in Ω and has locally a $C^{1,\beta}$ free boundary up to a set of zero $\mathcal{H}^{N-1}$ measure. In the case $N = 2$ the free boundary is locally a $C^{1,\beta}$ surface.*

Proof. Let u_ε be a minimizer of $\mathcal{J}_\varepsilon$. We already know that $\alpha_\varepsilon := |\{u_\varepsilon > 0\}| \le \alpha + C\varepsilon$ (see (4.1)). Let us see that $\alpha_\varepsilon \to \alpha$ as $\varepsilon \to 0$. If not, there exists a sequence $\varepsilon_j \to 0$ such that $\alpha_{\varepsilon_j} = |\{u_{\varepsilon_j} > 0\}| \le \theta < \alpha$. Let ϕ_0 be a minimizer of (P_α). By the strict monotonicity of $S(\alpha)$ (see [10, Remark 2.2]) we have

$$\begin{aligned}\mathcal{J}(\phi_0) = S(\alpha) < S(\theta) \le \mathcal{J}(u_{\varepsilon_j}) &= \mathcal{J}_{\varepsilon_j}(u_{\varepsilon_j}) - F_{\varepsilon_j}(\alpha_{\varepsilon_j}) \\ &\le \mathcal{J}_{\varepsilon_j}(\phi_0) - F_{\varepsilon_j}(\alpha_{\varepsilon_j}) = \mathcal{J}(\phi_0) - F_{\varepsilon_j}(\alpha_{\varepsilon_j}) \le \mathcal{J}(\phi_0) + C\varepsilon_j,\end{aligned}$$

a contradiction.

Now, taking $\phi_\varepsilon = u_\varepsilon$ we see that ϕ_ε is a minimizer of (P_{α_ε}). In fact, let v be an admissible function for (P_{α_ε}); then

$$\mathcal{J}(v) + F_\varepsilon(\alpha_\varepsilon) = \mathcal{J}_\varepsilon(v) \geq \mathcal{J}_\varepsilon(\phi_\varepsilon) = \mathcal{J}(\phi_\varepsilon) + F_\varepsilon(\alpha_\varepsilon)$$

and therefore

$$\mathcal{J}(v) \geq \mathcal{J}(\phi_\varepsilon).$$

The theorem is proved. □

Finally, we have the following result.

THEOREM 5.3. *Let u_ε be a minimizer of $\mathcal{J}_\varepsilon$ in $\mathcal{K}_1$. Then there exists $\phi_0 \in H^1(\Omega)$, a solution to (P_α) such that, up to a subsequence, $u_\varepsilon \to \phi_0$ in $H^1(\Omega)$.*

Proof. In the proof of Theorem 5.2 we showed that $|\{u_\varepsilon > 0\}| \to \alpha$ as $\varepsilon \to 0$.

It is easy to see that $\mathcal{J}_\varepsilon(u_\varepsilon)$ is bounded uniformly in ε and so u_ε is uniformly bounded in $H^1(\Omega)$. Therefore, passing to a subsequence if necessary, there exists $u_0 \in H^1(\Omega)$ such that

$$\begin{aligned} &u_\varepsilon \rightharpoonup u_0 \quad \text{weakly in } H^1(\Omega),\\ &u_\varepsilon \to u_0 \quad \text{strongly in } L^q(\partial\Omega),\\ &u_\varepsilon \to u_0 \quad \text{a.e. } \Omega. \end{aligned}$$

Thus,

$$\begin{aligned} &\|u_0\|_{L^q(\partial\Omega)} = 1,\\ &|\{u_0 > 0\}| \leq \alpha = \lim_{\varepsilon\to 0} |\{u_\varepsilon > 0\}|, \quad \text{and}\\ &\|u_0\|_{H^1(\Omega)} \leq \liminf_{\varepsilon\to 0} \|u_\varepsilon\|_{H^1(\Omega)}. \end{aligned}$$

Let us call $\phi_0 = u_0$ and let us see that ϕ_0 is a solution to (P_α). In fact, let $v \in H^1(\Omega)$ be such that $|\{v > 0\}| = \alpha$ and $\|v\|_{L^q(\partial\Omega)} = 1$. Then

$$\mathcal{J}(v) = \mathcal{J}_\varepsilon(v) \geq \mathcal{J}_\varepsilon(u_\varepsilon).$$

Now, since $\liminf_{\varepsilon\to 0} F_\varepsilon(|\{u_\varepsilon > 0\}|) \geq 0$, there holds that

$$\mathcal{J}(v) \geq \liminf_{\varepsilon\to 0} \mathcal{J}_\varepsilon(u_\varepsilon) \geq \liminf_{\varepsilon\to 0} \mathcal{J}(u_\varepsilon) \geq \mathcal{J}(\phi_0). \tag{5.2}$$

It remains to see that $|\{\phi_0 > 0\}| = \alpha$. Assume not; then $\alpha_1 := |\{\phi_0 > 0\}| < \alpha$. Thus, by the strict monotonicity of $S(\cdot)$, there holds that $S(\alpha) < S(\alpha_1)$, but

$$S(\alpha) = \inf_v \mathcal{J}(v) \geq \mathcal{J}(\phi_0) \geq S(\alpha_1),$$

a contradiction.

Now taking $v = \phi_0$ in (5.2),

$$\mathcal{J}(\phi_0) \leq \liminf_{\varepsilon\to 0} \mathcal{J}(u_\varepsilon) \leq \liminf_{\varepsilon\to 0} \mathcal{J}_\varepsilon(u_\varepsilon) \leq \mathcal{J}(\phi_0).$$

Hence, $\|\phi_0\|_{H^1(\Omega)} = \liminf_{\varepsilon\to 0} \|u_\varepsilon\|_{H^1(\Omega)}$ and thus, by taking a further subsequence if necessary, the convergence is actually strong. □

Remark 4. We believe that, as in the previous cases, the minimizers u_ε of $\mathcal{J}_\varepsilon$ in $\mathcal{K}_1$ will already be solutions to (P_α) for ε small. Nevertheless, despite the fact that the result of Theorem 5.3 does not give regularity of the minimizer ϕ_0, we believe that it could be of interest in numerical approximations of the solution to (P_α).

Acknowledgment. We wish to thank Prof. Luis Caffarelli for suggesting this approach to our optimization problem and for providing us with useful references.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Pure and Applied Mathematics, Vol. 65, Academic Press, New York, London, 1975.

[2] N. Aguilera, H. W. Alt, and L. A. Caffarelli, *An optimization problem with volume constraint*, SIAM J. Control Optim., 24 (1986), pp. 191–198.

[3] N. Aguilera, L. A. Caffarelli, and J. Spruck, *An optimization problem in heat conduction*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 14 (1988), pp. 355–387.

[4] H. W. Alt and L. A. Caffarelli, *Existence and regularity for a minimum problem with free boundary*, J. Reine Angew. Math., 325 (1981), pp. 105–144.

[5] H. W. Alt, L. A. Caffarelli, and A. Friedman, *A free boundary problem for quasilinear elliptic equations*, Ann. Scuola Norm. Sup. Pisa Cl., Sci. (4), 11 (1984), pp. 1–44.

[6] T. Aubin, *Equations différentielles non linéaires et le problème de Yamabe concernant la courbure scalaire*, J. Math. Pures Appl. (9), 55 (1976), pp. 269–296.

[7] O. Druet and E. Hebey, *The AB program in geometric analysis: Sharp Sobolev inequalities and related problems*, Mem. Amer. Math. Soc., 160 (761) (2002).

[8] H. Federer, *Geometric Measure Theory*, Grundlehren Math. Wiss. 153, Springer-Verlag, New York, 1969.

[9] J. Fernández Bonder and J. D. Rossi, *Asymptotic behavior of the best Sobolev trace constant in expanding and contracting domains*, Commun. Pure Appl. Anal., 1 (2002), pp. 359–378.

[10] J. Fernández Bonder, J. D. Rossi, and N. Wolanski, *On the best Sobolev trace constant and extremals in domains with holes*, preprint, University of Buenos Aires, Buenos Aires, 2004.

[11] C. Flores and M. del Pino, *Asymptotic behavior of best constants and extremals for trace embeddings in expanding domains*, Comm. Partial Differential Equations, 26 (2001), pp. 2189–2210.

[12] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Grundlehren Math. Wiss. 224, Springer-Verlag, Berlin, 1983.

[13] A. Henrot, *Minimization problems for eigenvalues of the Laplacian*, J. Evol. Equ., 3 (2003), pp. 443–461.

[14] B. Kawohl, *Rearrangements and convexity of level sets in PDE*, Lecture Notes in Math. 1150, Springer-Verlag, Berlin, 1985.

[15] C. Lederman, *An optimization problem in elasticity*, Differential Integral Equations, 8 (1995), pp. 2025–2044.

[16] C. Lederman, *A free boundary problem with a volume penalization*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 23 (1996), pp. 249–300.

[17] Y. Li and M. Zhu, *Sharp Sobolev trace inequalities on Riemannian manifolds with boundaries*, Comm. Pure Appl. Math., 50 (1997), pp. 449–487.

[18] E. H. Lieb and M. Loss, *Analysis*, Grad. Stud. Math. 14, 2nd ed., AMS, Providence, RI, 2001.

[19] M. W. Steklov, *Sur les problèmes fondamentaux en physique mathématique*, Ann. Sci. École Norm. Sup. (4), 19 (1902), pp. 455–490.

[20] E. Teixeira, *A nonlinear optimization problem in heat conduction*, Calc. Var. Partial Differential Equations, 24 (2005), pp. 21–46.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 5, pp. 1636–1649

L^∞-ESTIMATES FOR APPROXIMATED OPTIMAL CONTROL PROBLEMS*

C. MEYER† AND A. RÖSCH‡

Abstract. An optimal control problem for a two-dimensional elliptic equation is investigated with pointwise control constraints. This paper is concerned with discretization of the control by piecewise linear functions. The state and the adjoint state are discretized by linear finite elements. Approximation of order h in the L^∞-norm is proved in the main result.

Key words. linear-quadratic optimal control problems, error estimates, elliptic equations, numerical approximation, control constraints

AMS subject classifications. 49K20, 49M25, 65N30

DOI. 10.1137/040614621

1. Introduction.

This paper is concerned with the discretization of the two-dimensional elliptic optimal control problem

$$J(u) = F(y,u) = \frac{1}{2}\|y - y_d\|^2_{L^2(\Omega)} + \frac{\nu}{2}\|u\|^2_{L^2(\Omega)} \tag{1.1}$$

subject to the state equations

$$\begin{aligned} Ay + a_0 y &= u && \text{in } \Omega, \\ y &= 0 && \text{on } \Gamma \end{aligned} \tag{1.2}$$

and subject to the control constraints

$$a \le u(t,x) \le b \qquad \text{for a.a. } x \in \Omega, \tag{1.3}$$

where Ω is a bounded domain with boundary Γ; A denotes a second-order elliptic operator of the form

$$Ay(x) = -\sum_{i,j=1}^{2} D_i(a_{ij}(x) D_j y(x)),$$

where D_i denotes the partial derivative with respect to x_i, and a and b are real numbers. Moreover, $\nu > 0$ is a fixed positive number. We denote the set of admissible controls by U_{ad}:

$$U_{ad} = \{u \in L^2(\Omega) : \ a \le u \le b \ \text{ a.e. in } \Omega\}.$$

We discuss here the full discretization of the control and the state equations by a finite-element method. The asymptotic behavior of the discretized problem is studied.

*Received by the editors September 8, 2004; accepted for publication (in revised form) March 14, 2005; published electronically November 22, 2005. This work was supported by the DFG Research Center "Mathematics for Key Technologies" (FZT 86) in Berlin.
http://www.siam.org/journals/sicon/44-5/61462.html
†Technische Universität Berlin, Fakultät II Mathematik und Naturwissenschaften, Straße des 17. Juni 136, D-10623 Berlin, Germany (cmeyer@math.tu-berlin.de).
‡Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenbergerstraße 69, A-4040 Linz, Austria (arnd.roesch@oeaw.ac.at).

The approximation of the discretization for semilinear elliptic optimal control problems is discussed in Arada, Casas, and Tröltzsch [1]. The optimal control problem (1.1)–(1.3) is a linear-quadratic counterpart of such a semilinear problem.

The discretization of optimal control problems by piecewise constant functions is well investigated; we refer to Falk [7] and Geveci [8]. Piecewise constant and piecewise linear discretization in space are discussed for a parabolic problem in Malanowski [12]. Theory and numerical results for elliptic boundary control problems are contained in Casas and Tröltzsch [6] and Casas, Mateos, and Tröltzsch [5]. All of these papers are mainly focused on L^2-estimates. However, in Arada, Casas, and Tröltzsch [1] and Casas, Mateos, and Tröltzsch [5] we find also L^∞-estimates of order h for piecewise constant functions.

Piecewise linear control discretizations for elliptic optimal control problems were studied by Casas and Tröltzsch [6] and Casas [4]. In an abstract optimization problem, piecewise linear approximations were investigated in Rösch [15]. In these papers, the convergence was mainly discussed in the L^2-norm.

Error estimates in the L^∞-norm can also be obtained by other discretization concepts; see Hinze [10] and Meyer and Rösch [13].

The interest for L^∞-estimates is motivated by the following circumstances. The L^∞-space plays an important role in the theory of semilinear optimal control problems. Usually, sufficient second-order optimality conditions hold in an L^∞-neighborhood of the optimal solution. These optimality conditions are the main ingredients for the convergence theory of the SQP method. For numerical computations the linear-quadratic subproblems have to be solved with a sufficient accuracy in the L^∞-norm.

In this paper, we will show that also for piecewise linear controls the approximation order h can be obtained in the L^∞-norm. A result of this type cannot be obtained with one of the above mentioned methods. The L^∞-estimate is obtained in two main steps. We prove in the first step that the discretized solutions violate a pointwise projection formula only in an order h. The L^∞-estimates for grid points and later for arbitrary points are derived in the second step.

The paper is organized as follows: In section 2 the discretizations are introduced and the main results are stated. Section 3 contains auxiliary results. The proofs of the approximation result is placed in section 4. The paper ends with numerical experiments shown in section 5.

2. Discretization and main result. Throughout this paper, Ω denotes a convex bounded open subset in $\mathbb{R}^2$ of class $C^{1,1}$. The coefficients a_{ij} of operator A belong to $C^{0,1}(\bar{\Omega})$ and satisfy the ellipticity condition

$$m_0|\xi|^2 \le \sum_{i,j=1}^{2} a_{ij}(x)\xi_i\xi_j \qquad \text{for all } (\xi, x) \in \mathbb{R}^2 \times \bar{\Omega}, \quad m_0 > 0.$$

Moreover, we require $y_d \in L^p(\Omega)$ for some $p > 2$. For the function $a_0 \in L^\infty(\Omega)$, we assume $a_0 \ge 0$. Next, we recall a result of Grisvard [9, Theorem 2.4.2.5].

LEMMA 2.1 (see [9]). *For every $p > 2$ and every function $g \in L^p(\Omega)$, the solution y of*

$$Ay + a_0 y = g \qquad \text{in } \Omega, \quad y|_\Gamma = 0,$$

belongs to $H_0^1(\Omega) \cap W^{2,p}(\Omega)$. Moreover, there exists a positive constant c, independent of a_0, such that

$$\|y\|_{W^{2,p}(\Omega)} \le c\|g\|_{L^p(\Omega)}.$$

We introduce the adjoint equation

$$\begin{aligned} A^*p + a_0 p &= y - y_d && \text{in } \Omega, \\ p &= 0 && \text{on } \Gamma, \end{aligned} \tag{2.1}$$

where A^* denotes the formally adjoint operator. Due to Lemma 2.1, the state equation and the adjoint equation admit unique solutions in $H_0^1(\Omega) \cap W^{2,p}(\Omega)$ if $y_d \in L^p(\Omega)$ for $p > 2$. This space is embedded in $C^{0,1}(\bar{\Omega})$.

We call the solution y of (1.2) for a control u associated state to u and write $y(u)$. In the same way, we call the solution p of (2.1) corresponding to $y(u)$ associated adjoint state to u and write $p(u)$.

Introducing the projection

$$\Pi_{[a,b]}(f(x)) = \max(a, \min(b, f(x))),$$

we can formulate the necessary and sufficient first-order optimality condition for (1.1)–(1.3).

LEMMA 2.2. *A necessary and sufficient condition for the optimality of a control $\bar{u}$ with the corresponding state $\bar{y} = y(\bar{u})$ and adjoint state $\bar{p} = p(\bar{u})$, respectively, is that the equation*

$$\bar{u}(x) = \Pi_{[a,b]}\Big(-\frac{1}{\nu}\bar{p}(x)\Big) \tag{2.2}$$

holds.

Since the optimal control problem is strictly convex, we obtain the existence of a unique optimal solution. The optimality condition can be formulated as the variational inequality

$$(\nu\bar{u} + \bar{p}, u - \bar{u})_U \geq 0 \qquad \text{for all } u \in U_{ad},$$

where $(\cdot,\cdot)_U$ denotes the natural inner product in $U = L^2(\Omega)$. A standard pointwise a.e. discussion of this variational inequality leads to the above formulated projection formula; see [12]. Clearly, the Lipschitz continuity of $\bar{p}$ implies that also $\bar{u}$ is Lipschitz continuous.

We are now able to introduce the discretized problem. We define a finite-element–based approximation of the optimal control problem (1.1)–(1.3). To this aim, we consider a family of triangulations $(T_h)_{h>0}$ of $\bar{\Omega}$. With each element $T \in T_h$, we associate two parameters $\rho(T)$ and $\sigma(T)$, where $\rho(T)$ denotes the diameter of the set T, and $\sigma(T)$ is the diameter of the largest ball contained in T. The mesh size of the grid is defined by $h = \max_{T\in T_h} \rho(T)$. We suppose that the following regularity assumptions are satisfied.

(A1) There exist two positive constants ρ and σ such that

$$\frac{\rho(T)}{\sigma(T)} \leq \sigma, \qquad \frac{h}{\rho(T)} \leq \rho$$

hold for all $T \in T_h$ and all $h > 0$.

(A2) Let us define $\bar{\Omega}_h = \bigcup_{T\in T_h} T$, and let Ω_h and Γ_h denote its interior and its boundary, respectively. We assume that $\bar{\Omega}_h$ is convex and that the vertices of T_h placed on the boundary of Γ_h are points of Γ. From [14, estimate (5.2.19)], it is known that

$$|\Omega \setminus \Omega_h| \leq Ch^2,$$

where $|.|$ denotes the measure of the set.

Next, to every boundary triangle T of T_h we associate another triangle $\hat{T}$ with the curved boundary as follows: the edge between the two boundary nodes of T is substituted by the corresponding curved part of Γ. We denote by $\hat{T}_h$ the union of these curved boundary triangles with the interior triangles to Ω of T_h such that $\bar{\Omega} = \bigcup_{\hat{T}\in\hat{T}_h} \hat{T}$. Moreover, we set

$$V_h = \{y_h \in C(\bar{\Omega}) : \ y_h \in \mathcal{P}_1 \text{ for all } T \in T_h, \text{ and } y_h = 0 \text{ on } \bar{\Omega} \setminus \Omega_h\},$$

$$U_h = \{u_h \in C(\bar{\Omega}) : u_h \in \mathcal{P}_1 \text{ for all } T \in T_h, \text{ and } u_h = \Pi_{[a,b]}(0) \text{ on } \bar{\Omega} \setminus \Omega_h\},$$
$$U_h^{ad} = U_h \cap U_{ad},$$

where $\mathcal{P}_1$ is the space of polynomials of degree less than or equal to 1. The definition of the set U_h is motivated by the projection formula (2.2) and the homogeneous boundary condition (2.1) of the adjoint equation.

For each $u_h \in U_h$, we denote by $y_h(u_h)$ the unique element of V_h that satisfies

$$a(y_h(u_h), v_h) = \int_\Omega u_h v_h \, dx \qquad \text{for all } v_h \in V_h, \tag{2.3}$$

where $a : V_h \times V_h \to \mathbb{R}$ is the bilinear form defined by

$$a(y_h, v_h) = \int_\Omega \left(a_0(x) y_h(x) v_h(x) + \sum_{i,j=1}^{2} a_{ij}(x) D_i y_h(x) D_j v_h(x) \right) dx.$$

In other words, $y_h(u_h)$ is the approximated state associated with u_h. Because of $y_h = v_h = 0$ on $\bar{\Omega} \setminus \Omega_h$ the integrals over Ω can be replaced by integrals over Ω_h. The finite-dimensional approximation of the optimal control problem is defined by

$$\inf J(u_h) = \frac{1}{2}\|y_h(u_h) - y_d\|^2_{L^2(\Omega)} + \frac{\nu}{2}\|u_h\|^2_{L^2(\Omega)}, \qquad u_h \in U_h^{ad}. \tag{2.4}$$

The adjoint equation is discretized in the same way:

$$a^*(p_h(u_h), v_h) = \int_\Omega (y_h(u_h) - y_d) v_h \, dx \qquad \text{for all } v_h \in V_h \tag{2.5}$$

with

$$a^*(p_h(u_h), v_h) = \int_\Omega \left(a_0(x) y_h(x) v_h(x) + \sum_{i,j=1}^{2} a_{ji}(x) D_i y_h(x) D_j v_h(x) \right) dx.$$

Now, we are able to state the main result.

THEOREM 2.3. *Let $\bar{u}$ and u_h be the optimal solution of* (1.1) *and* (2.4)*, respectively. Then, there exists a positive constant C independent of h with*

$$\|\bar{u} - u_h\|_{L^\infty(\Omega)} \le Ch. \tag{2.6}$$

The proof of Theorem 2.3 is contained in section 4. Moreover, the constant C is specified in that section.

3. Auxiliary results. We start with an L^2-estimate corresponding to Theorem 2.3.

LEMMA 3.1. *Let $\bar{u}$ and u_h be the optimal solution of* (1.1) *and* (2.4)*, respectively. Then, the estimate*

$$\|\bar{u} - u_h\|_{L^2(\Omega)} \le C_2 h \tag{3.1}$$

holds true with a positive constant C_2.

This statement can be easily proved by the arguments of Casas and Tröltzsch [6]. It is a special case of a new general result of Casas [4].

This and the Sobolev imbeddings imply easily the following L^∞-estimate:

$$\|\bar{p} - p(u_h)\|_{L^\infty(\Omega)} \le c\|\bar{p} - p(u_h)\|_{H^2(\Omega)} \le ch. \tag{3.2}$$

LEMMA 3.2. *The inequality*

$$\|\bar{p} - p_h(u_h)\|_{L^\infty(\Omega)} \le \kappa h \tag{3.3}$$

is valid with a positive constant κ.

Proof. First, we recall an L^∞-estimate for the finite-element solution

$$\|p(u_h) - p_h(u_h)\|_{L^\infty(\Omega)} \le ch; \tag{3.4}$$

see Braess [3]. Using (3.2), we find

$$\|\bar{p} - p_h\|_{L^\infty(\Omega)} \le \|\bar{p} - p(u_h)\|_{L^\infty(\Omega)} + \|p(u_h) - p_h\|_{L^\infty(\Omega)} \le \kappa h. \qquad \square$$

Next, we introduce a new notation for the piecewise linear functions. Let E_i be an arbitrary vertex of the triangulation T_h. Then, we define a basis function $e_i \in U_h$ by

$$e_i(E_j) = \delta_{ij},$$

where δ_{ij} is the Kronecker symbol. Therefore, we can represent the functions u_h and $p_h(u_h)$ by

$$u_h(x) = \sum_{E_i} u_i e_i(x),$$
$$(p_h(u_h))(x) = \sum_{E_i} p_i e_i(x)$$

with $u_i = u_h(E_i)$ and $p_i = (p_h(u_h))(E_i)$.

We denote the set of neighboring vertices of E_i, i.e., $(e_i, e_j)_U \ne 0$ and $i \ne j$, by $N(E_i)$.

LEMMA 3.3. *For every j with $E_j \in N(E_i)$, we have*

$$\frac{1}{\nu}|p_i - p_j| \le Dh \tag{3.5}$$

with

$$D = \frac{L + 2\kappa}{\nu}, \tag{3.6}$$

where L denotes the Lipschitz constant of $\bar{p}$.

Proof. Because of Lemma 2.1, $\bar{p}$ belongs to $W^{2,p}(\Omega)$ for a certain $p > 2$. Therefore, $\bar{p}$ is Lipschitz, and we have

$$|\bar{p}(E_i) - \bar{p}(E_j)| \le Lh.$$

Combining this inequality with (3.3), we obtain

$$\begin{aligned} |p_i - p_j| &\le |p_i - \bar{p}(E_i)| + |\bar{p}(E_i) - \bar{p}(E_j)| + |\bar{p}(E_j) - p_j| \\ &\le \kappa h + Lh + \kappa h. \qquad \square \end{aligned}$$

Next, we recall a property concerning the mass matrix.

LEMMA 3.4. *For every basis function e_i*

$$(e_i, e_i)_U \ge \sum_{E_j \in N(E_i)} (e_i, e_j)_U \tag{3.7}$$

is valid.

Proof. The element mass matrix of the reference element T_r is given by

$$M_r = \frac{1}{24} \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix},$$

which has the desired property with equality. For an arbitrary triangle T_s we find

$$M_s = \frac{|T_s|}{|T_r|} M_r.$$

Consequently, every element mass matrix has this property. This holds also for the summation over all triangles. The inequality sign is obtained if the support of e_i contains at least one boundary point. $\square$

Next, we want to investigate the following quantity:

$$M := \max_i \left| u_i - \Pi_{[a,b]}\Big(-\frac{1}{\nu} p_i \Big) \right|. \tag{3.8}$$

Our main goal will be the proof of the inequality

$$M \le D \cdot h.$$

Of course for $M = 0$ such an inequality is automatically fulfilled. Therefore, we will now assume $M > 0$. In all what follows, the index i denotes a fixed vertex where this maximum is attained.

Equation (3.8) means that one of the following cases (A) and (B) occurs:

(A) $\quad M = u_i - \Pi_{[a,b]}(-\frac{1}{\nu} p_i)$,

(B) $\quad M = -(u_i - \Pi_{[a,b]}(-\frac{1}{\nu} p_i))$.

LEMMA 3.5. *Let $M > 0$ and i be an index where this maximum in* (3.8) *is attained. Then, we have*

$$\begin{aligned} M &\le u_i + \frac{1}{\nu} p_i \qquad \text{in case (A)}, \\ M &\le -\left(u_i + \frac{1}{\nu} p_i \right) \quad \text{in case (B)}. \end{aligned} \tag{3.9}$$

Moreover, there exists a constant $\varepsilon > 0$ such that $v_h = u_h - \varepsilon e_i$ is admissible in case (A) *and $v_h = u_h + \varepsilon e_i$ is admissible in case* (B).

Proof. We discuss here only case (A). Case (B) can be investigated in the same way.

Since M is positive and $\Pi_{[a,b]}(-\frac{1}{\nu}p_i) \in [a,b]$ by definition, we have

$$u_i > a.$$

Moreover, $u_i \le b$ and $M > 0$ imply $\Pi_{[a,b]}(-\frac{1}{\nu}p_i) < b$. Therefore, we have

$$\Pi_{[a,b]}\Big(-\frac{1}{\nu}p_i\Big) \ge -\frac{1}{\nu}p_i$$

and

$$M = u_i - \Pi_{[a,b]}\Big(-\frac{1}{\nu}p_i\Big) \le u_i + \frac{1}{\nu}p_i.$$

Because of $u_i > a$, there exists an $\varepsilon > 0$ such that

$$u_i - \varepsilon > a.$$

This means that the control $v_h = u_h - \varepsilon e_i$ is admissible. ∎

LEMMA 3.6. *Let $M > 0$, and let i be an index, where the maximum in* (3.8) *is attained. Then, we have*

$$u_i + \frac{1}{\nu}p_i \le \max_{E_j \in N(E_i)} -\Big(u_j + \frac{1}{\nu}p_j\Big) \quad \text{in case (A)}, \tag{3.10}$$

$$-(u_i + \frac{1}{\nu}p_i) \le \max_{E_j \in N(E_i)} \Big(u_j + \frac{1}{\nu}p_j\Big) \quad \text{in case (B)}. \tag{3.11}$$

Moreover, if equality holds in (3.10) *or* (3.11), *then we have*

$$u_i + \frac{1}{\nu}p_i = -\Big(u_j + \frac{1}{\nu}p_j\Big) \quad \text{for all } j \text{ with } E_j \in N(E_i).$$

Proof. Without loss of generality, we discuss only case (A). We start with the optimality condition for u_h:

$$(\nu u_h + p_h(u_h), v_h - u_h)_U \ge 0 \qquad \text{for all } v_h \in U_h^{ad}.$$

We test this inequality with $v_h = u_h - \varepsilon e_i$:

$$(\nu u_h + p_h(u_h), -\varepsilon e_i)_U \ge 0.$$

From this, we obtain

$$(\nu u_i + p_i)(e_i, e_i)_U \le \sum_{E_j \in N(E_i)} -(\nu u_j + p_j)(e_i, e_j)_U.$$

Using (3.7), we find

$$\begin{aligned}(\nu u_i + p_i)(e_i, e_i)_U &\le \max_{E_j \in N(E_i)} -\Big(u_j + \frac{1}{\nu}p_j\Big) \sum_{E_j \in N(E_i)} (e_i, e_j)_U \\ &\le \max_{E_j \in N(E_i)} -\Big(u_j + \frac{1}{\nu}p_j\Big)(e_i, e_i)_U.\end{aligned}$$

Division by $(e_i, e_i)_U$ yields (3.10). Since the scalar products $(e_i, e_j)_U$ are positive for all j with $E_j \in N(E_i)$, equality can only occur if

$$u_i + \frac{1}{\nu} p_i = -\Big(u_j + \frac{1}{\nu} p_j\Big) \qquad \text{for all } j \text{ with } E_j \in N(E_i). \qquad \square$$

Next, we denote an index where the maximum is attained in (3.10) for case (A) by k:

$$-\Big(u_k + \frac{1}{\nu} p_k\Big) = \max_{E_j \in N(E_i)} -\Big(u_j + \frac{1}{\nu} p_j\Big). \tag{3.12}$$

In case (B), an index k is defined by

$$u_k + \frac{1}{\nu} p_k = \max_{E_j \in N(E_i)} \Big(u_j + \frac{1}{\nu} p_j\Big). \tag{3.13}$$

LEMMA 3.7. *Assume $M > 0$. It holds that*

$$-\Big(u_k - \Pi_{[a,b]}\Big(-\frac{1}{\nu} p_k\Big)\Big) \le M \le u_i + \frac{1}{\nu} p_i \le -\Big(u_k + \frac{1}{\nu} p_k\Big) \tag{3.14}$$

in case (A) *and that*

$$u_k - \Pi_{[a,b]}\Big(-\frac{1}{\nu} p_k\Big) \le M \le -\Big(u_i + \frac{1}{\nu} p_i\Big) \le u_k + \frac{1}{\nu} p_k \tag{3.15}$$

in case (B).

Proof. Again, we discuss only case (A): combining (3.9), (3.10), and (3.12), we find

$$M \le u_i + \frac{1}{\nu} p_i \le -\Big(u_k + \frac{1}{\nu} p_k\Big). \tag{3.16}$$

Moreover, we have by definition of M

$$M \ge \Big|u_k - \Pi_{[a,b]}\Big(-\frac{1}{\nu} p_k\Big)\Big|. \tag{3.17}$$

Due to (3.12) and $M > 0$, $u_k + \frac{1}{\nu} p_k$ is negative. Hence, the expression $u_k - \Pi_{[a,b]}(-\frac{1}{\nu} p_k)$ is nonpositive: in the case $-\frac{1}{\nu} p_k \le b$, we get

$$\Pi_{[a,b]}\Big(-\frac{1}{\nu} p_k\Big) \ge -\frac{1}{\nu} p_k.$$

Consequently, we find

$$0 > u_k - \Big(-\frac{1}{\nu} p_k\Big) \ge u_k - \Pi_{[a,b]}\Big(-\frac{1}{\nu} p_k\Big).$$

In the other case, $-\frac{1}{\nu} p_k > b$, we have $\Pi_{[a,b]}(-\frac{1}{\nu} p_k) = b$, and now

$$0 \ge u_k - \Pi_{[a,b]}\Big(-\frac{1}{\nu} p_k\Big) \tag{3.18}$$

follows from $u_k \le b$.

Combining (3.17) and (3.18), we have

$$M \geq -\Big(u_k - \Pi_{[a,b]}\Big(-\frac{1}{\nu}p_k\Big)\Big). \tag{3.19}$$

This yields together with (3.16)

$$-\Big(u_k - \Pi_{[a,b]}\Big(-\frac{1}{\nu}p_k\Big)\Big) \leq M \leq u_i + \frac{1}{\nu}p_i \leq -\Big(u_k + \frac{1}{\nu}p_k\Big),$$

i.e., inequality (3.14). ∎

LEMMA 3.8. *Assume $M > 0$. Then, there exists an index i with*

$$M = \Big|u_i - \Pi_{[a,b]}\Big(-\frac{1}{\nu}p_i\Big)\Big|$$

and a corresponding index k, $E_k \in N(E_i)$ with

$$\Pi_{[a,b]}\Big(-\frac{1}{\nu}p_k\Big) \neq -\frac{1}{\nu}p_k. \tag{3.20}$$

Proof. Again, we discuss only case (A): first, we investigate the case where in inequality (3.14) at least one strong inequality occurs. Then, we have

$$-\Big(u_k - \Pi_{[a,b]}\Big(-\frac{1}{\nu}p_k\Big)\Big) < -\Big(u_k + \frac{1}{\nu}p_k\Big).$$

This implies directly

$$\Pi_{[a,b]}\Big(-\frac{1}{\nu}p_k\Big) < -\frac{1}{\nu}p_k, \tag{3.21}$$

and the assertion is proved for this case.

In the other case, we discuss as follows. Here, we know

$$M = -\Big(u_k - \Pi_{[a,b]}\Big(-\frac{1}{\nu}p_k\Big)\Big).$$

This means that the maximum M is also attained in the vertex E_k. Consequently, we have case (B) for the vertex E_k, and hence (3.15) holds with $i = k$ and a corresponding index m instead of k. For the case where at least one strong inequality occurs in (3.15), i.e.,

$$u_m - \Pi_{[a,b]}\Big(-\frac{1}{\nu}p_m\Big) < u_m + \frac{1}{\nu}p_m \qquad \text{with } m \in N(E_k),$$

we can proceed as in the first part of the proof. Hence, we have only to show that the equality case cannot occur for the index k, too: First, there exists at least one common neighboring vertex ($E_l \in N(E_i)$ and $E_l \in N(E_k)$). Next, we can apply Lemma 3.6 for the indices i and k. Therefore, we obtain the equations

$$\begin{aligned} u_i + \frac{1}{\nu}p_i &= -\Big(u_k + \frac{1}{\nu}p_k\Big), \\ u_i + \frac{1}{\nu}p_i &= -\Big(u_l + \frac{1}{\nu}p_l\Big), \\ u_k + \frac{1}{\nu}p_k &= -\Big(u_l + \frac{1}{\nu}p_l\Big), \end{aligned}$$

implying $u_i + \frac{1}{\nu}p_i = 0$. This is a contradiction to (3.9) and $M > 0$. Therefore, the equality case cannot occur for the index k, too. Consequently, the assertion is true. □

LEMMA 3.9. *Assume that*

$$Dh < b - a$$

is valid. Then, the estimate

$$M = \max_i \Big| u_i - \Pi_{[a,b]}\Big(-\frac{1}{\nu}p_i\Big)\Big| < Dh \tag{3.22}$$

holds true.

Proof. If $M = 0$, then (3.22) is automatically true. Therefore, we have only to discuss the case $M > 0$. Let us assume that the statement of Lemma 3.8 is true for an index i with case (A) and a corresponding index k. Case (B) can be discussed in the same way.

Inequality (3.21) implies directly

$$b = \Pi_{[a,b]}\Big(-\frac{1}{\nu}p_k\Big) < -\frac{1}{\nu}p_k. \tag{3.23}$$

From this and (3.5), we obtain

$$-\frac{1}{\nu}p_i > b - Dh.$$

By assumption, the value $b - Dh$ is greater than a. From (A),

$$u_i - \Pi_{[a,b]}\Big(-\frac{1}{\nu}p_i\Big) = M > 0,$$

and $u \le b$, we obtain

$$-\frac{1}{\nu}p_i \le b.$$

Consequently, we find

$$-\frac{1}{\nu}p_i = \Pi_{[a,b]}\Big(-\frac{1}{\nu}p_i\Big),$$

which implies

$$u_i + \frac{1}{\nu}p_i = u_i - \Pi_{[a,b]}\Big(-\frac{1}{\nu}p_i\Big) = M.$$

Using $u_i \le b$ and $\frac{1}{\nu}p_i < -(b - Dh)$, we find

$$u_i + \frac{1}{\nu}p_i < b - (b - Dh) = Dh.$$

Combining the last two inequalities, the assertion is proved. □

Let us briefly comment on the assumption $Dh < b - a$. First, this assumption is fulfilled for sufficiently small h. Second, in the cases $b - a \le Dh$ Theorem 2.3 holds trivially with $C = D$.

4. Proof of the main result. The proof of Theorem 2.3 is divided into two parts. In the next lemma we derive a corresponding estimate for the grid points. The estimate for arbitrary points is obtained in a second step.

LEMMA 4.1. *The estimate*

$$\max_i |u_h(E_i) - \bar{u}(E_i)| \le \Big(D + \frac{\kappa}{\nu}\Big)h$$

is valid.

Proof. For $b - a \le Dh$ the assertion is trivially fulfilled. Therefore, we have only to discuss the case $Dh < b - a$. From Lemma 3.9, we know

$$\max_i \Big|u_i - \Pi_{[a,b]}\Big(-\frac{1}{\nu}p_i\Big)\Big| \le Dh$$

or in other notation

$$\max_i \Big|u_h(E_i) - \Pi_{[a,b]}\Big(-\frac{1}{\nu}p_h(E_i)\Big)\Big| \le Dh.$$

From (3.3),

$$\|\bar{p} - p_h\|_{L^\infty(\Omega)} \le \kappa h,$$

and the Lipschitz continuity of the projection operator we deduce

$$\Big\|\Pi_{[a,b]}\Big(-\frac{1}{\nu}\bar{p}(E_i)\Big) - \Pi_{[a,b]}\Big(-\frac{1}{\nu}p_h(E_i)\Big)\Big\|_{L^\infty(\Omega)} \le \frac{\kappa}{\nu}h.$$

Using

$$\bar{u}(E_i) = \Pi_{[a,b]}\Big(-\frac{1}{\nu}\bar{p}(E_i)\Big)$$

and the triangle inequality, we end up with

$$\max_i |u_h(E_i) - \bar{u}(E_i)| \le \Big(D + \frac{\kappa}{\nu}\Big)h. \qquad \square$$

Now, we are able to prove Theorem 2.3.

Proof. A nongrid point $x \in T_i$ can be expressed by a convex linear combination of the vertices E_j of the corresponding triangle

$$x = \sum_{E_j \in T_i} \lambda_j E_j, \quad \sum_{E_j \in T_i} \lambda_j = 1.$$

Since u_h is linear on T_i, we get

$$\begin{aligned}
|u_h(x) - \bar{u}(x)| &= \Big|\sum_{E_j \in T_i} \lambda_j u_h(E_j) - \bar{u}(x)\Big| \\
&\le \sum_{E_j \in T_i} \lambda_j |u_h(E_j) - \bar{u}(E_j)| + \sum_{E_j \in T_i} \lambda_j |\bar{u}(x) - \bar{u}(E_j)| \\
&\le \Big(D + \frac{\kappa}{\nu}\Big)h + \sum_{E_j \in T_i} \lambda_j |\bar{u}(x) - \bar{u}(E_j)| \\
&\le \Big(D + \frac{\kappa}{\nu}\Big)h + \frac{L}{\nu}h.
\end{aligned}$$

In the final inequality we used the Lipschitz continuity of $\bar{u}$. Summarizing all results, we obtain

$$\|\bar{u} - u_h\|_{L^\infty(\Omega_h)} \le \Big(D + \frac{\kappa + L}{\nu}\Big)h.$$

Therefore, the assertion is true for every point $x \in T_i$ with

$$C = D + \frac{\kappa + L}{\nu}.$$

It remains to investigate the part $\Omega \setminus \Omega_h$. By definition, we have $u_h = \Pi_{[a,b]}(0)$ on this part. From (2.2), we obtain easily $\bar{u} = \Pi_{[a,b]}(0)$ on Γ. Let $x \in \Omega \setminus \Omega_h$ be an arbitrary point. From [14], we know that

$$\min_{x_\Gamma \in \Gamma} |x - x_\Gamma| \le c_\Gamma h^2$$

holds with a certain constant $c_\Gamma > 0$ independent of h. Therefore, we find for $x \in \Omega \setminus \Omega_h$

$$|u_h(x) - \bar{u}(x)| = |\Pi_{[a,b]}(0) - \bar{u}(x)| = |\bar{u}(x_\Gamma) - \bar{u}(x)| \le \frac{c_\Gamma L}{\nu} h^2. \qquad \square$$

5. Numerical example. We have tested the convergence theory by the following example:

$$\begin{aligned} -\Delta y &= u && \text{in } \Omega, \\ y &= 0 && \text{on } \Gamma \end{aligned} \tag{5.1}$$

with $\Omega = (0,1)\times(0,1)$. One can easily verify that this problem fulfills the assumptions mentioned in the beginning of section 2 except the boundary regularity. Although Γ is not of class $C^{1,1}$, the $W^{2,p}$-regularity of $\bar{p}$ (see Lemma 2.1) is obtained by a result of Grisvard [9] for convex polygonal domains.

In [13], we derived an exact solution to (5.1), which is also used here. For the reader's convenience, we recall this example.

The optimal state is defined by

$$\bar{y} = y_a - y_g$$

with an analytical part $y_a = \sin(\pi x_1)\sin(\pi x_2)$ and a less smooth function y_g. The function y_g represents the solution of

$$\begin{aligned} -\Delta y_g &= g && \text{in } \Omega, \\ y_g &= 0 && \text{on } \Gamma. \end{aligned}$$

Here, g is given by

$$g(x_1,x_2) = \begin{cases} \hat{u}(x_1,x_2) - a & \text{if } \hat{u}(x_1,x_2) < a, \\ 0 & \text{if } \hat{u}(x_1,x_2) \in [a,b], \\ \hat{u}(x_1,x_2) - b & \text{if } \hat{u}(x_1,x_2) > b \end{cases}$$

with $\hat{u}(x_1,x_2) = 2\pi^2 \sin(\pi x_1)\sin(\pi x_2)$. Due to the state equation (5.1), we obtain for the exact optimal control $\bar{u}$

$$\bar{u}(x_1,x_2) = \begin{cases} a & \text{if } \hat{u}(x_1,x_2) < a, \\ \hat{u}(x_1,x_2) & \text{if } \hat{u}(x_1,x_2) \in [a,b], \\ b & \text{if } \hat{u}(x_1,x_2) > b. \end{cases}$$

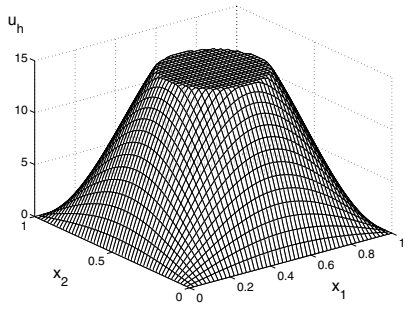


FIG. 5.1. *Optimal control* u_h.

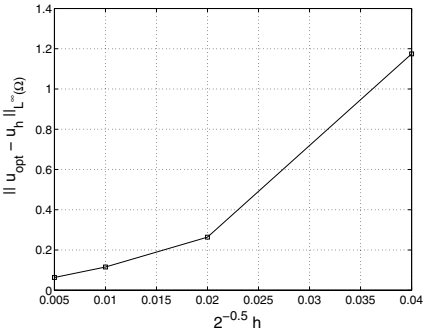


FIG. 5.2. $\|\bar{u} - u_h\|_{L^\infty(\Omega)}$.

TABLE 5.1

$h/\sqrt{2}$	0.04	0.02	0.01	0.005
$\|\bar{u} - u_h\|_{L^\infty(\Omega)}$	1.17450	0.26396	0.11536	0.06328

For the optimal adjoint state $\bar{p}$, we find

$$\bar{p}(x_1, x_2) = -2\,\pi^2\,\nu\,\sin(\pi x_1)\,\sin(\pi x_2).$$

To fulfill the necessary and sufficient first-order optimality conditions, the desired state y_d is defined by

$$y_d(x_1, x_2) = \bar{y} + \Delta\bar{p} = y_a - y_g + 4\,\pi^4\,\nu\,\sin(\pi x_1)\,\sin(\pi x_2).$$

The optimization problem was solved numerically by a primal-dual active set strategy; see [2], [11]. As mentioned in section 2, the state equation and the adjoint equation were discretized with linear finite elements. Here, uniform meshes were used. The resulting linear system of equations was solved with the conjugate gradient method.

To approximate the L^∞-norm $\|\bar{u} - u_h\|_{L^\infty(\Omega)}$, we evaluated $|\bar{u}(x) - u_h(x)|$ in the grid points, in the barycenters of the elements, and in the midpoints of the edges of the triangulation.

In a first test we chose $a = -15$ and $b = 15$. Here, we have $0 \in (a, b)$. Consequently the control is inactive near the boundary Γ.

Figure 5.1 shows the numerically calculated optimal control u_h for the mesh size $h/\sqrt{2} = 0.02$.

Figure 5.2 and Table 5.1 illustrate the convergence behavior for the first test. As one can see, the theoretical predictions are fulfilled, and one obtains linear approximation order for $\|\bar{u} - u_h\|_{L^\infty(\Omega)}$ (except on the coarsest grid).

In the second test we chose $a = 3$ and $b = 15$. Consequently, $0 \notin [a, b]$, i.e., the control is active near the boundary Γ.

Figure 5.3 shows again the numerically calculated optimal control u_h, for the mesh size $h/\sqrt{2} = 0.02$. Figure 5.4 and Table 5.2 illustrate the convergence behavior for the second test. The convergence behavior is similar to the first test and one again obtains linear convergence for $\|\bar{u} - u_h\|_{L^\infty(\Omega)}$.

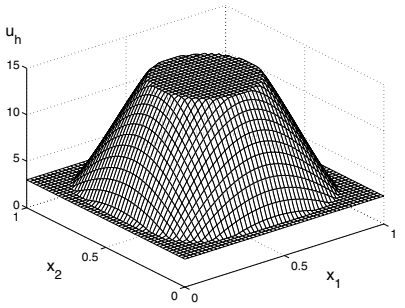


FIG. 5.3. *Optimal control* u_h.

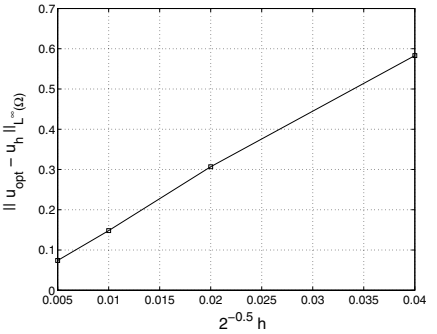


FIG. 5.4. $\|\bar{u} - u_h\|_{L^\infty(\Omega)}$.

TABLE 5.2

$h/\sqrt{2}$	0.04	0.02	0.01	0.005
$\|\bar{u} - u_h\|_{L^\infty(\Omega)}$	0.58292	0.30681	0.14813	0.07390

REFERENCES

[1] N. ARADA, E. CASAS, AND F. TRÖLTZSCH, *Error estimates for a semilinear elliptic optimal control problem*, Comput. Optim. Approx., 23 (2002), pp. 201–229.

[2] M. BERGOUNIOUX, K. ITO, AND K. KUNISCH, *Primal-dual strategy for constrained optimal control problems*, SIAM J. Control Optim., 37 (1999), pp. 1176–1194.

[3] D. BRAESS, *Finite Elemente*, Springer-Verlag, Berlin, 1992.

[4] E. CASAS, *Using piecewise linear functions in the numerical approximation of semilinear elliptic control problems*, submitted.

[5] E. CASAS, M. MATEOS, AND F. TRÖLTZSCH, *Error estimates for the numerical approximation of boundary semilinear elliptic control problems*, Comput. Optim. Appl., 31 (2005), pp. 193–219.

[6] E. CASAS AND F. TRÖLTZSCH, *Error estimates for linear-quadratic elliptic control problems*, in Analysis and Optimization of Differential Systems, V. Barbu, I. Lasiecka, D. Tiba, and C. Varsan, eds., Kluwer Academic, Boston, 2003, pp. 89–100.

[7] R. FALK, *Approximation of a class of optimal control problems with order of convergence estimates*, J. Math. Anal. Appl., 44 (1973), pp. 28–47.

[8] T. GEVECI, *On the approximation of the solution of an optimal control problem governed by an elliptic equation*, RAIRO Anal. Numer., 13 (1979), pp. 313–328.

[9] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[10] M. HINZE, *A variational discretization concept in control constrained optimization: The linear-quadratic case*, Comput. Optim. Appl., 30 (2005), pp. 45–61.

[11] K. KUNISCH AND A. RÖSCH, *Primal-dual active set strategy for a general class of constrained optimal control problems*, SIAM J. Optim., 13 (2002), pp. 321–334.

[12] K. MALANOWSKI, *Convergence of approximations vs. regularity of solutions for convex, control-constrained optimal control problems*, Appl. Math. Optim., 8 (1981), pp. 69–95.

[13] C. MEYER AND A. RÖSCH, *Superconvergence properties of optimal control problems*, SIAM J. Control Optim., 43 (2004), pp. 970–985.

[14] P. RAVIART AND J. THOMAS, *Introduction à l' Analyse Numérique des Équations aux Dérivées Partielles*, Masson, Paris, 1992.

[15] A. RÖSCH, *Error estimates for linear-quadratic control problems with control constraints*, Optim. Methods Softw., in print.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 5, pp. 1650–1676

OPTIMALITY OF AN (s, S) POLICY WITH COMPOUND POISSON AND DIFFUSION DEMANDS: A QUASI-VARIATIONAL INEQUALITIES APPROACH*

ALAIN BENSOUSSAN†, R. H. LIU‡, AND SURESH P. SETHI†

Abstract. We prove that an (s, S) policy is optimal in a continuous-review stochastic inventory model with a fixed ordering cost when the demand is a mixture of (i) a diffusion process and a compound Poisson process with exponentially distributed jump sizes, and (ii) a constant demand and a compound Poisson process. The proof uses the theory of impulse control. The Bellman equation of dynamic programming for such a problem reduces to a set of quasi-variational inequalities (QVI). An analytical study of the QVI leads to showing the existence of an optimal policy as well as the optimality of an (s, S) policy. Finally, the combination of a diffusion and a general compound Poisson demand is not completely solved. We explain the difficulties and what remains open. We also provide a numerical example for the general case.

Key words. stochastic inventory model, economic order quantity model, impulse control, quasi-variational inequalities, (s, S) policy, diffusion process, compound Poisson process

AMS subject classifications. 90B05, 93E20, 49N25, 49K15, 49K22, 49K45, 49L20

DOI. 10.1137/S0363012904443737

1. Introduction.

It is well known that the optimal ordering policies are of (s, S) type for a broad class of stochastic inventory models involving a fixed ordering cost, where s denotes the ordering point and $S \geq s$ is the order-up-to level. In other words, when the inventory level at time t is s or below, an order is issued at time t to bring the inventory level up to the level S. Ample literature exists. We review it briefly, while referring the readers to Presman and Sethi (2004) for a detailed review and a list of references on this subject.

One class of models that has been extensively studied in the literature is the continuous-review stochastic inventory models with a compound Poisson demand and fixed ordering cost. In this scenario, the demands arrive at random epochs governed by a Poisson process. At any epoch, the demand size is an independently and identically distributed (i.i.d.) random variable. Studies of (s, S) policies for this model are conducted by Richards (1975), Thompstone and Silver (1975), Archibald and Silver (1978), Feldman (1978), and Federgruen and Schechner (1983). Additionally, Tijms (1972), Sahin (1979, 1983), and Federgruen and Schechner (1983) consider the compound renewal demands in which the jump epochs are assumed to follow a renewal process. Zipkin (1986) uses a compound counting process to model demands. Some of these papers show that an (s, S) policy is optimal and others study the behavior of (s, S) policies without showing their optimality.

Another studied class of models is the so-called world-dependent demands or Markovian demands; see Beyer, Cheng, and Sethi (2006). Song and Zipkin (1993) consider a continuous-review model with world-dependent Poisson demands. They

*Received by the editors May 10, 2004; accepted for publication (in revised form) April 17, 2005; published electronically November 22, 2005.

http://www.siam.org/journals/sicon/44-5/44373.html

†School of Management, The University of Texas at Dallas, 2601 N. Floyd Rd., Richardson, TX 75080 (alain.bensoussan@utdallas.edu, sethi@utdallas.edu).

‡Department of Mathematics, University of Dayton, 300 College Park, Dayton, OH 45469 (ruihua.liu@notes.udayton.edu).

use a continuous-time Markov chain to model the "state of the world." To solve the problem, they invoke the standard uniformization procedure (Keilson (1979) and Van Dijk (1990)) to convert the continuous-time problem to a discrete-time problem and then use the discrete-time dynamic programming to obtain a state-dependent (s, S) policy. A verification theorem, usually required to prove optimality in such problems, is given in Beyer, Sethi, and Taksar (1998) for problems with Markovian demands and a fairly general surplus cost function. We should mention that in a continuous-review model with Poisson or compound Poisson demands, since demands arrive only at discrete epochs, the optimal ordering decisions can be restricted to these epochs without loss of optimality. This is the key observation that makes the uniformization procedure applicable.

In contrast with Poisson process models, the classical economic order quantity (EOQ) problem is a "purely continuous" inventory model in the sense that the demands arrive continuously at a fixed constant rate. A natural and interesting model to study is the one that combines the constant continuous demand and the compound Poisson demand. However, this work had not been done until a very recent paper by Presman and Sethi (2004). A major reason, as mentioned in their paper, is that the presence of a constant demand means that the optimal ordering decisions may not be restricted only at the jump epochs of the compound Poisson demands. As a result, the standard uniformization procedure does not work in this case; see Presman and Sethi (2004) for other difficulties arising in the analysis of such a model.

Presman and Sethi (2004) consider the demand process to be the sum of a constant demand rate and a compound Poisson process. They develop a new approach to prove the optimality of an (s, S) policy in the presence of a fixed ordering cost. This is also a unified approach in the sense that it deals with both the long-run average cost and the discounted-cost criteria. Their approach can be outlined as follows: it starts with an (s, S) policy and then finds the corresponding discounted-cost formula and a closely related modified cost function. By minimizing the modified cost function, a candidate optimal (s, S) policy is obtained. Using this candidate and the formula for the discounted cost, a potential function is constructed, which is shown to satisfy the dynamic programming equation associated with the problem.

The next step in the generalization of the model is to add a Wiener process to the demand process and to prove the optimality of an (s, S) policy. This is the subject of this paper. We study an inventory model with a fixed ordering cost and a general demand process that consists of a compound Poisson demand and a diffusion process. Here the drift of the diffusion process represents the constant part of the demand. This is a problem of impulse control, the theory of which has been developed by Bensoussan and Lions (1984). The general idea is briefly as follows. Under the framework of impulse control, the Bellman equation of dynamic programming for the inventory problem under consideration reduces to a quasi-variational inequality (QVI). Using the QVI approach, i.e., by solving the QVI and analyzing the properties of its solution, it is possible to prove the existence of an optimal impulse control, and hence of an optimal inventory policy.

Bensoussan and Tapiero (1982) formulate a stochastic demand model consisting of a diffusion process and a pure Poisson process. They consider a finite horizon problem and derive the associated QVI. While they give conditions for an (s, S) policy to be optimal, they do not solve the problem.

Constantinides and Richard (1978) treat an infinite horizon discounted-cost inventory problem with a diffusion demand. They prove the existence and the optimality of an (s, S) policy. Sulem (1986) provides an explicit solution for this problem as well

as for its average-cost version. Some related models are Bather (1966), Whitt (1973), Puterman (1975), and Beyer (1994). But they do not go as far, for our purposes, as Sulem (1986) does.

A model presented by Browne and Zipkin (1991) also involves a diffusion term in the demand but treats it in a different way. While our model directly includes the diffusion as a part of the demand, they assume that the demand process is state-dependent and the underlying "state of the world" is modeled either by a continuous-time Markov chain or by a diffusion process. They consider an (s, S) policy, which typically is not optimal for their model.

Our primary goal is to prove that an (s, S) policy solves the QVI for the general stochastic demand model under consideration. We achieve it in two important cases: when the demand is (i) a mixture of a diffusion process and a compound Poisson process with exponentially distributed jump sizes, and (ii) a mixture of a constant demand and a compound Poisson process. However, the combination of a diffusion process and a general compound Poisson process is not completely solved. We explain the difficulties that arise and what remains open.

The specific steps taken in our approach are as follows. (i) We solve the QVI (4.2) by postulating that the first inequality is an equation (4.4) to the right of a number s (to be determined), and the second one is an equation (4.5) to the left of s. The value of s is fixed by imposing that the solution of the QVI is C^1 on the real line. (ii) We analytically solve (4.4) and (4.5) together with the boundary conditions (4.6) and (4.7) and the smoothness condition (4.8). This is done in three steps: (a) We solve (4.4) with condition (4.6) and obtain a C^1 solution G_s for any given s. (b) Then we show that a minimum $S(s)$ of the function G_s exists and that condition (4.7) is satisfied. (c) We use (4.5) to determine a unique optimal s. (iii) We check that the function constructed in this way satisfies the original QVI (4.2) of the inventory problem. This method leads to a unique function, which is C^1, and a unique pair (s, S). It is also the value function over the class of all (s, S) policies.

We emphasize that our approach is different from the one used in Presman and Sethi (2004) even though both involve QVIs. Presman and Sethi (2004) start with an arbitrary (s, S) policy and consider the cost function obtained using that policy. They derive a specific function of s and S, which is closely related to the cost function. They minimize this function to obtain an optimal pair (s, S) within the class of all (s, S) policies. They prove that the cost function associated with such an optimal (s, S) pair satisfies the QVI for the inventory problem and, therefore, is the minimum cost with respect to all feasible ordering decisions. We, on the other hand, begin with the QVI associated with the inventory control problem. We solve the QVI, as indicated above, to obtain the value function. In summary, our method is analytical in nature, whereas Presman and Sethi (2004) use a probabilistic approach. Moreover, we extend the result on the optimality of an (s, S) policy to include a diffusion demand.

Specifically, our paper makes the following main contributions.

1. We formulate a fairly general stochastic inventory model and consider the appropriate QVI.

2. Our work generalizes the stochastic demand considered in Presman and Sethi (2004) by allowing a Brownian motion term in the demand process. However, we must note that our generalization comes at the expense of requiring the compound Poisson process to have exponentially distributed jump sizes.

3. We solve the QVI analytically to obtain a closed-form solution.

4. We prove that an (s, S) policy solves the QVI for two important special cases,

mentioned earlier and in the abstract.

5. We reveal an interesting relation between the optimal values of S and s; see (5.5) and Remark 5.5.

6. We explain the difficulties with the general model having a combined demand process involving both diffusion and compound Poisson demands.

The rest of the paper is organized as follows. In section 2, we provide a precise formulation of the stochastic inventory problem under consideration. We also present the QVI that must be satisfied by the value function of the problem. In section 3, we briefly review the general theory of impulse control and QVI. In section 4, we derive a set of equations for a given s and solve them to obtain a closed-form solution. Section 5 presents some properties of the solution that are important to the optimal (s, S) policy. The existence of the corresponding S for each given s is proved. In section 6, we study the optimal (s, S) policy using the solution and an additional condition that must be satisfied by the solution if an (s, S) policy is to be optimal, including its existence. In section 7, we deal with a special case in which the random jump size in the compound Poisson process is assumed to be exponentially distributed. Section 8 treats the nondiffusion case using a different approach for the optimality proof. In section 9, we provide some explanations for the difficulties with the general case. A numerical example is presented. Section 10 concludes the paper.

2. Problem formulation. In this section, we provide a precise formulation of the stochastic inventory problem under consideration. We present the QVI associated with the problem that is satisfied by the value function.

We first define the demand process. Let $(\Omega, \mathcal{F}, \mathcal{P})$ be the underlying probability space. The cumulative demand $y(t)$ in the interval $[0, t]$ is a stochastic process given by

$$y(t) = Dt + \sigma W(t) + N(t), \tag{2.1}$$

where $D \geq 0$ is the constant demand per unit time, $W(t)$ denotes the standard Brownian motion with $W(0) = 0$, and $\sigma \geq 0$. The process $Dt + \sigma W(t)$ is referred to as a diffusion process with drift D and volatility σ. $N(t)$ is a compound Poisson process defined next. Let $n(t)$ be a right-continuous Poisson process with $n(0) = 0$ and the intensity $\lambda \geq 0$. Let $\xi_i \geq 0, i = 1, 2, \ldots,$ be a sequence of i.i.d. nonnegative random variables having distribution density $\mu(\cdot)$. Then,

$$N(t) = \sum_{i \leq n(t)} \xi_i, \quad t \geq 0.$$

We note that the process $n(t)$ produces a sequence of jump times and that ξ_i denotes the size (random) of the demand at the ith jump. We shall sometimes simply use ξ for the jump size for convenience in exposition.

Assumption 2.1. The Wiener process $W(t)$, the Poisson process $n(t)$, and the jump sequence $\{\xi_i\}$ are all independent.

Next we define the class of admissible ordering policies. Let $\mathcal{F}_t, t > 0$, be the sigma algebra generated by $\{N(s), W(s), 0 < s \leq t\}$ and $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Let $\theta_i \geq 0, i = 1, 2, \ldots,$ be a strictly increasing sequence of stopping times with respect to the filtration $\{\mathcal{F}_{t+0}\}$, and let $u_i > 0$ be a positive random variable adapted to $\mathcal{F}_{\theta_i}$. An admissible ordering policy is defined by

$$U = (\theta_1, u_1, \theta_2, u_2, \ldots).$$

Here θ_i denotes the time of the ith order, and u_i denotes the amount ordered. We use $\mathcal{U}$ to denote the set of all admissible ordering policies.

The cumulative amount of orders from time 0 up to but not including t is then given by

$$M(t) = \sum_{\{i:\theta_i<t\}} u_i.$$

We assume the orders are delivered instantaneously. Then the surplus level $x^U(t)$ at time t under a policy $U \in \mathcal{U}$ is given by the equation

$$x^U(t) = x - Dt - \sigma W(t) - N(t) + M(t), \tag{2.2}$$

where $x^U(0) = x$ (a constant) is the initial surplus level at $t = 0$. Note the surplus $x^U(t)$ when positive means inventory and when negative means backlog.

Next we introduce the surplus and ordering costs. Let the surplus cost $f(x)$ be a nonnegative and piecewise continuously differentiable function with $f(0) = 0$. It gives the cost of holding inventory for $x > 0$ and the backlog cost for $x < 0$. As in Beyer, Sethi, and Taksar (1998), we make the following assumption on f.

Assumption 2.2. $f(x)$ has a polynomial growth rate.

Some additional properties of $f(x)$ will be specified later.

The cost $c(u)$ of ordering an amount u is given by

$$c(u) = \begin{cases} K + cu, & u > 0, \\ 0, & u = 0, \end{cases} \tag{2.3}$$

where $K > 0$ is the fixed set up cost of ordering, and c denotes the unit cost of each item ordered.

We consider in this paper a discounted cost objective function. Let $\rho > 0$ be the specified discount rate. For a given initial inventory level x and an ordering policy $U \in \mathcal{U}$, we define the discounted cost as

$$F(x, U) = E\left\{ \int_0^\infty f(x^U(t)) e^{-\rho t} dt + \sum_{i=1}^{\infty} c(u_i) e^{-\rho \theta_i} \right\}. \tag{2.4}$$

Define the value function associated with (2.4) as

$$F(x) = \inf_{U \in \mathcal{U}} F(x, U). \tag{2.5}$$

Our goal is to find a policy $U^* \in \mathcal{U}$ such that $F(x, U^*) = F(x)$.

Now we recall the notion of an (s, S) policy of ordering. For $-\infty < s < S < +\infty$, let $U^{s,S}$ denote the (s, S) policy given by the following function:

$$U^{s,S}(x) = \begin{cases} 0 & \text{if } x > s, \\ S - x & \text{if } x \le s, \end{cases}$$

where s is called the ordering level, and S is called the order-up-to level. Clearly $U^{s,S} \in \mathcal{U}$. Let $\mathcal{U}^{s,S}$ denote the subset of $\mathcal{U}$ containing all (s, S) policies.

Next we present the QVI associated with (2.2)–(2.5). To this end, we introduce the following operators. For function ϕ, let

$$\begin{cases} (A\phi)(x) = -\dfrac{\sigma^2}{2}\phi''(x) + D\phi'(x), \\ (B\phi)(x) = \lambda \displaystyle\int_0^\infty (\phi(x-\xi) - \phi(x))\mu(\xi)\, d\xi, \\ (M\phi)(x) = K + \inf\limits_{u\geq 0}(cu + \phi(x+u)). \end{cases}$$

The Bellman equation for the value function reduces to a set of inequalities:

$$\begin{cases} AV - BV + \rho V \leq f, \\ V \leq MV, \\ (AV - BV + \rho V - f)(V - MV) = 0. \end{cases} \tag{2.6}$$

These relations make sense a.e. x for any function $V(x)$ that is continuously differentiable. Our purpose therefore is to look for a C^1 solution of (2.6) with polynomial growth.

We make the following assumption on the random variable ξ.

Assumption 2.3. There is a $\bar{\pi} < 0$ such that

$$\int_0^\infty e^{-\pi\xi}\mu(\xi)d\xi < \infty \quad \text{for all } \pi > \bar{\pi}.$$

It should be noted that this assumption is stronger than that of a finite mean for the jump size ξ.

3. A brief review of the impulse control theory. The QVI (2.6) is a special case of the general QVI theory studied in Bensoussan and Lions (1984). They present results on the existence of solution and the probabilistic interpretation in the context of inventory control. In this section, we provide a brief review of related formulations and results from the general theory of QVIs for impulse control.

Let $\beta_\alpha(x) = \exp(-\alpha(1+x^2)^{\frac{1}{2}})$, $\alpha > 0$. Consider the Hilbert spaces L^2_α and H^1_α defined as follows:

$$L^2_\alpha = \left\{ v \,\middle|\, \int_{-\infty}^{+\infty} v^2\beta_\alpha^2\, dx < \infty \right\}, \quad H^1_\alpha = \{v | v \in L^2_\alpha, v' \in L^2_\alpha\}$$

with $(v,w)_\alpha = \int_{-\infty}^{+\infty} vw\beta_\alpha^2\, dx$, $v, w \in L^2_\alpha$. Define the bilinear form on H^1_α by

$$\begin{aligned} a_\alpha(v,w) = {} & \frac{\sigma^2}{2}\int_{-\infty}^{+\infty} v'w'\beta_\alpha^2\, dx + \int_{-\infty}^{+\infty} v'w\left(D - \frac{\alpha\sigma^2 x}{(1+x^2)^{\frac{1}{2}}}\right)\beta_\alpha^2\, dx \\ & + \rho\int_{-\infty}^{+\infty} vw\beta_\alpha^2\, dx - \int_{-\infty}^{+\infty}(Bv)w\beta_\alpha^2\, dx, \quad v, w \in H^1_\alpha. \end{aligned}$$

We first note some properties of Bv. These are

$$|Bv|_\alpha \leq \lambda(1 + C_\alpha)|v|_\alpha, \quad (Bv, v)_\alpha \leq \lambda(C_\alpha - 1)|v|^2_\alpha$$

with

$$C_\alpha = \left(\int_0^\infty \exp(2\alpha\xi)\mu(\xi)\, d\xi\right)^{\frac{1}{2}}, \quad 2\alpha < -\bar{\pi}.$$

These follow from

$$\int_{-\infty}^{+\infty} \beta_\alpha^2(x) \left(\int_0^{+\infty} v(x-\xi)\mu(\xi)\, d\xi \right)^2 dx \le \int_{-\infty}^{+\infty} \beta_\alpha^2(x) \left(\int_0^{+\infty} v^2(x-\xi)\mu(\xi)\, d\xi \right) dx$$

and

$$\beta_\alpha^2(x) \le \beta_\alpha^2(x-\xi)\exp(2\alpha\xi).$$

We also have

$$(Bv, v^+)_\alpha \le (Bv^+, v^+)_\alpha \text{ and } (Bv, v^-)_\alpha \ge -(Bv^-, v^-)_\alpha,$$

where $v^+ = \max(v, 0)$ and $v^- = \max(-v, 0)$. Set $\rho_\alpha = \alpha D + \alpha^2\sigma^2 + \lambda(C_\alpha - 1)$. Then we have

$$a_\alpha(v, v) \ge \frac{\sigma^2}{2} \int_{-\infty}^{+\infty} v'^2 \beta_\alpha^2\, dx + (\rho - \rho_\alpha) \int_{-\infty}^{+\infty} v^2 \beta_\alpha^2\, dx.$$

Note that for a sufficiently small α, we have $\rho_\alpha < \rho$.

We define the following variational inequality using the above notions:

$$\begin{cases} a_\alpha(v, w - v) \ge (f, w - v)_\alpha \\ \text{for any } w \in H_\alpha^1 \text{ such that } w \le Mv,\ v \le Mv, \text{ and } v \in H_\alpha^1. \end{cases} \tag{3.1}$$

Now let v^0 be defined as

$$v^0(x) = V_0 \exp(\alpha(1 + x^2)^{\frac{1}{2}}) = V_0 \beta_{-\alpha}(x),$$

where $V_0 > 0$ is a constant. We wish to find a V_0 such that

$$a_\alpha(v^0, w) \ge (f, w)_\alpha \quad \text{for any } w \ge 0. \tag{3.2}$$

Noting the fact that

$$|Bv^0(x)| \le \lambda v^0(x)\alpha e^\alpha \int_0^\infty \xi e^{\alpha\xi} \mu(\xi)\, d\xi,$$

we have

$$a_\alpha(v^0, w) \ge (\rho - \rho'_\alpha)(v^0, w)_\alpha,$$

where

$$\rho'_\alpha = D\alpha + \frac{1}{2}\sigma^2\alpha^2 + \frac{1}{2}\sigma^2\alpha + \lambda\alpha e^\alpha \int_0^\infty \xi e^{\alpha\xi} \mu(\xi)\, d\xi.$$

Therefore, it is sufficient for (3.2) to achieve

$$(\rho - \rho'_\alpha) V_0 \exp(\alpha(1 + x^2)^{\frac{1}{2}}) \ge f(x), \quad x \in (-\infty, +\infty).$$

Since f has a polynomial growth rate by Assumption 2.2, i.e., there exist positive constants f_0 and γ such that

$$f(x) \le f_0(1 + |x|^\gamma), \quad x \in (-\infty, +\infty),$$

and since

$$(\rho - \rho'_\alpha)V_0 \exp(\alpha(1+x^2)^{\frac{1}{2}}) \geq (\rho - \rho'_\alpha)V_0 \exp(\alpha|x|),$$

it is sufficient to pick a V_0 such that

$$\frac{V_0(\rho - \rho'_\alpha)}{f_0} \geq \max_{x>0}(1+x^\gamma)\exp(-\alpha x).$$

From the general result (Theorem 4.1, Chapter 4 in Bensoussan and Lions (1984)), it follows that the set of solutions of (3.1) satisfying $0 \leq v \leq v^0$ is nonempty, and it has a minimum solution $\underline{v}$ and a maximum solution $\bar{v}$. Moreover, one has the following probabilistic interpretation of the minimum solution $\underline{v}$ in the context of impulse control (Theorem 3.4, Chapter 6 in Bensoussan and Lions (1984)):

$$\underline{v}(x) = F(x) = \min_{U \in \mathcal{U}} F(x, U),$$

i.e., there exists an admissible control U such that the value function $F(x)$ is attained using this control, and $F(x)$ is indeed equal to the minimum solution $\underline{v}(x)$.

The C^1 solution of (2.6) we are aiming at will be a solution of (3.1). To check that it is the value function and the minimum solution of (3.1), one relies on the fact that such a solution satisfies

$$V(x) \leq F(x, U) \quad \text{for all } U \in \mathcal{U}.$$

These two statements are standard in the literature. For a proof of these in the pure diffusion case see section 1.4 in Chapter 6 of Bensoussan and Lions (1984).

For our purpose, it is sufficient to find a solution $V(x)$ of the QVI (2.6), which leads to an (s, S) policy. Then we can attach to this (s, S) policy an impulse control such that

$$V(x) = F(x, U^{s,S}).$$

This implies that the solution $V(x)$ obtained in this way is unique and is equal to $F(x)$.

Thus, the problem boils down to finding a solution $V(x)$ of (2.6), which is C^1 and of polynomial growth, and which corresponds to an (s, S) policy. Finding such a solution is, therefore, the main objective in the rest of the paper.

4. An (s, S) policy and QVI. Let F be a solution of (3.1), which is C^1 and of polynomial growth. Let $G(x) = F(x) + cx$ and

$$g(x) = f(x) + c\rho x. \tag{4.1}$$

Then G satisfies

$$\begin{cases} AG - BG + \rho G \leq g + cD + c\lambda\bar{\xi}, \\ G(x) \leq K + \inf_{u \geq 0} G(x+u), \\ (AG - BG + \rho G - g - cD - c\lambda\bar{\xi})(G(x) - K - \inf_{u \geq 0} G(x+u)) = 0, \end{cases} \tag{4.2}$$

where $\bar{\xi} = \int_0^\infty \xi\mu(\xi)\,d\xi$ is the mean of ξ.

Remark 4.1. If $\sigma = 0$ and $\lambda = 0$, then our model reduces to the classical EOQ model. In this case, the QVI (4.2) takes the special form

$$\begin{cases} DG' + \rho G \le g + cD, \\ G(x) \le K + \inf\limits_{u \ge 0} G(x+u), \\ (DG' + \rho G - g - cD)(G(x) - K - \inf\limits_{u \ge 0} G(x+u)) = 0. \end{cases} \tag{4.3}$$

Beyer and Sethi (1998) give a rigorous proof of the optimality of the well-known EOQ formula by using the QVI approach. It is easy to check that (4.3) is equivalent to the QVI (1) in their paper.

We make the following assumptions on the function g.

Assumption 4.2. There exists a number a such that g is increasing on (a, ∞) and decreasing and convex on $(-\infty, a)$. Furthermore, there exist a $c_0 > 0$ and an $x_1 \ge a$ such that $g'(x) \ge c_0$ for $x \ge x_1$.

Remark 4.3. Presman and Sethi (2004) do not require g to be convex on $(-\infty, a)$ in their study of the inventory model with demand consisting of only a constant term and a compound Poisson process.

Since we are interested in an (s, S) policy, we seek a solution of the QVI (4.2), which implies ordering up to S on the interval $(-\infty, s]$ and not ordering on the interval $(s, +\infty)$. This means that we are interested in finding an s such that G satisfies the following two equations:

$$AG - BG + \rho G = g + cD + c\lambda\bar{\xi} \quad \text{for } x \ge s, \tag{4.4}$$

$$G(x) = K + G(S) \quad \text{for } x \le s \tag{4.5}$$

with two boundary conditions

$$G'(s) = 0, \tag{4.6}$$

$$G'(S) = 0. \tag{4.7}$$

It follows from (4.4) to (4.7) that G is smooth, i.e.,

$$G \in C^1. \tag{4.8}$$

It will also have a polynomial growth rate. So our problem is reduced to finding a triple $\{s, S, G(x)\}$ which satisfies (4.4)–(4.7).

We will solve (4.4)–(4.7) in three steps. For s given, the first step finds a function $G = G_s$ which satisfies (4.4) and (4.6), which is constant on the left-hand side of s, and which is C^1. The function G_s will be obtained in Theorem 4.8 in this section. The second step is devoted to determining $S(s)$ satisfying (4.7); see Theorem 5.3 in section 5. The solution $\{s, S(s), G_s\}$ of (4.4), (4.6), and (4.7) will be unique by the analytic treatment pursued below. The third step will be to obtain s and to solve (4.5). This is accomplished in Theorem 6.1.

To proceed, it is convenient to introduce $H_s(x) = G_s'(x)$, which is the solution of

$$\begin{cases} -\dfrac{\sigma^2}{2} H_s''(x) + DH_s'(x) + (\rho + \lambda) H_s(x) - \lambda \displaystyle\int_0^{x-s} H_s(x-\xi)\mu(\xi)\, d\xi = g'(x), \\ H_s(s) = 0, \quad H_s \text{ has a polynomial growth, } x > s. \end{cases} \tag{4.9}$$

Note that $H_s(x) = 0$ for $x < s$. Set $Z_s(x) = H_s(x + s)$ and $g_s(x) = g(x + s)$, $x > 0$. It follows that

$$\begin{cases} -\dfrac{\sigma^2}{2} Z_s''(x) + DZ_s'(x) + (\rho + \lambda) Z_s(x) - \lambda \displaystyle\int_0^x Z_s(x - \xi)\mu(\xi)\, d\xi = g_s'(x), \\ Z_s(0) = 0, \quad Z_s \text{ has a polynomial growth, } x > 0. \end{cases} \tag{4.10}$$

Now we solve (4.10) by the transformation

$$Z_s(x) = \int_0^x \Gamma(x - \theta) Q_s(\theta)\, d\theta,$$

where $\Gamma(\theta)$ is a solution of

$$\begin{cases} -\dfrac{\sigma^2}{2} \Gamma''(\theta) + D\Gamma'(\theta) + (\rho + \lambda)\Gamma(\theta) - \lambda \displaystyle\int_0^\theta \Gamma(\theta - \xi)\mu(\xi)\, d\xi = 0, \\ \Gamma(0) = 1, \quad \Gamma(+\infty) = 0, \end{cases} \tag{4.11}$$

and $Q_s(x)$ is a solution of

$$-\frac{\sigma^2}{2} Q_s'(x) + \left(D - \frac{\sigma^2}{2} \Gamma'(0) \right) Q_s(x) = g_s'(x), \quad Q_s(+\infty) = 0. \tag{4.12}$$

Let us denote

$$\hat{\mu}(\pi) = \int_0^\infty e^{-\pi\xi} \mu(\xi)\, d\xi, \tag{4.13}$$

which by Assumption 2.3 is well defined for $\pi > \bar{\pi}$. We have the following results.

PROPOSITION 4.4. *There exists a unique solution $\Gamma(\theta) \geq 0$ of (4.11), whose Laplace transform $\hat{\Gamma}(\pi) = \int_0^\infty e^{-\pi\theta} \Gamma(\theta)\, d\theta$ is well defined for $\pi > \beta_1$, with $\bar{\pi} < \beta_1 < 0$, and it is given by*

$$\hat{\Gamma}(\pi) = \frac{\hat{\varphi}(\pi)}{\pi - \beta_1}, \tag{4.14}$$

where $\hat{\varphi}(\pi) > 0$ (to be defined next) for any $\pi > \bar{\pi}$.

Proof. By (4.11), the Laplace transform $\hat{\Gamma}(\pi)$ is given by

$$\hat{\Gamma}(\pi) = \frac{D - \frac{\sigma^2}{2}\pi - \frac{\sigma^2}{2}\Gamma'(0)}{\chi(\pi)} \tag{4.15}$$

with

$$\chi(\pi) = -\frac{\sigma^2}{2}\pi^2 + D\pi + \rho + \lambda - \lambda\hat{\mu}(\pi). \tag{4.16}$$

We have

$$\begin{aligned} &\chi(\bar{\pi}) = \chi(+\infty) = -\infty, \\ &\chi'(\pi) = -\sigma^2\pi + D + \lambda \int_0^\infty \xi e^{-\pi\xi} \mu(\xi)\, d\xi, \\ &\chi''(\pi) = -\sigma^2 - \lambda \int_0^\infty \xi^2 e^{-\pi\xi} \mu(\xi)\, d\xi < 0, \\ &\chi'(\bar{\pi}) = +\infty, \quad \chi'(+\infty) = -\infty, \quad \chi'(0) > 0. \end{aligned}$$

Hence, there exists a unique $\pi_0 > 0$ such that $\chi'(\pi_0) = 0$. Therefore, $\chi(\pi)$ has a unique maximum at π_0. Note that $\chi(0) = \rho > 0$. Then χ has two zeros β_1, β_2 with $\bar{\pi} < \beta_1 < 0 < \pi_0 < \beta_2$. We may then write (4.16) as

$$\chi(\pi) = -\frac{\frac{\sigma^2}{2}(\pi - \beta_1)(\pi - \beta_2)}{\hat{\varphi}(\pi)}, \quad \hat{\varphi}(\pi) > 0 \text{ for any } \pi > \bar{\pi}. \tag{4.17}$$

Then (4.15) becomes

$$\hat{\Gamma}(\pi) = -\frac{\left(D - \frac{\sigma^2}{2}\pi - \frac{\sigma^2}{2}\Gamma'(0)\right)\hat{\varphi}(\pi)}{\frac{\sigma^2}{2}(\pi - \beta_1)(\pi - \beta_2)}. \tag{4.18}$$

Since $\hat{\Gamma}(\pi)$ is well defined for any $\pi > 0$, we must get rid of $(\pi - \beta_2)$ in the denominator of (4.18). This implies that the numerator has a zero at β_2. As a result, we must have

$$\Gamma'(0) = \frac{2D}{\sigma^2} - \beta_2, \tag{4.19}$$

and accordingly, $\hat{\Gamma}(\pi)$ has the form (4.14). Additionally, since $\hat{\varphi}(0) = -\frac{\sigma^2\beta_1\beta_2}{2\rho} > 0$,

$$\Gamma(+\infty) = \lim_{\pi \to 0} \pi\hat{\Gamma}(\pi) = 0.$$

We now show that

$$\Gamma(\theta) \geq 0, \quad \theta \geq 0. \tag{4.20}$$

In fact, if (4.20) were not true, then Γ would have a negative minimum at $\theta_0 > 0$ such that $\Gamma(\theta_0) < 0$, $\Gamma'(\theta_0) = 0$, and $\Gamma''(\theta_0) > 0$. It follows that

$$\Gamma(\theta_0) - \int_0^{\theta_0} \Gamma(\theta_0 - \xi)\mu(\xi)\, d\xi \leq \Gamma(\theta_0)\left(1 - \int_0^{\theta_0} \mu(\xi)\, d\xi\right) < 0,$$

which is a contradiction with (4.11). This completes the proof. □

In addition to the proof, we have the following remark.

Remark 4.5. Since $\chi'(D/\sigma^2) > 0$, we know that $D/\sigma^2 < \pi_0 < \beta_2$. Hence, $\beta_2 > 2D/\sigma^2 - \beta_2$. Moreover,

$$\begin{aligned}\chi\left(\frac{2D}{\sigma^2} - \beta_2\right) &= -\frac{\sigma^2}{2}\beta_2^2 + D\beta_2 + \rho + \lambda - \lambda\hat{\mu}\left(\frac{2D}{\sigma^2} - \beta_2\right) \\ &= \lambda\hat{\mu}(\beta_2) - \lambda\hat{\mu}\left(\frac{2D}{\sigma^2} - \beta_2\right) < 0.\end{aligned}$$

Therefore, we have

$$\frac{2D}{\sigma^2} - \beta_2 < \beta_1 \text{ and } \Gamma'(0) < \beta_1.$$

PROPOSITION 4.6. *The solution* $\Gamma(\theta)$ *of* (4.11) *satisfies*

$$\exp(\beta_0\theta) \leq \Gamma(\theta) \leq \exp(\beta_1\theta), \quad \theta \geq 0, \tag{4.21}$$

where β_0 is the negative solution of the equation

$$-\frac{\sigma^2}{2}\beta_0^2 + D\beta_0 + \rho + \lambda = 0.$$

Proof. First we let $\Sigma(\theta) = \exp(\beta_1\theta)$. Then,

$$\begin{aligned}
&-\frac{\sigma^2}{2}\Sigma''(\theta) + D\Sigma'(\theta) + (\rho+\lambda)\Sigma(\theta) - \lambda\int_0^\theta \Sigma(\theta-\xi)\mu(\xi)\,d\xi \\
&= \exp(\beta_1\theta)\left(-\frac{\sigma^2}{2}\beta_1^2 + D\beta_1 + \rho + \lambda - \lambda\int_0^\theta \exp(-\beta_1\xi)\mu(\xi)\,d\xi\right) \\
&\geq \exp(\beta_1\theta)\left(-\frac{\sigma^2}{2}\beta_1^2 + D\beta_1 + \rho + \lambda - \lambda\int_0^\infty \exp(-\beta_1\xi)\mu(\xi)\,d\xi\right) = 0,
\end{aligned}$$

$\Sigma(0) = 1$, and $\Sigma(+\infty) = 0$. It follows that $\Gamma(\theta) \leq \exp(\beta_1\theta)$. Next, in terms of (4.11), we get $-\frac{\sigma^2}{2}\Gamma'' + D\Gamma' + (\rho+\lambda)\Gamma \geq 0$, which implies the first inequality of (4.21). This completes the proof. □

Using (4.19), the unique solution of (4.12) is given by

$$Q_s(x) = \frac{2}{\sigma^2}\int_x^\infty \exp(\beta_2(x-y))g_s'(y)dy.$$

Finally, we get the formula

$$H_s(x) = \int_s^x \Gamma(x-y)Q(y)dy, \tag{4.22}$$

where Q is given by

$$Q(x) = \frac{2}{\sigma^2}\int_x^\infty \exp(-\beta_2(y-x))g'(y)dy. \tag{4.23}$$

We have proved the following result.

PROPOSITION 4.7. *Let Assumptions* 2.1, 2.2, *and* 2.3 *hold. Then the function $H_s(x)$ defined in* (4.22) *is a solution of* (4.9).

We now obtain G_s using H_s. For a specified ordering level s, we set

$$G_s(x) = G_s(s) + \int_s^x H_s(y)\,dy. \tag{4.24}$$

Integrating (4.9) from s to $x > s$, we get

$$\begin{aligned}
&-\frac{\sigma^2}{2}(H_s'(x) - H_s'(s)) + DH_s(x) + (\rho+\lambda)\int_s^x H_s(y)\,dy \\
&\qquad - \lambda\int_s^x dy\int_0^{y-s} H_s(y-\xi)\mu(\xi)\,d\xi = g(x) - g(s).
\end{aligned}$$

Note that $H_s'(s) = Q(s)$. Hence,

$$\begin{aligned}
&-\frac{\sigma^2}{2}G_s''(x) + \frac{\sigma^2}{2}Q(s) + DG_s'(x) + (\rho+\lambda)(G_s(x) - G_s(s)) \\
&\qquad - \lambda\int_s^x dy\int_0^{y-s} H_s(y-\xi)\mu(\xi)\,d\xi = g(x) - g(s)
\end{aligned}$$

and

$$\int_s^x dy \int_0^{y-s} H_s(y-\xi)\mu(\xi)\,d\xi = \int_0^{x-s} G_s(x-\xi)\mu(\xi)\,d\xi - G_s(s) + G_s(s)\int_{x-s}^{\infty}\mu(\xi)\,d\xi.$$

Using these results and comparing with the right-hand side of (4.4), we obtain

$$G_s(s) = \frac{g(s) + \frac{\sigma^2}{2}Q(s) + cD + c\lambda\bar{\xi}}{\rho}. \tag{4.25}$$

We have proved the following theorem.

THEOREM 4.8. *Under Assumptions* 2.1, 2.2, *and* 2.3, *the function* $G_s(x)$ *defined by* (4.24) *and* (4.25) *is a solution of* (4.4) *and* (4.6).

5. Properties of the solution and finding S for any given s. This section further discusses the solution obtained above and derives some properties. These properties are important to finding an optimal (s, S) pair and completing the solution of (4.4)–(4.7). They are also important to proving the optimality of an (s, S) policy for our inventory model.

PROPOSITION 5.1. *There exists a unique number* $a_0 < a$ *such that* $Q(a_0) = 0$, $Q(x) < 0$ *if* $x < a_0$, *and* $Q(x) > 0$ *if* $x > a_0$. *Moreover,* $Q'(x) > 0$ *for* $x \le a$ *and* $Q(x) \ge 2c_0/\sigma^2\beta_2$ *for* $x \ge x_1$, *where* a, x_1, c_0 *are introduced in Assumption* 4.2.

Proof. In terms of (4.12), (4.19), and the relation $Q_s(x) = Q(x+s)$, we have

$$Q'(x) = \beta_2 Q(x) - \frac{2g'(x)}{\sigma^2}. \tag{5.1}$$

For $x < a$, from (4.23), and the convexity of g, we have

$$\begin{aligned} Q(x) &\ge \frac{2}{\sigma^2}\int_x^a \exp(-\beta_2(y-x))g'(y)\,dy \\ &\ge \frac{2}{\sigma^2}g'(x)\int_x^a \exp(-\beta_2(y-x))\,dy \\ &= \frac{2}{\sigma^2}g'(x)\frac{1-\exp(-\beta_2(a-x))}{\beta_2}, \end{aligned}$$

which gives

$$\beta_2 Q(x) - \frac{2g'(x)}{\sigma^2} \ge -\frac{2}{\sigma^2}g'(x)\exp(-\beta_2(a-x)) > 0.$$

Hence, $Q'(x) > 0$.

By (4.23), for $x < x_0 < a$, we have

$$Q(x) \le \frac{2}{\sigma^2}\int_x^{x_0} \exp(-\beta_2(y-x))g'(y)\,dy + \frac{2}{\sigma^2}\exp(-\beta_2(a-x))\int_a^{\infty}\exp(-\beta_2(y-a))g'(y)\,dy,$$

and again from the convexity of g, we have

$$Q(x) \le \frac{2}{\sigma^2}\frac{g'(x_0)}{\beta_2}(1-\exp(-\beta_2(x_0-x))) + \frac{2}{\sigma^2}\exp(-\beta_2(a-x))\int_a^{\infty}\exp(-\beta_2(y-a))g'(y)\,dy.$$

It can be seen that there exists an $x_0' < x_0$ such that

$$Q(x) < \frac{1}{\sigma^2\beta_2}g'(x_0) < 0 \quad \text{for } x < x_0'.$$

Since $Q(a) > 0$, it follows that there exists a unique $a_0 < a$ such that $Q(a_0) = 0$ and $Q(x) < 0$ if $x < a_0$, and $Q(x) > 0$ if $x > a_0$.

Finally, from Assumption 4.2, $g'(x) \geq c_0$ for $x \geq x_1 \geq a$, and thus, for $x > x_1$,

$$Q(x) \geq \frac{2c_0}{\sigma^2} \int_x^\infty \exp(-\beta_2(y-x))\, dy = \frac{2c_0}{\sigma^2\beta_2}.$$

This completes the proof. □

Remark 5.2. If g is convex on $(-\infty, \infty)$, then $Q'(x) \geq 0$ for any $x \in (-\infty, \infty)$. Indeed we have in this case,

$$Q(x) \geq \frac{2}{\sigma^2} g'(x) \int_x^\infty \exp(-\beta_2(y-x))\, dy = \frac{2}{\sigma^2} \frac{g'(x)}{\beta_2}.$$

Hence from (5.1), $Q'(x) \geq 0$.

Next we study the existence of an order-up-to level S. From Proposition 5.1, for $s > a_0$ and $y > s$, we see that $Q(y) > 0$. Hence, for $x > s > a_0$, we have $H_s(x) > 0$. Therefore, in particular,

$$G_s(x) > G_s(s) \quad \text{for } x > s > a_0. \tag{5.2}$$

For $s < x < a_0$, $H_s(x) < 0$. Hence, $G_s(x)$ decreases on $[s, a_0]$. Moreover, we know that for $x \geq x_1 \geq a$, we have $Q(x) \geq 2c_0/\sigma^2\beta_2$. Using (4.21), we have

$$\begin{aligned} H_s(x) &\geq \int_s^{a_0} \Gamma(x-\theta) Q(\theta)\, d\theta + \frac{2c_0}{\sigma^2\beta_2} \int_{x_1}^x \Gamma(x-\theta)\, d\theta \\ &\geq \int_s^{a_0} \exp(\beta_1(x-\theta)) Q(\theta)\, d\theta + \frac{2c_0}{\sigma^2\beta_2} \int_{x_1}^x \exp(\beta_0(x-\theta))\, d\theta \\ &= \exp(\beta_1(x-a_0)) \int_s^{a_0} \exp(\beta_1(a_0-\theta)) Q(\theta)\, d\theta - \frac{2c_0}{\sigma^2\beta_2\beta_0}(1 - \exp(\beta_0(x-x_0))). \end{aligned}$$

For x sufficiently large, we see that

$$H_s(x) \geq -\frac{c_0}{\sigma^2\beta_2\beta_0} > 0.$$

This implies that $G_s(x) \to \infty$ as $x \to \infty$.

Therefore, for $s < a_0$, $G_s(x)$ reaches its minimum on $[s, \infty)$. We denote by $S(s)$ the smallest minimum point. It is necessary to have $S(s) > a_0$, since $G_s(x)$ decreases on $[s, a_0]$. Furthermore,

$$H_s(S(s)) = 0.$$

For $s > a_0$, in view of (5.2), it is convenient to define $S(s) = s$. In summary, we have

$$G_s(S(s)) = \min_{x \geq s} G_s(x) \quad \text{for any } s. \tag{5.3}$$

Let us summarize the result we have just proved.

THEOREM 5.3. *Let Assumptions* 2.1, 2.2, 2.3, *and* 4.2 *hold. Then for any* s, *there exists an* $S(s)$ *such that* (5.3) *is satisfied. This, in turn, implies* (4.7).

Next we discuss some important properties of $S(s)$. For $s < a_0$, since $S(s)$ minimizes $G_s(x)$, we have $G_s''(S(s)) = H_s'(S(s)) > 0$, which implies that

$$Q(S(s)) + \int_s^{S(s)} \Gamma'(S(s) - \theta)Q(\theta)\, d\theta > 0, \quad s < a_0.$$

Note that for any s, $H_s(S(s)) = 0$, which gives

$$\int_s^{S(s)} \Gamma(S(s) - \theta)Q(\theta)\, d\theta = 0.$$

Therefore, $S(s)$ is differentiable and

$$S'(s)\left(Q(S(s)) + \int_s^{S(s)} \Gamma'(S(s) - \theta)Q(\theta)\, d\theta\right) - \Gamma(S(s) - s)Q(s) = 0. \tag{5.4}$$

It follows that

$$\begin{cases} S'(s) < 0 & \text{for } s < a_0, \\ S'(s) = 1 & \text{for } s > a_0. \end{cases} \tag{5.5}$$

Remark 5.4. Note that for $s < a_0$ and s close to a_0, we have $Q(s) \approx (s - a_0)Q'(a_0) = -\frac{2}{\sigma^2} g'(a_0)(s - a_0)$. Thus,

$$\int_s^{a_0} \Gamma(S(s) - \theta)(\theta - a_0)Q'(a_0)\, d\theta + \int_{a_0}^{S(s)} \Gamma(S(s) - \theta)(\theta - a_0)Q'(a_0)\, d\theta \approx 0,$$

and hence,

$$S(s) - a_0 \approx a_0 - s.$$

But we may also write (5.4) as

$$S'(s)\left(\Gamma(S(s) - s)Q(s) + \int_s^{S(s)} \Gamma(S(s) - \theta)Q'(\theta)\, d\theta\right) = \Gamma(S(s) - s)Q(s).$$

Hence, for $s < a_0$ and s close to a_0, we get

$$S'(s)\left(\Gamma(S(s) - s)Q(s) + \Gamma(S(s) - s)\int_s^{S(s)} Q'(\theta)\, d\theta\right) \approx \Gamma(S(s) - s)Q(s),$$

which implies that

$$S'(s) \approx \frac{Q(s)}{Q(S(s))} \approx \frac{s - a_0}{S(s) - a_0} \to -1 \quad \text{as } s \to a_0. \tag{5.6}$$

Remark 5.5. Relations (5.5) and (5.6) reveal the behavior of $S'(s)$ as s approaches the point a_0. An interesting question then arises: What is the behavior of $S'(s)$ and $S(s)$ as s decreases from a_0? The answer to this question would give us a deeper understanding of the optimal (s, S) policy. The answer may also enable us to deal with our model when no backlogging is allowed.

6. Optimal (s, S) policy as the solution of QVI. In this section we complete the solution of (4.4)–(4.7) by finding the value of s. In view of Theorems 4.8 and 5.3, the only remaining equation to be satisfied is (4.5). Moreover, it is enough to satisfy it at $x = s$, and this condition gives us the value of s. We then verify that the solution satisfies the QVI (4.2).

Consider the function $\gamma(s) = \int_s^{S(s)} H_s(x)\,dx$. Then,

$$\gamma'(s) = \int_s^{S(s)} \frac{\partial H_s}{\partial s}(x)\,dx = -Q(s)\int_s^{S(s)} \Gamma(x-s)\,dx.$$

We have

$$\begin{cases} \gamma'(s) > 0 & \text{for } s < a_0, \\ \gamma'(s) = 0 & \text{for } s \geq a_0. \end{cases}$$

Moreover, for $s < a_0$, we have by (4.21)

$$\begin{aligned} \gamma'(s) &\geq -Q(s)\int_s^{S(s)} \exp(\beta_0(x-s))\,dx \\ &= \frac{Q(s)}{\beta_0}(1 - \exp(\beta_0(S(s)-s))) \\ &\geq \frac{Q(s)}{\beta_0}(1 - \exp(\beta_0(a_0-s))). \end{aligned}$$

From Proposition 5.1, for $s < a_0$, we have $Q'(x) > 0$. Therefore, for $s < s_1 < a_0$, we have $Q(s) < Q(s_1) < 0$, and

$$\gamma'(s) \geq \frac{Q(s_1)}{\beta_0}(1 - \exp(\beta_0(a_0-s))).$$

Integrating the inequality from s to s_1, we get

$$\begin{aligned} \gamma(s_1) - \gamma(s) &\geq \frac{Q(s_1)}{\beta_0}\left(s_1 - s + \frac{1}{\beta_0}\exp(\beta_0(a_0-s_1)) - \frac{1}{\beta_0}\exp(\beta_0(a_0-s))\right) \\ &\geq \frac{Q(s_1)}{\beta_0}\left(s_1 - s + \frac{1}{\beta_0}\exp(\beta_0(a_0-s_1))\right). \end{aligned}$$

It follows that $\gamma(s) \to -\infty$ as $s \to -\infty$. As a result, we have shown that $\gamma(s)$ increases strictly from $-\infty$ to 0 as s goes from $-\infty$ to a_0. Therefore, there exists a unique $s < a_0$ such that $\gamma(s) = -K$, where K is the fixed component of the ordering cost function (2.3). This is the optimal s we have been looking for. Thus, we have the following result.

THEOREM 6.1. *Under the assumptions of Theorem* 5.3*, there exists one and only one s that satisfies the relation*

$$G_s(s) = K + G_s(S(s)). \tag{6.1}$$

Additionally, since $G_s(x)$ satisfies (4.4), we have

$$\begin{aligned} -\frac{\sigma^2}{2}G_s''(S(s)) + (\rho+\lambda)G_s(S(s)) - \lambda\int_0^{S(s)-s} G(S(s)-\xi)\mu(\xi)\,d\xi \\ - \lambda G_s(s)\int_{S(s)-s}^{\infty} \mu(\xi)\,d\xi = g(S(s)) + cD + c\lambda\bar{\xi}, \end{aligned}$$

which gives

$$(\rho+\lambda)G_s(S(s)) \geq \lambda G_s(S(s)) + g(S(s)) + cD + c\lambda\bar{\xi}.$$

Then,

$$\rho G_s(S(s)) \geq g(S(s)) + cD + c\lambda\bar{\xi}.$$

We also have from (4.25),

$$\rho G_s(s) = g(s) + \frac{\sigma^2}{2}Q(s) + cD + c\lambda\bar{\xi}. \tag{6.2}$$

Therefore, we get

$$-\rho K \geq g(S(s)) - g(s) - \frac{\sigma^2}{2}Q(s). \tag{6.3}$$

Remark 6.2. Presman and Sethi (2004) have shown that the ordering level s is unique when the demand consists of a compound Poisson process and a constant demand rate. Here we have proved the uniqueness of s, when the demand also includes a Wiener process.

We have completely solved (4.4)–(4.7). To finish, it remains to verify the QVI relations (4.2). The first relation in (4.2) is satisfied for $x \geq s$. For $x < s$, since $G_s(x) = G_s(s)$, we need to show that

$$\rho G_s(s) \leq g(x) + cD + c\lambda\bar{\xi}, \quad x < s.$$

This is done from (6.2), noting that $g(x)$ decreases for $x < s$ and $Q(s) < 0$.

The key issue is the second relation in (4.2), i.e., to verify the inequality

$$G_s(x) \leq K + \inf_{y \geq x} G_s(y). \tag{6.4}$$

It clearly holds for $x < s$, since $G_s(x) = G_s(s) = K + G_s(S(s))$. For $s < x < a_0$, we know that $H_s(x) < 0$. Then we have $G_s(x) \leq G_s(s)$, and (6.4) is satisfied.

There remains the case $x > a_0$, which turns out to be the most difficult one. First we prove the following result under a special condition. Then we treat two special cases in sections 7 and 8.

LEMMA 6.3. *If $S\,(:= S(s))$ is the unique zero of $H_s(x)$ on (s,∞), then (6.4) is satisfied for $x > a_0$.*

Proof. From the previous discussion on H_s, we have

$$\begin{cases} H_s(x) < 0 & \text{for } s < x < S, \\ H_s(x) > 0 & \text{for } x > S. \end{cases}$$

Therefore, $G_s(x)$ is decreasing on the interval (s,S) and increasing on (S,∞). Then,

$$G_s(S) < G_s(x) < G_s(s) \quad \text{for any } s < x < S$$

and

$$G_s(x) < G_s(y) \quad \text{for any } S < x < y.$$

It follows that for $s < x < S$,

$$G_s(x) < G_s(s) = K + G_s(S) < K + G_s(y) \qquad \text{for any } y \geq x,$$

and for $x > S$, $G_s(x) \leq G_s(y)$ for any $y \geq x$. Therefore (6.4) is satisfied. This completes the proof. □

With Lemma 6.3 in hand, we can state the main result of this section.

THEOREM 6.4. *Let Assumptions* 2.1, 2.2, 2.3*, and* 4.2 *hold. Let* $S(s)$ *be the unique zero of* $H_s(x)$ *on* (s, ∞)*. Then the triple* $\{s, S(s), G_s\}$ *defined by* (4.24), (4.25), (5.3)*, and* (6.1) *is a solution of the QVI* (4.2)*.*

In the next section, we give a special case where we can show that the condition of Lemma 6.3 holds, and thus by Theorem 6.4, we have a solution of the QVI (4.2). In section 8, we will see that in the nondiffusion case, we can obtain directly the result without relying on Lemma 6.3. This will be done by a different method, which fails when $\sigma > 0$.

7. Exponentially distributed jump size. We present a special case in which the random jump size in the compound Poisson process is assumed to be exponentially distributed. Additionally, we assume that the function g is convex on $(-\infty, \infty)$. Under these conditions, we prove the condition of Lemma 6.3, namely, the zero uniqueness of H_s. Thus by Theorem 6.4, we have obtained a solution of the QVI (4.2), which corresponds to an (s, S) policy. While the exponential assumption makes the analysis simpler, it is nevertheless an important case from the inventory modeling perspective.

Let $\mu(\xi) = e^{-\xi}$. Hence, $\bar{\pi} = -1$. Using (4.13) in (4.16), we have

$$\begin{aligned}\chi(\pi) &= -\frac{\sigma^2}{2}\pi^2 + D\pi + \rho + \lambda - \frac{\lambda}{\pi + 1} \\ &= \frac{-\frac{\sigma^2}{2}\pi^3 + (D - \frac{\sigma^2}{2})\pi^2 + (D + \rho + \lambda)\pi + \rho}{\pi + 1} = \frac{\zeta(\pi)}{\pi + 1},\end{aligned}$$

where

$$\zeta(\pi) = -\frac{\sigma^2}{2}\pi^3 + \left(D - \frac{\sigma^2}{2}\right)\pi^2 + (D + \rho + \lambda)\pi + \rho.$$

Since

$$\zeta(-\infty) = +\infty, \quad \zeta(-1) = -\lambda < 0, \quad \zeta(0) = \rho > 0, \text{ and } \zeta(+\infty) = -\infty,$$

ζ has three distinct zeros β_1, β_2, and β_3 with $\beta_3 < -1 < \beta_1 < 0 < \beta_2$. Therefore,

$$\chi(\pi) = \frac{-\frac{\sigma^2}{2}(\pi - \beta_3)(\pi - \beta_1)(\pi - \beta_2)}{\pi + 1}.$$

From (4.17), we get

$$\hat{\varphi}(\pi) = \frac{\pi + 1}{\pi - \beta_3},$$

and from (4.14), we have

$$\hat{\Gamma}(\pi) = \frac{\pi + 1}{(\pi - \beta_3)(\pi - \beta_1)} = \frac{1}{\beta_1 - \beta_3}\left(\frac{1 + \beta_1}{\pi - \beta_1} - \frac{1 + \beta_3}{\pi - \beta_3}\right).$$

It follows that

$$\Gamma(\theta) = \frac{1}{\beta_1 - \beta_3}((1+\beta_1)\exp(\beta_1\theta) - (1+\beta_3)\exp(\beta_3\theta)).$$

In terms of (4.22), we have

$$\begin{aligned} H_s(x) &= \frac{1}{\beta_1 - \beta_3}\int_s^x ((1+\beta_1)\exp(\beta_1(x-y)) - (1+\beta_3)\exp(\beta_3(x-y)))Q(y)\,dy \\ &= \frac{\exp(\beta_3 x)}{\beta_1 - \beta_3}\psi_s(x), \end{aligned}$$

where

$$\psi_s(x) = (1+\beta_1)\exp((\beta_1-\beta_3)x)\int_s^x \exp(-\beta_1 y)Q(y)\,dy - (1+\beta_3)\int_s^x \exp(-\beta_3 y)Q(y)\,dy.$$

Note that $\psi_s(a_0) < 0$, since $s < a_0$, and $Q(x) < 0$ for $x < a_0$. Additionally,

$$\begin{aligned} \psi_s'(x) &= (1+\beta_1)(\beta_1-\beta_3)\exp((\beta_1-\beta_3)x)\int_s^x \exp(-\beta_1 y)Q(y)\,dy \\ &\quad + (1+\beta_1)\exp((\beta_1-\beta_3)x)\exp(-\beta_1 x)Q(x) - (1+\beta_3)\exp(-\beta_3 x)Q(x) \\ &= (\beta_1-\beta_3)\exp((\beta_1-\beta_3)x)\left((1+\beta_1)\int_s^x \exp(-\beta_1 y)Q(y)\,dy + \exp(-\beta_1 x)Q(x)\right). \end{aligned}$$

Let

$$\vartheta(x) = (1+\beta_1)\int_s^x \exp(-\beta_1 y)Q(y)\,dy + \exp(-\beta_1 x)Q(x).$$

Then,

$$\vartheta'(x) = \exp(-\beta_1 x)(Q(x) + Q'(x)).$$

Recall Remark 5.2. Since g is convex on $(-\infty, \infty)$, we have $Q'(x) \geq 0$. Then $\vartheta'(x) > 0$ for $x > a_0$. Therefore,

$$\frac{\psi_s'(x)\exp(-(\beta_1-\beta_3)x)}{\beta_1-\beta_3}$$

increases strictly from $\vartheta(a_0) = (1+\beta_1)\int_s^{a_0}\exp(-\beta_1 y)Q(y)\,dy < 0$ to $+\infty$. Hence, there is a unique $S_0 > a_0$ such that $\psi_s'(S_0) = 0$ and $\psi_s'(S_0) < 0$ for $a_0 < x < S_0$, and $\psi_s'(S_0) > 0$ for $x > S_0$. Consequently, $\psi_s(x)$ decreases strictly on (a_0, S_0) and increases strictly on (S_0, ∞). Since $\psi_s(a_0) < 0$ and $\psi_s(+\infty) = +\infty$, it follows that there exists a unique $S(s)$ such that $\psi_s(S(s)) = 0$ on (a_0, ∞), and hence a unique $S := S(s)$ such that $H_s(S) = 0$ on (a_0, ∞).

Remark 7.1. If $\lambda = 0$ (pure diffusion), then it is not necessary to assume that g is convex everywhere. Indeed, in that particular case, $\chi(\pi)$ has only two zeros β_1, β_2 and $\beta_1 = \beta_0$, where β_0 is introduced in Proposition 4.6. Therefore, $H_s(S) = 0$ implies that

$$\int_s^S \exp(-\beta_0 y)Q(y)\,dy = 0$$

and so

$$-\int_s^{a_0}\exp(-\beta_0 y)Q(y)\,dy = \int_{a_0}^S \exp(-\beta_0 y)Q(y)\,dy.$$

It then follows that S is unique.

8. The nondiffusion case ($\sigma = 0$). In this section we drop the Wiener process from our demand, but we keep the general jump size distribution of the compound Poisson process. Thus the demand process we will consider is $y(t)$ given in (2.1) with $\sigma = 0$. We will not try to prove the zero uniqueness of H_s as we did in section 7 for the special exponential case. Instead, we develop a different analysis to ascertain the optimality of the obtained (s, S) policy. Recall that this model was treated by Presman and Sethi (2004) using a probabilistic method, and with a somewhat more general surplus cost function $f(x)$.

We state the following result.

THEOREM 8.1. *Under the assumptions of Theorem* 5.3 *and* $\sigma = 0$, *the triple* $\{s, S(s), G_s\}$ *defined by* (4.24), (4.25), (5.3), *and* (6.1) *is a solution of the QVI* (4.2).

Proof. From (4.15) and (4.16), we obtain

$$\hat{\Gamma}(\pi) = \frac{D}{D\pi + \rho + \lambda - \lambda\hat{\mu}(\pi)},$$

and from (4.11), Γ is the solution of

$$D\Gamma'(\theta) + (\rho + \lambda)\Gamma(\theta) - \lambda \int_0^\theta \Gamma(\theta - \xi)\mu(\xi)\, d\xi = 0, \quad \Gamma(0) = 1.$$

Hence, $\Gamma'(0) = -\frac{\rho+\lambda}{D}$. Therefore, from (4.19) we see that $\frac{\beta_2\sigma^2}{2} \to D$ as $\sigma \to 0$. Note that from (4.23), we get

$$\begin{aligned} Q(x) &= \frac{1}{D}\int_0^\infty \exp\left(-\frac{\beta_2\sigma^2}{2D}\theta\right) g'\left(x + \frac{\sigma^2\theta}{2D}\right) d\theta \\ &= \frac{1}{D}\int_0^\infty e^{-\theta} g'(x)\, d\theta = \frac{g'(x)}{D}. \end{aligned} \tag{8.1}$$

From the definition of a_0 in Proposition 5.1, we see that $a_0 = a$. Additionally, (4.25) becomes

$$G_s(s) = \frac{g(s) + cD + c\lambda\bar{\xi}}{\rho},$$

and (6.3) gives

$$-\rho K \geq g(S(s)) - g(s). \tag{8.2}$$

Now we let $L_s(x) = G_s(x) - G_s(s) = \int_s^x H_s(y)\, dy$. In terms of (4.22) and (8.1), we have

$$L_s(x) = \frac{1}{D}\int_s^x dy \int_s^y \Gamma(y - \theta)g'(\theta)\, dy = \frac{1}{D}\int_s^x \Gamma(x - \theta)(g(\theta) - g(s))\, d\theta.$$

Since g is decreasing on $(-\infty, a)$ and increasing on (a, ∞), we have $L_s(x) \leq 0$ for $s \leq x < a$. For $a < x < S$, we have

$$\begin{aligned} L_s(x) &= \frac{1}{D}\int_s^a \Gamma(x - \theta)(g(\theta) - g(s))\, d\theta + \frac{1}{D}\int_a^x \Gamma(x - \theta)(g(\theta) - g(s))\, d\theta \\ &\leq \frac{1}{D}\int_a^x \Gamma(x - \theta)(g(\theta) - g(s))\, d\theta \\ &\leq \frac{1}{D}(g(S) - g(s))\int_a^x \Gamma(x - \theta)\, d\theta \leq 0, \end{aligned}$$

where the last inequality holds because of (8.2). Moreover, $L_s(x)$ is the solution of

$$DL_s'(x) + (\rho+\lambda)L_s(x) - \lambda\int_0^{x-s} L_s(x-\xi)\mu(\xi)\,d\xi = g(x) - g(s),\ L_s(s) = 0. \tag{8.3}$$

Next, let $x_0 > S$ be the first value such that $L_s(x_0) = 0$ (note that $L_s(x) \to +\infty$ as $x \to +\infty$, so such an x_0 exists). We recall that $L_s(x) = 0$ for $x \le s$. Let $u \ge 0$. We can rewrite (8.3) as follows (even if $x_0 - u \le s$):

$$\begin{aligned} &DL_s'(x) + (\rho+\lambda)L_s(x) - \lambda\int_{x_0-u}^{x} L_s(\eta)\mu(x-\eta)\,d\eta \\ &\quad = (g(x)-g(s))\mathbb{1}_{x>s} + \lambda\int_s^{x_0-u} L_s(\eta)\mu(x-\eta)\,d\eta, \end{aligned} \tag{8.4}$$

which implies

$$DL_s'(x) + (\rho+\lambda)L_s(x) - \lambda\int_{x_0-u}^{x} L_s(\eta)\mu(x-\eta)\,d\eta \le (g(x)-g(s))\mathbb{1}_{x>s}. \tag{8.5}$$

Note that (8.4) and (8.5) hold for any x, and particularly for $x > x_0 - u$. Moreover, we have

$$L_s(x_0-u) \le 0. \tag{8.6}$$

Now we set

$$M_s(x) = L_s(x+u) + K. \tag{8.7}$$

It follows that

$$M_s(x_0-u) = K. \tag{8.8}$$

We rewrite (8.4) with x changed to $x+u > x_0$. Then,

$$DM_s'(x) + (\rho+\lambda)M_s(x) - (\rho+\lambda)K - \lambda\int_s^{x+u} L_s(\eta)\mu(x+u-\eta)\,d\eta = g(x+u) - g(s).$$

Since

$$\begin{aligned} \int_s^{x+u} L_s(\eta)\mu(x+u-\eta)\,d\eta &= \int_s^{x_0} L_s(\eta)\mu(x+u-\eta)\,d\eta + \int_{x_0}^{x+u} L_s(\eta)\mu(x+u-\eta)\,d\eta \\ &= \int_s^{x_0} L_s(\eta)\mu(x+u-\eta)\,d\eta + \int_{x_0-u}^{x} M_s(\eta)\mu(x-\eta)\,d\eta \\ &\quad -K\int_{x_0}^{x+u} \mu(x+u-\eta)\,d\eta, \end{aligned}$$

we have

$$
\begin{aligned}
DM_s'(x) + (\rho+\lambda)M_s(x) - \lambda \int_{x_0-u}^{x} M_s(\eta)\mu(x-\eta)\,d\eta \\
&= g(x+u) - g(s) + \rho K + \lambda K + \lambda \int_s^{x_0} L_s(\eta)\mu(x+u-\eta)\,d\eta \\
&\quad - \lambda K \int_{x_0}^{x+u} \mu(x+u-\eta)\,d\eta \\
&= g(x+u) - g(s) + \rho K + \lambda K + \lambda \int_s^{x_0} (L_s(\eta)+K)\mu(x+u-\eta)\,d\eta \\
&\quad - \lambda K \int_{s}^{x+u} \mu(x+u-\eta)\,d\eta \\
&\geq g(x+u) - g(s) + \rho K + \lambda K - \lambda K \int_s^{x+u} \mu(x+u-\eta)\,d\eta \\
&\geq g(x+u) - g(s) + \rho K.
\end{aligned} \tag{8.9}
$$

To proceed, we claim that the following relation holds:

$$
g(x+u) - g(s) + \rho K - (g(x) - g(s))\mathbb{1}_{x>s} \geq 0, \quad x > x_0 - u. \tag{8.10}
$$

In fact, we first note that $g(x_0) - g(s) > 0$, which follows from (8.5) written for $x = x_0$. Hence,

$$
g(x+u) - g(s) \geq g(x_0) - g(s) > 0.
$$

Now, if $x < s$, then (8.10) is obviously satisfied. If $x > s$, we have

$$
g(x+u) - g(s) + \rho K - (g(x) - g(s)) = g(x+u) - g(x) + \rho K.
$$

If $x > a$, we have $g(x+u) - g(x) > 0$, and if $s < x < a$, we have $g(x) < g(s)$ and $g(x+u) \geq g(x_0)$. Hence,

$$
g(x+u) - g(x) + \rho K \geq g(x_0) - g(s) + \rho K \geq 0.
$$

So in both cases we get (8.10).

Now setting $Y_s(x) = M_s(x) - L_s(x)$, $x \geq x_0 - u$, and combining (8.5)–(8.10), we get

$$
\begin{cases}
DY_s'(x) + (\rho+\lambda)Y_s(x) - \lambda \displaystyle\int_{x_0-u}^{x} Y_s(\eta)\mu(x-\eta)\,d\eta \geq 0, \\
Y_s(x_0-u) \geq 0.
\end{cases}
$$

Therefore, $Y_s(x) \geq 0$ for $x \geq x_0 - u$, and particularly $Y_s(x) \geq 0$ for $x \geq x_0$. It follows that

$$
L_s(x) \leq K + L_s(x+u), \quad x \geq x_0,\ u \geq 0,
$$

and then

$$
G_s(x) \leq K + G_s(x+u), \quad x \geq x_0,\ u \geq 0,
$$

i.e.,

$$
G_s(x) \leq K + \inf_{y \geq x} G_s(y), \quad x \geq x_0.
$$

For $s \le x \le x_0$, since $L_s(x) \le 0$, we have

$$G_s(x) \le G_s(s) = K + G_s(S) \le K + \inf_{y \ge x} G_s(y).$$

We have just completed the verification of the key relation (6.4) for $x > s$ and, therefore, have shown that the triple $\{s, S(s), G_s\}$ satisfies the QVI (4.2). This completes the proof. □

9. Remarks on the general case and numerical demonstration. We have pointed out in section 6 that the key issue in the analysis of the optimal (s, S) policy is (6.4). Whether it holds or not remains an open question for the general case, in which both diffusion and random jumps are present in the demand. In this section, we first remark the difficulty in trying to verify (6.4). Then we present a more complicated example to gain some insights into the general model.

Let, in general,

$$L_s(x) = G_s(x) - G_s(s) = \int_s^x H_s(y)\,dy = \int_s^x dy \int_s^y \Gamma(y-\theta)Q(\theta)\,d\theta.$$

Set $\zeta(x) = \frac{2}{\beta_2\sigma^2}(g(x) + \frac{\sigma^2}{2}Q(x))$. By (5.1), we have

$$\zeta'(x) = Q(x) \begin{cases} < 0 & \text{for } x < a_0, \\ > 0 & \text{for } x > a_0. \end{cases}$$

Hence,

$$L_s(x) = \int_s^x dy \int_s^y \Gamma(y-\theta)\zeta'(\theta)\,d\theta = \int_s^x \Gamma(x-y)(\zeta(y) - \zeta(s))\,dy.$$

We have shown that $L_s(x) \le 0$ for $s \le x \le a_0$. Then for $a_0 < x < S$, we have

$$\begin{aligned} L_s(x) &= \int_s^{a_0} \Gamma(x-y)(\zeta(y) - \zeta(s))\,dy + \int_{a_0}^x \Gamma(x-y)(\zeta(y) - \zeta(s))\,dy \\ &\le \int_{a_0}^x \Gamma(x-y)(\zeta(y) - \zeta(s))\,dy \\ &\le (\zeta(S) - \zeta(s)) \int_{a_0}^x \Gamma(x-y)\,dy. \end{aligned}$$

Remark 9.1. Here the difficulty is that we do not know if $\zeta(S) - \zeta(s) \le 0$ or not.

On the other hand, if we go back to see how we got (6.3), we can actually assert that

$$\begin{aligned} -\rho K &\ge \frac{\sigma^2}{2}H_s'(S) + g(S) - g(s) - \frac{\sigma^2}{2}Q(s) \\ &\ge \frac{\sigma^2}{2}(H_s'(S) - Q(S)) + g(S) - g(s). \end{aligned}$$

Remark 9.2. The difficulty here is that we do not know how $H_s'(S)$ compares with $Q(S)$.

In the rest of this section, we study a more general demand than that in section 7 and try to get some insights into this case. Let

$$\mu(\xi) = a_1 e^{-\alpha_1 \xi} + a_2 e^{-\alpha_2 \xi}$$

with $0 < \alpha_1 < \alpha_2$, $a_1 > 0$, $a_2 > 0$, $a_1\alpha_2 + a_2\alpha_1 = \alpha_1\alpha_2$. For $\pi > -\alpha_1 = \bar{\pi}$, we get

$$\hat{\mu}(\pi) = \frac{a_1}{\pi + \alpha_1} + \frac{a_2}{\pi + \alpha_2}$$

and

$$\chi(\pi) = -\frac{\sigma^2}{2}\pi^2 + D\pi + \rho + \lambda - \frac{\lambda a_1}{\pi + \alpha_1} - \frac{\lambda a_2}{\pi + \alpha_2} = \frac{\zeta(\pi)}{(\pi + \alpha_1)(\pi + \alpha_2)}.$$

Note that

$$\begin{aligned}
&\zeta(-\infty) = \zeta(\infty) = -\infty,\\
&\zeta(-\alpha_2) = \lambda a_2(\alpha_2 - \alpha_1) > 0,\\
&\zeta(-\alpha_1) = -\lambda a_1(\alpha_2 - \alpha_1) < 0,\\
&\zeta(0) = \rho\alpha_1\alpha_2 > 0.
\end{aligned}$$

Hence, ζ has four distinct zeros β_1, β_2, β_3, and β_4 with $\beta_4 < -\alpha_2 < \beta_3 < -\alpha_1 < \beta_1 < 0 < \beta_2$, and

$$\zeta(\pi) = -\frac{\sigma^2}{2}(\pi - \beta_4)(\pi - \beta_3)(\pi - \beta_1)(\pi - \beta_2).$$

It follows that

$$\hat{\Gamma}(\pi) = \frac{(\pi + \alpha_1)(\pi + \alpha_2)}{(\pi - \beta_1)(\pi - \beta_3)(\pi - \beta_4)} = \frac{A}{\pi - \beta_1} + \frac{B}{\pi - \beta_3} + \frac{C}{\pi - \beta_4},$$

where

$$A = \frac{(\beta_1 + \alpha_1)(\beta_1 + \alpha_2)}{(\beta_1 - \beta_3)(\beta_1 - \beta_4)}, \quad B = \frac{(\beta_3 + \alpha_1)(\beta_3 + \alpha_2)}{(\beta_3 - \beta_1)(\beta_3 - \beta_4)}, \quad C = \frac{(\beta_4 + \alpha_1)(\beta_4 + \alpha_2)}{(\beta_4 - \beta_1)(\beta_4 - \beta_3)},$$

and $A > 0$, $B > 0$, $C > 0$. Thus,

$$\Gamma(\theta) = A\exp(\beta_1\theta) + B\exp(\beta_3\theta) + C\exp(\beta_4\theta)$$

and

$$H_s(x) = \int_s^x (A\exp(\beta_1(x - \theta)) + B\exp(\beta_3(x - \theta)) + C\exp(\beta_4(x - \theta)))Q(\theta)\,dy.$$

We choose a surplus cost function $f(x)$ often used in the literature, i.e.,

$$f(x) = \begin{cases} hx & \text{for } x \geq 0,\\ -px & \text{for } x < 0, \end{cases}$$

where $h > 0$ is the unit holding cost per unit time, and $p > 0$ is the unit shortage cost per unit time. Then in terms of (4.1),

$$g(x) = \begin{cases} (h + c\rho)x & \text{for } x \geq 0,\\ -(p - c\rho)x & \text{for } x < 0. \end{cases}$$

Additionally, we assume $p > c\rho$ to avoid the trivial case in which the optimal s would be $-\infty$; see Presman and Sethi (2004).

Since $a = 0$ for the considered g function, we have

$$Q(x) = \begin{cases} \dfrac{2}{\sigma^2\beta_2}\left((p+h)e^{\beta_2 x} - (p-c\rho)\right) & \text{if } x < 0, \\ \dfrac{2}{\sigma^2\beta_2}(h+c\rho) & \text{if } x \geq 0. \end{cases}$$

Moreover,

$$a_0 = \frac{1}{\beta_2}\ln\left(\frac{p-c\rho}{p+h}\right) < 0.$$

Using the result for $Q(x)$, we can find the explicit expression for $H_s(x)$. Let

$$\begin{aligned} H_s^0(x) = & -(p-c\rho)\left(\frac{A}{\beta_1}e^{\beta_1(x-s)} + \frac{B}{\beta_3}e^{\beta_3(x-s)} + \frac{C}{\beta_4}e^{\beta_4(x-s)}\right) \\ & -(p+h)\left(\frac{A}{\beta_2-\beta_1}e^{\beta_1 x}e^{(\beta_2-\beta_1)s} + \frac{B}{\beta_2-\beta_3}e^{\beta_3 x}e^{(\beta_2-\beta_3)s}\right. \\ & \left. + \frac{C}{\beta_2-\beta_4}e^{\beta_4 x}e^{(\beta_2-\beta_4)s}\right). \end{aligned}$$

Then we have for $x < 0$,

$$\begin{aligned} H_s(x) = & \frac{2}{\sigma^2\beta_2}\left(H_s^0(x) + (p-c\rho)\left(\frac{A}{\beta_1} + \frac{B}{\beta_3} + \frac{C}{\beta_4}\right)\right. \\ & \left. + (p+h)\left(\frac{A}{\beta_2-\beta_1} + \frac{B}{\beta_2-\beta_3} + \frac{C}{\beta_2-\beta_4}\right)e^{\beta_2 x}\right), \end{aligned}$$

and for $x \geq 0$,

$$\begin{aligned} H_s(x) = & \frac{2}{\sigma^2\beta_2}\left(H_s^0(x) - (h+c\rho)\left(\frac{A}{\beta_1} + \frac{B}{\beta_3} + \frac{C}{\beta_4}\right)\right. \\ & \left. + (p+h)\beta_2\left(\frac{A}{\beta_1(\beta_2-\beta_1)}e^{\beta_1 x} + \frac{B}{\beta_3(\beta_2-\beta_3)}e^{\beta_3 x} + \frac{C}{\beta_4(\beta_2-\beta_4)}e^{\beta_4 x}\right)\right). \end{aligned}$$

We numerically test the zero of function $H_s(x)$. Set

$$\alpha_1 = 1, \quad \alpha_2 = 2, \quad a_1 = \frac{1}{2}, \quad a_2 = 2,$$

and choose

$$D = 0.1, \quad \lambda = 0.1, \quad \sigma = 0.2, \quad \rho = 0.05.$$

Using the Newton–Raphson procedure, the four zeros of ζ are calculated as

$$\beta_4 = -2.5136, \quad \beta_3 = -1.3112, \quad \beta_1 = -0.2497, \quad \beta_2 = 6.0746.$$

Set $p = 1$, $h = 1$, and $c = 5$. Then we obtain that $a_0 = -0.1615$.

From section 6 we know that the optimal value of s must satisfy the condition $s < a_0$. Therefore, we pick a number of values for s on the left of a_0 and numerically investigate the corresponding function $H_s(x)$. Figure 9.1 shows the graphs of $H_s(x)$ for seven different values of s, which are $s = -3$, -2, -1.5, -1, -0.75, -0.5, and -0.3. These graphs suggest that $H_s(x)$ has a unique zero S on (s, ∞).

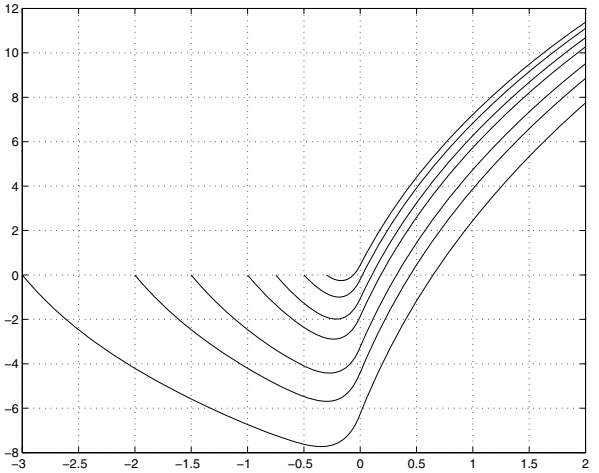


FIG. 9.1. *Graph of $H_s(x)$ for various values of s.*

10. Conclusions. We consider a continuous-review stochastic inventory model with a demand consisting of a compound Poisson process and a diffusion process. We formulate the inventory problem as an impulse control problem, which allows us to use the QVI approach to study it.

We prove the optimality of an (s, S) policy in two cases. (i) When the demand is a mixture of a diffusion process and a compound Poisson process with exponentially distributed jump sizes, and (ii) when the demand is a mixture of a constant demand and a compound Poisson process. However, the combination of a diffusion process and a general compound Poisson demand is not completely solved. We explain the difficulties that arise in the course of trying to prove the optimality (Remarks 9.1 and 9.2). We also present a condition in Lemma 6.3 that might be the key if the optimal decision is indeed of (s, S) type, which we believe to be the case. Although we have verified this condition for the special case of exponentially distributed jump demand and our numerical results in section 9 suggest that it holds for a more complicated example, the optimality of the general model remains an open and interesting problem for further study.

More generally, it would be of interest to extend the analysis of this paper to problems including Lévy processes having nonfinite Lévy measures. This extension is potentially important for problems in finance, where stock prices are often modeled by Lévy processes and there are fixed transaction costs incurred in buying and selling of stocks.

Further generalizations include inventory models with the Markovian demand as discussed in Beyer, Cheng, and Sethi (2006) in the discrete-time framework and with information delays as discussed in Bensoussan, Cakanyildirim, and Sethi (2005) also in the discrete-time framework. Extensions of our results to inventory models with lead times, lost sales case (Remark 5.5), nonstationary problems, and inventory problems with multiple products could also be considered. One could also investigate the use of the vanishing discount approach to obtain a proof of the optimality of an (s, S) policy in the average-cost case.

To conclude, we should mention that there is a large amount of literature treating the various proposed problems in the discrete-time framework. To ascertain the optimality of (s, S) policies in these cases when time is continuous and demand is more

general than a compound Poisson process is a challenging research agenda for future research on optimal control models of inventory problems.

Acknowledgment. We thank the reviewers whose comments have improved the exposition of this paper.

REFERENCES

B. Archibald and E. Silver (1978), (s, S) *policies under continuous review and discrete compound Poisson demands*, Manag. Sci., 24, pp. 899–909.

J. Bather (1966), *A continuous time inventory model*, J. Appl. Probab., 3, pp. 538–549.

A. Bensoussan and J.-L. Lions (1984), *Impulse Control and Quasi-Variational Inequalities,* Gauthier-Villars, Paris, France.

A. Bensoussan and C. S. Tapiero (1982), *Impulsive control in management: Prospects and applications*, J. Optim. Theory Appl., 37, pp. 419–442.

A. Bensoussan, M. Cakanyildirim, and S. P. Sethi (2005), *Optimality of Base Stock and* (s, S) *Policies for Inventory Problems with Information Delays*, Working Paper SOM200549, The University of Texas at Dallas, Richardson, TX.

D. Beyer (1994), *An inventory model with Wiener demand process and positive lead time*, Optimization, 29, pp. 181–193.

D. Beyer and S. P. Sethi (1998), *A proof of the EOQ formula using quasi-variational inequalities*, Internat J. Systems Sci., 29, pp. 1295–1299.

D. Beyer, F. Cheng, and S. P. Sethi (2006), *Markovian Demand Inventory Models,* Springer, New York, forthcoming.

D. Beyer, S. P. Sethi, and M. Taksar (1998), *Inventory models with Markovian demands and cost functions of polynomial growth*, J. Optim. Theory Appl., 98, pp. 281–323.

S. Browne and P. Zipkin (1991), *Inventory models with continuous stochastic demands*, Ann. Appl. Probab., 1, pp. 419–435.

G. Constantinides and S. Richard (1978), *Existence of optimal simple policies for discounted-cost inventory and cash management in continuous time*, Oper. Res., 26, pp. 620–636.

A. Federgruen and Z. Schechner (1983), *Cost formulas for continuous review inventory models with fixed delivery lags*, Oper. Res., 31, pp. 957–965.

R. Feldman (1978), *A continuous review* (s, S) *inventory system in a random environment*, J. Appl. Probab., 15, pp. 654–659.

J. Keilson (1979), *Markov Chain Models: Rarity and Exponentiality,* Springer-Verlag, New York.

E. Presman and S. P. Sethi (2004), *Stochastic Inventory Models with Continuous and Poisson Demands and Discounted and Average Costs*, Working Paper, The University of Texas at Dallas, Richardson, TX.

M. L. Puterman (1975), *A diffusion process model for a storage system*, in Logistics, M. A. Geisler, ed., North-Holland, Amsterdam, pp. 143–159.

F. R. Richards (1975), *Comments on the distribution of inventory position in a continuous-review* (s, S) *inventory system*, Oper. Res., 23, pp. 366–371.

I. Sahin (1979), *On the stationary analysis of continuous review* (s, S) *inventory systems with constant lead times*, Oper. Res., 27, pp. 717–730.

I. Sahin (1983), *On the continuous review* (s, S) *inventory model under compound renewal demand and random lead times*, J. Appl. Probab., 20, pp. 213–219.

J.-S. Song and P. Zipkin (1993), *Inventory control in a fluctuating demand environment*, Oper. Res., 41, pp. 351–370.

A. Sulem (1986), *A solvable one-dimensional model of a diffusion inventory system*, Math. Oper. Res., 11, pp. 125–133.

R. Thompstone and E. Silver (1975), *A coordinated inventory control system for compound Poisson demand and zero lead time*, Int. J. Prod. Res., 13, pp. 581–602.

H. C. Tijms (1972), *Analysis of* (s, S) *Inventory Models,* Math. Centre Tracts 40, Mathematisch Centrum, Amsterdam.

N. van Dijk (1990), *On a simple proof of uniformization for continuous and discrete-state continuous time Markov chains*, Adv. in Appl. Probab., 22, pp. 749–750.

W. Whitt (1973), *Diffusion Models for Inventory and Production Systems*, preprint, Yale University.

P. Zipkin (1986), *Stochastic leadtimes in continuous time inventory models*, Naval Res. Logist. Quart., 33, pp. 763–774.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 5, pp. 1677–1690

APPLICATIONS OF LEFSCHETZ NUMBERS IN CONTROL THEORY*

PETER SAVELIEV†

Abstract. We develop some applications of techniques of the Lefschetz coincidence theory in control theory. The topics are existence of equilibria and their robustness, and controllability and its robustness.

Key words. control system, equilibrium, controllability, robustness, fixed point theory, Lefschetz number, coincidence point, homology theory, algebraic topology

AMS subject classifications. 93B, 55M20, 55H25

DOI. 10.1137/S0363012904442240

1. Introduction.

The goal of this paper is to provide examples of what Lefschetz coincidence theory can contribute to control theory. We discuss existence of equilibria and their robustness, and controllability and its robustness.

We develop some topological techniques, already available in dynamics, in the control theoretic setting. A (discrete) dynamical system on a manifold M is simply a map $f : M \to M$. Then $x \in M$ and $f(x)$ are the current and next states of the system, respectively. An equilibrium of the system is a fixed point of f. The problem of detecting equilibria can be treated via the more general coincidence problem [2, sect. VI.14], [4], [35, Ch. 7], [15]: "Given two maps $f, g : N \to M$ between two n-dimensional manifolds, what can be said about the coincidence set C of all x such that $f(x) = g(x)$?" Indeed, the equilibrium set $C = \{x \in M : f(x) = x\}$ is the coincidence set of f and the identity map $g : M \to M$. The famous Lefschetz coincidence theorem states that if the Lefschetz number λ_{fg} is not equal to zero, then there is at least one coincidence, i.e., $C \neq \varnothing$. Using this and other invariants, one can find out whether a dynamical system has an equilibrium or a periodic point.

In the case of a *controlled* dynamical system, the next state $f(x, u)$ depends not only on the current state, $x \in M$, but also on the *input*, $u \in U$. A discrete time control system is given by the space of inputs U, the space of states M, the "state-input" space $N = M \times U$, a map $f : N = M \times U \to M$, and the projection $g : N = M \times U \to M$ (in general, N is a fiber bundle and $g : N \to M$ is the bundle projection). Then, just as above, the equilibrium set $C = \{x \in M : f(x, u) = x \text{ for some } u \in U\}$ of the system is the coincidence set of the pair (f, g). However, since the dimensions of N and M are no longer equal, the Lefschetz *number* is replaced with the Lefschetz *homomorphism* [31], which does a better job of detecting coincidences.

Another application of the coincidence theory approach is controllability. A system is called controllable if any state can be reached from any other state; i.e., for each pair of states $x, y \in M$ there are inputs $u_0, \dots, u_r \in U$ such that $x_1 = f(u_0, x), x_2 = f(u_1, x_1), \dots, y = x_{r+1} = f(u_r, x_r)$. Therefore controllability is equivalent to surjectivity of the composition of several iterations of f. On the other hand, a map is surjective if it has a coincidence with any constant map.

*Received by the editors March 22, 2004; accepted for publication (in revised form) May 8, 2005; published electronically November 22, 2005.
http://www.siam.org/journals/sicon/44-5/44224.html
†Department of Mathematics, Marshall University, Huntington, WV 25755-2560 (saveliev@member.ams.org).

The state space M is often a manifold, as opposed to a Euclidean space, when it appears in robotics. For example, $M = \mathbf{T}^n = (\mathbf{S}^1)^n$, the n-dimensional torus, is the space of all possible states of a robotic arm with n revolving joints [27, p. 1]; or $M = \mathbf{R}^3 \times SO(3)$ is the space of positions of a rigid body [25, Ch. 2]. Typically, we have $N = M \times U$. However, nontrivial bundles are also common. For example, consider a spherical pendulum with a gas jet control which is always directed in the tangent space. Then its state space is $M = \mathbf{S}^2$, the 2-sphere, while the state-input space N is the tangent bundle $T\mathbf{S}^2$ of $\mathbf{S}^2$, which is an $\mathbf{R}^2$-bundle over M not isomorphic to $M \times \mathbf{R}^2$ [27, p. 17]. In spite of the abundance of such examples [6], [25], [27], topological techniques have not thus far found broad applications in control theory. The only recent examples known to the author are [18], [19], [20], [21], [22].

The topological approach provides the following advantages. Consider a control system as a triple (M, N, f) of topological spaces M, N and a continuous map f as described above. Since our knowledge of the model is inevitably imprecise, we have to deal with perturbations of the system. As perturbations may be understood as variations of unknown parameters of the system, their effect on the behavior of the system is also unknown. However, if the system depends continuously on these parameters, the change of M, N, and f is also continuous. This means that we are to consider spaces homeomorphic to M, N and maps homotopic to f. An appropriate tool to deal with this degree of generality is homology theory. Indeed, the homology groups $H_*(M), H_*(N)$ of M, N and the homology homomorphism $f_* : H_*(N) \to H_*(M)$ of f remain constant under homeomorphisms of M, N and homotopies of f. They can also be rigorously and effectively computed [26], [19].

Further, the perturbations of f are normally assumed "small" (in particular, this is the basis of the notion of structural stability). However, unless actual estimates are available, we do not know how "small" the perturbations of the real system are. Therefore, in order to take into account the "worst possible scenario," we consider large, but still continuous, perturbations of the system. As an example, a constant external force, such as gravitation, in any of the above robotic systems may be treated as such a perturbation. Thus the use of homology theory provides answers with a new, for control theory, degree of robustness. Providing results of this nature is the first objective of this paper. We apply Lefschetz coincidence theory to prove existence of equilibria (Theorem 6.1) and controllability (Theorem 7.2) for systems determined by maps homotopic to f.

The second objective of this paper is to study robustness of these properties under arbitrarily small perturbations because sometimes they produce a dramatic change in the properties of the system. This change may be the loss of an equilibrium (Theorem 6.4) or the loss of controllability (Theorem 7.3).

The paper is organized as follows. Some preliminaries from algebraic topology are outlined in section 2. In section 3 we review the classical theory of Lefschetz numbers and show its inadequacy for control theory. In section 4 we consider the necessary generalization, the Lefschetz homomorphism, of the Lefschetz number and state several relevant results about existence of coincidences. In section 5 we state some results about removability of coincidences. In section 6 we provide sufficient conditions of existence of equilibria of a discrete system and their robustness. In section 7 we provide sufficient conditions of controllability of a discrete system and its robustness. In section 8 we discuss how our coincidence results can be applied to existence of equilibria and controllability of continuous time control systems. Notions of control theory are defined as needed; for details see [27], [29], [34].

2. Preliminaries from algebraic topology. The terminology we use is standard [2]. Suppose N is a topological space and $A \subset N$ is a subspace. The singular *homology group* $H_k(N, A)$ of N relative to A over $\mathbf{Q}$ or any other field is defined as follows. If Δ_k is the standard k-simplex, $k = 0, 1, 2, \dots$, any map $\sigma : \Delta_k \to N$ is called a singular k-simplex in N. We let $C_k(N, A)$ be the vector space over $\mathbf{Q}$ generated by all singular k-simplices of N whose images are not completely in A. Then the boundary operator $\partial_k : C_k(N, A) \to C_{k-1}(N, A)$ is defined in the natural way, and we let $H_k(N, A) = \ker \partial_k / \operatorname{Im} \partial_k$. Further, let $C^k(N, A)$ be the dual of $C_k(N, A)$, i.e., the vector space of all linear functions from $C_k(N, A)$ to $\mathbf{Q}$. Then ∂_k generates the coboundary operator $\partial^k : C^k(N, A) \to C^{k+1}(N, A)$, and we let $H^k(N, A) = \ker \partial^k / \operatorname{Im} \partial^k$ be the *cohomology group* of N relative to A. Also $H_k(N) = H_k(N, \varnothing)$, $H^k(N) = H^k(N, \varnothing)$. If (N, A) is a simplicial complex, its simplicial homology and cohomology are defined in the same way starting with $C_k(N, A)$ generated by all simplices of (N, A). The homology and cohomology groups $H_k(N, A; G)$, $H^k(N, A; G)$ over any group G can be defined in a similar fashion.

Homology and cohomology groups $\{H_k(N, A) : k = 0, 1, 2, \dots\}$, $\{H^k(N, A) : k = 0, 1, 2, \dots\}$ over fields are (graded) vector spaces with the following properties. The Betti numbers, $b_k = \dim H_k(N)$, for $k = 0, 1, 2$, are the numbers of path components, "tunnels," and "voids" of N, respectively. In the case of a path connected N, the identities of $H_0(N) = H^0(N) = \mathbf{Q}$ are denoted by 1. If N is contractible, it is *acyclic*; i.e., $H_k(N) = H^k(N) = 0$ for $k > 0$. If N is an n-dimensional simplicial complex, $H_k(N) = 0$ for all $k > n$. If M is a compact connected orientable n-dimensional manifold with boundary ∂M, then $H_n(M, \partial M) = H^n(M, \partial M) = \mathbf{Q}$. The identities of these two groups are the *fundamental classes* O_M and $\overline{O}_M$ of M, respectively. Further, there is the *Poincaré duality* isomorphism $D_M : H^k(M, \partial M) \to H_{n-k}(M)$ given by the cap product with the fundamental class O_M. The *cap product* is the homomorphism $\frown : H^k(N, A) \otimes H_m(N, A) \to H_{m-k}(N)$ given by $x \frown a = (1 \times x)\Delta a$, where Δ is a diagonal approximation. Then $a \in H_k(N, A)$ and $x \in H^k(N, A)$ are called *dual* if $x \frown a = \langle x, a \rangle = x(a) = 1$. In particular, O_M and $\overline{O}_M$ are dual. By the Künneth theorem, $H_k(M \times U) = \sum_{i+j=k} H_i(M) \otimes H_j(U)$, $k = 0, 1, 2, \dots$.

Suppose B is a subspace of the topological space M and $f : N \to M$ is a map; then $f : (N, A) \to (M, B)$ is a *map of pairs* if $f(A) \subset B$. In this case, f generates the natural homomorphism from $C_k(N, A)$ to $C_k(M, B)$. This homomorphism generates $f_* : H_k(N, A) \to H_k(M, B)$, the *homology homomorphism* of f, and $f^* : H^k(M, B) \to H^k(N, A)$, the *cohomology homomorphism* of f. Two maps $f, g : (N, A) \to (M, B)$ are called *homotopic* if f can be continuously "deformed" into g; i.e., there is a map $F : [0, 1] \times (N, A) \to (M, B)$ such that $F(0, \cdot) = f$ and $F(1, \cdot) = g$. If f and g are homotopic, then $f_* = g_*$. In particular, if f is homotopic to a constant map, then f_* is trivial; i.e., $f_* : H_k(N, A) \to H_k(M, B)$ is zero for $k = 1, 2, \dots$, or simply $f_* = 0$. In the case of n-manifolds, the homomorphism $f_* : H_n(N, \partial N) \to H_n(M, \partial M)$ is the multiplication by $\deg f$, the degree of f. The kth homotopy group $\pi_k(N)$ of N is the group of homotopy classes of maps of k-spheres to N.

3. Review of Lefschetz theory. In this section, M and N are orientable compact connected manifolds with boundaries $\partial M, \partial N$, and $\dim M = \dim N = n$.

Consider the fixed point problem: "If $f : M \to M$ is a map, what can be said about the set of points $x \in M$ such that $f(x) = x$?" Applications of fixed point theorems (Kakutani, Banach, etc.) to control problems are abundant [1], [7], [8], [16], [23], [28]. However, the methods we suggest in this paper go far beyond those.

One may associate with f an integer λ_f called the Lefschetz number [3]:

$$\lambda_f = \sum_k (-1)^k \mathrm{Trace}(f_{*k}),$$

where $f_{*k} : H_k(M) \to H_k(M)$ is induced by f. The *Lefschetz fixed point theorem* states that if $\lambda_f \neq 0$, then f has a fixed point.

The coincidence problem is concerned with a similar question about two maps $f, g : N \to M$ and their coincidences $x \in N, f(x) = g(x)$. One of the main tools is the Lefschetz coincidence number λ_{fg} defined similarly to λ_f as the alternating sum of traces of a certain endomorphism on the homology group of M. Algebraically, if $h : E_* \to E_*$ is a (degree 0) endomorphism of a finitely generated graded vector space $E_* = \{E_k\}$, given by $h_k : E_k \to E_k$, then its Lefschetz number is $L(h) = \sum_k (-1)^k \mathrm{Trace}(h_k)$. To apply this formula in the topological setting we let $E_* = H_*(M)$; then the Lefschetz number is defined as $\lambda_{fg} = L(g_* D_N f^* D_M^{-1})$, where $D_M : H^k(M, \partial M) \to H_{n-k}(M)$, $D_N : H^k(N, \partial N) \to H_{n-k}(N)$ are the Poincaré duality isomorphisms. Observe that for $f^* : H^k(M, \partial M) \to H^k(N, \partial N)$ to be well defined, the map f has to be boundary preserving, $f : (N, \partial N) \to (M, \partial M)$.

A *Lefschetz-type coincidence theorem* states that if $\lambda_{fg} \neq 0$, then the pair (f, g) (and any pair homotopic to them) has a coincidence. The converse is false in general. When $\lambda_{fg} = 0$, the maps f, g may have coincidences, but under certain circumstances they can be removed by homotopies of f, g [5].

Until the 1990's, such theorems have been mostly considered in the following two settings. First [2, sect. VI.14], [35, Ch. 7], $f : N \to M$ is a map between two n-manifolds as above. This way the Lefschetz theorem can be applied to detect equilibria of a dynamical system, but it does not apply to an even simplest control system because the dimensions of $N = M \times U$ and M have to be equal. Second [15], $f : X \to M$ is a map from an arbitrary topological space X to an open subset of $\mathbf{R}^n$, and all fibers $f^{-1}(y)$ are acyclic. Here the dimensions are also equal in the sense that $H_*(X) = H_*(M)$ (Vietoris theorem). Thus neither case is broad enough to cover control systems, the input spaces U of which have nonzero dimension.

As an example from dynamics, consider the problem of existence of closed orbits of a flow. The flow is given by a map $f : M \times [0, \infty) \to M$ so that the initial position is $f(0, x) = x$ and $f(t, x)$ is the position at time t. Closed orbits correspond to coincidences of f and the projection $p : M \times [0, \infty) \to M$. More generally, one considers $f : M \times X \to M$, where X is a topological space. This situation was studied in [24], [12], [13], [11] under the name "parametrized fixed point theory." These results can be applied to detect equilibria (section 6), but the setting is not general enough to study controllability (section 7). The author [30], [31] extended some of the results of [13] to the general case of two arbitrary maps $f, g : N \to M$ from an arbitrary topological space to a manifold. The content of these papers is briefly outlined in the next section.

4. Detecting coincidences. In this section, N is an arbitrary topological space, $A \subset N$, M is an orientable compact connected manifold with boundary ∂M, $\dim M = n$, and $f : (N, A) \to (M, \partial M)$, $g : N \to M$ are maps.

The generalization of the Lefschetz number is based on the fact that since the finitely generated graded vector space $E = H_*(M)$ is equipped with the cap product $\frown : E^* \otimes E_* \to E_*$, one can define the Lefschetz class $L(h) \in E_*$ of a graded endomorphism h given by $h_k : E_k \to E_{k+m}$ of any degree m, not just of degree 0 as in the classical case.

DEFINITION 4.1 (see [31, Proposition 2.2]). *If $h : H_k(M) \to H_{k+m}(M)$, $k = 0, 1, 2, \dots$, is a graded homomorphism of degree m, then the Lefschetz class $L(h) \in H_m(M)$ is defined as*

$$L(h) = \sum_k (-1)^{k(k+m)} \sum_j x_j^k \frown h(a_j^k),$$

where $\{a_1^k, \dots, a_{m_k}^k\}$ is a basis for $H_k(M)$ and $\{x_1^k, \dots, x_{m_k}^k\}$ the corresponding dual basis for $H^k(M)$.

It is easy to see that if the degree m of h is zero, $L(h) = \sum_k (-1)^k \mathrm{Trace}(h_k)$.

For a given $z \in H_s(N, A)$, suppose the homomorphism h_{fg}^z is defined as the composition

$$H_i(M) \xrightarrow{D_M^{-1}} H^{n-i}(M, \partial M) \xrightarrow{f^*} H^{n-i}(N, A) \xrightarrow{\frown z} H_{s-n+i}(N) \xrightarrow{g_*} H_{s-n+i}(M),$$

i.e.,

$$h_{fg}^z(x) = g_*((f^* D_M^{-1}(x)) \frown z).$$

Its degree is $m = s - n$.

DEFINITION 4.2. *The Lefschetz homomorphism $\Lambda_{fg} : H_s(N, A) \to H_{s-n}(M)$, $k = 0, 1, \dots$, of the pair (f, g) is defined by*

$$\Lambda_{fg}(z) = L(h_{fg}^z).$$

The degree of the homomorphism h_{fg}^z is zero if $z \in H_n(N, A)$. If, moreover, N is an orientable compact connected manifold of dimension n, we have $H_n(N, \partial N) = \mathbf{Q}$. Its identity is the fundamental class $O_N \in H_n(N, \partial N)$ of N. Since $D_N(x) = x \frown O_N$, we recover the classical *Lefschetz number*, $\lambda_{fg} = \Lambda_{fg}(O_N)$.

THEOREM 4.3 (see [31, Theorem 6.1]) (existence of coincidences). *If $\Lambda_{fg} \neq 0$, then any pair of maps f', g' homotopic to f, g has a coincidence.*

Especially important for the control theory applications are the following two corollaries. They are applied to existence of equilibria (section 6) and controllability (section 7), respectively. Observe that the second corollary is about a map of pairs and the first is not.

COROLLARY 4.4 (existence of fixed points) (cf. [13]). *Let $g : M \times U \to M$ be a map. Given $v \in H_s(U)$, suppose the homomorphism $g_v : H_i(M) \to H_{i+s}(M)$, $i = 0, 1, \dots$, of degree s is defined by*

$$g_v(x) = (-1)^{(n-i)s} g_*(x \otimes v),$$

$x \in H_i(M)$. Then, if

$$L(g_v) \neq 0 \text{ for some } v \in H_s(U),$$

then any map $g' : M \times U \to M$ homotopic to g has a fixed point x, $g'(x, u) = x$ for some u.

Proof. Let $(N, A) = (M, \partial M) \times U$, and apply Theorem 4.3 to the pair p, g, where $p : (M, \partial M) \times U \to (M, \partial M)$ is the projection. Also, according to Corollary 5.7 in [31], $\Lambda_{pg}(O_M \otimes v) = L(g_v)$. ∎

COROLLARY 4.5 (sufficient condition of surjectivity). *If*

$$f_* : H_n(N, A) \to H_n(M, \partial M) = \mathbf{Q} \text{ is nonzero,}$$

then any map $f' : (N, A) \to (M, \partial M)$ homotopic to f is onto.

Proof. Apply Theorem 4.3 to the pair f, c, where c is any constant map (as in section 5 in [30] and Proposition 6.8 in [31]). □

In the case of manifolds of equal dimensions, the condition of this corollary is equivalent to the nonvanishing of the degree $\deg f$ [2, p. 186] of f.

5. Removing coincidences. In this section, M is a compact orientable connected manifold with boundary ∂M, $\dim M = n$, N is a manifold, and $f, g : N \to M$ are maps.

When $\dim N = \dim M = n > 2$, the vanishing of the Lefschetz number λ_{fg} implies that the coincidence set can be removed by homotopies of f, g [5]. If $\dim N = n + m, m > 0$, this is no longer true even if λ_{fg} is replaced with Λ_{fg}. Some progress has been made for $m = 1$. In this case the secondary obstruction to the removability of a coincidence set was considered in [10], [9], [17]. These results can be used to study removability of equilibria when the dimension of the input space is 1. However, the conditions on f and g are hard to verify. Necessary conditions of the global removability for arbitrary m were considered in [14, section 5] with N a torus and M a nilmanifold. For some $m > 1$, a partial converse of Theorem 4.3 is provided by the author [32]. A version of this theorem is given below.

Suppose F is an isolated subset of the coincidence set of f, g and $f(F) = g(F) = \{x\}$, $x \in M \backslash \partial M$. Let D be a open neighborhood of x such that $D \cap \partial M = \varnothing$. Choose a neighborhood W of F in N with no coincidences such that $f(W) \subset D$ and $g(W) \subset D$. Suppose $V \subset \overline{V} \subset W$ is another neighborhood of F; then there is an open neighborhood $B \subset \overline{B} \subset D$ of x such that $f(W \backslash V) \subset D \backslash B$. Therefore $f : (W, W \backslash V) \to (D, D \backslash B)$ is a map of pairs.

THEOREM 5.1 (local removability of coincidences). *Suppose the following property is satisfied:*

$$(*)\ H^{k+1}(W, W \backslash V; \pi_k(\mathbf{S}^{n-1})) = 0 \text{ for } k \geq n+1.$$

Suppose also that

$$f_* : H_n(W, W \backslash V) \to H_n(D, D \backslash B) = \mathbf{Q} \text{ is zero.}$$

Then there is a homotopy of f constant on the complement of V to a map f' such that the new pair has no coincidences in V.

Since D is arbitrary we can say that the homotopy can be chosen *arbitrarily small.*

Proof. According to the proof of Theorem 2 in [32] the coincidence subset F can be removed by a homotopy of f constant on $N \backslash V$, provided the local cohomology index $I^W_{fg}(\tau)$ vanishes. This index is defined as follows. Since $F \subset V$ is the set of all coincidences in W, the map $(f, g) : (W, W \backslash V) \to D^\times = (D \times D, D \times D \backslash d(D))$, where $d(D)$ is the diagonal of $D \times D$, is well defined. Therefore the homomorphisms $(f, g)_* : H_k(W, W \backslash V) \to H_k(D^\times)$ and $(f, g)^* : H^k(D^\times) \to H^k(W, W \backslash V)$ are also well defined. Now let I_{fg} be the homology coincidence homomorphism defined by $I_{fg} = (f, g)_* : H_k(W, W \backslash V) \to H_k(D^\times)$. Let $I^W_{fg}(\tau) = (f, g)^*(\tau) \in H^n(W, W \backslash V)$ be the cohomology coincidence index [32, section 2], where τ is the identity of $H^n(D^\times) = \mathbf{Q}$. By Theorem 6.1 in [31], $\Lambda_{fg}(z) = \pi_*(\tau \frown I_{fg}(z))$, where $\pi : D \times D \to D$ is the projection on the first factor. Then, for any $z \in H_n(W, W \backslash V)$,

$$\begin{aligned} \Lambda_{fg}(z) &= \pi_*(\tau \frown (f, g)_*(z)) = \pi_*(f, g)_*((f, g)^*(\tau) \frown z) \\ &= \langle (f, g)^*(\tau), z \rangle = \left\langle I^W_{fg}(\tau), z \right\rangle. \end{aligned}$$

Therefore $I^W_{fg}(\tau) = 0$ if and only if $\Lambda_{fg}(z) = 0$ for all $z \in H_n(W, W \backslash V)$. Finally, observe that $g|_W$ is homotopic to a constant map. Therefore $f_* = 0$ if and only if $\Lambda_{fg}(z) = 0$ for all $z \in H_n(N, A)$ (section 5 in [30]). □

Condition (*) ensures that only the primary obstruction to removability, i.e., the Lefschetz number, can be nonzero. Further investigation of necessary conditions of removability of coincidences will require computing higher order obstructions. The case when $f(F)$ is not a single point is best addressed in the context of Nielsen theory via Wecken-type theorems [33]. In general, the homotopy of f cannot be always chosen arbitrarily small.

Especially important for the control theory applications are the following corollaries. They are applied to the disappearance under perturbations of equilibria (section 6) and controllability (section 7), respectively.

COROLLARY 5.2 (removability of fixed points). *Suppose the conditions of Theorem* 5.1 *are satisfied for $N = M \times U$, where U is a manifold, $x \in M \backslash \partial M$ is an isolated fixed point of $f : M \times U \to M$ (i.e., $f(x,u) = x$ for some $u \in U$), and $F = \{x\} \times \{u \in U : g(x,u) = x\}$. Then there is a homotopy of f to a map f' such that f' has no fixed points in a neighborhood of F. The homotopy can be chosen arbitrarily small and constant on the complement of a neighborhood of F.*

Proof. If $g : M \times U \to M$ is the projection, then F is the coincidence set of f, g. □

COROLLARY 5.3 (necessary condition of surjectivity). *Suppose the conditions of Theorem* 5.1 *are satisfied for $F = f^{-1}(x)$ of $f : N \to M$. Then there is a homotopy of f to a map f' which is not onto; specifically, $x \notin f'(N)$. The homotopy can be chosen arbitrarily small and constant on the complement of a neighborhood of F.*

Proof. If g is the constant map, then F is the coincidence set of f, g. □

These two corollaries are partial converses of Corollaries 4.4 and 4.5, respectively.

A submanifold F of N satisfies condition (*) if one of the following three conditions holds [32, section 4]:

- (a1) *M is a surface, i.e., $n = 2$; or*
- (a2) *F is acyclic, i.e., $H_k(F) = 0$ for $k = 1, 2, \ldots$; or*
- (a3) *every component of F is a homology m-sphere, i.e., $H_k(F) = 0$ for $k \neq 0, m$, for the following values of m and n:*
  (1) *$m = 4$ and $n \geq 6$;*
  (2) *$m = 5$ and $n \geq 7$;*
  (3) *$m = 12$ and $n = 7, 8, 9$, or $n \geq 14$.*

6. Existence of equilibria. In this section, M is a compact orientable connected manifold with boundary ∂M, $\dim M = n$, and U is a topological space.

A *discrete time control system* D_g is given by a map $g : M \times U \to M$, with U the *space of inputs* and M the *space of states* of the system.

We say that $D_{g'}$ is a *perturbation* of D_g if g' homotopic to g. To justify this definition, recall that a system $D_{g'}$ is normally called a perturbation of D_g if g' is "close enough" to g in terms of the distance between $g(z)$ and $g'(z)$. However, if g' is a simplicial approximation of g [2, p. 251], then g and g' are homotopic. Thus we permit large but continuous perturbations of the system. Properties preserved under such perturbations may be called *strongly robust*.

As before, suppose $\{a_1^k, \ldots, a_{m_k}^k\}$ is a basis for $H_k(M)$ and $\{x_1^k, \ldots, x_{m_k}^k\}$ the corresponding dual basis for $H^k(M)$.

THEOREM 6.1 (existence of equilibria). *If*

$$L(g_v) = (-1)^{ns} \sum_k (-1)^k \sum_j x_j^k \frown g_*(a_j^k \otimes v) \neq 0 \text{ for some } v \in H_s(U),$$

then every perturbation of the discrete time system D_g has an equilibrium.

Proof. In light of Corollary 4.4 we need only to show that the above formula for the Lefschetz class $L(g_v)$ of $g_v(x) = (-1)^{(n-i)s} g_*(x \otimes v)$, $x \in H_i(M)$, is true. Since the degree of g_v is s and $a_j^k \in H_k(M)$, we substitute $m = s$ and $i = k$ in Definition 4.1:

$$\begin{aligned} L(g_v) &= \sum_k (-1)^{k(k+s)} \sum_j x_j^k \frown (-1)^{(n-k)s} g_*(a_j^k \otimes v) \\ &= \sum_k (-1)^{k^2+ns} \sum_j x_j^k \frown g_*(a_j^k \otimes v), \end{aligned}$$

and the formula follows. □

The following is a generalization of a well-known theorem about dynamical systems.

COROLLARY 6.2. *Suppose D_g is a perturbation of the constant system D_p; i.e., $p(x,u) = x$ for all u. If the Euler characteristic of M is nonzero, $\chi(M) \neq 0$, then D_g has an equilibrium.*

Proof. Since $p_*(a_j^k \otimes v) = a_j^k$ if $v = 1$ and 0 otherwise, we have

$$\begin{aligned} L(g_v) &= \sum_k (-1)^k \sum_j x_j^k \frown p_*(a_j^k \otimes v) \\ &= \sum_k (-1)^k \sum_j 1 \\ &= \sum_k (-1)^k m_k \\ &= \chi(M). \end{aligned}$$ □

COROLLARY 6.3. *Suppose $M = \mathbf{S}^n$, and suppose one of the following conditions is satisfied:*

(1) *$g_*(d \otimes 1) \neq (-1)^{n+1} d$, where d is the identity of $H_n(\mathbf{S}^n)$; or*

(2) *$g_*(1 \otimes v) \neq 0$ for some $v \in H_n(U)$.*

Then every perturbation of the discrete time system D_g, $g : \mathbf{S}^n \times U \to \mathbf{S}^n$, has an equilibrium.

Proof. Let us compute $L(g_v)$ for an arbitrary $v \in H_s(U)$. As $a_j^k \in H_k(M)$, we have $a_j^k \otimes v \in H_{k+s}(M \times U)$ and $g_*(a_j^k \otimes v) \in H_{k+s}(M)$. Since $H_i(M) = H_i(\mathbf{S}^n) = 0$ for all $i \neq 0, n$, we have $g_*(a_j^k \otimes v) = 0$, except for the following two cases. (1) Choose $v = 1 \in H_0(U), s = 0$; then either $k = 0, a_j^0 = 1, x_j^0 = 1$, or $k = n, a_j^n = d, x_j^n = \overline{d}$. (2) Choose $v \in H_n(U), s = n$; then $k = 0, a_j^0 = 1, x_j^0 = 1$. Here $\overline{d}$ is the dual of d, $\overline{d} \frown d = 1$. Thus we have

$$\begin{aligned} (1)\ L(g_1) &= (-1)^{n0}(1 \frown g_*(1 \otimes 1) + (-1)^n \overline{d} \frown g_*(d \otimes 1)) \\ &= 1 + (-1)^n \overline{d} \frown g_*(d \otimes 1); \\ (2)\ L(g_v) &= (-1)^{nn}(1 \frown g_*(1 \otimes v)) \\ &= (-1)^n g_*(1 \otimes v). \end{aligned}$$

Now, if either $L(g_1)$ or $L(g_v)$ is nonzero, then D_g has an equilibrium by Theorem 6.1. □

Condition (1) means that the degree of $\overline{g}(\cdot) = g(\cdot, u_0) : \mathbf{S}^n \to \mathbf{S}^n$ is not equal to $(-1)^{n+1}$.

If $U = M$ is a compact Lie group and $g : M \times M \to M$ is the multiplication, then D_g has a equilibrium [13, Ex. 2.3]. For more examples, see [13], [30], [31].

In the control setting, Corollary 5.2 reads as follows.

THEOREM 6.4 (removability of equilibria). *Suppose U is a manifold, and suppose $x \in M\backslash\partial M$ is an isolated equilibrium of D_g. Suppose condition $(*)$ is satisfied for $F = \{x\} \times \{u \in U : g(x,u) = x\}$ and*

$$f_* : H_n(W, W\backslash V) \to H_n(D, D\backslash B) = \mathbf{Q} \text{ is zero,}$$

where $V \subset \overline{V} \subset W$ and $B \subset \overline{B} \subset D \subset M\backslash\partial M$ are neighborhoods of F and x, respectively. Then this equilibrium can be removed by an arbitrarily small perturbation restricted to a neighborhood of F.

7. Controllability. In this section, M is a compact orientable connected manifold with boundary ∂M, $\dim M = n$, and U is a topological space.

Suppose a discrete system D_f is given by $f : M \times U \to M$. The system D_f is called *controllable* [34] if any state can be reached from any other state by means of f; i.e., for each pair of states $x, y \in M$ there are inputs $u_0, \dots, u_r \in U$ such that $x_1 = f(u_0, x), x_2 = f(u_1, x_1), \dots, y = x_{r+1} = f(u_r, x_r)$, notation $x \rightsquigarrow_f y$.

Below, this notion is generalized in three nontypical, but topologically appropriate, ways. First, we consider the possibility of an arbitrary state reached not from any given state but from a state in a particular subset L of M. Second, as before, we permit arbitrary, not necessarily small, perturbations of f. Third, instead of looking into controllability of a new, perturbed system D_g, where g is homotopic to f, we allow for consecutive applications of possibly different maps each homotopic to f.

DEFINITION 7.1. *Given $L \subset M$, let $f' : L \times U \to M$ be the restriction of f. Then the system is called* strongly robustly controllable from L *if for any map f_0 homotopic to f', any maps $f_1, \dots, f_r$ homotopic to f, and for each $y \in M$ there is $x \in L$ and inputs $u_0, \dots, u_r \in U$ such that*

$$x_1 = f_0(x, u_0), x_2 = f_1(x_1, u_1), \dots, y = x_{r+1} = f_r(x_r, u_r).$$

Then the system is controllable if it is controllable from any point.

It is clear that controllability is equivalent to surjectivity of several iterations of f. To deal with surjectivity we apply Corollary 4.5, which requires f to be a map of pairs. For this purpose, in this section we make the following assumption about D_f. If the initial state lies at the boundary ∂M of M, then the next state, regardless of the input, lies within a certain neighborhood of ∂M. For simplicity we make a topologically equivalent assumption,

$$f(\partial M \times U) \subset \partial M.$$

Next, let U' be the set of controls that take any given state to the boundary of M, i.e.,

$$U' = \{u \in U : f(x,u) \in \partial M \text{ for all } x \in M\}.$$

Then $f(M \times U') \subset \partial M$. Combining this with the above assumption we conclude that f is a map of pairs, $f : (M, \partial M) \times (U, U') \to (M, \partial M)$. Let $L' = L \cap \partial M$; then $f' : (L, L') \times (U, U') \to (M, \partial M)$ is also a map of pairs.

The following theorem translates the above "reachability" condition into the language of homology: any element of $H_n(M, \partial M) = \mathbf{Q}$ can be reached from some $a_0 \in H_*(L, L')$ by means of f_*.

THEOREM 7.2 (sufficient condition of robust controllability). *Suppose that there are* $a_0 \in H_p(L, L')$, $v_0 \in H_{s_0}(U, U'), \dots, v_r \in H_{s_r}(U, U')$ *such that*

$$a_1 = f'_*(a_0 \otimes v_0), a_2 = f_*(a_1 \otimes v_1), \dots, a_{r+1} = f_*(a_r \otimes v_r) \in H_n(M, \partial M) \backslash \{0\}.$$

Then the discrete time system D_f *is strongly robustly controllable from* L.

Here, if $a_i \in H_{n_i}(M, \partial M)$, $i = 0, 1, 2, \dots, r$, then $n_0 = p, n_1 = p + s_0, n_2 = n_1 + s_1, \dots, n_{r+1} = n_r + s_r = n$. Thus we have a sequence of homology classes $a_0, \dots, a_r$ of $(M, \partial M)$ "climbing" dimensions from p to n.

Proof. The result of consecutive applications of f is defined as a map $F : (L, L') \times (U, U')^{r+1} \to (M, \partial M)$ given by

$$F(x, u_0, \dots, u_r) = f(\dots f(f'(x, u_0), u_1), \dots, u_r);$$

i.e., it is given by the composition

$$\begin{aligned} F : (L, L') \times (U, U') \times \dots \times (U, U') &\xrightarrow{f' \times Id} \\ (M, \partial M) \times (U, U') \times \dots \times (U, U') &\xrightarrow{f \times Id} \dots. \end{aligned}$$

Then $x \rightsquigarrow_f F(x, u_0, \dots, u_r)$. Suppose a map f_0 is homotopic to f' and maps $f_1, \dots, f_r$ are homotopic to f. The result of consecutive applications of $f_0, \dots, f_r$ is defined as a map $G : (L, L') \times (U, U')^{r+1} \to (M, \partial M)$ given by

$$G(x, u_0, \dots, u_r) = f_r(\dots f_1(f_0(x, u_0), u_1), \dots, u_r).$$

Therefore strong robust controllability from L means that $G : L \times U^{r+1} \to M$ is onto. By Corollary 4.5, if

$$F_* : H_n((L, L') \times (U, U') \times \dots \times (U, U')) \to H_n(M, \partial M) = \mathbf{Q}$$

is nonzero, then every map homotopic to F is onto. Since G is clearly homotopic to F, all we need to prove is that F_* is nonzero. By the Künneth theorem, F_* is given by the composition

$$\begin{aligned} F_* : H_*(L, L') \otimes H_*(U, U') \otimes \dots \otimes H_*(U, U') &\xrightarrow{f'_* \otimes Id} \\ H_*(M, \partial M) \otimes H_*(U, U') \otimes \dots \otimes H_*(U, U') &\xrightarrow{f_* \otimes Id} \dots. \end{aligned}$$

Now the condition of Theorem 7.2 implies that $f_*(\dots f_*(f'_*(a_0 \otimes v_0) \otimes v_2) \otimes \dots \otimes v_r) \neq 0$ for some $a_0 \in H_p(L, L')$ and some $v_0 \in H_{s_1}(U, U'), \dots, v_r \in H_{s_r}(U, U')$ such that $p + s_1 + \dots + s_r = n$. Therefore $F_*(a_0 \otimes v_0 \otimes v_2 \otimes \dots \otimes v_r) \neq 0$. □

Moreover, it is clear that what we have is the "finite time reachability"; i.e., every state can be reached in a finite number, $r + 1$, of steps, and that number is common for all states.

Theorem 7.2 involves multiple iterations of f_*, while it is preferable to have a condition involving only f_* itself. Let us consider a case when this is possible.

Consider first a simple example, where $U = \mathbf{S}^1$, $U' = \varnothing$, $M = \mathbf{T}^n = (\mathbf{S}^1)^n$, and $f : \mathbf{S}^1 \times \mathbf{T}^n \to \mathbf{T}^n$ is given by $f(u, x_1, \dots, x_n) = (u, x_1, \dots, x_{n-1})$. This may serve as a model for a robotic arm with n joints where only the first joint can be controlled directly and the next state of a joint is "read" from the current state of the previous joint. The system is obviously controllable. Indeed, after n iterations with inputs $u_1, \dots, u_n$ the system's state is $(u_n, \dots, u_1)$. Whether the system is robustly

controllable is not as obvious. The affirmative answer is provided by Theorem 7.2 as follows. Let L be a point, $p = 0$. Now, with d the identity of $H_1(\mathbf{S}^1)$ we choose

$$\begin{aligned} &v_0 = v_1 = \cdots = v_n = d \in H_1(\mathbf{S}^1), \text{ and} \\ &a_0 = 1 \in H_0(\mathbf{T}^n), \\ &a_1 = d \in H_1(\mathbf{T}^n), \\ &a_2 = d \otimes d \in H_2(\mathbf{T}^n), \\ &\quad \cdots \\ &a_n = d \otimes \cdots \otimes d \in H_n(\mathbf{T}^n). \end{aligned}$$

More generally, suppose the state space M has the product structure, $M = K_1 \times \cdots \times K_s$, where K_i are manifolds of dimensions k_i. Suppose $f = (h_1, \dots, h_s)$, where $h_i : U \times M \to K_i$. For $i = 1, \dots, s$, define maps $h_i^a : K_{i-1} \to K_i$, where $K_0 = U$, by $h_i^a(x_{i-1}) = h_i(a_0, \dots, a_{i-2}, x_{i-1}, a_i, \dots, a_s)$. If all h_i^a are onto, then the system is controllable. According to Corollary 4.5 it suffices to require that all $h_{i*}^a : H_{k_i}(K_{i-1}) \to H_{k_i}(K_i)$ are nonzero, $i = 1, \dots, s$.

Theorem 7.2 can be informally understood as follows. If there are some submanifolds $M_1, \dots, M_r$, $\dim M_i = n_i$, of M such that $M_0 = L, M_1 = f(M_0 \times U)$, $M_2 = f(M_1 \times U), \dots, M = f(M_r \times U)$, then the system is controllable. It means that the restrictions $f_0 : L \times U \to M_1, f_1 : M_1 \times U \to M_2, \dots, f_r : M_r \times U \to M$ of f are surjective. This holds, provided $f_{i*}(O_{M_i} \otimes O_U) = q_i O_{M_{i+1}}$, where $O_{M_i} \in H_{n_i}(M_i)$ is the fundamental class of M_i, for some $q_i \in \mathbf{Q}$. Since each O_{M_i} corresponds to $a_i = J_{i*}(O_{M_i}) \in H_{n_i}(M)$, where $J_i : M_i \to M$ is the inclusion, we arrive at the requirement of Theorem 7.2. The robustness of each of these surjectivity conditions can be tested by means of Corollary 5.3. As a special case we have the following.

THEOREM 7.3 (necessary condition of robust controllability). *Suppose U is a manifold and there is a fiber $F = f^{-1}(x), x \in M$, of f satisfying condition $(*)$ and*

$$f_* : H_n(W, W \backslash V) \to H_n(D, D \backslash B) = \mathbf{Q} \text{ is zero,}$$

where $V \subset \overline{V} \subset W$ and $B \subset \overline{B} \subset D \subset M \backslash \partial M$ are neighborhoods of F and x, respectively. Then there is an arbitrarily small perturbation restricted to a neighborhood of F of the system D_f which is not controllable from M; specifically, x is unreachable from any point.

Proof. Corollary 5.3 implies that there is g homotopic to f such that $x \notin g(M \times U)$. □

8. Continuous systems. In this section we outline, in fewer details than above, the possibilities of applying Lefschetz numbers to continuous systems.

In this section, M is a compact orientable connected smooth manifold with boundary ∂M, and $\dim M = n$. Let TM be the tangent bundle of M; then $\dim TM = 2n$.

A *continuous time control system* C_h [27, p. 16] is defined as a commutative diagram

$$\begin{array}{ccc} Q & \xrightarrow{h} & TM, \\ \downarrow p & \swarrow \pi_M & \\ M & & \end{array}$$

where $p : Q \to M$ is a fiber bundle over M and π_M is the projection. Thus C_h is a parametrized vector field on M.

We say that $x \in M$ is an *equilibrium* of this system if there is $y \in Q$ such that $h(y) = (x, 0) \in TM, x = p(y) \in M$. Detecting an equilibrium can be restated as a coincidence problem. Suppose $i : M \to TM$ is the inclusion and $p_1 : Q \times M \to Q$, $p_2 : Q \times M \to M$ are the projections. Define the maps $f, g : Q \times M \to TM$ by $f = hp_1, g = ip_2$. Then a coincidence of the pair f, g is an equilibrium of the system C_h. Therefore equilibria can be detected by means of the coincidence results in section 4, and their robustness can be studied by means of the results of section 5.

We have a simpler coincidence problem when M is parallelizable; i.e., TM is isomorphic to $M \times \mathbf{R}^n$. For example, $\mathbf{S}^1$, $\mathbf{S}^3$, $\mathbf{S}^7$ are parallelizable. Let $q : TM \simeq M \times \mathbf{R}^n \to M$ be the projection. Then a coincidence of the pair qh, p is an equilibrium of the system C_h, and we can use Theorem 4.3 to detect equilibria and Theorem 5.1 to study their robustness. In fact, D_{qh} is a discrete control system associated with the continuous system C_h. In particular, when $Q = M \times U$, the results of sections 6 and 7 can be applied to study equilibria and controllability of C_h.

For a general M a discrete system D_f associated with the continuous system C_h may be constructed as follows.

Let $\mathcal{A}$ be the topological space of *admissible controls* associated with C_h, i.e., a set of functions $z : [0, d] \to Q$, for all $d \in \mathbf{R}$. A map $c_z : [0, d] \to M$ is called a *trajectory* of the control system if there exists a control $z \in \mathcal{A}$ satisfying $pz = c_z$ and $\frac{d}{dt} c_z = hz$.

We assume that $Q = M \times U$, where U is the topological space of all possible inputs, and $p : Q = M \times U \to M$ is the projection. Then $\mathcal{A}$ is the set of pairs (c, p), where $c : [0, d] \to M$ is a trajectory and $p : [0, d] \to U$ is a function representing the input. To simplify things even further we consider only constant inputs. First, we assume that the system C_h satisfies the following existence and uniqueness property: for every $x \in M$ and any *constant* input $p(t) = u \in U$ there is a unique trajectory c such that $c(0) = x$ and $(c, p) \in \mathcal{A}$. Then the following end point map, $f_d : M \times U \to M$, is well defined. We let $f_d(x, u) = c(d)$, where $c : [0, d] \to M$ is the above trajectory. Assume also that the map $f = f_d$ is continuous. Then for each $d \geq 0$ we have a discrete time control system D_f.

Next, the system C_h is called *controllable* if any state can be reached from any other state; i.e., for each pair of states $x, y \in M$ there is a trajectory $c : [0, d] \to M$ such that $x = c(0), y = c(d)$.

We make the same assumption about D_f as in section 7: if the initial state lies at the boundary ∂M of M, then the next state, regardless of the input, lies within a certain neighborhood W of ∂M, or, alternatively, $f(\partial M \times U) \subset \partial M$. In particular, this condition is satisfied if $h(x, u)$ is tangent to ∂M for all $x \in \partial M$. Let U' be the set of controls that take any given state to the boundary ∂M, i.e.,

$$U' = \{u \in U : f(x, u) \in \partial M \text{ for all } x \in M\}.$$

Then f is a map of pairs, $f : (M, \partial M) \times (U, U') \to (M, \partial M)$. Given a subset L of M, let $L' = L \cap \partial M$, and let $f' : (L, L') \times (U, U') \to (M, \partial M)$ be the restriction of f.

THEOREM 8.1 (sufficient condition of controllability). *Suppose that there are* $a_0 \in H_p(L, L')$, $v_0 \in H_{s_0}(U, U'), \dots, v_r \in H_{s_r}(U, U')$ *such that*

$$a_1 = f'_*(a_0 \otimes v_0), a_2 = f_*(a_1 \otimes v_1), \dots, a_{r+1} = f_*(a_r \otimes v_r) \neq 0.$$

Then the continuous time system C_h *is controllable from* L *by means of piecewise constant controls.*

Proof. The discrete system D_f is controllable from L by Theorem 7.2. ∎

It follows also that if for a small enough $\varepsilon > 0$ a map $k : Q \to TM$ satisfies $d(k(z), h(z)) < \varepsilon$ for all $z \in Q$, where d is the distance on TM, and the system C_k satisfies all of the above assumptions, then C_k is also controllable. We can say then that C_h is *robustly controllable*.

Consider the applicability of Theorem 8.1 to local controllability or controllability in a Euclidean space. In either case, M is the n-ball. Then $H_i(M, \partial M)$ is nontrivial only in dimension n. As a result the above "chain" of homology classes $a_1, a_2, \dots, a_{r+1}$ has to have only one "link," $a_1 = f'_*(a_0 \otimes v_0) \in H_n(M, \partial M) \backslash \{0\}$. Thus Theorem 8.1 reduces to the claim of one-step controllability, provided f'_{*n} is nonzero. As a result the similarity between the homology reachability condition of Theorem 8.1 and the Lie bracket condition [27, section 3.1] does not materialize. The author believes, however, that a generalization of Theorem 7.2 will provide a necessary connection.

Observe also that if $\partial M = \varnothing$, then $f = f_d$ is homotopic to the constant map f_0 under the homotopy $H(t, x, u) = f_t(x, u)$, and hence $f_* = 0$. Therefore the condition of Theorem 8.1 is never satisfied.

Here is another approach to controllability. Let $\mathcal{A}'$ be the set of controls whose trajectories have one of the end points at the boundary of M, i.e.,

$$\mathcal{A}' = \{z : [0, d] \to Q, z \in \mathcal{A}, c_z(0) \in \partial M \text{ or } c_z(d) \in \partial M\}.$$

Define $G(u) = (c_z(0), c_z(d))$, the end points of the trajectory $c_z = pz : [0, d] \to M$ corresponding to z. Then $G : (\mathcal{A}, \mathcal{A}') \to (M \times M, \partial(M \times M))$ is a well-defined map of pairs.

THEOREM 8.2 (sufficient condition of controllability). *If*

$$G_* : H_{2n}(\mathcal{A}, \mathcal{A}') \to H_{2n}(M \times M, \partial(M \times M)) = \mathbf{Q} \text{ is nonzero,}$$

then the continuous time system C_h is controllable.

Proof. By Corollary 4.5, G is onto. □

A similar condition is found in [28], where a boundary operator $l : AC([0, 1], \mathbf{R}^n) \times L^\infty([0, 1], \mathbf{R}^n) \to \mathbf{R}^p$ is considered instead of G. One of the conditions of controllability is $\deg l_0 \neq 0$, where l_0 is the restriction of l to some p-dimensional subspace and $\deg l_0$ its topological degree.

Acknowledgment. The author thanks the referees for their comments that have helped significantly improve this paper.

REFERENCES

[1] K. BALACHANDRAN AND J. P. DAUER, *Controllability of nonlinear systems via fixed-point theorems*, J. Optim. Theory Appl., 53 (1987), pp. 345–352.
[2] G. E. BREDON, *Topology and Geometry*, Springer-Verlag, New York, 1993.
[3] R. F. BROWN, *The Lefschetz Fixed Point Theorem*, Scott-Foresman, Chicago IL, 1971.
[4] R. F. BROWN, *Fixed point theory*, in History of Topology, North–Holland, Amsterdam, 1999, pp. 271–299,
[5] R. F. BROWN AND H. SCHIRMER, *Nielsen coincidence theory and coincidence-producing maps for manifolds with boundary*, Topology Appl., 46 (1992), pp. 65–79.
[6] F. BULLO AND A. D. LEWIS, *Geometric Control of Mechanical Systems: Modeling, Analysis, and Design for Simple Mechanical Control Systems*, Springer-Verlag, New York, 2004.
[7] N. CARMICHAEL AND M. D. QUINN, *Fixed-point methods in nonlinear control*, IMA J. Math. Control Inform., 5 (1988), pp. 41–67.
[8] G. CONTI, P. NISTRI, AND P. ZECCA, *Controllability problems via set-valued maps*, in Recent Advances in Mathematical Theory of Systems, Control, Networks and Signal Processing, II (Kobe, 1991), Mita, Tokyo, 1992, pp. 253–258.

[9] D. Dimovski and R. Geoghegan, *One-parameter fixed point theory*, Forum Math., 2 (1990), pp. 125–154.
[10] F. B. Fuller, *The homotopy theory of coincidences*, Ann. of Math., 59 (1954), pp. 219–226.
[11] R. Geoghegan, *Nielsen fixed point theory*, in Handbook of Geometric Topology, R. Daverman and R. Sher, eds, North–Holland, Amsterdam, 2002, pp. 499–521.
[12] R. Geoghegan and A. Nicas, *Trace and torsion in the theory of flows*, Topology, 33 (1994), pp. 683–719.
[13] R. Geoghegan, A. Nicas, and J. Oprea, *Higher Lefschetz traces and spherical Euler characteristics*, Trans. Amer. Math. Soc., 348 (1996), pp. 2039–2062.
[14] D. Gonçalves, J. Jezierski, and P. Wong, *Obstruction Theory and Coincidences in Positive Codimension*, preprint, Bates College, Lewiston, ME, 2002.
[15] L. Górniewicz, *Topological Fixed Point Theory of Multivalued Mappings*, Math. Appl. 495., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999.
[16] D. Idczak, *Applications of the fixed point theorem to problems of controllability*, Bull. Soc. Sci. Lett. Lódz, 39 (1989).
[17] J. Jezierski, *One codimensional Wecken type theorems*, Forum Math., 5 (1993), pp. 421–439.
[18] E. A. Jonckheere, *Algebraic and Differential Topology of Robust Stability*, Oxford University Press, New York, 1997.
[19] T. Kaczynski, K. Mischaikow, and M. Mrozek, *Computational Homology*, Springer-Verlag, New York, 2004.
[20] E. Kappos, *The Conley index and global bifurcations.* I: *Concepts and theory*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 5 (1995), pp. 937–953.
[21] E. Kappos, *The role of Morse-Lyapunov functions in the design of nonlinear global feedback dynamics*, in Variable Structure and Lyapunov Control, Lecture Notes in Control and Inform. Sci. 193, A. S. I. Zinober, ed., Springer-Verlag, Berlin, 1994, pp. 249–267.
[22] E. Kappos, *Necessary conditions for the design of global feedback dynamics*, in Proceedings of the International Symposium on Nonlinear Theory and Its Applications, Las Vegas, NV, 1995.
[23] J. Klamka, *Schauder's fixed-point theorem in nonlinear controllability problems*, Control Cybernet., 29 (2000), pp. 153–165.
[24] R. J. Knill, *On the homology of the fixed point set*, Bull. Amer. Math. Soc., 77 (1971), pp. 184–190.
[25] J.-C. Latombe, *Robot Motion Planning*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1991.
[26] K. Mischaikow, *Topological techniques for efficient rigorous computations in dynamics*, Acta Numer., 11 (2003), pp. 435–477.
[27] H. Nijmeijer and A. van der Schaft, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, 1990.
[28] P. Nistri, *On a general notion of controllability for nonlinear systems*, Boll. Un. Mat. Ital. C (6), 5 (1986), 383–403 (1987).
[29] S. Sastry, *Nonlinear Systems. Analysis, Stability, and Control*, Interdiscip. Appl. Math. 10, Springer-Verlag, New York, 1999.
[30] P. Saveliev, *A Lefschetz-type coincidence theorem.* Fund. Math., 162 (1999), pp. 65–89.
[31] P. Saveliev, *The Lefschetz coincidence theory for maps between spaces of different dimensions*, Topology Appl., 116 (2001), pp. 137–152.
[32] P. Saveliev, *Removing coincidences of maps between manifolds of different dimensions*, Topol. Methods Nonlinear Anal., 22 (2003), pp. 105–114.
[33] P. Saveliev, *Higher order Nielsen numbers*, Fixed Point Theory Appl., 1 (2005), pp. 47–66.
[34] E. D. Sontag, *Mathematical Control Theory. Deterministic Finite Dimensional Systems*, 2nd ed., Texts Appl. Math. 6, Springer-Verlag, New York, 1998.
[35] J. W. Vick, *Homology Theory, An Introduction to Algebraic Topology*, Springer-Verlag, New York, 1994.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 5, pp. 1691–1711

INTRINSIC OBSERVER-BASED STABILIZATION FOR SIMPLE MECHANICAL SYSTEMS ON LIE GROUPS*

D. H. S. MAITHRIPALA†, W. P. DAYAWANSA‡, AND J. M. BERG†

Abstract. This paper presents a dynamic observer for a class of simple mechanical systems on Lie groups. This observer provides velocity estimates based on configuration measurements. The observer is *intrinsic*, so its performance does not depend on the choice of coordinates, and it is *coordinate free*, in the sense that the equations may be written explicitly without specifying coordinates for the configuration space. Our main result is obtained by specializing a previous result of Aghannan and Rouchon concerning velocity estimation of simple mechanical systems on Riemannian manifolds to such systems on Lie groups. This specialization is nonobvious and extremely powerful. Further we extend the original result to include velocity-dependent external forces. This estimator, combined with a coordinate-free formulation of passivity-based state-feedback control, allows the construction of a coordinate-free, intrinsic dynamic output feedback compensator. This is, to our knowledge, the first time such a result has been reported. Explicit expressions are computed for the Lie groups $SO(3)$ and $SE(3)$, allowing easy specialization to practical problems of rigid body motion. The theory is illustrated via application to the axisymmetric top and to a six-degrees-of-freedom microelectromechanical system.

Key words. nonlinear observers, mechanical systems, Lie groups, MEMS

AMS subject classifications. 93B29, 93B51, 93C41

DOI. 10.1137/S0363012904439891

1. Introduction.

The traditional approach to nonlinear control has been to extend the extremely successful concepts developed for linear systems. This tactic has led to notable success, but it is inherently limited by the great variety of nonlinear phenomena. An alternative is to exploit the structural properties of specific classes of nonlinear systems. In particular, the systematic geometric study of *mechanical control systems* has received much attention. Formally, a holonomic simple mechanical system consists of (i) a smooth manifold corresponding to the configuration space of the system, (ii) a smooth Lagrangian corresponding to kinetic energy minus potential energy, and (iii) a set of external forces or one-forms [1]. When some of these forces may be used for control, we refer to a simple mechanical *control* system [6, 24]. The study of mechanical systems from a modern geometric point of view can be found, for example, in the excellent texts of Abraham and Marsden [1], Arnold [5], Bloch et al. [12], Bullo and Lewis [15], and Marsden and Ratiu [29]. Certain important nonlinear optimal control problems naturally lead to the consideration of such systems. The relationship between optimal control and mechanics is explored in essays by Bloch and Crouch [8] and Jurdjevic [22]. Work by Bloch, Leonard, and Marsden [9], Bloch and Leonard [11], Bloch [12], and the references therein, develops the notion of controlled Lagrangian systems, in which Lagrangian systems are stabilized by symmetry-preserving kinetic energy shaping and damping injection in such a way

*Received by the editors January 15, 2004; accepted for publication (in revised form) June 24, 2005; published electronically November 22, 2005. This work was supported by National Science Foundation grants ECS0218345 and ECS0220314.

http://www.siam.org/journals/sicon/44-5/43989.html

†Department of Mechanical Engineering, Texas Tech University, Lubbock, TX 79409 (sanjeeva.maithripala@ttu.edu, jordan.berg@ttu.edu).

‡Department of Mathematics and Statistics, Texas Tech University, Lubbock, TX 79409 (wijesuriya.dayawansa@ttu.edu).

that the closed-loop system retains a Lagrangian structure [7, 9, 11]. These results have been extended to the case where the uncontrolled mechanical system has no underlying symmetry [10]. A parallel development for controller synthesis in Hamiltonian systems is the port controlled Hamiltonian approach [17, 33, 34, 40]. Both the Lagrangian and Hamiltonian methods make extensive use of structural features such as the Riemannian metric or the Poisson structure of the system. Symmetry-preserving tracking controls have been developed recently for general control systems admitting symmetry by Martin, Rouchon, and Rudolph [30] using the geometric notion of Cartan's moving frame method. The underlying Riemannian structure is exploited by Bullo and Murray [14] to derive intrinsic tracking controls for fully actuated simple mechanical systems on a general Riemannian manifold. The tools of passivity-based control have also been extended, through a generalization of LaSalle's invariance theorem, to a class of simple mechanical systems on Riemannian manifolds [32]. In a recent paper [2], Aghannan and Rouchon present an intrinsic observer that, given measurements of the configuration variables of a simple mechanical system on a Riemannian manifold, provides estimates of the states (both configurations and velocities). This is accomplished with a reformulation of the Luenberger observer, where the observation error is defined intrinsically by the geodesic distance between the actual and estimated configuration variables. These powerful results concerning the dynamics and control of systems on manifolds are intrinsic, implying that the performance will not depend on the choice of coordinates. However, on a general Riemannian manifold, their explicit expression requires coordinates. On a Lie group, however, due to the natural identification of tangent spaces and neighborhoods of a point by left-translation, coordinate-free explicit expressions may be feasible.

Simple mechanical control systems on Lie groups provide a rich source of control problems. Some examples include underwater vehicles, satellites, surface vessels, airships, hovercraft, and robots [3, 4, 11, 13, 41, 42]. Simple mechanical systems on Lie groups are also interesting as *subsets* of more complex interconnected systems. An example is a model of an electrostatically actuated microelectromechanical system (MEMS) with a mechanical subsystem represented as a simple mechanical system on the Lie group $SE(3)$, and an electrical subsystem without such an additional structure [25]. Systems on Lie groups exhibiting symmetry can be further exploited. When the Lagrangian of the system is invariant under the left action of the group on itself, and the external forces acting on the system are left invariant, a reduction of dynamics to the Lie algebra of the Lie group is immediate [11, 13, 29]. In this case no coordinates need to be introduced on the Lie group to express the system and control strategies. Open-loop, coordinate-free, motion planning algorithms using small amplitude forcing have been developed for underactuated systems with such symmetry [13]. One of the main contributions of the present work is to show how, with left-invariant kinetic energy but without additional symmetry constraints, explicit formulas for intrinsic, dynamic output feedback controllers may be obtained without introducing coordinates on the Lie group. For example, there is a wide class of mechanical systems on Lie groups for which the forces are dissipative or nonleft invariant. Examples include MEMS, robot manipulators, land vehicles, and general three-dimensional motion in a gravitational field with or without damping.

Specifically we specialize to Lie groups two intrinsic results for control [32] and estimation [2] of simple mechanical systems on general Riemannian manifolds. We give coordinate-free explicit expressions of these results valid on any simple mechanical system on a Lie group with left-invariant kinetic energy. For a given Lie group the key computations required are the Levi-Civita connection, the Riemannian curvature,

and an approximation to the associated distance function and parallel transport. The computation of these quantities require only a choice of coordinates for the Lie algebra of the Lie group. Once these quantities are computed they can be used for any simple mechanical system with left-invariant kinetic energy, merely by specifying the particular kinetic energy tensor and external forces. Unless the external forces are left invariant, the expression of the force terms may require coordinates on the Lie group.

In section 2 below we briefly review the necessary mathematical background. In section 3 we first employ the results of [32] to derive passivity-based full state-feedback control for uncoupled simple mechanical systems on Lie groups. We then derive similar results for simple mechanical systems on Lie groups that are coupled to a system on a general manifold by generalizing the results of [16].

Often in applications, configuration and velocity are not both easily measured. In some cases the velocities are available [4, 31, 36], while in others, it is the configuration [3, 37, 39]. In section 4 we consider the latter case, where the configuration variables are measured and the velocities must be estimated. Section 4 presents a coordinate-free explicit expression of the results of [2]. We apply feedback passivation followed by damping injection. Since the original result in [2] assumes all forces to be only configuration dependent, we require an extension of that result to accommodate velocity- and configuration-dependent forces. The extension is presented in the appendix.

Section 4 concludes with a statement of a separation principle for the dynamic output feedback law resulting from the combination of this estimator with the uncoupled passivity-based controller of section 3.

In section 5 we specifically compute and apply the dynamic output feedback controller to representative systems on two Lie groups of special interest, namely, the rotation group $SO(3)$ and the Euclidean motion group $SE(3)$. The expressions given in this section may be applied to many problems of practical significance arising from rigid body motion by specializing only the inertia tensor and the external force terms. The classical axisymmetric top problem is used to demonstrate the construction and performance of the observer on $SO(3)$. An example on $SE(3)$ models setpoint control in the presence of a saddle-node bifurcation for a MEMS. Simulation results show excellent performance.

2. Mathematical background. This section briefly describes the notation and several geometric notions that will be used in the rest of the paper. For additional details the reader is referred to the texts of [1, 15, 12, 18, 20, 21, 29, 35]. Let G be a connected Lie group and let $\mathcal{G} \simeq T_eG$ be its Lie algebra. The left-translation of $\zeta \in \mathcal{G}$ to T_gG will be denoted $g \cdot \zeta = DL_g\,\zeta$. The Lie bracket on $\mathcal{G}$ for any two $\zeta, \eta \in \mathcal{G}$ will be denoted $[\zeta, \eta] = ad_\zeta\,\eta$, and the dual of the ad operator will be denoted ad^*. Any smooth vector field $X(g)$ on G has the form $g \cdot \zeta(g)$ for some smooth $\zeta : G \mapsto \mathcal{G}$. Let $\{e_i\}$ be any basis for the Lie algebra $\mathcal{G}$ and let $\{E_i(g) = g \cdot e_i\}$ be the associated left-invariant basis vector field on G. Now $[e_i, e_j] = C_{ij}^k e_k$, where C_{ij}^k are the structure constants of the Lie algebra $\mathcal{G}$ ($C_{ij}^k = -C_{ji}^k$), and $[E_i, E_j] = C_{ij}^k E_k$.

2.1. The Riemannian structure. Consider a left-invariant metric $\langle\langle \cdot, \cdot \rangle\rangle$ on G. Such a metric induces a unique inner product $\langle\langle \cdot\,, \cdot \rangle\rangle_{\mathcal{G}}$ on $\mathcal{G}$ by the restriction of $\langle\langle \cdot\,, \cdot \rangle\rangle$ to T_eG. Define the isomorphism $I : \mathcal{G} \mapsto \mathcal{G}^*$ by the relation $\langle I\zeta\,,\, \eta \rangle = \langle\langle \zeta\,,\, \eta \rangle\rangle_{\mathcal{G}}$. Here $\langle \cdot, \cdot \rangle$ denotes the usual pairing between a vector and a covector. Let the matrix I be defined by $I_{ij} = \langle\langle e_i\,,\, e_j \rangle\rangle_{\mathcal{G}}$ and let I^{ij} be its inverse. I is symmetric and positive definite. In similar fashion such an I induces a unique left-invariant metric on G by the relation $\langle\langle g \cdot \zeta\,,\, g \cdot \eta \rangle\rangle = \langle I\zeta\,,\, \eta \rangle$.

The presentation that follows is based on the texts of [20, 21, 35]. Associated with any metric is a unique connection that is torsion free and metric called the Levi-Civita connection. For a vector field $X = X^k E_k$ and a vector $v = v^k E_k$ the Levi-Civita connection is given by

$$\nabla_v X = (dX^k(v) + \omega_{ij}^k(g) v^i X^j) E_k \,, \tag{2.1}$$

where $\omega_{ij}^k(g)$ are the connection coefficients in the frame $\{E_k\}$. If the metric is left invariant, then the connection coefficients are constant, given by

$$\omega_{ij}^k = \frac{1}{2}\left(C_{ij}^k - I^{ks}(I_{ir}C_{js}^r + I_{jr}C_{is}^r)\right). \tag{2.2}$$

Note that since in general E_k are not coordinate vector fields, ω_{ij}^k are not the Christoffel symbols. In the case of a left-invariant metric, the coefficients of the Riemannian curvature two-forms R_{jab}^k are also constant and can be shown to be [26]

$$R_{jab}^k = (-\omega_{rj}^k C_{ab}^r + 2\omega_{ar}^k \omega_{bj}^r). \tag{2.3}$$

We remark that they are in general different from the usual curvature coefficients that one would obtain in a coordinate frame field. The Riemannian curvature is then

$$R(\zeta\,,\,\eta)\xi = \{R_{jab}^k \xi^j (\zeta^a\,\eta^b - \zeta^b\,\eta^a) - \omega_{ij}^k C_{ab}^i \zeta^a\,\eta^b\,\xi^j\} e_k. \tag{2.4}$$

These derivations are based on Cartan's structural equations as presented in sections 9.3b–9.3e of [20].

2.1.1. The local distance function on a Lie group. Given any two points g and $\tilde{g}$ on a Riemannian manifold $(G, \langle\langle\cdot,\cdot\rangle\rangle)$, define the set of curves,

$$\Lambda(g,\tilde{g}) := \{\gamma : [0\ 1] \mapsto G \mid \gamma \text{ is piecewise smooth and } \gamma(0) = g, \gamma(1) = \tilde{g}\}. \tag{2.5}$$

Then the distance between g and $\tilde{g}$ is defined as

$$d(g,\tilde{g}) := \inf\{l(\gamma) : \gamma \in \Lambda(g,\tilde{g})\} \tag{2.6}$$

and defines a metric on the Riemannian manifold $(G, \langle\langle\cdot,\cdot\rangle\rangle)$ [21, 35]. If a C^1 curve $\gamma \in \Lambda(g,\tilde{g})$ exists such that $d(g,\tilde{g}) = l(\gamma)$, then it is referred to as a *segment*. It is known that segments are always geodesics and that any two sufficiently close points can be connected by a unique segment. In fact, since Lie groups are geodesically complete from the Hopf–Rinow theorem [21, 35] it follows that any two points on a Lie group can be joined by a geodesic.

For g and $\tilde{g}$ sufficiently close there exists a unique $\zeta_e \in \mathcal{G}$ such that

$$e := g^{-1}\tilde{g} = \exp \zeta_e. \tag{2.7}$$

Recall that $\exp s\zeta_e = e(s)$ is the one-parameter subgroup generated by ζ_e with respect to left-translation with $e(0) = id$ and $e(1) = e$. The inverse of this exponential map (2.7) defines local coordinates around g, commonly referred to as logarithmic coordinates [18]. For a fixed $g \in G$, define the function

$$f(\tilde{g}) := ||\zeta_e||_{\mathcal{G}}. \tag{2.8}$$

Since $e(s) = \exp s\zeta_e$ is a one-parameter subgroup, $f(\tilde{g})$ is the length of this curve and hence for g fixed $d(g, \tilde{g}) \leq f(\tilde{g})$. Equality holds if the metric is bi-invariant. In logarithmic coordinates, $f(\tilde{g}) = \sqrt{\zeta_e^T I \zeta_e}$ and up to order-two terms in ζ_e, $d(g, \tilde{g}) = \sqrt{\zeta_e^T I \zeta_e}$. Thus up to order-two terms, the geodesic distance between g and $\tilde{g}$ is explicitly given by the function (2.8) and is referred to as a local distance function.

The function $F(\tilde{g}) := \frac{1}{2} d^2(g, \tilde{g})$ plays a crucial role in the observer to be presented in section 4. From the above discussion it follows that up to third order $F(\tilde{g}) = \frac{1}{2} f^2(\tilde{g})$. Thus in logarithmic coordinates, up to second order, it follows that grad $F(\tilde{g}) = \tilde{g} \cdot \zeta_e$. The approximation arguments are intrinsic since smooth coordinate changes will not reduce the order of the neglected higher-order terms.

2.2. Simple mechanical control systems on Lie groups. A simple mechanical control system evolving on a Lie group G, equipped with a left-invariant metric $\langle\langle \cdot\,, \cdot \rangle\rangle$, is defined as a system with kinetic energy $E(\dot{g}) = \frac{1}{2}\langle\langle \dot{g}\,, \dot{g} \rangle\rangle$, and Lagrangian $L(g, \dot{g}) = E(\dot{g}) - U(g)$ for some smooth function $U(g)$ on G [15, 32]. A function with all nondegenerate critical points is referred to as a Morse function. For convenience in this paper we assume that $U(g)$ is a globally defined Morse function. Let $I : \mathcal{G} \mapsto \mathcal{G}^*$ be the isomorphism associated with the kinetic energy metric. Then the Euler–Lagrange equations of motion are given by

$$\dot{g} = g \cdot \zeta, \tag{2.9}$$

$$\nabla_{\dot{g}}\, \dot{g} = g \cdot I^{-1} \left(f^c(g) + f^d(g, \zeta) + \sum_i^m u_i f^i(g) \right) = g \cdot S(g, \zeta), \tag{2.10}$$

where $f^c(g)$, $f^d(g, \zeta)$, $f^i(g) \in \mathcal{G}^*$, and $u_i \in \mathcal{R}$. The conservative force $f^c(g)$ and damping force $f^d(g, \zeta)$ satisfy the conditions $\langle dU, g \cdot \xi \rangle = -\langle f^c(g), \xi \rangle$ and $\langle f^d(g, \xi), \xi \rangle \leq 0$ for any $\xi \in \mathcal{G}$. Here the $f^i(g)$ denote the control directions and are assumed to be linearly independent. The u_i are the magnitude of the forces and are the controls of the system. If $m < \dim(G)$, then the system is said to be *underactuated*, and if $m = \dim(G)$, the system is said to be *fully actuated*.

Equation (2.9) is the kinematic equation and (2.10) is the Euler–Lagrange equation of the system. These equations can also be expressed as

$$\dot{g} = g \cdot \zeta\,, \tag{2.11}$$

$$\dot{\zeta} = I^{-1} \left(ad^*_\zeta\, I\zeta + f^c(g) + f^d(g, \zeta) + \sum_i^m u_i f^i(g) \right), \tag{2.12}$$

where now (2.11)–(2.12) define a dynamical system on $G \times \mathcal{G}$, the left trivialization of TG. This formulation does not require coordinates on the Lie group G. In the case where the forcing terms do not depend on the configuration variable g, (2.12) represents a complete reduction of dynamics to $\mathcal{G}$ and is referred to as the Euler–Poincare equation. The kinematic equation (2.11) can be integrated to recover the configuration once the velocities have been solved for and hence is referred to as the reconstruction equation.

3. Passivity-based control for simple mechanical systems.

3.1. Uncoupled simple mechanical systems. The uncontrolled equilibrium points of the system are of the form $(\bar{g}, 0)$, where $\bar{g}$ is a critical point of $U(g)$. Assume that $\bar{g}$ is a local minimum. Since $U(g)$ is assumed to be a Morse function, $\bar{g}$ is in fact an

isolated local minimum. This implies that the equilibrium $(\bar{g}, 0)$ of the uncontrolled system is stable. If any of these criteria are not satisfied by the natural potential energy, for example if the desired equilibrium is not a local minima of $U(g)$, the methods of [10, 32] may be used to shape the potential energy. If the damping forces satisfy $\langle f^d(g,\zeta),\zeta\rangle < 0$, then the equilibrium $(\bar{g}, 0)$ is locally asymptotically stable. If this inequality is not strict, then the equilibrium is guaranteed only to be stable. Here we wish to enforce convergence to the equilibrium via "damping injection" control. Similarly, if the natural damping is insufficient, we may use this strategy to augment it.

For convenience assume that the damping force $f^d(g,\zeta)$ is of Rayleigh type. That is, $f^d(g,\zeta) = -R(g)\zeta$, where $R(g) : \mathcal{G} \mapsto \mathcal{G}^*$ is a map smooth in g so that the relation $\langle R(g)\zeta,\eta\rangle = \langle\langle\zeta,\eta\rangle\rangle_D$ defines a degenerate inner product on $\mathcal{G}$. This means that in a matrix representation $R(g)$ is symmetric and positive semidefinite. Let $\mathrm{Im}(B(g)) := span\{f^1(g),\dots,f^m(g)\}$.

To be *passive* a system must have a storage function satisfying the dissipation inequality with supply rate $y^T u$, where y is the system output [40]. For simple mechanical control systems the output that is compatible with passivity is completely determined and is given intrinsically by $y_i = \langle f^i(g)\,,\,\zeta\rangle$ or in a matrix representation by $y = B(g)^T\zeta$. Consider the storage function

$$H(g,\zeta) = \frac{1}{2}\langle\langle\zeta\,,\,\zeta\rangle\rangle_{\mathcal{G}} + U(g), \tag{3.1}$$

$$\dot{H} = -\langle R(g)\zeta,\zeta\rangle + \sum_i^m u_i\langle f^i(g)\,,\,\zeta\rangle \le \sum_i^m u_i\langle f^i(g)\,,\,\zeta\rangle = y^T u. \tag{3.2}$$

Thus (2.11)–(2.12) are passive with storage function H. Now consider the *damping injection* control

$$u_i = -y_i = -\langle f^i(g)\,,\,\zeta\rangle. \tag{3.3}$$

In matrix representations, (3.3) can also be written as $u = -B(g)^T\zeta$ and (3.2) as $\dot{H} = -\zeta^T R(g)\zeta - \zeta^T B(g)B(g)^T\zeta$.

Recall that Lie groups are complete metric spaces [35]. Thus, using this control, the generalized LaSalle invariance theorem of [32] guarantees that the trajectories of (2.11)–(2.12) converge to the largest invariant set of (2.11)–(2.12) contained in $\mathcal{S} := \{(g,\zeta) \mid \dot{H} = 0\}$. Let $\mathcal{N}(R(g))$ be the null space of the degenerate inner product $\langle\langle\cdot,\cdot\rangle\rangle_D$ and define $[\mathrm{Im}(B(g))]^\perp := \{\zeta \in \mathcal{G} \mid \langle f^i(g),\zeta\rangle = 0 \quad \text{for} \quad i = 1,\dots,m\}$. If at every $g \in G$, $\mathcal{N}(R(g)) \cap [\mathrm{Im}(B(g))]^\perp = \{0\}$, then $\mathcal{S} = \{(g,\zeta) \mid \zeta = 0\}$ and the largest invariant set contained in $\mathcal{S}$ consists of only the equilibrium points of the system. The equilibrium points of the system are given by the critical points of the potential energy function $U(g)$, and by assumption $(\bar{g}, 0)$ is a local nondegenerate minimum. Thus the damping control (3.3) locally asymptotically stabilizes $(\bar{g}, 0)$. In terms of a matrix representation the condition $\mathcal{N}(R(g)) \cap [\mathrm{Im}(B(g))]^\perp = \{0\}$ implies that the symmetric matrix $R(g) + B(g)B^T(g)$ is positive definite. From this point on we will assume this condition is satisfied. This is trivially the case for fully actuated systems.

We say an equilibrium is almost globally asymptotically stable if its region of attraction is an open and dense set. In particular the stabilization results are almost global if the potential energy function $U(g)$ is a globally defined smooth proper Morse function with a unique minimum at the desired equilibrium configuration $\bar{g}$ [23]. We return to this point in section 4 when we consider almost global performance of the dynamic output feedback compensator.

3.2. Coupled simple mechanical systems. Consider the product space $\mathcal{M} = \mathcal{Q} \times TG$ for some smooth manifold Q, and the class of systems on $\mathcal{M}$ of the form

$$\dot{q} = s^0(q, g, \zeta) + \sum_{i=1}^{m} s^i(q, g, \zeta) u_i, \tag{3.4}$$

$$\dot{g} = g \cdot \zeta \ , \tag{3.5}$$

$$\dot{\zeta} = I^{-1} \left(ad^*_\zeta I\zeta + f^c(g) + f^d(g, \zeta) + \sum_{i=m}^{m} f^i(g, \zeta, y) y_i \right), \tag{3.6}$$

$$y_i = h_i(q) \quad \text{for} \quad i = 1, 2, \dots, m, \tag{3.7}$$

where $q \in \mathcal{Q}$ and $s^i : \mathcal{M} \mapsto T\mathcal{Q}$ are smooth maps such that $\pi_Q \circ s^i = \pi_M$, where $\pi_Q : T\mathcal{Q} \mapsto \mathcal{Q}$ is the projection of $T\mathcal{Q}$ onto $\mathcal{Q}$, and $\pi_M : \mathcal{M} \mapsto \mathcal{Q}$ is the projection of $\mathcal{M}$ on to $\mathcal{Q}$. Furthermore let $y := [y_1\ y_2\ \cdots\ y_m]^T = [h_1\ h_2\ \cdots\ h_m]^T := h \in \mathcal{R}^m$ and $s := [s^1\ s^2\ \cdots\ s^m]$ be such that $D\,h$ is onto at every $q \in \mathcal{Q}$ and $\dim(span\{s^i\}_{i=1}^m) = m$ uniformly. If the matrix $L_s\, h := (L_{s^i} h_j)$ for $i, j = 1, 2, 3 \dots, m$ is nonsingular for all $q \in \mathcal{Q}, g \in G, \zeta \in \mathcal{G}$, then the interconnected system (3.4)–(3.7) has uniform relative degree $\{1, 1, \dots 1\}$ with respect to the outputs y_i. The uniform relative degree of the system implies that the feedback law

$$u = -[L_s\, h]^{-1} (L_{s^0}\, h - \nu) \tag{3.8}$$

is globally smooth. Thus the state feedback (3.8) input-output linearizes the system. Since $D\,h$ is full rank at each $q \in \mathcal{Q}$, the set $h^{-1}(y) \subset \mathcal{Q}$ is a smooth embedded submanifold of $\mathcal{Q}$ for each $y \in \mathcal{R}^m$. Introducing local coordinates (y, z) on $\mathcal{Q}$, the system (3.4)–(3.7) together with the control (3.8) can be expressed as

$$\dot{y} = \nu, \tag{3.9}$$

$$\dot{g} = g \cdot \zeta \ , \tag{3.10}$$

$$\dot{\zeta} = I^{-1} \left(ad^*_\zeta I\zeta + f^c(g) + f^d(g, \zeta) + \sum_{i=m}^{m} f^i(g, \zeta, y) y_i \right), \tag{3.11}$$

$$\dot{z} = N(z, g, \zeta, \nu, y). \tag{3.12}$$

The zero dynamics of the system are given by

$$\dot{g} = g \cdot \zeta \ , \tag{3.13}$$

$$\dot{\zeta} = I^{-1} \left(ad^*_\zeta I\zeta + f^c(g) + f^d(g, \zeta) \right), \tag{3.14}$$

$$\dot{z} = N(z, g, \zeta, 0, 0). \tag{3.15}$$

Consider the following candidate storage function for the input-output linearized system (3.9)–(3.11):

$$V(y, g, \zeta) = \frac{1}{2} \langle \langle \zeta, \zeta \rangle \rangle_{\mathcal{G}} + U(g) + \frac{1}{2} \sum_{i=1}^{m} {y_i}^2, \tag{3.16}$$

where the potential energy of the mechanical system is $U(g)$ and $U(g) \geq 0$ is Morse. Then specializing the results of [16, 38, 40] on passivity of interconnected subsystems, it can be shown that the control $\nu_i = -\langle f^i(g, \zeta, y)\ ,\ \zeta \rangle + w_i$ renders the input-output linearized system (3.9)–(3.11) passive with respect to the input-output pair (w, y)

and storage function V. Thus if $\bar{g}$ is a nondegenerate local minimum of $U(g)$ and $\langle f^d(g,\zeta)\ ,\ \zeta\rangle < 0$ for all $\zeta \in \mathcal{G}$, the control $w = -y$ locally asymptotically stabilizes the equilibrium $(0,\bar{g},0)$ of the input-output linearized system (3.9)–(3.11). Explicitly this control is given by

$$\nu_i = -\langle f^i(g,\zeta,y)\ ,\ \zeta\rangle - y_i. \tag{3.17}$$

Furthermore if the equilibrium of (3.15) is locally asymptotically stable with $g \equiv \bar{g}$ and $\zeta = 0$, then the equilibrium $(0,\bar{g},0)$ of the whole system (3.4)–(3.6) is locally asymptotically stable. The stability result of the composite system (3.4)–(3.6) is almost global if additionally $U(g)$ is a smooth proper Morse function with a unique minimum at the equilibrium configuration $\bar{g}$ and N is vacuous or satisfies some additional requirements given by Theorem 4.7 of [38]. Observe that if $\dim(\mathcal{Q}) = m$, then N is vacuous.

4. Intrinsic observer for velocity estimation. The controls (3.3) and (3.17) involve the feedback of both the configuration g and the velocity ζ. In the case where the configuration is available for measurement, but the velocity is not, we propose an intrinsic observer to estimate the velocity variable. This is based on the work reported in [2]. There a velocity estimate based on a configuration measurement is presented for a general Riemannian manifold and expressed in coordinates. We specialize that result to a Lie group equipped with a left-invariant metric. Our reformulation avoids the need to introduce coordinates on the Lie group and includes a proof that the external forcing can exhibit velocity dependence in addition to configuration dependence.

Consider the system given by (2.9)–(2.10). Let $(\tilde{g},\tilde{\zeta})$ be the estimated value of (g,ζ) and let $\alpha,\beta > 0$ be constant. Then it is shown in [2] that the following observer converges locally exponentially if the initial observer configuration error is sufficiently small:

$$\dot{\tilde{g}} = \tilde{g}\cdot\tilde{\zeta} - 2\alpha \operatorname{grad} F(\tilde{g}), \tag{4.1}$$

$$\nabla_{\dot{\tilde{g}}}\ \tilde{g}\cdot\tilde{\zeta} = \tilde{g}\cdot\Gamma(S) - \tilde{g}\cdot R(\tilde{\zeta},\tilde{g}^{-1}\operatorname{grad} F(\tilde{g}))\tilde{\zeta} - \beta \operatorname{grad} F(\tilde{g}), \tag{4.2}$$

where $F(\tilde{g}) = \frac{1}{2}d(g,\tilde{g})^2$ and $\Gamma(S)$ is the parallel transport of the resultant external force S at g to $\tilde{g}$ along the geodesic joining the two points. In [2] it is pointed out that replacing $\Gamma(S)$ and grad F by their respective first-order approximations will not affect the local convergence properties of the observer.

Although convergence is proved in [2] assuming that $S = S(g)$, the same basic argument holds when $S = S(g,\zeta)$, where we now use $\tilde{\zeta}$ instead of ζ in the parallel transport term $\Gamma(S(g,\tilde{\zeta}))$. In [2], the first variation of the observer dynamics is constructed, then contraction analysis is used to prove local exponential convergence of the observer. In the appendix we show that the first variation of the observer dynamics does not change when S is allowed to depend on the velocity ζ. Thus the contraction argument of [2] applies without modification and the local exponential convergence of the observer (4.1)–(4.2) follows even when $S = S(g,\zeta)$.

It was shown in section 2.1.1 that up to second order grad $F = \tilde{g}\zeta_e$. Therefore the first-order approximation of the observer (4.1)–(4.2) can be expressed as

$$\dot{\tilde{g}} = \tilde{g}\cdot(\tilde{\zeta} - 2\alpha\zeta_e), \tag{4.3}$$

$$\nabla_{\dot{\tilde{g}}}\ \tilde{g}\cdot\tilde{\zeta} = \tilde{g}\cdot\Gamma(S) - \tilde{g}\cdot R(\tilde{\zeta},\zeta_e)\tilde{\zeta} - \beta\ \tilde{g}\cdot\zeta_e. \tag{4.4}$$

Expanding (4.4), the first-order approximation of the observer is explicitly given by

$$\dot{\tilde{g}} = \tilde{g} \cdot (\tilde{\zeta} - 2\alpha\zeta_e), \tag{4.5}$$

$$\dot{\tilde{\zeta}} = I^{-1}\Big(ad^*_{\tilde{\zeta}} I\tilde{\zeta} - \alpha(ad^*_{\zeta_e} I\tilde{\zeta} + ad^*_{\tilde{\zeta}} I\zeta_e)\Big) + \alpha[\zeta_e, \tilde{\zeta}]_{\mathcal{G}} + \Gamma(S) - R(\tilde{\zeta}, \zeta_e)\tilde{\zeta} - \beta\zeta_e, \tag{4.6}$$

where up to order-two terms,

$$\Gamma(S) = \Big(S^k(g, \tilde{\zeta}) - \omega^k_{ij} S^i(g, \tilde{\zeta})\zeta^j_e\Big)\ e_k. \tag{4.7}$$

This coordinate-free formulation of the observer clearly shows its structure. For instance, the terms in (4.6) involving the gain α are the corrections to the inertial forces of the observer that are needed to compensate for curvature effects, $\Gamma(S)$ is the intrinsic model of the external forces of the observed system, and $R(\tilde{\zeta}, \zeta_e)\tilde{\zeta}$ is the curvature term that is needed to correct for the effects of possible divergence of nearby geodesics. The $2\alpha\zeta_e$ and $\beta\zeta_e$ terms are the error feedback that ensure convergence. The coordinate-free formulation also shows the versatility of the expressions (4.5)–(4.6). Specifically it is readily applicable to any simple mechanical system on the Lie group G. Depending on the specific problem all the control designer needs to do is specify the kinetic energy tensor I and the external forces S.

Using the observer (4.5)–(4.6), the control (3.3) can be implemented with velocity estimates replacing velocity measurements as

$$u_i = -\langle f^i(g)\ ,\ \tilde{\zeta}\rangle. \tag{4.8}$$

It is natural to ask whether the dynamic output feedback control (4.8) preserves the stability properties of the state-feedback control (3.3)—that is, whether a separation principle holds. In the appendix we show, using results of [28], that it does. In particular, if (3.3) is almost globally stabilizing (resp., asymptotically stabilizing), then so is (4.8).

5. Examples. In this section we demonstrate the preceding constructions for two cases of practical significance in which the configuration space of the simple mechanical systems are Lie groups—first $SO(3)$, and then $SE(3)$. Since these groups arise in many practical problems involving rigid body motions we include here explicit expressions for the Riemannian connection, Riemannian curvature, and the approximate local distance functions. To implement the observer in a specific application, now only the inertia tensor I and the external force S need to be changed. The effectiveness of the observer is demonstrated in $SO(3)$ for the axisymmetric top and in $SE(3)$ for a model of an electrostatically actuated MEMS.

5.1. The rotation group $SO(3)$. The rotation group, $SO(3)$, is the group of matrices $R \in GL(3, \mathcal{R})$ that satisfy the conditions $R\,R^T = R^T\,R = I$ and $\det(R) = 1$. Euler's theorem states that any given $R \in SO(3)$ is a rotation about some axis n by an angle ψ, that is, $R = \exp(\psi\hat{n})$, where

$$\psi\hat{n} = \frac{\psi}{2\sin\psi}(R - R^T), \tag{5.1}$$

and $\cos\psi = (\mathrm{tr}(R) - 1)/2$ for $|\psi| < \pi$.

The Lie algebra $so(3)$ of $SO(3)$ is the set of traceless skew symmetric 3×3 matrices. The Lie algebra $so(3)$ is identified with $\mathcal{R}^3$ by the isomorphism

$$\xi \in \mathcal{R}^3 \mapsto \hat{\xi} = \begin{bmatrix} 0 & -\xi^3 & \xi^2 \\ \xi^3 & 0 & -\xi^1 \\ -\xi^2 & \xi^1 & 0 \end{bmatrix} \in so(3), \tag{5.2}$$

where $\xi = [\xi^1\ \xi^2\ \xi^3]^T$. We will use both ξ and $\hat{\xi}$ to mean the same element of $so(3)$.

The isomorphism $I : so(3) \simeq \mathcal{R}^3 \mapsto so(3)^* \simeq \mathcal{R}^3$ defined by the positive definite matrix I induces a left-invariant metric on $SO(3)$ by the relation $\langle\langle R\cdot\xi\,,\, R\cdot\psi\rangle\rangle = \langle\langle \xi\,,\,\psi\rangle\rangle_{so(3)} = I\xi\cdot\psi$ for any two elements $R\cdot\xi, R\cdot\psi\ \in T_RSO(3)$. The Lie bracket on $so(3)$ is $[\xi\,,\,\psi]_{so(3)} = ad_\xi\,\psi = \xi\times\psi$ and the dual of the ad operator is given by $ad^*_\xi\,\Pi = \Pi\times\xi$, where $\Pi\in so(3)^*\simeq\mathcal{R}^3$.

From (2.11)–(2.12), a simple mechanical control system on $SO(3)$ takes the form

$$\dot{R} = R\hat{\zeta}, \tag{5.3}$$

$$\dot{\zeta} = I^{-1}\Big(I\zeta\ \times\ \zeta + \tilde{S}(R,\zeta)\Big), \tag{5.4}$$

where $\tilde{S}(R,\zeta) = f^c(R) + f^d(R,\zeta) + \sum_i^m u_i f^i(R)$. The passivity-based damping injection (3.3) takes the form

$$u_i = -\langle f^i(R)\,,\,\zeta\rangle. \tag{5.5}$$

The intrinsic observer (4.5)–(4.6) takes the form

$$\dot{\tilde{R}} = \tilde{R}(\hat{\tilde{\zeta}} - 2\alpha\hat{\zeta_e}), \tag{5.6}$$

$$\dot{\tilde{\zeta}} = I^{-1}\Big(I\tilde{\zeta}\times\tilde{\zeta} - \alpha\,(I\tilde{\zeta}\times\zeta_e + I\zeta_e\times\tilde{\zeta})\Big) + \alpha\zeta_e\times\tilde{\zeta} + \Gamma(S) - R_c(\tilde{\zeta},\zeta_e)\tilde{\zeta} - \beta\zeta_e, \tag{5.7}$$

where ζ_e satisfies $\exp(\zeta_e) = R^T\tilde{R}$ and is given by (5.1) as

$$\zeta_e = \frac{\psi}{2\sin\psi}(R^T\tilde{R} - \tilde{R}^TR), \tag{5.8}$$

where $\cos\psi = (\mathrm{tr}(R^T\tilde{R}) - 1)/2$ for $|\psi| < \pi$. The parallel transport term $\Gamma(S)$ is calculated from (4.7), where $S(R,\zeta) = I^{-1}\tilde{S}(R,\zeta)$, and the curvature term $R_c(\tilde{\zeta},\zeta_e)\tilde{\zeta}$ is calculated from (2.4).

If the potential energy $U(R)$ of the mechanical system is a globally defined smooth Morse function with a unique minimum at the equilibrium configuration $\bar{R}$, then the control (5.5) almost globally stabilizes the equilibrium $(\bar{R},0)$. Furthermore since $SO(3)$ is compact, from Corollary A.2 in the appendix it also follows that (5.5) implemented with the velocity observer also almost globally stabilizes the equilibrium $(\bar{R},0)$ if the initial observer configuration error is sufficiently small.

In the canonical basis the nonzero structure constants C^k_{ij} on $so(3)\simeq\mathcal{R}^3$ are

$$C^3_{12} = 1, \quad C^2_{13} = -1, \quad C^1_{23} = 1.$$

In the special case of axisymmetric rigid bodies, $I = \mathrm{diag}(I_x, I_y, I_z)$. For such examples using (2.2) the nonzero connection coefficients ω^k_{ij} are calculated to be

$$\omega^1_{23} = \frac{I_x - I_y + I_z}{2I_x}, \quad \omega^1_{32} = \frac{-I_x - I_y + I_z}{2I_x}, \quad \omega^2_{13} = \frac{I_x - I_y - I_z}{2I_y},$$

$$\omega^2_{31} = \frac{I_x + I_y - I_z}{2I_y}, \quad \omega^3_{12} = \frac{-I_x + I_y + I_z}{2I_z}, \quad \omega^3_{21} = \frac{-I_x + I_y - I_z}{2I_z}.$$

The nonzero curvature coefficients are too numerous to be listed here.

5.1.1. Angular velocity estimation for the axisymmetric top. In this section we demonstrate the effectiveness of the observer (5.6)–(5.7) by means of simulation. Consider the classical problem of an axisymmetric top in a gravitational field. Let $P = \{P_1, P_2, P_3\}$ be an inertial frame fixed at the fixed point of the top and let $e = \{e_1, e_2, e_3\}$ be a body-fixed orthonormal frame with the origin coinciding with that of P. At $t = 0$, the two frames coincide. Then let the coordinates of a point p in the inertial frame P be given by x and in the body frame, e, let the coordinates of the point p be given by X. They are related by $x(t) = R(t)X$, where $R(t) \in SO(3)$. Let $-P_3$ be the direction of gravity and let I be the inertia matrix of the axisymmetric top about the fixed point. The kinetic energy of the top is $K = I\zeta \cdot \zeta/2$, where ζ is the body angular velocity and the potential energy is $U(R) = mgl\, Re_3 \cdot P_3$. Here m is the mass of the top, g is the gravitational constant, and l is the distance along the e_3 axis to the center of mass. For simplicity we assume the top to be symmetric about the e_3 axis. The generalized potential forces $f^c(R)$ in the body frame will be given by the relation $\langle f^c(R), \zeta\rangle = -\langle dU, R \cdot \zeta\rangle = -mgl\, R\hat{\zeta}e_3 \cdot P_3$ for any $\zeta \in so(3)$, which yields that $f^c(R) = mgl\ R^T P_3 \times e_3$. The metric induced on $SO(3)$ by the kinetic energy is left invariant, and the system is a simple mechanical system on $SO(3)$. Thus the equations of motion on $SO(3) \times so(3)$ are given by (5.3)–(5.4), where $\tilde{S}(R) = mgl\ R^T P_3 \times e_3$. Since it is assumed that the top is symmetric about the e_3 axis, the inertia matrix is diagonal with $I_1 = I_2$, that is, $I = \mathrm{diag}(I_1, I_1, I_3)$.

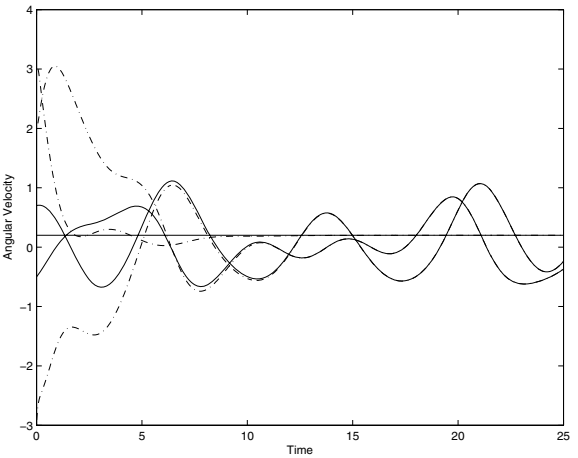


FIG. 5.1. *Angular velocity estimates versus true values in axisymmetric top simulation. The true values are the solid lines while the dotted lines are the estimated values.*

Substituting $I = \mathrm{diag}(1, 1, 2)$ and $S = I^{-1}(R^T P_3 \times e_3)$ in the observer (5.6)–(5.7) with $\alpha = \beta = 10$, we estimate the angular velocities of the axisymmetric top. The simulation results are shown in Figure 5.1. The initial body angular velocities of the axisymmetric top are $[.7\ \ -.5\ \ .2]$, the initial observer angular velocity is $[-3\ \ 2\ \ 3]$, while the initial observer configuration error corresponds to a $\pi/10$ radian rotation about the $P_2 = [0\ \ 1\ \ 0]^T$ axis.

5.2. The special Euclidean motion group $SE(3)$. The special Euclidean motion group $SE(3)$ is the semidirect product $SO(3) \times_s \mathcal{R}^3$. As a matrix group, an

element $A \in SE(3)$ and its inverse A^{-1} can be represented by

$$A = \begin{bmatrix} R & b \\ 0 & 1 \end{bmatrix}, \qquad A^{-1} = \begin{bmatrix} R^T & -R^T b \\ 0 & 1 \end{bmatrix},$$

where $R \in SO(3)$ and $b \in \mathcal{R}^3$.

The Lie algebra of $SE(3)$, denoted by $se(3)$, is the set of matrices

$$\zeta = \begin{bmatrix} \hat{\xi} & v \\ 0 & 0 \end{bmatrix},$$

where $\hat{\xi} \in so(3)$ and $v \in \mathcal{R}^3$. Then $se(3) \simeq \mathcal{R}^3 \times \mathcal{R}^3$ by identifying $\zeta \in se(3)$ with $(\xi\,,\, v) \in \mathcal{R}^3 \times \mathcal{R}^3$.

Let the inner product between the two elements $(\xi, v), (\psi, u) \in se(3)$ on $se(3)$, $\langle\langle \cdot\,,\, \cdot \rangle\rangle_{se(3)}$ be defined as $\langle\langle (\xi, v), (\psi, u) \rangle\rangle_{se(3)} = I_b \xi \cdot \psi + M v \cdot u$, where I_b is a positive definite matrix. This inner product on $se(3)$ defines a left-invariant metric on $SE(3)$ in the usual way. The Lie bracket on $se(3)$ is given by

$$[(\xi\,,\, v), (\psi\,,\, u)]_{se(3)} = ad_{(\xi\,,\, v)}(\psi\,,\, u) = (\xi \times \psi\,,\, \xi \times u - \psi \times v), \tag{5.9}$$

and the dual of the ad operator is given by

$$ad^*_{(\xi\,,\, v)} \begin{bmatrix} \Pi \\ \mu \end{bmatrix} = \begin{bmatrix} \Pi \times \xi + \mu \times v \\ \mu \times \xi \end{bmatrix}, \tag{5.10}$$

where $(\Pi, \mu) \in se(3)^* \simeq \mathcal{R}^3 \times \mathcal{R}^3$.

From (2.11)–(2.12), a simple mechanical control system on $SE(3)$ takes the form

$$\begin{bmatrix} \dot{R} & \dot{b} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} R & b \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{\xi} & v \\ 0 & 0 \end{bmatrix} \tag{5.11}$$

$$\begin{bmatrix} \dot{\xi} \\ \dot{v} \end{bmatrix} = I^{-1} \left(\begin{bmatrix} I_b \xi \times \xi \\ M v \times \xi \end{bmatrix} + \tilde{S}(R, b, \xi, v) \right), \tag{5.12}$$

where $\tilde{S}(R, b, \xi, v) = f^c(R, b) + f^d(R, b, \xi, v) + \sum_i^m u_i f^i(R, b)$.

The passivating control (3.3) takes the form

$$u_i = -\langle f^i(R, b)\,,\, (\xi, v) \rangle. \tag{5.13}$$

The intrinsic observer (4.5)–(4.6) takes the form

$$\begin{bmatrix} \dot{\tilde{R}} & \dot{\tilde{b}} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \tilde{R} & \tilde{b} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{\tilde{\xi}} - 2\alpha\, \hat{\xi}_e & \tilde{v} - 2\alpha\, v_e \\ 0 & 0 \end{bmatrix} \tag{5.14}$$

$$\begin{bmatrix} \dot{\tilde{\xi}} \\ \dot{\tilde{v}} \end{bmatrix} = \begin{bmatrix} I_b^{-1} \left(I_b \tilde{\xi} \times \tilde{\xi} - \alpha (I_b \tilde{\xi} \times \xi_e + I_b \xi_e \times \tilde{\xi}) \right) + \alpha\, \xi_e \times \tilde{\xi} \\ \tilde{v} \times \tilde{\xi} - 2\alpha\, \tilde{v} \times \xi_e \end{bmatrix}$$

$$+ \Gamma(S) - R_c(\tilde{\zeta}, \zeta_e)\tilde{\zeta} - \beta \zeta_e, \tag{5.15}$$

where $\zeta_e = (\Omega_e, V_e)$ satisfies $\exp(\zeta_e) = A^{-1}\tilde{A}$ and is explicitly given by

$$\Omega_e = \frac{\psi}{2\sin\psi}(R^T\tilde{R} - \tilde{R}^T R), \tag{5.16}$$

$$V_e = W^{-1}(R^T\tilde{b} - \tilde{R}^T b), \tag{5.17}$$

where $\cos\psi = (\mathrm{tr}(R^T\tilde{R}) - 1)/2$ for $|\psi| < \pi$ [29] and

$$W = I_{3\times 3} + \frac{(1-\cos\psi)}{\psi^2}\Omega_e + \frac{(\psi - \sin\psi)}{\psi^3}\Omega_e^2.$$

The parallel transport term $\Gamma(S)$ is calculated from (4.7), where $S(R, b, \xi, v) = I^{-1}\tilde{S}(R, b, \xi, v)$, and the curvature term $R_c(\tilde{\zeta}, \zeta_e)\tilde{\zeta}$ is calculated from (2.4). In the canonical basis the nonzero structure constants C_{ij}^k on $se(3) \simeq \mathcal{R}^6$ are

$$\begin{aligned} &C_{12}^3 = 1, \quad C_{13}^2 = -1, \quad C_{15}^6 = 1, \quad C_{16}^5 = -1, \quad C_{23}^1 = 1, \\ &C_{24}^6 = -1, \quad C_{26}^4 = 1, \quad C_{34}^5 = 1, \quad C_{35}^4 = -1. \end{aligned}$$

In the case where $I_b = \mathrm{diag}(I_x, I_y, I_z)$ and M is a positive scalar, using (2.2) the nonzero connection coefficients ω_{ij}^k are shown to be

$$\begin{aligned} &\omega_{23}^1 = \frac{I_x - I_y + I_z}{2I_x}, \quad \omega_{32}^1 = \frac{-I_x - I_y + I_z}{2I_x}, \quad \omega_{13}^2 = \frac{I_x - I_y - I_z}{2I_y}, \\ &\omega_{31}^2 = \frac{I_x + I_y - I_z}{2I_y}, \quad \omega_{12}^3 = \frac{-I_x + I_y + I_z}{2I_z}, \quad \omega_{21}^3 = \frac{-I_x + I_y - I_z}{2I_z}, \\ &\omega_{15}^6 = \omega_{26}^4 = \omega_{34}^5 = 1, \quad \omega_{16}^5 = \omega_{24}^6 = \omega_{35}^4 = -1. \end{aligned}$$

The coefficients of the curvature tensor R_{jab}^k can be calculated using (2.3).

5.2.1. Stabilization of an electrostatically actuated six-degrees-of-freedom micromirror. In this section, our observer-based stabilization approach is applied to an electrostatically actuated MEMS. An intrinsic and geometric model of a six-degrees-of-freedom, electrostatically actuated micromirror is developed in [25]. Such devices may be used as steerable micromirrors, for example, to guide an optical beam into one of a number of output fiberoptics or to simultaneously correct the phase and amplitude of an optical signal. For further physical motivation and applications the reader is referred to [25, 27] and the references therein. The device consists of a fixed bottom plate and a rigid top plate. The top plate is free to rotate and translate, subject to the constraint that each side is connected to a support structure through flexible cantilevers. The bottom plate is segmented into four drive electrodes. The system is actuated by a voltage difference between each electrode and the grounded top plate. Figure 5.2 is a schematic of the movable top plate. The points q_i denote the locations at which the external spring and damping forces act on the system.

The device is modeled as a mechanical subsystem coupled to an electrical subsystem via the electrostatic actuation forces. The configuration space of the mechanical system is the group of three-dimensional Euclidean motions represented by $SE(3)$. Let $e = (e_1, e_2, e_3)$ be a body-fixed orthonormal coordinate frame centered at the center of mass of the top plate, and let $P = (P_1, P_2, P_3)$ be an inertially fixed orthonormal coordinate frame coinciding with e when the system is in equilibrium with no actuation. At a given time t the orientation of the top plate with respect to the inertial frame is given by $R(t)$, while the displacement of the center of mass of the top

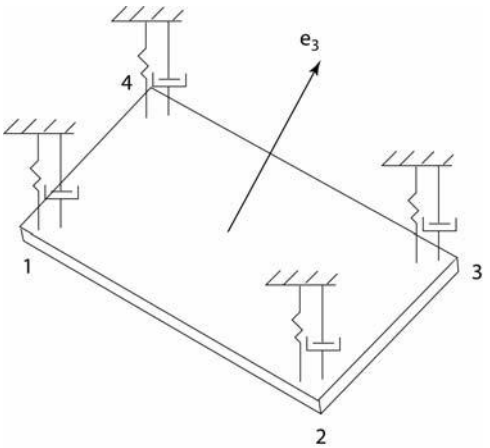


FIG. 5.2. *Schematic diagram of a rigid three-dimensional microactuator.*

plate with respect to the inertial frame is given by $b(t)$. The body angular velocity of the top plate is denoted by $\xi(t)$. The velocity of the center of mass of the top plate in the inertial coordinates is denoted by $\dot{b} = R(t)v(t)$, where $v(t)$ is the velocity of the center of mass in the body frame. The nonactuated equilibrium gap between the center of mass of the top plate and the bottom plate is d. The resistance in the pth capacitor circuit is r_p. The permittivity of the dielectric medium between the electrodes is ϵ. The electrode areas are each assumed to be equal, and denoted A. The charge stored in the pth capacitor is Q_p. The voltage control supplied to the pth capacitor is u_p. The position vector in the body-fixed coordinates of the ith point is q_i. The inertia matrix of the top plate is denoted by I. The total mass of the top plate is m.

The spring forces exerted by the cantilever beams are assumed to be linear in the absolute displacement. This assumption implies that the spring forces $F_p^C(R, b)$ are given by

$$\begin{bmatrix} F_i^C(R,b) \\ 0 \end{bmatrix} = - \begin{bmatrix} K_i & 0 \\ 0 & 0 \end{bmatrix} \left(\begin{bmatrix} R & b \\ 0 & 1 \end{bmatrix} \begin{bmatrix} q_i \\ 1 \end{bmatrix} - \begin{bmatrix} q_i \\ 1 \end{bmatrix} \right).$$

The structural dissipative forces $F_i^D(R, b, \xi, v)$ of the system are assumed to be of Rayleigh type and are given by

$$\begin{bmatrix} F_i^D(R,b,\xi,v) \\ 0 \end{bmatrix} = - \begin{bmatrix} C_i & 0 \\ 0 & 0 \end{bmatrix} \left(\begin{bmatrix} R & b \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{\xi} & v \\ 0 & 0 \end{bmatrix} \begin{bmatrix} q_i \\ 1 \end{bmatrix} \right).$$

The 3×3 matrices K_i and C_i are positive semidefinite. Computing the torques about the center of mass of the movable electrode in the body coordinates, the generalized body forces due to the stiffness and structural damping of the cantilever are expressed by

$$f_i^c = \begin{bmatrix} \hat{q}_i R^T F_i^c \\ R^T F_i^C \end{bmatrix}, \quad f_i^d = \begin{bmatrix} \hat{q}_i R^T F_i^D \\ R^T F_i^D \end{bmatrix}.$$

Neglecting parasitics, allowing fringing, but assuming that the electrostatic field generated by each individual electrode does not interact with the others (these are standard

assumptions in the modeling of multielectrode electrostatic devices [19]), the generalized body electrostatic forces are given by $BW(Q)$, where $Q = [Q_1 \;\; Q_2 \;\; Q_3 \;\; Q_4]^T$,

$$W(Q) = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \frac{1}{4\epsilon A} \begin{bmatrix} 1 & -1 & -1 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 \end{bmatrix} \begin{bmatrix} Q_1^2 \\ Q_2^2 \\ Q_3^2 \\ Q_4^2 \end{bmatrix},$$

$$B = \begin{bmatrix} -l_y\, e_1 & -l_x\, e_2 & 0 \\ 0 & 0 & 2e_3 \end{bmatrix}.$$

Using these generalized forces, in [25] the governing equation of the MEMS is shown to be

$$\dot{Q}_p = -\frac{1}{r_p}\left(V_{d_p} - u_p\right) \quad \text{for } p = 1, 2, 3, 4, \tag{5.18}$$

$$\begin{bmatrix} \dot{R} & \dot{b} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} R & b \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{\xi} & v \\ 0 & 0 \end{bmatrix} \tag{5.19}$$

$$\begin{bmatrix} \dot{\xi} \\ \dot{v} \end{bmatrix} = I^{-1}\left(\begin{bmatrix} I_b\xi \times \xi \\ Mv \times \xi \end{bmatrix} + \sum_{i=1}^{4} f_i^c(R, b) + \sum_{i=1}^{4} f_i^d(R, b, \xi, v) + B\; w(Q) \right), \tag{5.20}$$

$$y_{m_p} = V_{d_p} \quad \text{for } p = 1, 2, 3, 4, \tag{5.21}$$

$$y_p = Q_p \quad \text{for } p = 1, 2, 3, 4, \tag{5.22}$$

where g, y_{m_p}, and y_p are the measured outputs. The system has relative degree $\{1, 1, 1, 1\}$ with respect to the output $y = Q$, and stable zero dynamics.

For a given $\bar{Q}$, the corresponding equilibrium points of (5.19)–(5.20) are given by $\bar{\xi} = 0$, $\bar{v} = 0$, and

$$0 = \sum_{i=1}^{4} f_i^c(\bar{R}, \bar{b}) + Bw(\bar{Q}). \tag{5.23}$$

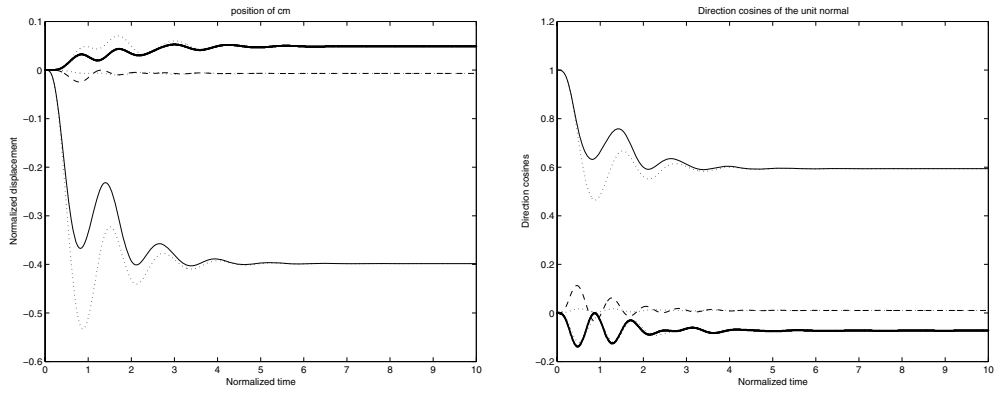
Using the results of section 3.2 it can be shown that the feedback law

$$u = \begin{bmatrix} V_{d_1} \\ V_{d_2} \\ V_{d_3} \\ V_{d_4} \end{bmatrix} + \begin{bmatrix} r_1(l_y\xi_1 - l_x\xi_2 + 2v_3)(Q_1 + \bar{Q}_1)/4\epsilon A \\ r_2(-l_y\xi_1 - l_x\xi_2 + 2v_3)(Q_2 + \bar{Q}_2)/4\epsilon A \\ r_3(-l_y\xi_1 + l_x\xi_2 + 2v_3)(Q_3 + \bar{Q}_3)/4\epsilon A \\ r_4(l_y\xi_1 + l_x\xi_2 + 2v_3)(Q_4 + \bar{Q}_4)/4\epsilon A \end{bmatrix} - \alpha \begin{bmatrix} r_1(Q_1 - \bar{Q}_1) \\ r_2(Q_2 - \bar{Q}_2) \\ r_4(Q_3 - \bar{Q}_3) \\ r_4(Q_4 - \bar{Q}_4) \end{bmatrix} \tag{5.24}$$

with $\alpha > 0$ locally asymptotically stabilizes the corresponding given equilibrium $(0, \bar{Q}, \bar{R}, \bar{b}, 0, 0)$.

The control (5.24) involves angular and linear velocity measurements. Making these measurements in situ on the MEMS is infeasible. Thus assuming that the configuration variables (R, b) are available for measurement (see [25] for a discussion of how to do this), we estimate the angular and linear velocities of the mirror in the body frame using the intrinsic observer of section 5.2, where now $S(R, b, \xi, v) = I^{-1}(\sum_{i=1}^{4} f_i^c(R, b) + \sum_{i=1}^{4} f_i^d(R, b, \xi, v) + B\; w(Q))$. MATLAB simulation results are shown in Figure 5.3. For comparison purposes the performance of both the full state-feedback controller and the dynamic output feedback controller are plotted on the

same figures for identical initial conditions of the MEMS. In the case of the dynamic output feedback control the initial observer configuration error corresponds to a $\pi/10$ rotation about the [1 1 0] axis and a translation of [3 3 3]. The initial body angular velocity error is [9 12 -9] and the initial velocity error of the center of mass in the body coordinates is [6 9 -3].



(a) Center of Mass (b) Direction Cosines

FIG. 5.3. *The position of the center of mass and the direction cosines of the unit normals of the movable electrode versus time. In the case of dynamic output feedback the thin solid curve corresponds to the P_3 direction, the thick solid curve corresponds to the P_1 direction, and the dashed curve corresponds to the P_2 direction. The dotted curves correspond to full state feedback.*

6. Conclusion. We present an intrinsic observer-based approach to the stabilization of a class of simple mechanical control systems on a Lie group. Specifically, we consider systems with left-invariant kinetic energy and a measured configuration variable. The result is obtained by specializing two general formulations on arbitrary Riemannian manifolds, namely, passivity-based control and intrinsic velocity estimation. This specialization is noteworthy because it results in drastically simplified explicit expressions for the controller that can be readily applied once the kinetic energy tensor and the external forces are specified. The observer is explicitly computed for two special cases of particular importance, namely, the rotation group $SO(3)$ and the Euclidean motion group $SE(3)$. The expressions given in those sections may be applied to the many problems of practical significance arising from rigid body motion by specializing only the inertia tensor and the external forces. Here the results are applied to estimation of the velocities of the axisymmetric top and to the stabilization of an electrostatically actuated MEMS model. Simulations show excellent performance.

Appendix.

A.1. First variation equations of the observer. If a separation principle is to hold in the presence of velocity-dependent control, then the observer must converge even when velocity terms are allowed in the external forces. Following [2] we first construct the first variation of the observer and then use contraction analysis to prove local exponential convergence of the observer. In what follows we show that including such velocity terms in S does not change the first variation equations obtained in [2]. Thus the contraction analysis of [2] holds without modification, and the observer converges when $S = S(q, v)$.

Consider the simple mechanical system on a Riemannian manifold $(M, \langle\langle \cdot, \cdot \rangle\rangle)$ given by

$$\dot{q} = v, \tag{A.1}$$

$$\nabla_{\dot{q}}\, v = S(q, v), \tag{A.2}$$

and the observer

$$\dot{\tilde{q}} = X(\tilde{q}, \tilde{v}) \;=\; \tilde{v} - \alpha \operatorname{grad} F(\tilde{q}), \tag{A.3}$$

$$\nabla_{\dot{\tilde{q}}}\, \tilde{v} = Y(\tilde{q}, \tilde{v}) \;=\; \Gamma(S(q, \tilde{v})) - R(\tilde{v}, \operatorname{grad} F(\tilde{q}))\tilde{v} - \beta \operatorname{grad} F(\tilde{q}), \tag{A.4}$$

where $R(\cdot, \cdot)\cdot$ is the curvature and $\Gamma(S(q, \tilde{v}))$ is the parallel transport of the external forces $S(q, \tilde{v})$ to $\tilde{q}$ along the unique geodesic joining q to $\tilde{q}$ (for q and $\tilde{q}$ sufficiently close). Observe that in the parallel transport term of $\Gamma(S(q, \tilde{v}))$ we use $\tilde{v}$ instead of v.

Let $(q(t), v(t))$ be a solution of (A.3)–(A.4) with initial condition (q_0, v_0) and $(\tilde{q}(t), \tilde{v}(t))$ be a solution of (A.3)–(A.4) with initial conditions $(\tilde{q}(0), \tilde{v}(0)) \neq (q_0, v_0)$. Let $s \mapsto \gamma(s) \in M$ be a smooth curve on M such that $\gamma(0) = q_0$ and $\gamma(1) = \tilde{q}_0$ and let $s \mapsto \tau(s) \in T_{\gamma(s)}M$ be a smooth vector field defined along $\gamma(s)$ such that $\tau(0) = v_0$ and $\tau(1) = \tilde{v}_0$. Let $(\tilde{q}(s, t), \tilde{v}(s, t))$ be a solution of (A.3)–(A.4) with initial conditions $(\gamma(s), \tau(s))$. Define

$$\frac{\partial \tilde{q}}{\partial s}(s, t) = J_q(s, t) \in T_{\tilde{q}(s,t)}M, \tag{A.5}$$

$$\frac{\partial \tilde{v}}{\partial s}(s, t) = \nabla_{J_q} \tilde{v} = J_v(s, t) \in T_{\tilde{q}(s,t)}M. \tag{A.6}$$

In coordinates these equations correspond exactly to those given by (6) in [2]. By construction it follows that $[J_q, \dot{\tilde{q}}] = 0$ and thus the first variation of (A.3)–(A.4) can be intrinsically computed as follows:

$$\frac{\partial \dot{\tilde{q}}}{\partial s} = \nabla_{J_q} \dot{\tilde{q}} = \nabla_{\dot{\tilde{q}}} J_q = \nabla_{J_q} X, \tag{A.7}$$

$$\frac{\partial \nabla_{\dot{\tilde{q}}}\, \tilde{v}}{\partial s} = \nabla_{J_q} \nabla_{\dot{\tilde{q}}}\, \tilde{v} = \nabla_{\dot{\tilde{q}}} \nabla_{J_q}\, \tilde{v} + R(J_q, \dot{\tilde{q}})\tilde{v} = \nabla_{J_q} Y, \tag{A.8}$$

where the second equality in (A.7) follows from $[J_q, \dot{\tilde{q}}] = 0$ and the second equality in (A.8) follows from

$$\nabla_{J_q} \nabla_{\dot{\tilde{q}}}\, \tilde{v} - \nabla_{\dot{\tilde{q}}} \nabla_{J_q}\, \tilde{v} = R(J_q, \dot{\tilde{q}})\tilde{v}.$$

Thus from (A.6), (A.7), and (A.8) we have the first variation equations of (A.3)–(A.4) as

$$\nabla_{\dot{\tilde{q}}} J_q = \nabla_{J_q} X, \tag{A.9}$$

$$\nabla_{\dot{\tilde{q}}} J_v = -R(J_q, \dot{\tilde{q}})\tilde{v} + \nabla_{J_q} Y. \tag{A.10}$$

When $X(\tilde{q}, \tilde{v}) = \tilde{v}$ and $Y(\tilde{q}, \tilde{v}) = 0$ we recover Jacobi's equation of geodesic variation $\nabla^2_{\dot{\tilde{q}}} J_q = -R(J_q, \tilde{v})\tilde{v}$.

Substituting for X and Y from (A.3)–(A.4) in (A.9) and (A.10) we have

$$\nabla_{\dot{\tilde{q}}} J_q = J_v - \alpha \nabla_{J_q} \text{grad}\, F, \tag{A.11}$$

$$\nabla_{\dot{\tilde{q}}} J_v = -R(J_q, \dot{\tilde{q}})\tilde{v} + \nabla_{J_q}\Gamma - \beta \nabla_{J_q} \text{grad}\, F - \nabla_{J_q} R(\tilde{v}, \text{grad}\, F(\tilde{q}))\tilde{v}. \tag{A.12}$$

From [35]

$$\begin{aligned}\nabla_{J_q} R(\tilde{v}, \text{grad}\, F(\tilde{q}))\tilde{v} = (\nabla_{J_q} R)\left((\tilde{v}, \text{grad}\, F(\tilde{q}))\tilde{v}\right) + R(\nabla_{J_q}\tilde{v}, \text{grad}\, F(\tilde{q}))\tilde{v} \\ + R(\tilde{v}, \nabla_{J_q} \text{grad}\, F(\tilde{q}))\tilde{v} + R(\tilde{v}, \text{grad}\, F(\tilde{q}))\nabla_{J_q}\tilde{v}.\end{aligned}$$

Then the first variation equations of the observer (A.3)–(A.4) are

$$\nabla_{\dot{\tilde{q}}} J_q = J_v - \alpha \nabla_{J_q} \text{grad}\, F, \tag{A.13}$$

$$\begin{aligned}\nabla_{\dot{\tilde{q}}} J_v = &-R(J_q, \dot{\tilde{q}})\tilde{v} + \nabla_{J_q}\Gamma - \beta \nabla_{J_q} \text{grad}\, F \\ &-(\nabla_{J_q} R)\left((\tilde{v}, \text{grad}\, F(\tilde{q}))\tilde{v}\right) - R(\nabla_{J_q}\tilde{v}, \text{grad}\, F(\tilde{q}))\tilde{v} \\ &-R(\tilde{v}, \nabla_{J_q} \text{grad}\, F(\tilde{q}))\tilde{v} - R(\tilde{v}, \text{grad}\, F(\tilde{q}))\nabla_{J_q}\tilde{v}.\end{aligned} \tag{A.14}$$

These equations correspond exactly with (9) of [2]. Note our curvature convention follows [20, 21, 35] which results in a sign difference from [2].

When $\tilde{q}$ and q are sufficiently close (i.e., up to second order)

$$\text{grad}\, F = 0, \quad \nabla_{J_q} \text{grad}\, F = J_q, \quad \nabla_{J_q}\Gamma(S(q, \tilde{v})) = 0.$$

Thus for sufficiently close $\tilde{q}$ and q the first variation equations reduce to

$$\nabla_{\dot{\tilde{q}}} J_q = J_v - \alpha J_q, \tag{A.15}$$

$$\nabla_{\dot{\tilde{q}}} J_v = -\beta J_q. \tag{A.16}$$

These equations correspond exactly with equation (10) of [2].

A.2. Separation principle. The following lemma proved in [28] provides the basis for a separation principle when the Lie group G is compact. Consider the system

$$\dot{g} = g \cdot \zeta, \tag{A.17}$$

$$\dot{\zeta} = I^{-1}\left(ad^*_\zeta \tilde{I}\zeta + f(g, \zeta)\right) + \psi(g, \zeta, q), \tag{A.18}$$

with $(g, \zeta) \in G \times \mathcal{G}$ and $q \in \mathcal{R}^n$ and $V = U(g) + \frac{1}{2}\langle\langle \zeta, \zeta \rangle\rangle_{\mathcal{G}}$, where $U(g)$ is a smooth globally defined Morse function and is the potential energy of the system. Also consider the following assumptions.

Assumption A.1. *The point* $(\bar{g}, 0)$ *is an almost globally stable equilibrium of* (A.17)–(A.18) *with* $\psi \equiv 0$ *and furthermore* $\langle\langle g^{-1}\text{grad}\, U, \zeta\rangle\rangle_{\mathcal{G}} + \langle\langle I^{-1} f, \zeta\rangle\rangle_{\mathcal{G}} \leq 0$.

The condition $\langle\langle g^{-1}\text{grad}\, U, \zeta\rangle\rangle_{\mathcal{G}} + \langle\langle I^{-1} f, \zeta\rangle\rangle_{\mathcal{G}} \leq 0$ is satisfied by any simple mechanical system with potential energy $U(g)$ and Rayleigh-type dissipation. The equilibrium $(\bar{g}, 0)$ is an almost globally stable equilibrium if $\bar{g}$ is a unique minimum of $U(g)$.

Assumption A.2. *The function* $q(t) \in \mathcal{R}^n$ *satisfies*

$$||q(t)|| \leq c||q(0)||e^{-\lambda t} \tag{A.19}$$

for some $c > 0, \lambda > 0$, *and all* $t > 0$.

ASSUMPTION A.3. *The interconnection term satisfies* $\psi(g,\zeta,0)\equiv 0$ *and the linear growth conditions*

$$||\psi|| \le \gamma_1(||q||)||\zeta|| + \gamma_2(||q||) \tag{A.20}$$

for two class $\mathcal{K}_\infty$ *functions* $\gamma_1(\cdot),\gamma_2(\cdot)$.

LEMMA A.1. *If the Lie group* G *is compact and if Assumptions* A.1–A.3 *are satisfied, then the equilibrium* $(\bar{g},0)$ *of the system* (A.17)–(A.18) *is almost globally stable. Convergence is asymptotic if the inequality in Assumption* A.1 *is strict.*

COROLLARY A.2. *For a compact Lie group* G *if the state-feedback control* (3.3) *almost globally stabilizes* $(\bar{g},0)$*, then the dynamic output feedback control* (4.8) *almost globally stabilizes* $(\bar{g},0)$*. Convergence is asymptotic if the control* (3.3) *ensures asymptotic convergence.*

Proof of Corollary A.2. Define the observation error of the velocities by $\zeta_{oe} := \tilde{\zeta} - \zeta$. From Theorem 1 of [2] we have that for sufficiently small initial observer configuration error,

$$(d(g(t),\tilde{g}(t)) + ||\zeta_{oe}(t)||_{\mathcal{G}}) < (d(g(0),\tilde{g}(0)) + ||\zeta_{oe}(0)||_{\mathcal{G}})\, e^{-\lambda t} \tag{A.21}$$

for some $\lambda > 0$. From (A.21) it follows that for nonzero initial velocity error,

$$||\zeta_{oe}(t)||_{\mathcal{G}} < c\,||\zeta_{oe}(0)||_{\mathcal{G}}\; e^{-\lambda t} \tag{A.22}$$

for some $c > 0$.

Substituting $\tilde{\zeta} = \zeta + \zeta_{oe}$ in the controls (4.8) we obtain a simple mechanical system in the form of (A.17)–(A.18), where the interconnection term $\psi(g,\zeta,\zeta_{oe}) = B(g)B(g)^T\zeta_{oe}$ with $B(g) = [f^1(g)\, f^2(g)\, \cdots\, f^m(g)]$. If the potential energy $U(g)$ of the system is a globally defined smooth Morse function with a unique minimum at $\bar{g}$ and the damping injection $u = -B^T\zeta$ almost globally stabilizes the equilibrium $(\bar{g},0)$, then it can be verified that Assumption A.1 is satisfied. From (A.22) Assumption A.2 is satisfied. Since $\psi(g,\zeta,\zeta_{oe}) = B(g)B(g)^T\zeta_{oe}$, if G is compact, it can be easily verified that Assumption A.3 is satisfied. Thus from Lemma A.1 the dynamic feedback control (4.8) almost globally stabilizes the equilibrium $(\bar{g},0)$ provided that the initial observer configuration error is sufficiently small. Convergence is asymptotic if the inequality in Assumption A.1 is strict. This is guaranteed if (3.3) ensures asymptotic convergence.

REFERENCES

[1] R. ABRAHAM AND J. E. MARSDEN, *Foundations of Mechanics*, 2nd ed., Benjamin/Cummings, Reading, MA, 1978.

[2] N. AGHANNAN AND P. ROUCHON, *An intrinsic observer for a class of Lagrangian systems*, IEEE Trans. Automat. Control, 48 (2003), pp. 936–945.

[3] M. R. AKELLA, *Rigid body attitude tracking without angular velocity feedback*, Systems Control Lett., 42 (2001), pp. 321–326.

[4] M. R. AKELLA, J. T. HALBERT, AND G. R. KOTAMRAJU, *Rigid body attitude control with inclinometer and low-cost gyro measurements*, Systems Control Lett., 49 (2003), pp. 151–159.

[5] V. I. ARNOLD, *Mathematical Methods of Classical Mechanics*, 2nd ed., Springer-Verlag, New York, 1989.

[6] J. BAILLIEUL, *The geometry of controlled mechanical systems*, in Mathematical Control Theory, Springer, New York, 1999, pp. 322–354.

[7] A. M. BLOCH AND J. E. MARSDEN, *Stabilization of rigid body dynamics by the energy-casimir method*, Systems Control Lett., 14 (1991), pp. 341–346.

[8] A. M. Bloch and P. E. Crouch, *Optimal control, optimization, and analytical mechanics*, in Mathematical Control Theory, Springer, New York, 1999, pp. 268–321.

[9] A. M. Bloch, N. E. Leonard, and J. E. Marsden, *Controlled Lagrangians and the stabilization of mechanical systems* I: *The first matching theorem*, IEEE Trans. Automat. Control, 45 (2000), pp. 2253–2270.

[10] A. M. Bloch, D. E. Chang, N. E. Leonard, and J. E. Marsden, *Controlled Lagrangians and the stabilization of mechanical systems* II: *Potential shaping*, IEEE Trans. Automat. Control, 46 (2001), pp. 1556–1571.

[11] A. M. Bloch and N. E. Leonard, *Symmetries, conservation laws, and control*, in Geometry, Mechanics, and Dynamics, Springer-Verlag, New York, 2002, pp. 431–460.

[12] A. M. Bloch, J. Baillieul, P. Crouch, and J. E. Marsden, *Nonholonomic Mechanics and Control,* Springer-Verlag, New York, 2003.

[13] F. Bullo, N. E. Leonard, and A. D. Lewis, *Controllability and motion algorithms for underactuated Lagrangian systems on Lie groups*, IEEE Trans. Automat. Control, 45 (2000), pp. 1437–1454.

[14] F. Bullo and R. M. Murray, *Tracking for fully actuated mechanical systems: A geometric framework*, Automatica J. IFAC, 35 (1999), pp. 17–34.

[15] F. Bullo and A. D. Lewis, *Geometric Control of Mechanical Systems: Modeling, Analysis, and Design for Simple Mechanical Control Systems,* Springer-Verlag, New York, 2004.

[16] C. I. Byrnes, A. Isidori, and J. C. Willems, *Passivity, feedback equivalence, and the global stabilization of minimum phase nonlinear systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 1228–1240.

[17] D. E. Chang, A. M. Bloch, N. E. Leonard, J. E. Marsden, and C. A. Woolsey, *The equivalence of controlled Lagrangian and controlled Hamiltonian systems*, ESAIM Control Optim. Calc. Var., 8 (2002), pp. 393–422.

[18] J. J. Duistermaat and J. A. C. Kolk, *Lie Groups*, Springer-Verlag, Berlin, Heidelberg, 2000.

[19] D. Elata, O. Bochobza-Degni, and Y. Nemirovsky, *Analytical approach and numerical alpha-line method for pull-in hyper-surface extraction of electrostatic actuators with multiple uncoupled voltage source*s, J. Microelectromechanical Systems, 12 (2003), pp. 681–691.

[20] T. Frankel, *The Geometry of Physics. An Introduction,* Cambridge University Press, Cambridge, UK, 1997.

[21] J. Jost, *Riemannian Geometry and Geometric Analysis,* Springer-Verlag, Berlin, 2002.

[22] V. Jurdjevic, *Optimal control, geometry, and mechanics*, in Mathematical Control Theory, Springer-Verlag, New York, 1999, pp. 268–321.

[23] D. E. Koditschek, *The application of total energy as a Lyapunov function for mechanical control systems*, in Dynamics and Control of Multibody Systems, Contemp. Math. 97, J. E. Marsden, P. S. Krishnaprasad, and J. C. Simo, eds., AMS, Providence, RI, 1989, pp. 131–157.

[24] A. D. Lewis and R. M. Murray, *Configuration controllability of simple mechanical control systems*, SIAM J. Control Optim., 35 (1997), pp. 766–790.

[25] D. H. S. Maithripala, R. O. Gale, M. W. Holtz, J. M. Berg, and W. P. Dayawansa, *Nano-precision control of micromirrors using output feedback*, in Proceedings of the CDC, Maui, HI, 2003, pp. 2652–2657.

[26] D. H. S. Maithripala, J. M. Berg, and W. P. Dayawansa, *An Intrinsic Observer for a Class of Simple Mechanical Systems on Lie Groups*, in Proceedings of the ACC, Boston, 2004, pp. 1546–1551.

[27] D. H. S. Maithripala, J. M. Berg, and W. P. Dayawansa, *Control of an electrostatic MEMS using static and dynamic output feedback*, ASME J. Dynam. Systems Measurement Control, 41 (2005), pp. 443–450.

[28] D. H. S. Maithripala, J. M. Berg, and W. P. Dayawansa, *Almost-global tracking of simple mechanical systems on a general class of Lie groups*, IEEE Trans. Automat. Control, to appear.

[29] J. E. Marsden and T. S. Ratiu, *Introduction to Mechanics and Symmetry*, 2nd ed., Springer-Verlag, New York, 1999.

[30] F. Martin, P. Rouchon, and J. Rudolph, *Invariant tracking*, ESAIM Control Optim. Calc. Var., 10 (2004), pp. 1–13.

[31] F. Mazenc and A. Astolfi, *Robust output feedback stabilization of the angular velocity of a rigid body*, Systems Control Lett., 39 (2000), pp. 203–210.

[32] M. C. Muñoz-Lecanda and F. J. Yániz-Fernández, *Dissipative control of mechanical systems*: *A geometric approach*, SIAM J. Control Optim., 40 (2002), pp. 1505–1516.

[33] R. Ortega, A. J. van der Schaft, I. Mareels, and B. Maschke, *Putting energy back in control*, IEEE Control Systems Magazine, 21 (2001), pp. 18–33.

[34] R. Ortega, A. van der Schaft, B. Maschke, and G. Escobar, *Interconnection and damping assignment passivity-based control of port-controlled Hamiltonian systems*, Automatica J. IFAC, 38 (2002), pp. 585–596.
[35] P. Petersen, *Riemannian Geometry,* Springer-Verlag, New York, 1998.
[36] H. Rehbinder and X. Hu, *Nonlinear state estimation for rigid body motion with low pass sensors*, Systems Control Lett., 40 (2000), pp. 183–191.
[37] S. Salcudean, *A globally convergent angular velocity observer for rigid body motion*, IEEE Trans. Automat. Control, 36 (1991), pp. 1493–1497.
[38] R. Sepulchre, M. Jankovic, and P. Kokotovic, *Constructive Nonlinear Control,* Springer-Verlag, London 1997.
[39] G. V. Smirnov, *Attitude determination and stabilization of a spherically symmetric rigid body in a magnetic field*, Internat. J. Control, 74 (2001), pp. 341–347.
[40] A. J. van der Schaft, *L_2-Gain and Passivity Techniques in Nonlinear Control,* Springer-Verlag, London, 2000.
[41] C. A. Woolsey and N. E. Leonard, *Moving mass control for underwater vehicles*, in Proceedings of the American Control Conference, Anchorage, AK, 2001, pp. 2824–2830.
[42] M. Zefran, V. Kumar, and C. B. Croke, *On the generation of smooth three-dimensional rigid body motions*, IEEE Trans. Robotics and Automation, 14 (1998), pp. 576–589.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 5, pp. 1712–1724

EXPONENTIAL STABILITY OF NONLINEAR SINGULARLY PERTURBED DIFFERENTIAL EQUATIONS*

G. GRAMMEL†

Abstract. Nonlinear singularly perturbed differential equations with two natural time scales are under consideration. It is shown that the exponential stability of both the boundary layer systems and the reduced system imply the exponential stability of the fully coupled system for small perturbation parameters. The asymptotic behavior of gain and decay rates is investigated. Moreover, it turns out that the achieved exponential stability is robust with respect to multivalued regular perturbations and small delays. The method of proof does not rely on (converse) Lyapunov theorems but on an appropriate version of Tychonov's theorem and manages without the smoothness of the vector fields involved.

Key words. exponential stability, singular perturbations, nonlinear differential equations with delays

AMS subject classifications. 34E15, 34D15, 34A60, 34K20

DOI. 10.1137/040614591

1. Introduction.

On $\mathbb{R}^n \times \mathbb{R}^m$, we consider the singularly perturbed differential equation

$$\begin{aligned} \dot{x}_\epsilon(t) &= f(x_\epsilon(t), y_\epsilon(t)), \quad x_\epsilon(0) = x^0, \\ \epsilon\, \dot{y}_\epsilon(t) &= g(x_\epsilon(t), y_\epsilon(t)), \quad y_\epsilon(0) = y^0, t \geq 0, \end{aligned} \tag{1}$$

where $\epsilon > 0$ is a small perturbation parameter reflecting the presence of two natural time scales. An important aim in singular perturbation theory is the decomposition of the coupled system into two unperturbed subsystems and to draw conclusions from certain properties of the unperturbed subsystems. In connection with exponential stability properties, this decomposition method works very well. The following fact on finite-dimensional smooth systems is well known; see [10, section 7, Corollary 2.3]. If both the unperturbed fast (boundary layer) systems and the unperturbed slow (reduced) system are exponentially stable with uniform gain and decay rates, then the singularly perturbed system is exponentially stable as well, at least for sufficiently small perturbation parameters. Suitable upper bounds for the perturbation parameter are obtained in [14]. In [4] the asymptotic behavior of the decay rates is investigated. All these works mentioned rely on a sophisticated Lyapunov theory, involving appropriate (converse) Lyapunov theorems. Consequently, the results are presented in a smooth, finite-dimensional ODE framework.

A different approach using the approximating properties of an averaged slow limit system has been presented in [2], where under mild regularity conditions the (near) asymptotic stability of the slow subsystem is investigated. In [7, 8] a similar method has been applied in order to characterize the exponential stability of (multivalued) differential equations with a fast time variable. Here, the method of proof is based on a suitable time discretization along with a computation of the distance between the

*Received by the editors September 7, 2004; accepted for publication (in revised form) May 15, 2005; published electronically November 22, 2005.
http://www.siam.org/journals/sicon/44-5/61459.html
†Center for Mathematics, M6, Technical University Munich, Boltzmann St. 3, 85747 Garching, Germany (grammel@ma.tum.de).

perturbed and the unperturbed trajectories. The present paper aims at an adaption of this method to the coupled singularly perturbed system (1). The method provides several features listed below.

- The regularity of the vector fields involved is reduced to Lipschitz continuity.
- Estimates for both gain and decay rates are obtained.
- The results can be extended to differential equations in Banach spaces.
- Multivalued regular perturbations are taken into account.
- Small delays are considered.

The paper is organized as follows. In section 2 we present the setting and the main result, Theorem 2.5. This result is compared to the present literature on related problems. A version of Tychonov's theorem (see [15] or [3]) on the approximation of singularly perturbed differential equations along with important tools for the proof of Theorem 2.5 are provided in section 3. The proof of Theorem 2.5 is carried out in section 4. In section 5 an application to small delays is worked out in detail.

2. Setting and main result.

Assumption 2.1 (regularity). The mappings $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ and $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ are Lipschitz continuous with constant $L \geq 0$. Moreover, the origin $(0,0) \in \mathbb{R}^n \times \mathbb{R}^m$ is an equilibrium point of (f,g), i.e., $f(0,0) = 0$ and $g(0,0) = 0$.

The decomposition of the singularly perturbed differential equation (1) into two unperturbed subsystems works as follows. Introducing the new time variable $s := \epsilon t$, the dynamic equations of (1) become

$$\begin{aligned} \dot{x}_\epsilon(s) &= \epsilon\, f(x_\epsilon(s), y_\epsilon(s)), \\ \dot{y}_\epsilon(s) &= g(x_\epsilon(s), y_\epsilon(s)), \quad s \geq 0. \end{aligned} \tag{2}$$

Setting $\epsilon = 0$ yields the so-called boundary layer systems, where the slow state $x \in \mathbb{R}^n$ is fixed.

Assumption 2.2 (exponential stability of the boundary layer systems). The boundary layer systems

$$\dot{y}_x(s) = g(x, y_x(s)), \quad y_x(0) = y^0, \quad s \geq 0,$$

are exponentially stable. In particular, there are constants $D \geq 1$, $\beta > 0$ such that for any $x \in \mathbb{R}^n$ there is a $\phi(x) \in \mathbb{R}^m$ with

$$\|y_x(s) - \phi(x)\| \leq De^{-\beta s}\|y^0 - \phi(x)\|$$

for all $s \geq 0$ and $y^0 \in \mathbb{R}^m$.

The graph of the mapping $\phi : \mathbb{R}^n \to \mathbb{R}^m$ defines the so-called equilibrium manifold in $\mathbb{R}^n \times \mathbb{R}^m$, whose regularity is investigated in the next section. Plugging for each $x \in \mathbb{R}^n$ the corresponding equilibrium $\phi(x) \in \mathbb{R}^m$ into the dynamic equations for the slow variable in (1) yields the so-called reduced system, which defines an approximating limit system for the x-trajectories, as $\epsilon \to 0$; see Theorem 3.5.

Assumption 2.3 (exponential stability of the reduced system). The reduced system

$$\dot{x}_0(t) = f(x_0(t), \phi(x_0(t))), \quad x_0(0) = x^0, \quad t \geq 0,$$

is exponentially stable. In particular, there are constants $C \geq 1$, $\alpha > 0$ such that

$$\|x_0(t)\| \leq Ce^{-\alpha t}\|x^0\|$$

for all $t \geq 0$ and $x^0 \in \mathbb{R}^n$.

These assumptions guarantee the exponential stability not only of the singularly perturbed differential equation (1) but also of a particular multivalued inflation of (1).

DEFINITION 2.4 (multivalued perturbation). *For a nondecreasing continuous function $\kappa : [0, \infty) \to [0, \infty)$ with $\kappa(0) = 0$ we define the multivalued perturbed vector fields*

$$F(x, y, \epsilon) := f(x, y) + \kappa(\epsilon)(\|x\| + \|y - \phi(x)\|)\mathbf{B}_{\mathbb{R}^n},$$

$$G(x, y, \epsilon) := g(x, y) + \kappa(\epsilon)(\|x\| + \|y - \phi(x)\|)\mathbf{B}_{\mathbb{R}^m},$$

where $\mathbf{B}_{\mathbb{R}^n}$ (respectively, $\mathbf{B}_{\mathbb{R}^m}$) denotes the closed unit ball of $\mathbb{R}^n$ (respectively, $\mathbb{R}^m$).

The multivalued perturbed vector fields produce a singularly perturbed differential inclusion

$$\begin{aligned} \dot{x}_\epsilon(t) &\in F(x_\epsilon(t), y_\epsilon(t), \epsilon), \quad x_\epsilon(0) = x^0, \\ \epsilon\, \dot{y}_\epsilon(t) &\in G(x_\epsilon(t), y_\epsilon(t), \epsilon), \qquad y_\epsilon(0) = y^0, \quad t \geq 0, \end{aligned} \tag{3}$$

which can be considered as a multivalued inflation of (1). We are now in a position to formulate the main result on the uniform exponential stability of nonlinear singularly perturbed systems.

THEOREM 2.5. *Let Assumptions 2.1, 2.2, and 2.3 be effective. There is a constant $E > 0$ such that for any $\delta > 0$ there is an $\epsilon_\delta > 0$ such that for any solution to (3) we can estimate*

$$\begin{aligned} \|x_\epsilon(t)\| &\leq (C + \delta)e^{-(\alpha-\delta)t}\|x^0\| + \epsilon E e^{-(\alpha-\delta)t}\|y^0 - \phi(x^0)\|, \\ \|y_\epsilon(t) - \phi(x_\epsilon(t))\| &\leq (D + \delta)e^{-(\beta-\delta)\frac{t}{\epsilon}}\|y^0 - \phi(x^0)\| + \epsilon E e^{-(\alpha-\delta)t}\|x^0\| \\ &\quad + \epsilon^2 E e^{-(\alpha-\delta)t}\|y^0 - \phi(x^0)\| \end{aligned}$$

for all initial values $x^0 \in \mathbb{R}^n$, $y^0 \in \mathbb{R}^m$, all times $t \geq 0$, and all perturbation parameters $\epsilon \in (0, \epsilon_\delta]$.

Note that the gain and decay of the leading terms in the estimations above approximate the exponential characteristics of the reduced system and the boundary layer systems. Surprisingly, the higher order terms are linear and quadratic, despite the fact that the order of the multivalued regular perturbation, determined by $\kappa(\epsilon)$, might be arbitrary.

What follows is a short discussion on related results. First, it is mentionable that in the pioneering work of Tychonov on the approximation of singularly perturbed differential equations (see [15] or also the survey [16] and the references therein) only (uniform) asymptotic stability properties of the fast subsystem are used. Moreover, as elaborated in [13] using Lyapunov's matrix function method, under certain conditions the additional asymptotic stability of the reduced system forces the asymptotical stability of the whole system for sufficiently small perturbation parameters. Related results can be found in [12]. The interest in exponential stability properties, as investigated in [14, 10, 4], seems to originate from the possibility of verifying it via linearization. In the present paper, we focus on additional multivalued perturbations. Hence it is essential to concentrate on exponential stability, since it is well known that mere asymptotic stability is not preserved even under small linearly bounded regular perturbations.

Second, we emphasize that the particular structure of the singularly perturbed differential inclusion (3) is essential in order to prove Theorem 2.5. The main tool, Tychonov's theorem, is valid for multivalued differential equations only under additional suppositions on the fast subsystems that are natural for single-valued differential equations but quite restrictive for multivalued differential equations. The point is that, in general, the reduced system does not produce all slow limit trajectories, as the perturbation parameter tends to zero. In other words, it is difficult to achieve upper semicontinuity at $\epsilon = 0$ of the solution mapping assigning to $\epsilon > 0$ the (slow) solution set to a singularly perturbed differential inclusion and to $\epsilon = 0$ the solution set to the reduced system. In [5] this problem is solved via a restriction to Lipschitz continuous solutions that excludes oscillations of the fast trajectories. Another approach is attempted in [17, 18]. Here, forward invariance and attractivity of a set of fast equilibrium states is assumed in order to achieve upper semicontinuity of the solution mapping. Recently, this idea has been enhanced in [19], where problems on an infinite time horizon are under investigation. However, for singularly perturbed differential inclusions not possessing a particular structure of the fast subinclusion, the use of averaging techniques (see [1] and also [6] for singularly perturbed nonlinear control systems) is indispensable in order to construct a sufficiently rich limit system.

3. Approximation. In this section we prove a version of Tychonov's theorem on the approximating properties of the reduced system, Theorem 3.5. This theorem applies to system (3) on a time interval $t \in [0, T]$, where $T > 0$ can be considered as a discretization step of the slow variable.

We proceed as follows. First, we show the Lipschitz continuity of the equilibrium manifold in order to obtain unique trajectories for the reduced system. Then we present a simple lemma on perturbed linear recursions which is used to obtain estimations for the slow variables on time intervals $[kT, (k+1)T]$ for $k \in \mathbb{N}$.

In what follows we consider the singularly perturbed system (3) in the natural time scale of the fast subsystem on the time interval $s \in \left[0, \frac{T}{\epsilon}\right]$. We introduce a step size of the fast variable, $H > 0$, and show boundedness of the trajectories for $s \in [0, H]$ uniformly in $\epsilon > 0$. An iteration of this result gives bounds for the trajectories on the whole time interval $s \in \left[0, \frac{T}{\epsilon}\right]$. Again, for the iteration the lemma on perturbed linear recursions is needed to show the exponential decay of the fast variables on time intervals $[kH, (k+1)H]$ for $k \in \mathbb{N}$.

3.1. Lipschitz continuity of the equilibrium manifold. The exponential stability of the boundary layer systems forces the Lipschitz continuity of the equilibrium manifold.

LEMMA 3.1. *Let Assumptions* 2.1 *and* 2.2 *be effective. The mapping* $\phi : \mathbb{R}^n \to \mathbb{R}^m$, $x \mapsto \phi(x)$ *is Lipschitz continuous with Lipschitz constant* $K = 2LSe^{LS}$, *where* $S = \log(2D)/\beta$.

Proof. For $x \in \mathbb{R}^n$, we define a mapping $A_x : \mathbb{R}^m \to \mathbb{R}^m$ by $A_x(y^0) = y_x(S)$ (in which $y_x(\cdot)$ is the solution to the corresponding boundary layer system with $y_x(0) = y^0$), where $S \geq 0$ is chosen such that

$$De^{-\beta S} = \frac{1}{2}.$$

For $x_1, x_2 \in \mathbb{R}^n$, using Assumption 2.2, we conclude that

$$\begin{aligned}\|\phi(x_1) - \phi(x_2)\| &= \|A_{x_1}(\phi(x_1)) - A_{x_2}(\phi(x_2))\| \\ &\leq \|A_{x_1}(\phi(x_1)) - A_{x_1}(\phi(x_2))\| + \|A_{x_1}(\phi(x_2)) - A_{x_2}(\phi(x_2))\| \\ &\leq De^{-\beta S}\|\phi(x_1) - \phi(x_2)\| + \|x_1 - x_2\| LSe^{LS}.\end{aligned}$$

We obtain

$$\|\phi(x_1) - \phi(x_2)\| \leq 2\|x_1 - x_2\| LSe^{LS},$$

and the proof is finished. ∎

3.2. Perturbed linear recursions. For two matrices $M, N \in \mathbb{R}^{2\times 2}$ we write $M \leq N$ if $M_{ij} \leq N_{ij}$ for all $1 \leq i, j \leq 2$.

LEMMA 3.2. *Let $a, b, c > 0$ be real constants with $a \neq b$ and $M_\epsilon \in \mathbb{R}^{2\times 2}$ defined by*

$$M_\epsilon := \begin{pmatrix} a + \epsilon c & \epsilon c \\ \epsilon c & b + \epsilon c \end{pmatrix}$$

for any $\epsilon \geq 0$. Then there is an $\epsilon_0 > 0$ such that we can estimate

$$0 \leq M_\epsilon^k \leq \begin{pmatrix} (a+\epsilon 2c)^k + \epsilon^2 (2d)^2 (b+\epsilon 2c)^k & \epsilon 2d(a+\epsilon 2c)^k + \epsilon 2d(b+\epsilon 2c)^k \\ \epsilon 2d(a+\epsilon 2c)^k + \epsilon 2d(b+\epsilon 2c)^k & (b+\epsilon 2c)^k + \epsilon^2 (2d)^2 (a+\epsilon 2c)^k \end{pmatrix}$$

for $k = 0, 1, 2, \ldots$, $\epsilon \in (0, \epsilon_0]$, where $d = \frac{c}{|a-b|}$.

Proof. The first inequality is obvious; we concentrate on the second one. The unperturbed matrix M_0 is diagonal with eigenvalues $\lambda_1(0) = a$, $\lambda_2(0) = b$ and eigenvectors $v_1(0) = (1, 0)^T$, $v_2(0) = (0, 1)^T$. Hence, for $\epsilon > 0$ sufficiently small, the characteristic polynomial of the perturbed matrix M_ϵ has two different zeros as well. A straightforward calculation yields

$$\left|\frac{\mathrm{d}\lambda_1}{\mathrm{d}\epsilon}(0)\right| = c, \qquad \left|\frac{\mathrm{d}\lambda_2}{\mathrm{d}\epsilon}(0)\right| = c.$$

Hence, for $\epsilon > 0$, we obtain the estimations $|\lambda_1(\epsilon) - a| \leq 2c\epsilon$, $|\lambda_2(\epsilon) - b| \leq 2c\epsilon$ for the eigenvalues of M_ϵ. For $\epsilon > 0$ we are looking for eigenvectors of the form $v_1(\epsilon) = (1, w(\epsilon))$, $v_2(\epsilon) = (u(\epsilon), 1)$. Straightforward computations yield

$$\left|\frac{\mathrm{d}w}{\mathrm{d}\epsilon}(0)\right| = \frac{c}{|a-b|}, \qquad \left|\frac{\mathrm{d}u}{\mathrm{d}\epsilon}(0)\right| = \frac{c}{|a-b|},$$

which allows us to estimate $|w(\epsilon)| \leq \frac{2c}{|a-b|}$, $|u(\epsilon)| \leq \frac{2c}{|a-b|}$ for sufficiently small $\epsilon > 0$. We conclude that

$$M_\epsilon^k \leq \begin{pmatrix} 1 & \epsilon\frac{2c}{|a-b|} \\ \epsilon\frac{2c}{|a-b|} & 1 \end{pmatrix} \begin{pmatrix} a+\epsilon 2c & 0 \\ 0 & b+\epsilon 2c \end{pmatrix}^k \begin{pmatrix} 1 & \epsilon\frac{2c}{|a-b|} \\ \epsilon\frac{2c}{|a-b|} & 1 \end{pmatrix}$$

for $\epsilon > 0$ sufficiently small, and the claim easily follows. ∎

3.3. Boundedness of the trajectories. In what follows, we consider the singularly perturbed differential inclusion (3) in the natural time scale of the fast subsystem, $s = \epsilon t$. Then the system becomes

$$\begin{aligned} \dot{x}_\epsilon(s) &\in \epsilon\, F(x_\epsilon(s), y_\epsilon(s), \epsilon), \quad x_\epsilon(0) = x^0, \\ \dot{y}_\epsilon(s) &\in G(x_\epsilon(s), y_\epsilon(s), \epsilon), \quad y_\epsilon(0) = y^0, \quad s \geq 0. \end{aligned} \tag{4}$$

Tychonov's theorem on the approximation applies to a time horizon that is bounded in the natural time scale of the slow subsystem. Hence, for a fixed $T \geq 0$,

we are interested in the system above for times $s \in [0, \frac{T}{\epsilon}]$. To this end, we consider the system above first on a time horizon $s \in [0, H]$, where $0 < H \leq \frac{T}{\epsilon}$. Since the estimations obtained are linear in the initial values, we can iterate the procedure $\left[\frac{T}{H\epsilon}\right]$ times (here, we set $[x] := \max\{k \in \mathbb{N} : \ k \leq x\}$ for real numbers $x \in \mathbb{R}$) and obtain estimations on the whole interval $s \in [0, \frac{T}{\epsilon}]$.

For an arbitrary $H > 0$ and $\epsilon > 0$ we set

$$P_H^\epsilon := \max_{(x_\epsilon, y_\epsilon)} \max_{s \in [0,H]} \|F(x_\epsilon(s), y_\epsilon(s), \epsilon)\|,$$

where the first maximum is taken over all trajectories (x_ϵ, y_ϵ) of (4).

LEMMA 3.3. *Let Assumptions* 2.1 *and* 2.2 *be effective. For any $H > 0$ there is an $\epsilon_H > 0$ such that*

$$P_H^\epsilon \leq 2(L + LK + 2)\|x^0\| + 4LD\|y^0 - \phi(x^0)\|$$

for all $\epsilon \in (0, \epsilon_H]$.

Proof. We set $M_s := \max_{0 \leq s' \leq s} \|y_\epsilon(s') - \phi(x_\epsilon(s'))\|$ and $N_s := \max_{0 \leq s' \leq s} \|x_\epsilon(s')\|$. Using the Gronwall lemma, for $s \in [0, H]$, we calculate

$$\begin{aligned}
&\|y_\epsilon(s) - \phi(x_\epsilon(s))\| \\
&\leq \|y_\epsilon(s) - y_{x^0}(s, y^0)\| + \|y_{x^0}(s, y^0) - \phi(x^0)\| + \|\phi(x^0) - \phi(x_\epsilon(s))\| \\
&\leq Lse^{Ls}\left(s\epsilon P_H^\epsilon + \kappa(\epsilon)N_s + \kappa(\epsilon)M_s\right) + De^{-\beta s}\|y^0 - \phi(x^0)\| + Ks\epsilon P_H^\epsilon \\
&\leq (Lse^{Ls}s\epsilon + Ks\epsilon)P_H^\epsilon + Lse^{Ls}\kappa(\epsilon)N_s + Lse^{Ls}\kappa(\epsilon)M_s + De^{-\beta s}\|y^0 - \phi(x^0)\|. \qquad (5)
\end{aligned}$$

We fix an arbitrary $\eta \in (0, \frac{1}{2}]$. Then, for $\epsilon_H > 0$ fulfilling

$$LHe^{LH}\kappa(\epsilon_H) \leq \eta \qquad (6)$$

and $\epsilon \in (0, \epsilon_H]$ we obtain

$$M_H \leq 2(LHe^{LH}H\epsilon + KH\epsilon)P_H^\epsilon + N_H + 2D\|y^0 - \phi(x^0)\|.$$

For $\epsilon_H > 0$ fulfilling (6) and

$$\kappa(\epsilon_H) \leq 1 \qquad (7)$$

we calculate

$$\begin{aligned}
P_H^\epsilon &\leq LN_H + L\max_{0 \leq s \leq H}\|y_\epsilon(s)\| + \kappa(\epsilon)N_H + \kappa(\epsilon)M_H \\
&\leq (L+1)N_H + L\max_{0 \leq s \leq H}\|\phi(x_\epsilon(s))\| + (L+1)M_H \\
&\leq (L + LK + 1)N_H + (L+1)M_H \\
&\leq (L + LK + 2)(\|x^0\| + \epsilon H P_H^\epsilon) + 2(L+1)(LHe^{LH}H\epsilon + KH\epsilon)P_H^\epsilon \\
&\quad + 2LD\|y^0 - \phi(x^0)\|
\end{aligned}$$

for all $\epsilon \in (0, \epsilon_H]$. Hence, for $\epsilon_H > 0$ fulfilling (6), (7), and

$$(L + LK + 2)\epsilon_H H + 2(L+1)(LHe^{LH}H\epsilon_H + KH\epsilon_H) \leq \frac{1}{2} \qquad (8)$$

we obtain the required estimation. ∎

LEMMA 3.4 (boundedness of the trajectories). *Let Assumptions* 2.1 *and* 2.2 *be effective. For any* $\delta \in (0,\beta)$ *there is an* $\epsilon_0 > 0$ *such that for any* $T > 0$ *there is a constant* $C_T \geq 0$ *with*

$$\begin{aligned}\|x_\epsilon(s)\| &\leq C_T\|x^0\| + \epsilon C_T\|y^0 - \phi(x^0)\|,\\ \|y_\epsilon(s) - \phi(x_\epsilon(s))\| &\leq \epsilon C_T\|x^0\| + (D+\delta)e^{-(\beta-\delta)s}\|y^0 - \phi(x^0)\| + \epsilon^2 C_T\|y^0 - \phi(x^0)\|\end{aligned}$$

for $\epsilon \in (0,\epsilon_0]$, $s \in \left[0, \frac{T}{\epsilon}\right]$. *Moreover, for* $\epsilon \in (0,\epsilon_0]$ *and* $s = \frac{T}{\epsilon}$, *the improved estimation*

$$\left\| y_\epsilon\left(\frac{T}{\epsilon}\right) - \phi\left(x_\epsilon\left(\frac{T}{\epsilon}\right)\right)\right\| \leq \epsilon C_T\|x^0\| + e^{-(\beta-\delta)\frac{T}{\epsilon}}\|y^0 - \phi(x^0)\| + \epsilon^2 C_T\|y^0 - \phi(x^0)\|$$

is valid.

Proof. Let $\delta \in (0,\beta)$. We choose $H \geq 1$ such that

$$e^{-\delta H} \leq \frac{1}{2D} \quad \text{and} \quad e^{-(\beta-\delta)H} \leq \frac{1}{2},$$

and then $\eta \in (0, \frac{1}{2}]$ and $\epsilon_0 \in (0, \epsilon_H]$, where $\epsilon_H > 0$ is given by (6), (7), and (8) such that

$$\frac{\epsilon_0 2D}{\epsilon_H L} + 4LD\epsilon_0 H + \eta 2D \leq \min\left\{\frac{e^{-(\beta-\delta)H}}{2}, \delta e^{-(\beta-\delta)H}\right\}.$$

From Lemma 3.3 and (5) we obtain

$$\begin{aligned}&\|y_\epsilon(s) - \phi(x_\epsilon(s))\|\\ &\leq \frac{\epsilon P_H^\epsilon}{\epsilon_H 4L} + \eta N_s + \eta M_s + De^{-\beta s}\|y^0 - \phi(x^0)\|\\ &\leq \frac{\epsilon P_H^\epsilon}{\epsilon_H 4L} + \eta N_H + \eta\left(\frac{\epsilon P_H^\epsilon}{\epsilon_H 2L} + N_H + 2D\|y^0 - \phi(x^0)\|\right) + De^{-\beta s}\|y^0 - \phi(x^0)\|\\ &\leq \frac{\epsilon P_H^\epsilon}{\epsilon_H 4L} + 2\eta\epsilon H P_H^\epsilon + \eta\left(\frac{\epsilon P_H^\epsilon}{\epsilon_H 2L} + 2D\|y^0 - \phi(x^0)\|\right) + De^{-\beta s}\|y^0 - \phi(x^0)\|\\ &\leq \left(2(L + LK + 2)\|x^0\| + 4LD\|y^0 - \phi(x^0)\|\right)\left(\frac{\epsilon}{\epsilon_H 4L} + 2\eta\epsilon H + \frac{\eta\epsilon}{\epsilon_H 2L}\right)\\ &\quad + \eta 2D\|y^0 - \phi(x^0)\| + De^{-\beta s}\|y^0 - \phi(x^0)\|. \qquad (9)\end{aligned}$$

Furthermore, we have

$$\|x_\epsilon(s)\| \leq \|x^0\| + \epsilon H\left(2(L + LK + 2)\|x^0\| + 4LD\|y^0 - \phi(x^0)\|\right)$$

for all $s \in [0, H]$, where $\epsilon_H > 0$ is given by (6) and (8). In particular, for $s = H$, we obtain

$$\begin{aligned}\|y_\epsilon(H) - \phi(x_\epsilon(H))\| &\leq (L + LK + 2)\left(\frac{\epsilon}{\epsilon_H L} + \epsilon 2H\right)\|x^0\|\\ &\quad + \left(\frac{\epsilon 2D}{\epsilon_H L} + 4LD\epsilon H + \eta 2D + De^{-\beta H}\right)\|y^0 - \phi(x^0)\|\\ &\leq (L + LK + 2)\left(\frac{\epsilon}{\epsilon_H L} + \epsilon 2H\right)\|x^0\|\\ &\quad + e^{-(\beta-\delta)H}\|y^0 - \phi(x^0)\|\end{aligned}$$

and

$$\|x_\epsilon(H)\| \le (1 + \epsilon H 2(L + LK + 2))\,\|x^0\| + \epsilon H 2LD\|y^0 - \phi(x^0)\|.$$

For $k = 0, 1, \dots, k_H := \left[\frac{T}{\epsilon H}\right]$ we set

$$X_k := \|x_\epsilon(kH)\|, \qquad Y_k := \|y_\epsilon(kH) - \phi(x_\epsilon(kH))\|.$$

Then we obtain the linear recursive relation

$$\begin{aligned} X_{k+1} &\le (1 + \epsilon C_H)X_k + \epsilon C_H Y_k, \\ Y_{k+1} &\le \epsilon C_H X_k + e^{-(\beta-\delta)H} Y_k, \end{aligned}$$

where

$$C_H := \max\left\{(L + LK + 2)\left(\frac{1}{\epsilon_H L} + 2H\right), 2H(L + LK + 2), 2HLD\right\}.$$

The corresponding matrix is

$$M_\epsilon := \begin{pmatrix} 1 + \epsilon C_H & \epsilon C_H \\ \epsilon C_H & e^{-(\beta-\delta)H} \end{pmatrix}.$$

According to Lemma 3.2 we can estimate

$$0 \le M_\epsilon^k \le \begin{pmatrix} (1 + \epsilon 2C_H)^k + \epsilon^2(2d)^2 & \epsilon 2d(1 + \epsilon 2C_H)^k + \epsilon 2d \\ \epsilon 2d(1 + \epsilon 2C_H)^k + \epsilon 2d & e^{-(\beta-\delta)kH} + \epsilon^2(2d)^2(1 + \epsilon 2C_H)^k \end{pmatrix}$$

with $d = 2C_H$. Since

$$(1 + \epsilon 2C_H)^k \le (1 + \epsilon 2C_H)^{\frac{T}{\epsilon H}} \le e^{\frac{2C_H T}{H}} \le e^{2C_H T}$$

for $k = 0, 1, \dots, k_H$, the estimation for the slow variable follows easily. As for the fast variable we have

$$\|y_\epsilon(kH) - \phi(x_\epsilon(kH))\| \le \epsilon C_T\|x^0\| + e^{-(\beta-\delta)kH}\|y^0 - \phi(x^0)\| + \epsilon^2 C_T\|y^0 - \phi(x^0)\|$$

for any $k = 0, 1, \dots, k_H$. For times $s \in [kH, (k+1)H]$, the required estimation follows from (9) and from the choice of η and ϵ_0. □

3.4. A version of Tychonov's theorem. Now we are in a position to prove an appropriate version of Tychonov's theorem on the uniform approximation of the slow trajectories. The interesting aspect of this version is the linearity of the estimations with respect to the initial conditions. The theorem is formulated and proven in the natural time scale of the fast subsystem $s = \epsilon t$; hence $x_\epsilon(\cdot)$ denotes the solution to the singularly perturbed system (4), whereas $x_0(\cdot)$ is the solution to the reduced system

$$\dot{x}_0(s) = \epsilon\, f(x_0(s), \phi(x_0(s))), \quad x_0(0) = x^0, \quad s \in \left[0, \frac{T}{\epsilon}\right].$$

Note that this version of Tychonov's theorem does not use Assumption 2.3, and hence gives information on the approximation of the slow variable on a time horizon that is bounded in the natural time scale of the slow variable. Under exponential stability

suppositions on the reduced system, the result can be extended to an infinite time interval; see [9].

THEOREM 3.5 (Tychonov). *Let Assumptions* 2.1 *and* 2.2 *be effective. There is an* $\epsilon_0 > 0$ *such that for any* $T > 0$ *there is a* $B_T \geq 0$ *such that we can estimate*

$$\max_{0 \leq s \leq \frac{T}{\epsilon}} \|x_\epsilon(s) - x_0(s)\| \leq (\epsilon + \kappa(\epsilon)) B_T \|x^0\| + \epsilon B_T \|y^0 - \phi(x^0)\|$$

for all initial values $x^0 \in \mathbb{R}^n$, $y^0 \in \mathbb{R}^m$ *and all perturbation parameters* $\epsilon \in (0, \epsilon_0]$.

Proof. We fix an arbitrary $\delta \in (0, \beta)$. Let $\epsilon_0 > 0$ be the corresponding upper bound for the perturbation parameter given by Lemma 3.4. Using Lemma 3.4, we calculate, for $s \in \left[0, \frac{T}{\epsilon}\right]$,

$$\begin{aligned}
&\|x_\epsilon(s) - x_0(s)\| \\
&\leq \int_0^s \epsilon \|f(x_\epsilon(s'), y_\epsilon(s')) - f(x_0(s'), \phi(x_0(s')))\| ds' \\
&\quad + \int_0^s \epsilon\kappa(\epsilon) \left(\|x_\epsilon(s')\| + \|y_\epsilon(s') - \phi(x_\epsilon(s'))\|\right) ds' \\
&\leq \int_0^s \epsilon L \left(\|x_\epsilon(s') - x_0(s')\| + \|y_\epsilon(s') - \phi(x_0(s'))\|\right) ds' \\
&\quad + \int_0^s \epsilon\kappa(\epsilon) \left(\|x_\epsilon(s')\| + \|y_\epsilon(s') - \phi(x_\epsilon(s'))\|\right) ds' \\
&\leq \int_0^s \epsilon L \left(\|x_\epsilon(s') - x_0(s')\| + \|y_\epsilon(s') - \phi(x_\epsilon(s'))\| + \|\phi(x_\epsilon(s')) - \phi(x_0(s'))\|\right) ds' \\
&\quad + \int_0^s \epsilon\kappa(\epsilon) \left(\|x_\epsilon(s')\| + \|y_\epsilon(s') - \phi(x_\epsilon(s'))\|\right) ds' \\
&\leq \int_0^s \epsilon L(1+K) \|x_\epsilon(s') - x_0(s')\| + \epsilon(L + \kappa(\epsilon)) \|y_\epsilon(s') - \phi(x_\epsilon(s'))\| + \epsilon\kappa(\epsilon) \|x_\epsilon(s')\| ds' \\
&\leq \int_0^s \epsilon L(1+K) \|x_\epsilon(s') - x_0(s')\| ds' \\
&\quad + \int_0^s \epsilon(L + \kappa(\epsilon)) \left(\epsilon C_T \|x^0\| + (D + \delta) e^{-(\beta - \delta)s'} \|y^0 - \phi(x^0)\|\right) ds' \\
&\quad + \int_0^s \epsilon(L + \kappa(\epsilon)) \epsilon^2 C_T \|y^0 - \phi(x^0)\| ds' \\
&\quad + \int_0^s \epsilon\kappa(\epsilon) \left(C_T \|x^0\| + \epsilon C_T \|y^0 - \phi(x^0)\|\right) ds'
\end{aligned}$$

for all $\epsilon \in (0, \epsilon_0]$. Hence, the Gronwall lemma yields the required estimate. ∎

4. Proof of Theorem 2.5. We fix a $\delta \in (0, \min\{\alpha, \beta\})$ and choose $T > 0$ such that

$$e^{-\delta T} \leq \frac{1}{2C}.$$

Let $\epsilon_0 > 0$ be the corresponding upper bound for the perturbation parameter given by Lemma 3.4. Then we choose $\epsilon_\delta \in (0, \epsilon_0]$ such that

$$(\epsilon_\delta + \kappa(\epsilon_\delta)) B_T \leq \min\left\{ \frac{e^{-(\alpha - \delta)T}}{2}, \delta e^{-(\alpha - \delta)T} \right\}.$$

For $s \in \left[0, \frac{T}{\epsilon}\right]$ we calculate

$$\begin{aligned} \|x_\epsilon(s)\| &\le \|x_\epsilon(s) - x_0(s)\| + \|x_0(s)\| \\ &\le (\epsilon + \kappa(\epsilon)) B_T \|x^0\| + \epsilon B_T \|y^0 - \phi(x^0)\| + C e^{-\alpha\epsilon s} \|x^0\|. \end{aligned} \tag{10}$$

In particular, for $s = \frac{T}{\epsilon}$, we obtain

$$\left\| x_\epsilon\left(\frac{T}{\epsilon}\right) \right\| \le e^{-(\alpha-\delta)T} \|x^0\| + \epsilon B_T \|y^0 - \phi(x^0)\|.$$

Moreover, by Lemma 3.4 we have

$$\left\| y_\epsilon\left(\frac{T}{\epsilon}\right) - \phi\left(x_\epsilon\left(\frac{T}{\epsilon}\right)\right) \right\| \le \epsilon C_T \|x^0\| + e^{-(\beta-\delta)\frac{T}{\epsilon}} \|y^0 - \phi(x^0)\| + \epsilon^2 C_T \|y^0 - \phi(x^0)\|.$$

For $k = 0, 1, 2, \ldots$, we set

$$X_k := \left\| x_\epsilon\left(\frac{kT}{\epsilon}\right) \right\|, \qquad Y_k := \left\| y_\epsilon\left(\frac{kT}{\epsilon}\right) - \phi\left(x_\epsilon\left(\frac{kT}{\epsilon}\right)\right) \right\|.$$

Then we obtain the linear recursive relation

$$\begin{aligned} X_{k+1} &\le e^{-(\alpha-\delta)T} X_k + \epsilon C' Y_k, \\ Y_{k+1} &\le \epsilon C' X_k + e^{-(\beta-\delta)\frac{T}{\epsilon}} Y_k + \epsilon C' Y_k, \end{aligned}$$

where

$$C' := \max\{B_T, C_T\}.$$

The corresponding matrix becomes

$$M_\epsilon := \begin{pmatrix} e^{-(\alpha-\delta)T} & \epsilon C' \\ \epsilon C' & e^{-(\beta-\delta)\frac{T}{\epsilon}} + \epsilon C' \end{pmatrix}.$$

As for the fast variables, the claim follows directly from Lemmas 3.2 and 3.4. For the slow variables we have

$$\left\| x_\epsilon\left(\frac{kT}{\epsilon}\right) \right\| \le e^{-(\alpha-\delta)kT} \|x^0\| + \epsilon E e^{-(\alpha-\delta)kT} \|y^0 - \phi(x^0)\|$$

for $k \in \mathbb{N}$. For times $s \in [\frac{kT}{\epsilon}, \frac{(k+1)T}{\epsilon}]$ the required estimation follows from (10) and from the choice of ϵ_δ. □

Remark 4.1. The proof of Theorem 2.5 is based on Gronwall-type estimations only; in particular no compactness arguments are used. Hence, the result is valid for systems in arbitrary real Banach spaces, and the proofs remain unchanged.

Remark 4.2. An inspection of the proof of Theorem 2.5 shows that the condition $\kappa(0) = 0$ in Definition 2.4 is only needed in order to achieve that the gain and exponential decay rates of the singularly perturbed multivalued differential equation (3) approximate the corresponding values of the reduced system and the boundary layer systems. If we are merely interested in the exponential stability of the singularly perturbed multivalued differential equation (3) for small perturbation parameters, then we need only that $\kappa(0) \ge 0$ is sufficiently small.

5. An application to small delays. An important application of decomposition results such as Theorem 2.5 is the construction of feedback controls. Assumptions 2.2 and 2.3 along with Theorem 2.5 allow a significant simplification of this problem, since the decoupled stabilization takes place in lower dimensional spaces. According to Theorem 2.5, one has to construct the feedback in a way that the boundary layer systems and the reduced system are exponentially stable. This construction is described in detail in [10] via Lyapunov functions.

Still the question arises to what extent this feedback is robust with respect to delays. In other words, for what kind of mappings $h : [0,\infty) \to [0,\infty)$ is the system with delays

$$\begin{aligned} \dot{x}_\epsilon(t) &= \tilde{f}(x_\epsilon(t), y_\epsilon(t), x_\epsilon(t-h(\epsilon)), y_\epsilon(t-h(\epsilon))), \quad t>0,\\ \epsilon\,\dot{y}_\epsilon(t) &= \tilde{g}(x_\epsilon(t), y_\epsilon(t), x_\epsilon(t-h(\epsilon)), y_\epsilon(t-h(\epsilon))), \quad t>0,\\ x_\epsilon(t) &= x^0, \quad t\in[-h(\epsilon),0],\\ y_\epsilon(t) &= y^0, \quad t\in[-h(\epsilon),0], \end{aligned} \tag{11}$$

exponentially stable for $\epsilon > 0$ sufficiently small? Since Theorem 2.5 applies to multivalued perturbations as given by (3), we can answer this question. To this end we formulate the delay as a regular perturbation, which is automatically contained in the differential inclusion (3). A similar idea has been used in [11] in order to show that exponentially stabilizing feedback controls are robust with respect to small delays.

THEOREM 5.1. *Let the mappings $\tilde{f} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, $\tilde{g} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be Lipschitz continuous with Lipschitz constant $L \geq 0$. Let $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ given by*

$$f(x,y) := \tilde{f}(x,y,x,y), \quad g(x,y) := \tilde{g}(x,y,x,y)$$

satisfy Assumptions 2.1, 2.2, *and* 2.3. *Let $h : [0,\infty) \to [0,\infty)$ satisfy*

$$\frac{h(\epsilon)}{\epsilon} \to 0, \quad \text{as} \quad \epsilon \to 0.$$

Then there is a constant $E > 0$ such that for any $\delta > 0$ there is an $\epsilon_\delta > 0$ such that for any solution to the delayed feedback system (11) *we can estimate*

$$\begin{aligned} \|x_\epsilon(t)\| &\leq (C+\delta)e^{-(\alpha-\delta)t}\|x^0\| + \epsilon E e^{-(\alpha-\delta)t}\|y^0 - \phi(x^0)\|,\\ \|y_\epsilon(t) - \phi(x_\epsilon(t))\| &\leq (D+\delta)e^{-(\beta-\delta)t/\epsilon}\|y^0 - \phi(x^0)\| + \epsilon E e^{-(\alpha-\delta)t}\|x^0\|\\ &\quad + \epsilon^2 E e^{-(\alpha-\delta)t}\|y^0 - \phi(x^0)\| \end{aligned}$$

for all initial values $x^0 \in \mathbb{R}^n$, $y^0 \in \mathbb{R}^m$, all times $t \geq 0$, and all perturbation parameters $\epsilon \in (0, \epsilon_\delta]$.

Proof. Let (x_ϵ, y_ϵ) be a solution to (11). Without loss of generality we can assume that $0 < \epsilon \leq 1$. Then the right-hand side of the differential equation for x_ϵ is of order $\frac{1}{\epsilon}$ as well, and the x_ϵ trajectory can be treated as the y_ϵ trajectory. For short we write $z_\epsilon := (x_\epsilon, y_\epsilon)$ and use the norm $\|z_\epsilon(t)\| = \|x_\epsilon(t)\| + \|y_\epsilon(t)\|$. We claim that

$$\max_{t-h(\epsilon)\leq s\leq t} \|z_\epsilon(s)\| \leq 2\|z_\epsilon(t)\| \tag{12}$$

for all $t \geq 0$ and all sufficiently small perturbation parameters $\epsilon > 0$. We prove (12) by induction. First, let $t \in [0, h(\epsilon)]$ and $t^* \in [0,t]$ with

$$\|z_\epsilon(t^*)\| = \max_{t-h(\epsilon)\leq s\leq t} \|z_\epsilon(s)\|.$$

Then we can estimate

$$\begin{aligned}\|z_\epsilon(t)\| &\geq \|z_\epsilon(t^*)\| - L\frac{h(\epsilon)}{\epsilon}\max_{t-h(\epsilon)\leq s\leq t}\|z_\epsilon(s)\| - L\frac{h(\epsilon)}{\epsilon}\|z^0\| \\ &\geq \left(1 - 2L\frac{h(\epsilon)}{\epsilon}\right)\max_{t-h(\epsilon)\leq s\leq t}\|z_\epsilon(s)\|,\end{aligned}$$

which yields the estimation (12) for sufficiently small $\epsilon > 0$. In the next step we let $T \geq h(\epsilon)$ and assume that the estimation (12) is valid for $t = T - h(\epsilon)$. Let $T^* \in [T - h(\epsilon), T]$ such that

$$\|z_\epsilon(T^*)\| = \max_{T-h(\epsilon)\leq s\leq T}\|z_\epsilon(s)\|.$$

Then we can write

$$\begin{aligned}\|z_\epsilon(T)\| &\geq \|z_\epsilon(T^*)\| - L\frac{h(\epsilon)}{\epsilon}\max_{T-h(\epsilon)\leq s\leq T}\|z_\epsilon(s)\| - L\frac{h(\epsilon)}{\epsilon}\max_{T-2h(\epsilon)\leq s\leq T-h(\epsilon)}\|z_\epsilon(s)\| \\ &\geq \max_{T-h(\epsilon)\leq s\leq T}\|z_\epsilon(s)\| - L\frac{h(\epsilon)}{\epsilon}\max_{T-h(\epsilon)\leq s\leq T}\|z_\epsilon(s)\| - L\frac{h(\epsilon)}{\epsilon}2\|z_\epsilon(T-h(\epsilon))\| \\ &\geq \left(1 - L\frac{h(\epsilon)}{\epsilon} - 2L\frac{h(\epsilon)}{\epsilon}\right)\max_{T-h(\epsilon)\leq s\leq T}\|z_\epsilon(s)\|,\end{aligned}$$

which yields the estimation (12) for $t = T$ and sufficiently small $\epsilon > 0$.

In what follows we make use of the estimate

$$\max_{t-2h(\epsilon)\leq s\leq t}\|z_\epsilon(s)\| \leq 4\|z_\epsilon(t)\|, \tag{13}$$

which is valid for all $t \geq h(\epsilon)$ and sufficiently small $\epsilon > 0$. With Assumption 2.1 and (13) we can estimate

$$\|x_\epsilon(t) - x_\epsilon(t-h(\epsilon))\| \leq 4h(\epsilon)L(\|x_\epsilon(t)\| + \|y_\epsilon(t)\|)$$

and

$$\|y_\epsilon(t) - y_\epsilon(t-h(\epsilon))\| \leq 4\frac{h(\epsilon)}{\epsilon}L(\|x_\epsilon(t)\| + \|y_\epsilon(t)\|).$$

We are in a position to formulate the delay in (11) as a regular perturbation. To this end we write

$$\begin{aligned}&\left\|f(x_\epsilon(t), y_\epsilon(t)) - \tilde{f}(x_\epsilon(t), y_\epsilon(t), x_\epsilon(t-h(\epsilon)), y_\epsilon(t-h(\epsilon)))\right\| \\ &\quad\leq 8\frac{h(\epsilon)}{\epsilon}L^2(\|x_\epsilon(t)\| + \|y_\epsilon(t)\|) \\ &\quad\leq 8\frac{h(\epsilon)}{\epsilon}L^2(\|x_\epsilon(t)\| + \|y_\epsilon(t) - \phi(y_\epsilon(t))\| + \|\phi(y_\epsilon(t))\|) \\ &\quad\leq 8\frac{h(\epsilon)}{\epsilon}L^2(1+K)(\|x_\epsilon(t)\| + \|y_\epsilon(t) - \phi(y_\epsilon(t))\|)\end{aligned}$$

and compute in the same way

$$\begin{aligned}&\|g(x_\epsilon(t), y_\epsilon(t)) - \tilde{g}(x_\epsilon(t), y_\epsilon(t), x_\epsilon(t-h(\epsilon)), y_\epsilon(t-h(\epsilon)))\| \\ &\quad\leq 8\frac{h(\epsilon)}{\epsilon}L^2(1+K)(\|x_\epsilon(t)\| + \|y_\epsilon(t) - \phi(y_\epsilon(t))\|).\end{aligned}$$

Hence, the system (11) is included in the differential inclusion (3) with

$$\kappa(\epsilon) = 8\frac{h(\epsilon)}{\epsilon}L^2(1+K),$$

and the proof is finished. □

REFERENCES

[1] Z. Artstein, *Invariant measures of differential inclusions applied to singular perturbations*, J. Differential Equations, 152 (1999), pp. 289–307.

[2] Z. Artstein, *Stability in the presence of singular perturbations*, Nonlinear Anal., 34 (1998), pp. 817–827.

[3] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw–Hill, New York, 1955.

[4] M. Corless and L. Glielmo, *On the exponential stability of singularly perturbed systems*, SIAM J. Control Optim., 30 (1992), pp. 1338–1360.

[5] A. L. Dontchev, T. Donchev, and I. Slavov, *A Tychonov-type theorem for singularly perturbed differential inclusions*, Nonlinear Anal., 26 (1996), pp. 1547–1554.

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

[7] G. Grammel, *Exponential stability via the averaged system*, J. Dynam. Control Systems, 7 (2001), pp. 327–338.

[8] G. Grammel and I. Maizurna, *Exponential stability and partial averaging*, J. Math. Anal. Appl., 283 (2003), pp. 276–286.

[9] F. C. Hoppensteadt, *Analysis and Simulation of Chaotic Systems*, Springer, New York, 1993.

[10] P. V. Kokotovic, H. K. Khalil, and J. O'Reilly, *Singular Perturbation Methods in Control: Analysis and Design*, Academic Press, London, 1986.

[11] X. Mao, *Exponential stability of nonlinear differential delay equations*, Systems Control Lett., 28 (1996), pp. 159–165.

[12] A. A. Martynyuk, *Stability by Lyapunov's Matrix Function Method with Applications*, Pure Appl. Math. 214, Marcel Dekker, New York, 1998.

[13] A. A. Martynyuk, *Uniform asymptotic stability of a singularly perturbed system via the Lyapunov matrix function*, Nonlinear Anal., 11 (1987), pp. 1–4.

[14] A. Saberi and H. Khalil, *Quadratic type Lyapunov functions for singularly perturbed systems*, IEEE Trans. Automat. Control, 29 (1984), pp. 542–550.

[15] A. N. Tychonov, *Systems of differential equations containing small parameters in the derivatives*, Mat. Sb., 31 (1952), pp. 575–586.

[16] A. B. Vasilieva, *On the development of singular perturbation theory at Moscow State University and elsewhere*, SIAM Rev., 36 (1994), pp. 440–452.

[17] V. Veliov, *A Generalization of the Tikhonov theorem for singularly perturbed differential inclusions*, J. Dynam. Control Systems, 3 (1997), pp. 291–319.

[18] V. Veliov, *Convergence of the solution set of singularly perturbed differential inclusions*, Nonlinear Anal., 30 (1997), pp. 5505–5514.

[19] F. Watbled, *On singular perturbations for differential inclusions on the infinite interval*, J. Math. Anal. Appl., 310 (2005), pp. 362–378.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 5, pp. 1725–1747

A BEHAVIORAL APPROACH TO TIME-VARYING LINEAR SYSTEMS. PART 1: GENERAL THEORY*

ACHIM ILCHMANN† AND VOLKER MEHRMANN‡

Abstract. We develop a behavioral approach to linear, time-varying, differential-algebraic systems. The analysis is "almost everywhere" in the sense that the statements hold on $\mathbb{R} \setminus \mathbb{T}$, where $\mathbb{T}$ is a discrete set. Controllability, observability, and autonomy are introduced and related to the behavior of the system. Classical results on the behavior of time-invariant systems are studied in the context of time-varying systems.

Key words. time-varying linear systems, behavioral approach, controllability, observability, autonomous system, adjoint system, latent variables

AMS subject classifications. 93B11, 93B40, 93B36

DOI. 10.1137/S0363012904442239

Notation.

I_d	$:= \operatorname{diag}[1, \dots, 1] \in \mathbb{R}^{d\times d}$
0_d	$:= (0, \dots, 0)^T \in \mathbb{R}^d$
$\mathcal{A}$	the ring of real analytic functions $f : \mathbb{R} \to \mathbb{R}$
$\mathcal{M}$	the field of real meromorphic functions
$\mathcal{A}[D]$, $\mathcal{M}[D]$	the skew polynomial ring of differential polynomials with coefficients in $\mathcal{A}$, $\mathcal{M}$, respectively, indeterminate D, and multiplication rule $Df = fD + \dot{f}$
$\mathcal{C}^N(M, \mathbb{R}^q)$	the real vector space of N-times differentiable functions $f : M \to \mathbb{R}^q$, $M \subset \mathbb{R}$ an open set, $N \in \mathbb{N} \cup \{\infty\}$
$\mathcal{C}^\omega(\mathbb{I}, \mathbb{R}^q)$	the real vector space of real analytic functions $f : \mathbb{I} \to \mathbb{R}^q$, $\mathbb{I} \subset \mathbb{R}$ an open interval
$\mathcal{C}^\infty_{\mathrm{pw}}(\mathbb{R}^q)$	$:= \{w \in \mathcal{C}^\infty(\mathbb{R} \setminus \mathbb{T}, \mathbb{R}^q) \mid \mathbb{T} \subset \mathbb{R} \text{ discrete}\}$
$\mathcal{C}^\infty_t(\mathbb{R}^q)$	$:= \{w \in \mathcal{C}^\infty(\mathbb{I}, \mathbb{R}^q) \mid \mathbb{I} \subset \mathbb{R} \text{ an open interval with } t \in \mathbb{I}\}$, $t \in \mathbb{R}$
$\operatorname{im}_t M$	$:= \{w \in \mathcal{C}^\infty_t(\mathbb{R}^q) \mid \exists\, l \in \mathcal{C}^\infty_t(\mathbb{R}^m)$ for all $\tau \in \operatorname{dom} w \cap \operatorname{dom} l : w(\tau) = M(\frac{\mathrm{d}}{\mathrm{d}\tau})l(\tau)\}$, $t \in \mathbb{R}$, $M(D) \in \mathcal{M}[D]^{q\times m}$
$\operatorname{im} M$	$:= \{w \in \mathcal{C}^\infty_{\mathrm{pw}}(\mathbb{R}^q) \mid \exists\, l \in \mathcal{C}^\infty_{\mathrm{pw}}(\mathbb{R}^m)$ for a.a. $\tau \in \operatorname{dom} w \cap \operatorname{dom} l : w(\tau) = M(\frac{\mathrm{d}}{\mathrm{d}\tau})l(\tau)\}$, $M(D) \in \mathcal{M}[D]^{q\times m}$
$\ker_t R$	$:= \{w \in \mathcal{C}^\infty_t(\mathbb{R}^q) \mid R(\frac{\mathrm{d}}{\mathrm{d}\tau})w(\tau) = 0$ for all $\tau \in \operatorname{dom} w\}$, $t \in \mathbb{R}$, $R(D) \in \mathcal{M}[D]^{g\times q}$
$\ker R$	$:= \{w \in \mathcal{C}^\infty_{\mathrm{pw}}(\mathbb{R}^q) \mid R(\frac{\mathrm{d}}{\mathrm{d}\tau})w(\tau) = 0$ for almost all $\tau \in \mathbb{R}\}$, $R(D) \in \mathcal{M}[D]^{g\times q}$
$\operatorname{dom} w$	the domain of a function w

*Received by the editors March 22, 2004; accepted for publication (in revised form) March 8, 2005; published electronically November 23, 2005.
http://www.siam.org/journals/sicon/44-5/44223.html
†Institut für Mathematik, Technische Universität Ilmenau, Weimarer Straße 25, 98693 Ilmenau, Germany (achim.ilchmann@tu-ilmenau.de).
‡Institut für Mathematik, MA 4-5, TU Berlin, Straße des 17. Juni 136, D-10623 Berlin, Germany (mehrmann@math.tu-berlin.de). This author's research was supported by DFG Research Center MATHEON *Mathematics for key technologies* in Berlin.

1. Introduction.

1.1. An algebraic approach and solution spaces. The aim of the present paper is to develop a behavioral approach to linear time-varying systems described by differential-algebraic equations of the form

$$R(\tfrac{\mathrm{d}}{\mathrm{d}t})w = 0, \tag{1.1}$$

where $R(D)$ is a $g\times q$ polynomial matrix in the indeterminate D with real meromorphic coefficient matrices belonging to $\mathcal{M}^{g\times q}$; we use the notation $R(D)\in\mathcal{M}^{g\times q}[D]$.

Instead of considering real meromorphic coefficients of $R(D)$ on the whole time axis $\mathbb{R}$, we also could develop the theory on some open interval $\mathbb{I}\subset\mathbb{R}$; this is omitted.

The ring $\mathcal{M}[D]$ is endowed with the multiplication rule

$$Df = fD + \dot{f}. \tag{1.2}$$

This is a consequence of assuming the associative rule $(Df)g = D(fg)$ for all differentiable functions f, g which yields $(Df)(g) = \frac{\mathrm{d}}{\mathrm{d}t}f\cdot g + f\cdot\frac{\mathrm{d}}{\mathrm{d}t}g = (\frac{\mathrm{d}}{\mathrm{d}t}f + fD)(g)$. The noncommutativity of $\mathcal{M}[D]$, in contrast to the commutative ring $\mathbb{R}[D]$ in the time-invariant case, is crucial in the following.

Note that we distinguish between the algebraic indeterminate D and the differential operator $\frac{\mathrm{d}}{\mathrm{d}t}$; for

$$R(D) = \sum_{i=0}^{n} R_i D^i \in \mathcal{M}[D]^{g\times q} \cong \mathcal{M}^{g\times q}[D],$$

equality in (1.1) means

$$\sum_{i=0}^{n} R_i(t)w^{(i)}(t) = 0 \qquad \text{for almost all } t\in\mathbb{R}.$$

Skew polynomial rings are, for example, treated in the monograph [6]; the ring $\mathcal{M}[D]$ was introduced in [14] to study linear time-varying systems. We are interested in the behavior introduced by all solutions of (1.1). Since the coefficients of $R(D)$ are meromorphic functions, we can only expect solutions which are defined "almost globally" (see subsection 1.3). To be more precise, we allow for the solution space

$$\mathcal{C}^\infty_{\mathrm{pw}}(\mathbb{R}^q) = \big\{w\in\mathcal{C}^\infty(\mathbb{R}\setminus\mathbb{T},\mathbb{R}^q)\;\big|\;\mathbb{T}\subset\mathbb{R}\text{ discrete}\big\}$$

of piecewise $\mathcal{C}^\infty$-functions (see the notation) defined almost everywhere on $\mathbb{R}$, and the set

$$\mathcal{C}^\infty_t(\mathbb{R}^q) = \big\{w\in\mathcal{C}^\infty(\mathbb{I},\mathbb{R}^q)\;\big|\;\mathbb{I}\subset\mathbb{R}\text{ an open interval with } t\in\mathbb{I}\big\},\qquad t\in\mathbb{R},$$

of $\mathcal{C}^\infty$-solution pieces on some open interval including t.

For $R(D)\in\mathcal{M}[D]^{g\times q}$, we study the *almost global behavior* given by the kernel representation

$$\ker R = \{w\in\mathcal{C}^\infty_{\mathrm{pw}}(\mathbb{R}^q)\,|\,R(\tfrac{\mathrm{d}}{\mathrm{d}\tau})w(\tau) = 0 \text{ for almost all } \tau\in\mathbb{R}\}$$

and the *local behavior*

$$\ker_t R = \{w\in\mathcal{C}^\infty_t(\mathbb{R}^q)\,|\,R(\tfrac{\mathrm{d}}{\mathrm{d}\tau})w(\tau) = 0 \text{ for all } \tau\in\operatorname{dom} w\},\qquad t\in\mathbb{R}.$$

1.2. Examples of system classes. Our approach generalizes results on the following subclasses of systems:

(a) Time-varying state space systems of the form

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}t}x(t) &= A(t)x(t) + B(t)u(t), \\ y(t) &= C(t)x(t) + F(t)u(t), \end{aligned} \tag{1.3}$$

with real analytic matrices $A \in \mathcal{A}^{n\times n}$, $B \in \mathcal{A}^{n\times m}$, $C \in \mathcal{A}^{p\times n}$, and $F \in \mathcal{A}^{p\times m}$, are well studied; see, for example, the standard monograph [30].

(b) Time-varying descriptor systems of the form

$$\begin{aligned} E(t)\,\tfrac{\mathrm{d}}{\mathrm{d}t}x(t) &= A(t)x(t) + B(t)u(t), \\ y(t) &= C(t)x(t) + F(t)u(t), \end{aligned} \tag{1.4}$$

with $A \in \mathcal{A}^{\ell\times n}$, $B \in \mathcal{A}^{\ell\times m}$, $C \in \mathcal{A}^{p\times n}$, $F \in \mathcal{A}^{p\times m}$, where $E \in \mathcal{A}^{\ell\times n}$ is allowed to be singular in the sense that $\operatorname{rk} E(t) < \min\{\ell, n\}$ for some $t \in \mathbb{R}$, have been studied by different authors. In [5] controllability and observability were studied in terms of derivative arrays. In [2] a first behavior-like approach to systems (1.4) with analytic coefficients was discussed. A more general approach that allows for larger classes of coefficients and that can be implemented also numerically was introduced in [20] and generalized partially to the nonlinear case in [18]. A completely different approach results from the study of differential-algebraic equations introduced in [1, 9, 19]. A general solvability theory for nonsquare linear time-varying systems was first given in [16] and analyzed for control problems in a behavioral context in [2, 20, 26]; see also [18] for the general nonlinear case.

(c) In [14] time-varying polynomial systems of the form

$$\begin{aligned} P(\tfrac{\mathrm{d}}{\mathrm{d}t})z(t) &= Q(\tfrac{\mathrm{d}}{\mathrm{d}t})u(t), \\ y(t) &= V(\tfrac{\mathrm{d}}{\mathrm{d}t})z(t) + W(\tfrac{\mathrm{d}}{\mathrm{d}t})u(t), \end{aligned} \tag{1.5}$$

where $P(D)$, $Q(D)$, $V(D)$, and $W(D)$ are matrices of size $r \times r$, $r \times m$, $p \times r$, and $p \times m$, respectively, over $\mathcal{M}[D]$ are studied under the following assumptions:

- $P(D)$ represents a so-called *full* operator, i.e., if z is a real analytic solution of $P(\frac{\mathrm{d}}{\mathrm{d}t})z = 0$ on some interval $\mathbb{I} \subset \mathbb{R}$, then this solution can be analytically extended to the whole of $\mathbb{R}$.
- For every $u \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^m)$ with bounded support to the left, there exist some $z \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^r)$ and $y \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^p)$ so that (1.5) is satisfied.

Time-invariant polynomial (so-called Rosenbrock) systems of the form (1.5)—i.e., $P(D)$, $Q(D)$, $V(D)$, and $W(D)$ are matrices over $\mathbb{R}[D]$ and $\det P(\cdot) \neq 0$—were introduced in [27] and are well studied; see, for example, [11, 39].

(d) Time-invariant polynomial systems in the so-called kernel representation

$$R(\tfrac{\mathrm{d}}{\mathrm{d}t})w(t) = 0, \qquad R(D) \in \mathbb{R}[D]^{g\times q} \tag{1.6}$$

were introduced by Willems in [35]; see also [36, 37, 38] and the monograph [24].

It is easy to see that time-varying descriptor systems (1.4) or, if $E = I_n$ and $n = \ell$, state space systems (1.3) are special cases of time-varying Rosenbrock systems

(1.5). Furthermore, time-varying Rosenbrock systems of the form (1.5) are a special case of systems in kernel representation (1.1): set $w = [z^T, u^T, y^T]^T$ and

$$R(D) = [R_1(D), R_2(D)], \quad R_1(D) = \begin{bmatrix} P(D) \\ V(D) \end{bmatrix}, \quad R_2(D) = \begin{bmatrix} -Q(D), & 0 \\ W(D), & -I_p \end{bmatrix}. \tag{1.7}$$

1.3. Examples of time-varying scalar differential equations. In the following, we present some prototypical scalar differential equations which illustrate how time-varying coefficients may affect the solutions in very different ways. Set, for $r(D) \in \mathcal{M}[D]$ and $\mathcal{W}$ a suitable solution space to be specified,

$$\ker_{\mathcal{W}} r(\tfrac{\mathrm{d}}{\mathrm{d}t}) := \{w \in \mathcal{W} \mid r(\tfrac{\mathrm{d}}{\mathrm{d}t})w = 0\}.$$

(i) Let $r(D) = tD + 1$. Then the function $t \mapsto w(t) = t^{-1}$ is a meromorphic solution of $r(\frac{\mathrm{d}}{\mathrm{d}t})w = t\frac{\mathrm{d}}{\mathrm{d}t}w + w = 0$.
The point 0 is the only zero of the leading coefficient $t \mapsto t$ of $r(D)$, and 0 is also a pole of $t \mapsto w(t)$. Therefore,

$$\ker_{\mathcal{A}} r(\tfrac{\mathrm{d}}{\mathrm{d}t}) = \ker_{\mathcal{C}^\infty(\mathbb{R},\mathbb{R})} r(\tfrac{\mathrm{d}}{\mathrm{d}t}) = \{0\},$$

but, for every interval $\mathbb{I} \subset \mathbb{R}$ with $0 \notin \mathbb{I}$,

$$\dim \ker_{\mathcal{M}} r(\tfrac{\mathrm{d}}{\mathrm{d}t}) = \dim \ker_{\mathcal{A}_{|_{\mathbb{I}}}} r(\tfrac{\mathrm{d}}{\mathrm{d}t}) = 1 = \deg r(D).$$

In this example, in the meromorphic case the dimension of the solution space equals the degree of $r(D)$. This is not true in general, as illustrated by the following example.

(ii) Let $r(D) = t^2D + 1$. Then the function $t \mapsto w(t) = e^{1/t}$ solves $r(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$. The point 0 is again the only zero of the leading coefficient $t \mapsto t^2$ of $r(D)$, and 0 is also a pole of $t \mapsto w(t)$. But w is not meromorphic and the singularity at $t = 0$ differs from (i) as follows: no matter whether the solution w in (i) approaches 0 from the left or right, the limit at $t = 0$ does not exist; whereas, for the solution w in the present example, we have $\lim_{t\to 0-} w(t) = 0$ and $\lim_{t\to 0+} w(t) = \infty$. Hence,

$$\ker_{\mathcal{M}} r(\tfrac{\mathrm{d}}{\mathrm{d}t}) = \{0\}.$$

For every interval $\mathbb{I} \subset \mathbb{R}$ with $0 \notin \mathbb{I}$ we have

$$\dim \ker_{\mathcal{M}_{|_{\mathbb{I}}}} r(\tfrac{\mathrm{d}}{\mathrm{d}t}) = 1 = \deg r(D).$$

(iii) Let $r(D) = tD - 1$. Then the function $t \mapsto w(t) = t$ solves $r(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ and

$$\dim \ker_{\mathcal{A}} r(\tfrac{\mathrm{d}}{\mathrm{d}t}) = 1 = \deg r(D).$$

Note that again the point $t = 0$ is the only zero of the leading coefficient $t \mapsto t$ of $r(D)$, but this time the zero does not produce a pole of the solution, the solution w is even a real analytic function on $\mathbb{R}$. However, the solution is not as arbitrary as for time-invariant systems, since $w(0) = 0$ is the only value at $t = 0$.

(iv) Let $r(D) = 2tD - 1$. Then the functions $t \mapsto w_+(t) = \sqrt{t}$ and $t \mapsto w_-(t) = \sqrt{-t}$ solve $r(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ on $(0, \infty)$ and $(-\infty, 0)$, respectively. For every interval $\mathbb{I} \subset \mathbb{R}$ with $0 \notin \mathbb{I}$, we have

$$\dim \ker_{\mathcal{A}_{|\mathbb{I}}} r(\tfrac{\mathrm{d}}{\mathrm{d}t}) = 1 = \deg r(D).$$

However,

$$\ker_{\mathcal{M}} r(\tfrac{\mathrm{d}}{\mathrm{d}t}) = \{0\}.$$

The real analytic solution w_+ on $(0, \infty)$ cannot be continued to $(-\varepsilon, \infty)$ for any $\varepsilon > 0$.

This also proves that the attempt to connect real analytic solutions between critical points by cutting the neighborhood and going into the complex sphere, as suggested by Ilchmann et al. [13], does not work.[1]

(v) Let $r(D) = (1 - t^2)^2 D + 2t$. Then the function

$$t \mapsto w(t) = \begin{cases} \exp\left\{-(1-t^2)^{-1}\right\}, & t \in (-1, 1), \\ 0, & t \in \mathbb{R} \setminus (-1, 1), \end{cases}$$

satisfies $w \in \ker_{\mathcal{C}^\infty} r(\frac{\mathrm{d}}{\mathrm{d}t})$, is not real analytic, and has compact support. This is impossible for time-invariant scalar differential equations.

1.4. An example of a mobile manipulator. Systems of differential-algebraic equations play an important role in modeling multibody systems, electric circuits, or coupled systems of partial differential equations; see [1, 10]. We present an application which first shows that modeling does not necessarily lead to a state space system; second, it illustrates a simple system where the notion of input, output, and state is not a priori clear; and third, the example serves to illustrate the concepts introduced in the following sections. Consider a simplified, linearized model of a two-dimensional, three-link constrained mobile manipulator [12] as depicted in Figure 1.

The Lagrangian equations of motion take the form

$$\begin{aligned} M(\theta)\,\ddot{\theta} + D(\theta, \dot{\theta})\,\dot{\theta} + K(\theta) &= u + F^T(\theta)\mu, \\ \psi(\theta) &= 0, \end{aligned} \tag{1.8}$$

where $\theta = [\theta_1, \theta_2, \theta_3]^T$ is the vector of joint displacements, $u \in \mathbb{R}^3$ is the vector of control torques applied at the joints, and the maps $M : \mathbb{R}^3 \to \mathbb{R}^{3\times 3}$, $D : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^{3\times 3}$, and $K : \mathbb{R}^3 \to \mathbb{R}^3$ model the mass, centrifugal and Coriolis forces, gravity, respectively. $l_1, l_2, l_3, l > 0$ are the lengths of the robot arms. The nonlinear constraint function is $\psi : \mathbb{R}^3 \to \mathbb{R}^2$, $F = \frac{\partial \psi}{\partial \theta}$, and $\mu \in \mathbb{R}^2$ represents the Lagrange multipliers and $F^T(\theta)\mu$ is the generalized constraint force. We are interested in the behavior, i.e., local solutions $t \mapsto [\theta(t)^T, u(t)^T]$ of (1.8). It can be shown that $u(\cdot)$ is a latent variable; for its definition, see [24, sect. 6.2]. Under suitable smoothness assumptions of the involved functions, it can be shown (see, for example, [25, p. 62]) that there exists a local (possibly global) solution $\theta(\cdot)$ of (1.8) on some open interval $\mathbb{I}$. Linearizing along this trajectory [4] and rewriting the system in Cartesian coordinates yields a model of the form

$$\begin{aligned} M_0(t)\,\ddot{z}(t) + D_0(t)\,\dot{z}(t) + K_0(t)\,z(t) &= S_0\,u(t) + F_0^T(t)\,\mu, \\ F_0(t)\,z(t) &= 0, \end{aligned}$$

[1] We are indebted to the anonymous referee of an earlier version of the present paper for pointing out this example to us.

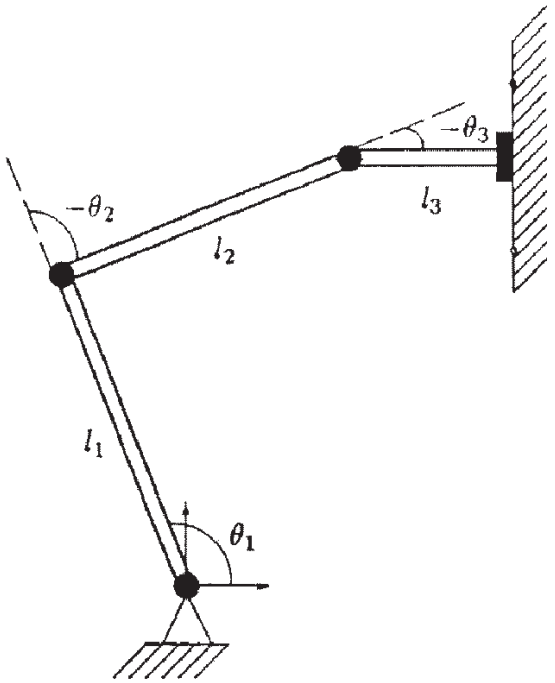


FIG. 1. *Three-link constrained mobile manipulator.*

where $M_0, D_0, K_0 \in \mathcal{C}^\omega(\mathbb{I}, \mathbb{R}^{3\times 3})$ and $S_0 \in \mathbb{R}^{3\times 3}$, $F_0^T \in \mathbb{R}^{3\times 2}$ with S_0 having full row rank. Introducing the eight-dimensional variable $x(t) = [z(t)^T, \dot{z}(t)^T, \mu(t)^T]^T$ results in the equivalent descriptor system description of the form

$$\begin{aligned} E(t)\,\tfrac{\mathrm{d}}{\mathrm{d}t}x(t) &= A(t)x(t) + B(t)u(t), \\ y(t) &= C(t)x(t), \end{aligned} \tag{1.9}$$

where

$$E(t) := \begin{bmatrix} I_3 & 0 & 0 \\ 0 & M_0(t) & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A(t) := \begin{bmatrix} 0 & I_3 & 0 \\ -K_0(t) & -D_0(t) & F_0^T(t) \\ F_0(t) & 0 & 0 \end{bmatrix}, \quad B(t) := \begin{bmatrix} 0 \\ S_0 \\ 0 \end{bmatrix},$$

and $C(\cdot)$ denotes a matrix with appropriate format; see [12] for explicit data. Actually, in this example F_0 does not depend on t.

1.5. Literature survey. The crucial difference between time-varying and time-invariant ordinary, linear differential equations is that the solutions behave qualitatively considerably different. Whereas any local solution of a time-invariant system is always extendable to a global analytic solution, solutions of time-varying systems may have finite escape times. Simple examples have been presented in subsection 1.3. All algebraic contributions to time-varying systems struggle with this difficulty.

Early algebraic contributions to time-varying systems in polynomial descriptions are given in [15, 40, 41]; however, the assumptions on the system classes are rather restrictive.

In [7], matrices over the ring of linear differential operators $k[D]$ are considered, where k is a differential field. Linear dynamics are finitely generated left $k[D]$-modules. This contribution is rather on the algebraic side; the solution space is not specified. In [29] contributions to duality of systems in the setup of [7] for systems in generalized state space representation are given; however, the solution space is not specified either.

An important contribution by Fröhler and Oberst [8] has the following background: Consider the simple examples given in subsection 1.3. It can be shown that the local solution $(t \mapsto 1/t) \in \ker(t\frac{\mathrm{d}}{\mathrm{d}t} + 1)$ can be extended to a distribution belonging to $\mathcal{D}'(\mathbb{R}, \mathbb{R})$; however, $(t \mapsto \exp(1/2t^2)) \in \ker(t^3\frac{\mathrm{d}}{\mathrm{d}t} + 1)$ cannot be extended to a distribution belonging to $\mathcal{D}'(\mathbb{R}, \mathbb{R})$. Hence enlarging the solution space to allow for distributions on $\mathbb{R}$ does not necessarily resolve the problem, even in the simple case when the coefficients of the time-varying systems are polynomials. However, if the solution space is enlarged even further to allow for Sato's hyperfunctions, i.e., generalized distributions introduced in [31, 32], then [8] considers systems of the form (1.1), respectively, behavior in the kernel representation $\ker R$, where the coefficient matrices of the polynomial $R(D)$ are defined over rational analytic functions

$$\frac{f(\cdot)}{g(\cdot)} \quad \text{for } f, g \in \mathbb{C}[t] \quad \text{with} \quad g(t) \neq 0 \text{ for all } t \in \mathbb{I}.$$

Note that by multiplication with a least common multiple of all denominators of the coefficients, the coefficients of $R(D)$ are polynomials. Based on the seminal paper [22], where an algebraic analytic approach is developed to show a categorical duality between the solution spaces of linear partial differential equations with constant coefficients and certain polynomial modules associated to them, a generalization to time-varying but ordinary differential equations is achieved in [8].

The skew polynomial ring $\mathcal{M}[D]$ was first exploited by [14] to describe time-varying linear systems. This ring is nice in the sense that it is simple (i.e., the only two-sided ideals are the trivial ones) and admits right- and left-Euclidean division. Hence matrices over the ring can be transformed into the Teichmüller–Nakayama normal form; see section 2. The latter is the essential tool in [14] to study time-varying Rosenbrock systems of the form (1.5). The solution space is the set of $\mathcal{C}^\infty$-functions on the whole time axis; this is ensured by the assumption that $\operatorname{im} Q(\frac{\mathrm{d}}{\mathrm{d}t}) \subset \operatorname{im} P(\frac{\mathrm{d}}{\mathrm{d}t})$ and, most importantly, that $P(D)$ is a "full" operator, i.e., every local analytic solution of $P(\frac{\mathrm{d}}{\mathrm{d}t})z = 0$ is extendable to a global analytic solution on the whole of $\mathbb{R}$. The latter is a rather restrictive assumption. To overcome this assumption, in [13] a first approach in the spirit of the present paper was presented for scalar systems. A behavioral approach to a certain class of time-varying systems was presented in [3].

A completely different approach results from the study of differential-algebraic equations introduced in [1, 9]. A general solvability theory for nonsquare linear time-varying systems was first given in [16] and analyzed for control problems in a behavioral context in [2, 20, 26]; see also [18] for the general nonlinear case, and a latest monograph [19].

This paper is organized as follows. In section 2, the algebraic tools, such as the Teichmüller–Nakayama normal form, and some facts on the behavior are collected. In section 3, we introduce and characterize algebraically the concept of controllable behavior for the kernel and image representation. The relationship between behavior, controllable, and autonomous behavior is investigated in section 4. In section 5, observability is defined, it is related via the adjoint of the kernel representation to the controllable behavior, and it is characterized algebraically. Finally, in section 6 we

investigate the elimination of latent variables.

2. Behavior. In this section we present the Teichmüller–Nakayama normal form for matrices over $\mathcal{M}[D]$. This will be the main tool for analyzing $\ker_t R$. To this end we recall some results on matrices over the skew polynomial ring $\mathcal{M}[D]$; a standard reference for this is [6]. $\mathcal{M}[D]$ is *simple*, i.e., the only ideals which are right and left ideals at the same time are the trivial ones; the rank of a matrix over $\mathcal{M}[D]$ is unambiguous, since column rank and row rank coincide; the Teichmüller–Nakayama normal form is the analogue of the Smith normal form for matrices over the commutative ring $\mathbb{R}[D]$; it is simpler for matrices over $\mathcal{M}[D]$, since the class of transformations is larger. $W(D) \in \mathcal{M}[D]^{n\times n}$ is called *unimodular* if and only if there exists some $W(D)^{-1} \in \mathcal{M}[D]^{n\times n}$ such that $W(D)\,W(D)^{-1} = I_n$; two elements $q_1, q_2 \in \mathcal{M}[D]$ are *similar* if and only if $q_1 a = b q_2$ for some $a, b \in \mathcal{M}[D]$ for which q_1 and b (q_2 and a) are left (right) coprime. For example, $a(D) = D$ and $b(D) = D - 1/t$ are similar: $[D + (t^2-1)/t]a(D) = b(D)\,[D+t]$ and $D + (t^2-1)/t$, $b(D)$ are right coprime, $a(D)$, $D + t$ are left coprime. Moreover, this example shows that a unique factorization of the ring elements cannot be expected. However, Ore [23] has shown that the degree of similar polynomials coincide. The latter property is crucial for determining dimensions of solution spaces.

A proof and an interesting historical description of the following normal form can be found in [6, Chap. 8]. The proof is constructive, using elementary matrices and Euclidean division. So if the coefficients consist of real polynomials $\mathbb{R}[t]$, then it is possible to calculate a normal form by means of computer algebra.

THEOREM 2.1 (Teichmüller–Nakayama normal form). *Any $R(D) \in \mathcal{M}[D]^{g\times q}$ with $\mathrm{rk}_{\mathcal{M}[D]} R(D) = \ell$ can be factorized into*

$$R(D) = U(D)^{-1} \begin{bmatrix} I_{\ell-1} & 0 & 0 \\ 0 & r(D) & 0 \\ 0 & 0 & 0_{(g-\ell)\times(q-\ell)} \end{bmatrix} V(D)^{-1}, \tag{2.1}$$

where $U(D)$ and $V(D)$ are $\mathcal{M}[D]$-unimodular matrices of sizes g and q, respectively, and $r(D) \in \mathcal{M}[D]$ is nonzero, unique up to similarity, and of unique degree.

Remark 1. Let $R(D) \in \mathcal{M}[D]^{g\times q}$ and consider the factorization (2.1).

(i) Then we have, for almost all $t \in \mathbb{R}$,

$$\text{for all } w \in \mathcal{C}_t^\infty(\mathbb{R}^q): \left[w \in \ker_t R \iff w \in \ker_t \left(\left[\begin{array}{ccc|c} I_{\ell-1} & & & \\ & r & & 0_{\ell\times q} \end{array} \right] V^{-1} \right) \right].$$

Hence we may assume, without restriction of generality, that $R(D)$ has full row rank.

(ii) The set $\ker_t R$ becomes a *real vector space* if endowed, for $w_1, w_2 \in \ker_t R$, with addition

$$(w_1 + w_2)(\tau) := w_1(\tau) + w_2(\tau) \qquad \text{for all } \tau \in \operatorname{dom} w_1 \cap \operatorname{dom} w_2$$

and obvious scalar multiplication. The dimension of this vector space is defined as

$$\dim \ker_t R := \sup\left\{ k \in \mathbb{N} \;\middle|\; \exists\, w_1, \dots, w_k \in \ker_t R \text{ linearly independent on } \bigcap_{i=1}^{k} \operatorname{dom} w_i \right\}.$$

Furthermore,

$$\dim \ker_t R = \begin{cases} \deg r(D) & \text{for almost all } t \in \mathbb{R} \quad \text{if } \operatorname{rk} R(D) = q, \\ \infty & \text{for all } t \in \mathbb{R} \quad \text{if } \operatorname{rk} R(D) < q. \end{cases}$$

The latter is a simple consequence of (2.1) and the fact that the set of t where $r(\frac{\mathrm{d}}{\mathrm{d}t})\varphi(t) = 0$ does not have a solution is a subset of $\{t \in \mathbb{R} \mid r_N(t) = 0\}$, where $r(D) = \sum_{i=0}^{N} r_i(t) D^i$, $r_N \not\equiv 0$. To see this, use the canonical transformation to a vector-valued differential equation of first order; see, for example, [34, Chap. IV].

(iii) Let $\mathbb{T} = \mathbb{T}(R, U, V, r)$ denote the union of all zeros and poles of the meromorphic coefficients in all nonzero entries of $U(D)$, $U(D)^{-1}$, $V(D)$, $V(D)^{-1}$, and $r(D)$. Certainly, $\mathbb{T}$ is a discrete set which depends on the factorization and hence is not unique. $\mathbb{T}$ encompasses all possible critical points where a finite escape may occur (see the examples in subsection 1.3); however, $\mathbb{T}$ might be much larger. We gain system theoretic information from the normal form but may also hide information: consider, for example, a state space system of the form (1.3). Then this system does not have any critical points; however, taking it into a normal form may introduce a possibly nonempty set $\mathbb{T}$. It is an open problem to determine an algorithm for the transformation into the Teichmüller–Nakayama normal form which produces a "minimal" set $\mathbb{T}$.
However, there are situations where it is possible to determine a set including all critical points without invoking algebraic transformations, as in the Teichmüller–Nakayama normal form: For general linear and nonlinear descriptor systems, it has been shown in [16, 17, 18, 20] that for sufficiently often differentiable coefficient functions there exist invariants (corresponding to ranks of submatrices) which are independent of the choice of transformation matrices, and the set of points where these quantities jump includes all critical points.

(iv) If $R(D)$ is not left invertible, then the set of points where the local behavior is nontrivial, i.e., $\{t \in \mathbb{R} \mid \ker_t R \neq \{0\}\}$, is discrete.

Remark 2. Suppose that $R(D)$ has constant coefficients, i.e., $R(D) \in \mathbb{R}[D]^{g \times q}$.

(i) If the class of unimodular transformations for the computation of the normal form (2.1) is restricted to $\mathbb{R}[D]$-unimodular matrices, then we arrive at the Smith normal form

$$R(D) = U(D)^{-1} \begin{bmatrix} \operatorname{diag}\{r_1(D), \dots, r_l(D)\} & 0_{\ell \times (q-\ell)} \\ 0_{(g-\ell)\times \ell} & 0_{(g-\ell)\times(q-\ell)} \end{bmatrix} V(D)^{-1}, \tag{2.2}$$

where $U(D)$ and $V(D)$ are $\mathbb{R}[D]$-unimodular matrices of sizes g and q, respectively, and $r_i(D) \in \mathbb{R}[D]$ are nonzero monic polynomials with $r_i | r_{i+1}$, $i = 1, \dots, \ell - 1$, where $\ell = \operatorname{rk}_{\mathbb{R}[D]} R(D)$ and $r_i(D) = \psi_i(D)/\psi_{i-1}(D)$, $\psi_0(\cdot) \equiv 1$, and $\psi_i(D)$ is the greatest common divisor of minors of order i of $R(D)$; see, for example, [28, pp. 91–93].
Note that due to the smaller class of transformations, the Smith normal form is less simple than the Teichmüller–Nakayama normal form.

(ii) Suppose in addition that $\mathrm{rk}_{\mathbb{R}[D]} R(D) = q$. Then every local solution $w \in \mathcal{C}_t^N(\mathbb{R}^q)$ of $R(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$, where N is sufficiently large depending on $\deg R(D)$ and the degrees of the transformation matrices, can be continued to a global solution on $\mathbb{R}$ and is even real analytic. This follows immediately from the Smith normal form (2.2) and the theory of linear time-invariant differential equations. Therefore, we may identify $\ker_t R = \ker R$ for any $t \in \mathbb{R}$, and it follows that $\dim \ker_t R = \sum_{i=1}^{\ell} \deg r_i(D)$ for all $t \in \mathbb{R}$.

Remark 3. Suppose that $R(D) \in \mathcal{M}[D]^{g\times q}$ has full rank and $g \le q$. Let $R(D)$ be factorized as in (2.1) and differently into

$$R(D) = \bar{U}(D)^{-1} \left[\begin{array}{cc|c} I_{g-1} & & \\ & \bar{r}(D) & 0_{g\times(q-g)} \end{array}\right] \bar{V}(D)^{-1}. \tag{2.3}$$

Then a simple algebraic manipulation shows that

$$\bar{V}(D)^{-1} V(D) = \begin{bmatrix} W_1(D) & 0 \\ W_3(D) & W_4(D) \end{bmatrix}, \tag{2.4}$$

where $W_1(D) \in \mathcal{M}[D]^{g\times g}$ and $W_4(D) \in \mathcal{M}[D]^{(q-g)\times(q-g)}$ are unimodular, and where $W_3(D) \in \mathcal{M}[D]^{(q-g)\times g}$.

3. Controllability. In this section we introduce, study, and characterize the concept of controllability of systems (1.1). This is a generalization of the behavioral concept introduced by Willems [35]; see also [24].

DEFINITION 3.1. *Let $R(D) \in \mathcal{M}[D]^{g\times q}$ and $t \in \mathbb{R}$. A* local subbehavior $\mathfrak{B}_t$ *of* $\ker_t R$*, i.e., a subset $\mathfrak{B}_t \subset \ker_t R$, is called* locally controllable at $t \in \mathbb{R}$ *if and only if for every $w^1, w^2 \in \mathfrak{B}_t$ and every $t_0 \in (-\infty, t) \cap \mathrm{dom}\, w^1$ there exist $t_1 \in \mathrm{dom}\, w^2 \cap (t, \infty)$ and $w \in \mathfrak{B}_t$ such that*

$$w(t) = \begin{cases} w^1(t), & t \in (-\infty, t_0] \cap \mathrm{dom}\, w^1, \\ w^2(t), & t \in [t_1, \infty) \cap \mathrm{dom}\, w^2. \end{cases}$$

A behavior $\mathfrak{B} = \bigcup_{t\in\mathbb{R}} \mathfrak{B}_t$, $\mathfrak{B}_t \subset \ker_t R$, is called controllable almost everywhere *if and only if $\mathfrak{B}_t$ is locally controllable for almost all $t \in \mathbb{R}$. Since $\ker_t R$ is a real vector space by Remark* 1(ii)*, the family of its linear subspaces may be partially ordered by inclusion, and thus constitutes a lattice with respect to $+$ and $\cap$. Hence $\ker_t^{\mathrm{contr}} R \subset \ker_t R$ as* largest controllable local behavior *of $\ker_t R$ is well defined. The set $\ker^{\mathrm{contr}} R = \bigcup_{t\in\mathbb{R}} \ker_t^{\mathrm{contr}} R$ is called the* largest controllable behavior *of $\ker R$.*

This concept is illustrated in Figure 2.

Remark 4.

(i) Loosely speaking, controllability means that any two trajectories $w^1, w^2 \in \ker_t R$ can be connected by another trajectory $w \in \ker_t R$ so that in finite time w^1 moves via w into w^2. A similar notion of controllability via trajectories was introduced in [11] for time-invariant Rosenbrock systems of the form (1.5). For time-invariant state space systems of the form (1.3), the concept of controllability coincides with the one introduced in [24, sect. 5.2].

(ii) Since $\ker_t R$ is a linear subspace, the trajectory w^2 in Definition 3.1 may be replaced, without restriction of generality, by $w^2 = 0$.

We are now in position to prove the main theorem of this section, which characterizes controllability in algebraic terms. Recall that $R(D)$ is called *right invertible* if and only if there exists some $R^{\#}(D) \in \mathcal{M}[D]^{q\times g}$ such that $R(D)R^{\#}(D) = I_g$. In view of Remark 1(i) we assume that $R(D)$ has full row rank.

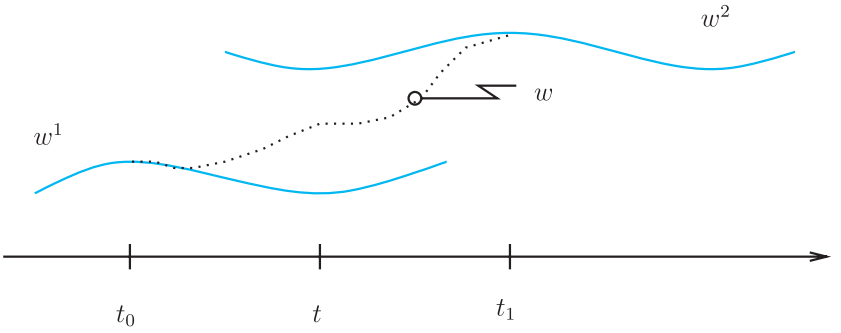


FIG. 2. *Local controllability at t.*

THEOREM 3.2. *Let $R(D) \in \mathcal{M}[D]^{g\times q}$ have full row rank. Then the behavior $\ker R$ is controllable almost everywhere if and only if $R(D)$ is right invertible.*

Proof. Suppose that $R(D)$ is factorized as in (2.1) and let $\mathbb{T} = \mathbb{T}(R, U, V, r)$ denote the discrete set given in Remark 1(iii). Then it remains to show that $\ker_t R$ is locally controllable at $t \in \mathbb{R} \setminus \mathbb{T}$ if and only if $r(D)$ is a nonzero meromorphic function.

"$\Rightarrow$": Suppose that $\deg r(D) \geq 1$ and $t \in \mathbb{R} \setminus \mathbb{T}$. By [34, Chap. IV] there exists an open interval $\mathbb{I} \subset \mathbb{R} \setminus \mathbb{T}$ with $t \in \mathbb{I}$ and some nonzero real analytic solution $\varphi : \mathbb{I} \to \mathbb{R}$ which solves $r(\frac{\mathrm{d}}{\mathrm{d}t})\varphi = 0$. By the construction of $\mathbb{T}$ and letting e_g denote the gth canonical basis vector in $\mathbb{R}^q$, it follows that

$$\hat{w}^1 := V(\tfrac{\mathrm{d}}{\mathrm{d}t})\, \varphi\, e_g \in \mathcal{C}^\omega(\mathbb{I}, \mathbb{R}^q)$$

and solves $R(\frac{\mathrm{d}}{\mathrm{d}t})\hat{w}^1 = 0$.

Seeking a contradiction, suppose that $\ker_t R$ were locally controllable at $t \in \mathbb{R}$. Let $t_0 \in (-\infty, t) \cap \mathbb{I}$. Then there exist $t_1 \in (t, \infty)$ and $w \in \ker_t R$ such that

$$w(t) = \begin{cases} w^1(t), & t \in (-\infty, t_0] \cap \operatorname{dom} w^1, \\ 0, & t \in [t_1, \infty). \end{cases} \tag{3.1}$$

Therefore,

$$\operatorname{diag}\big\{1, \ldots, 1, r(\tfrac{\mathrm{d}}{\mathrm{d}t}), 0, \ldots, 0\big\} V(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1} w = 0 \qquad \text{for all } t \in \operatorname{dom} w,$$

which yields

$$V(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1} w =: [0, \ldots, 0, \varphi_g, \ldots, \varphi_q]^T \in \mathcal{C}^\omega(\operatorname{dom} w, \mathbb{R}^q)$$

and $r(\frac{\mathrm{d}}{\mathrm{d}t})\varphi_g(t) = 0$ for all $t \in \operatorname{dom} w$. By (3.1) we have $\varphi_g(t) = 0$ for all $t \in [t_1, \infty)$, and since φ_g is real analytic, the identity property of real analytic functions gives $\varphi \equiv \varphi_g \equiv 0$, which is a contradiction.

"$\Leftarrow$": Let $t \in \mathbb{R} \setminus \mathbb{T}$, let $r(D)$ be meromorphic and nonzero, and let $w^1 \in \ker_t R$. Then there exists some open interval $\mathbb{I} := (\tau_0, \tau_1) \subset (\mathbb{R} \setminus \mathbb{T}) \cap \operatorname{dom} w^1$ with $t \in \mathbb{I}$ such that

$$w^1 =: V(\tfrac{\mathrm{d}}{\mathrm{d}t})\, [0, \ldots, 0, \varphi_{g+1}, \ldots, \varphi_q]^T \in \mathcal{C}^\infty(\mathbb{I}, \mathbb{R}^q).$$

Choose $\delta \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$ such that

$$\delta(t) = \begin{cases} 1, & t \leq \tau_0, \\ 0, & t \geq \tau_1. \end{cases}$$

Then

$$w := V(\tfrac{\mathrm{d}}{\mathrm{d}t})\,\delta\,[0,\dots,0,\varphi_{g+1},\dots,\varphi_q]^T \in \mathcal{C}^\infty(\mathbb{I},\mathbb{R}^q)$$

satisfies $R(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ and

$$w(t) = \begin{cases} w^1(t), & t \le \tau_0, \\ 0, & t \ge \tau_1. \end{cases}$$

This completes the proof. □

For time-invariant systems (1.1), Theorem 3.2 is derived differently in [24, Thm. 5.2.10].

Remark 5. For time-varying systems (1.3) or (1.5), it is well known that controllability of the system yields that it can be controlled in arbitrary short time. The proof of Theorem 3.2, in particular the choice of (τ_0,τ_1) and δ, shows that this is also valid for the behavior $\ker_t R$: If $\ker_t R$ is controllable, then $t_0 < t$ and $t_1 > t$ in Definition 3.1 can be replaced by any $t_0' < t < t_1'$ arbitrary close to t.

In the following remark we recall the classical concept of controllability for time-varying state space systems and clarify the set of admissible input functions.

Remark 6. Controllability for state space systems (1.3) means (see, for example, [33, Def. 3.1.6]) that for any $x^0, x^1 \in \mathbb{R}^n$ and $t_0 \in \mathbb{R}$, there exist $t_1 > t_0$ and a continuous function $u : [t_0,t_1] \to \mathbb{R}^m$ such that

$$x(t) = (Lu)(t) := \Phi(t,t_0)x^0 + \int_{t_0}^{t} \Phi(t,\tau)B(\tau)u(\tau)d\tau, \qquad t \in [t_0,t_1],$$

satisfies $x(t_1) = x^1$. Here Φ denotes the transition matrix of the homogeneous system $\dot{x} = Ax$.

Using the fact that the set of $\mathcal{C}^\infty$-functions with support in $[t_0,t_1]$ lies dense, with respect to the $\mathcal{L}^1$-norm, in the set of piecewise continuous functions with support included in $[t_0,t_1]$, it follows from a straightforward modification of the proof of Lemma A2 in [14] that, for all $t \in (t_0,t_1)$,

$$\begin{aligned}&\big\{(Lu)(t) \;\big|\; u \in \mathcal{C}^\infty((t_0,t_1),\mathbb{R}^m)\big\} \\ &\quad = \big\{(Lu)(t) \;\big|\; u : [t_0,t_1] \to \mathbb{R}^m \text{ piecewise continuous with } \operatorname{supp} u \subset [t_0,t_1]\big\}.\end{aligned}$$

Therefore, although in the original definition u is required to be continuous, we may choose, without any restriction of generality, $u \in \mathcal{C}^\infty(\mathbb{R},\mathbb{R}^m)$ with $\operatorname{supp} u \subset [t_0,t_1]$.

In the following proposition, it is shown how controllability encompasses other well-established controllability concepts.

PROPOSITION 3.3. *Consider a time-varying Rosenbrock system of the form* (1.5) *with corresponding* $R(D)$ *as defined in* (1.7)*, and suppose that* $R(D)$ *has full row rank. Then the following conditions are equivalent:*

(i) $\ker R$ *is controllable almost everywhere.*
(ii) $[P(D), -Q(D)]$ *is right invertible.*
(iii) $\ker[P,Q]$ *is controllable almost everywhere.*
(iv) (1.5) *is controllable in the sense defined in* [14].
(v) *If* $R(D)$ *represents a time-invariant Rosenbrock system* (1.5)*, then* (1.5) *is controllable in the sense defined in* [11].

(vi) *If $R(D)$ represents a state space system* (1.3) *with corresponding $R(D)$ as defined in* (1.7)*, then* (1.3) *is controllable in the classical sense as, for example, given in* [33, *Def.* 3.1.6].

Proof. The equivalences "(i) $\Leftrightarrow$ (ii) $\Leftrightarrow$ (iii)" follow from Theorem 3.2 and simple algebraic manipulations; "(ii) $\Leftrightarrow$ (iv)" follows from [14, Thm. 6.4]. "(ii) $\Leftrightarrow$ (v)" follows from [11, Cor. 7.3]. It remains to prove that the classical concept of controllability as given in Remark 6 is encompassed in the behavioral setup. It is easy to see that (iii) implies (vi), and we omit the proof. To prove the converse, suppose that (vi) holds. Then for given

$$(x^i, u^i) \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^n) \times \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^m) \quad \text{such that } \tfrac{\mathrm{d}}{\mathrm{d}t}x^i(t) = A(t)x^i(t) + B(t)u^i(t), \quad i = 1, 2,$$

and given $t_0 \in \mathbb{R}$, we need to find

$$(x, u) \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^n) \times \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^m), \quad \text{so that } \tfrac{\mathrm{d}}{\mathrm{d}t}x(t) = A(t)x(t) + B(t)u(t),$$

and $t_1 > t_0$ such that

$$(x(t), u(t)) = \begin{cases} \big(x^1(t), u^1(t)\big) & \text{for all } t \le t_0, \\ \big(x^2(t), u^2(t)\big) & \text{for all } t \ge t_1. \end{cases} \tag{3.2}$$

Let $\bar{x}^1 = x^1(t_0)$ and, for arbitrary but fixed $t_1 > t_0$, let $\bar{x}^2 = x^2(t_1)$. Then by (vi) we may choose $\widehat{u} \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^m)$ with $\operatorname{supp} \widehat{u} \subset [t_0, t_1]$ such that

$$x(t) = \Phi(t, t_0)\bar{x}_1 + \int_{t_0}^{t} \Phi(t, \tau)B(\tau)\widehat{u}(\tau)d\tau \qquad \text{satisfies} \quad x(t_2) = \bar{x}^2.$$

Define, for all $t \in \mathbb{R}$,

$$u(t) = \begin{cases} u^1(t) & \text{for all } t \le t_0, \\ \widehat{u}(t) & \text{for all } t \in (t_0, t_1), \\ u^2(t) & \text{for all } t \ge t_1 \end{cases} \quad \text{and} \quad x(t) = \Phi(t, t_0)\bar{x}_1 + \int_{t_0}^{t} \Phi(t, \tau)B(\tau)u(\tau)d\tau.$$

Then (x, u) satisfies $\dot{x} = Ax + Bu$ and (3.2). The function u is in general not infinitely many times differentiable at t_0 or at t_1, but applying Remark 6, one may replace $\widehat{u}$ so that $u \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^m)$. This completes the proof. ∎

Next we study, for $R(D) \in \mathcal{M}[D]^{g\times q}$, the relationship between the local kernel representation $\ker_t R$ and the *local image representation* at $t \in \mathbb{R}$, i.e., for some $M(D) \in \mathcal{M}[D]^{q\times m}$, the real vector space

$$\operatorname{im}_t M := \{w \in \mathcal{C}_t^\infty(\mathbb{R}^q) \mid \exists\, l \in \mathcal{C}_t^\infty(\mathbb{R}^m) \text{ for all } \tau \in \operatorname{dom} w \cap \operatorname{dom} l : w(\tau) = M(\tfrac{\mathrm{d}}{\mathrm{d}t})\, l(\tau)\}.$$

PROPOSITION 3.4. *Let $R(D) \in \mathcal{M}[D]^{g\times q}$ have full row rank.* $\ker R$ *is controllable almost everywhere if and only if there exist $m \in \mathbb{N}$ and $M(D) \in \mathcal{M}[D]^{q\times m}$ such that* $\ker_t R = \operatorname{im}_t M$ *for almost all $t \in \mathbb{R}$.*

Proof. Suppose $R(D)$ is factorized as in (2.1) and let $\mathbb{T}$ denote the discrete set given in Remark 1. By Theorem 3.2 it remains to show that $r(D)$ is a nonzero meromorphic function if and only if $\ker_t R = \operatorname{im}_t M$ for all $t \in \mathbb{R} \setminus \mathbb{T}$.

"⇒": Set

$$M(D) := V(D) \begin{bmatrix} 0_{g\times(q-g)} \\ I_{q-g} \end{bmatrix}.$$

Then $\operatorname{im}_t M \subset \ker_t R$ for all $t \in \mathbb{R} \setminus \mathbb{T}$ is immediate. If $w \in \ker_t R$ for $t \in \mathbb{R} \setminus \mathbb{T}$, then $r(D)$ being nonzero and meromorphic yields

$$\left[I_g \;\middle|\; 0_{g\times(q-g)}\right] V(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1} w(t) = 0 \qquad \text{for all } t \in \operatorname{dom} w \cap (\mathbb{R} \setminus \mathbb{T}),$$

and so there exists $l \in \mathcal{C}^\infty_t(\mathbb{R}^m)$ such that

$$V(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1} w = \begin{bmatrix} 0_{g\times(q-g)} \\ I_{(q-g)} \end{bmatrix} l.$$

"$\Leftarrow$": Let $t \in \mathbb{R} \setminus \mathbb{T}$ and choose an open interval $\mathbb{I} \subset (\mathbb{R} \setminus \mathbb{T})$ with $t \in \mathbb{I}$. Seeking a contradiction, by Theorem 3.2 one may assume that $\deg r(D) \geq 1$. Comparing the gth components of the identical vector spaces

$$\left\{ w \in \mathcal{C}^\infty(\mathbb{I}, \mathbb{R}^q) \;\middle|\; \left[\begin{array}{cc|c} I_{g-1} & & \\ & r(\frac{\mathrm{d}}{\mathrm{d}t}) & 0_{g\times(q-g)} \end{array}\right] V(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1} w = 0 \right\}$$

and $\{w \in \mathcal{C}^\infty(\mathbb{I}, \mathbb{R}^q) \mid \exists\, l \in \mathcal{C}^\infty(\mathbb{I}, \mathbb{R}^m) : w = M(\frac{\mathrm{d}}{\mathrm{d}t}) l\}$ yields that

$$\begin{aligned} \dim\{(V(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1} w(t))_g \mid w \in \mathcal{C}^\infty(\mathbb{I}, \mathbb{R}^q) \wedge r(\tfrac{\mathrm{d}}{\mathrm{d}t})(V(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1} w(t))_g = 0\} \\ = \dim\{(V(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1} M(\tfrac{\mathrm{d}}{\mathrm{d}t}) l(t))_g \mid l \in \mathcal{C}^\infty(\mathbb{I}, \mathbb{R}^m)\}. \end{aligned}$$

However, the former has finite dimension $\deg r(D) \geq 1$, while the latter is zero-dimensional or has infinite dimension. This is a contradiction, and hence the proof of the proposition is complete. □

Proposition 3.4 is known for time-invariant systems; see [24, Thm. 6.6.1]. However, the different proof presented here might also be of interest in the time-invariant case.

In the following proposition we show how to present the largest controllable behavior in terms of the nonunique factorization (2.1).

PROPOSITION 3.5. *If $R(D) \in \mathcal{M}[D]^{g\times q}$ is factorized as in* (2.1)*, then we have*

$$\ker_t^{\mathrm{contr}} R = \{w \in \ker_t R \,|\, [I_g, 0_{g\times(q-g)}] V(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1} w = 0\} \quad \textit{for almost all } t \in \mathbb{R}.$$

Proof. Since $[I_g, 0_{g\times(q-g)}]\, V(D)^{-1}$ is right invertible, it follows from Theorem 3.2 that

$$\ker_t^{\mathrm{c}} R := \{w \in \ker_t R \,|\, [I_g, 0] V(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1} w = 0\}$$

is a controllable behavior almost everywhere. Therefore, we have to show that $\ker_t^{\mathrm{contr}} R \subset \ker_t^{\mathrm{c}} R$ almost everywhere. Let $\mathbb{T}$ denote the union of all zeros and poles of the meromorphic coefficients in all entries of $U(D)$, $U(D)^{-1}$, $V(D)$, $V(D)^{-1}$, $r(D)$, $\bar{U}(D)$, $\bar{U}(D)^{-1}$, $\bar{V}(D)$, $\bar{V}(D)^{-1}$, and $\bar{r}(D)$. Then $\mathbb{T}$ is a discrete set. Let $w \in \ker_t^{\mathrm{contr}} R$ for $t \in \mathbb{R} \setminus \mathbb{T}$. Choose an open interval $\mathbb{I} \subset \mathbb{T}$ with $t \in \mathbb{I}$. Then

$$V(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1} w =: [0, \ldots, 0, \varphi_g, \ldots, \varphi_q]^T \in \mathcal{C}^\infty(\mathbb{I}, \mathbb{R}^q) \quad \text{and} \quad r(\tfrac{\mathrm{d}}{\mathrm{d}t})\varphi_g = 0.$$

The function φ_g, as a solution of a linear ordinary differential equation with real analytic coefficients on $\mathbb{I}$, is real analytic on $\mathbb{I}$ itself. Therefore, the normal form (2.1) and the identity property of analytic function yields $\varphi_g \equiv 0$. This proves $w(t) = V(\frac{\mathrm{d}}{\mathrm{d}t})\,[0, \ldots, 0, \varphi_{g+1}, \ldots, \varphi_q]^T \in \ker_t^c R$.

If $R(D)$ is factorized as in (2.3), then by Remark 3 one concludes that

$$[I_g, 0]\,\bar{V}(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1}w = [I_g, 0]\begin{bmatrix} W_1(\frac{\mathrm{d}}{\mathrm{d}t}) & 0 \\ W_3(\frac{\mathrm{d}}{\mathrm{d}t}) & W_4(\frac{\mathrm{d}}{\mathrm{d}t}) \end{bmatrix} V(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1}w = [W_1(\tfrac{\mathrm{d}}{\mathrm{d}t}), 0]V(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1}w,$$

and the result follows, since $W_1(D)$ is unimodular. This completes the proof. □

Example 1. Revisiting example (1.9), we now can show that this system is locally controllable almost everywhere.

Without loss of generality, we may assume that the coordinate system for the Lagrange multipliers is such that $F_0 = [F_1\, 0]$ with nonsingular $F_1 \in \mathbb{R}^{2\times 2}$, and we partition

$$-K_0 = \begin{bmatrix} K_{11}(t) & K_{12}(t) \\ K_{21}(t) & K_{22}(t) \end{bmatrix}, \quad M_0 = \begin{bmatrix} M_{11}(t) & M_{12}(t) \\ M_{21}(t) & M_{22}(t) \end{bmatrix}, \quad -D_0 = \begin{bmatrix} D_{11}(t) & D_{12}(t) \\ D_{21}(t) & D_{22}(t) \end{bmatrix},$$

and

$$S_0 = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix},$$

with $K_{11}(t), M_{11}(t), D_{11}(t), \in \mathbb{R}^{2\times 2}$, $S_1 \in \mathbb{R}^{2\times 3}$, and all other formats accordingly. Then the system (1.9), for $t \in \mathbb{I}$, may be written as

$$\begin{bmatrix} I_2 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & M_{11}(t) & M_{12}(t) & 0 \\ 0 & 0 & M_{21}(t) & M_{22}(t) & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \\ \dot{x}_5 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 & I_2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ K_{11}(t) & K_{12}(t) & D_{11}(t) & D_{12}(t) & F_1^T \\ K_{21}(t) & K_{22}(t) & D_{21}(t) & D_{22}(t) & 0 \\ F_1 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ S_1 \\ S_2 \\ 0 \end{bmatrix} u.$$

Since F_1 is nonsingular and S_1, S_2 are constant matrices of full row rank, it follows that $x_1 = 0$ and $\dot{x}_1 = 0$, whence $x_3 = 0$. Therefore, (1.9) is equivalent to

$$\begin{bmatrix} D & -1 & 0_{2\times 1} & 0 \\ -K_{12}(t) & M_{12}(t) & -F_1 & S_1 \\ -K_{22}(t) & M_{22}(t) & 0 & S_2 \end{bmatrix} \begin{bmatrix} x_2 \\ x_4 \\ x_5 \\ u \end{bmatrix} = 0,$$

with corresponding right invertible matrix $R(D)$. By Theorem 3.2, the system (1.9) is locally controllable almost everywhere on $\mathbb{I}$.

4. Autonomous behavior. In this section we show that the local behavior (in the sense almost everywhere) can be decomposed into the direct sum of the controllability subspace and an autonomous subspace.

DEFINITION 4.1. *Let $R(D) \in \mathcal{M}[D]^{g\times q}$ and $t \in \mathbb{R}$. A local subbehavior $\mathfrak{B}_t \subset \ker_t R$ is called* autonomous *if and only if for any $w^1, w^2 \in \mathfrak{B}_t$ with $w^1 \equiv w^2$ on some open interval $\mathbb{I} \subset \operatorname{dom} w^1 \cap \operatorname{dom} w^2$ with $t \in \mathbb{I}$ it follows that $w^1 \equiv w^2$ on $\operatorname{dom} w^1 \cap \operatorname{dom} w^2$.*

A behavior $\mathfrak{B} = \bigcup_{t\in\mathbb{R}} \mathfrak{B}_t \subset \ker_t R$ *is called* autonomous *if and only if* $\mathfrak{B}_t$ *is autonomous for almost all* $t \in \mathbb{R}$.

The above definition is a generalization of autonomous subbehavior of time-invariant systems as, for example, defined in [24, p. 67].

PROPOSITION 4.2. *Consider* $R(D) \in \mathcal{M}[D]^{g\times q}$ *with factorization* (2.1) *and* $\operatorname{rk} R(D) = g$. *Then for any autonomous behavior* $\ker^{\mathrm{aut}} R$, *the following properties hold:*

(i) $\ker_t^{\mathrm{aut}} R \cap \ker_t^{\mathrm{contr}} R = \{0\}$ *for almost all* $t \in \mathbb{R}$.

(ii) *If* $w \in \ker_t^{\mathrm{aut}} R$, *then*

$$\begin{bmatrix} I_{g-1} & \\ & r(\frac{\mathrm{d}}{\mathrm{d}t}) \end{bmatrix} V(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1} w = 0.$$

(iii)

$$\left\{ w \in \mathcal{C}_t^\infty(\mathbb{R}^q) \;\middle|\; \begin{bmatrix} I_{g-1} & \\ & r(\frac{\mathrm{d}}{\mathrm{d}t}) \end{bmatrix} V(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1} w = 0 \right\}$$

is an autonomous behavior for almost all $t \in \mathbb{R}$.

(iv) *The behavior* $\ker R$ *is autonomous if and only if* $R(D)$ *has full column rank.*

Proof. (i) If $w \in \ker^{\mathrm{aut}} R$ and $w \not\equiv 0$, then it cannot belong to the controllable behavior; otherwise Definition 4.1 would be violated.

(ii) By (i) and Proposition 3.5, any $w \in \ker_t^{\mathrm{aut}} R$ satisfies $[0_{(q-g)\times g}, I_{q-g}] V(\frac{\mathrm{d}}{\mathrm{d}t})^{-1} w = 0$. Hence (ii) follows from (2.1).

(iii) Let $\mathbb{T}$ denote the discrete set given in Remark 1 and let $t \in \mathbb{R}\setminus\mathbb{T}$. If $w \in \ker_t R$ and satisfies

$$\begin{bmatrix} I_{g-1} & \\ & r(\frac{\mathrm{d}}{\mathrm{d}t}) \end{bmatrix} V(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1} w = 0,$$

then (2.1) yields that w is of the form

$$w = V(\tfrac{\mathrm{d}}{\mathrm{d}t})\, [0, \ldots, 0, \varphi_g, 0, \ldots, 0]^T$$

for some $\varphi_g \in \mathcal{C}_t^\infty(\mathbb{R})$ with $r(\frac{\mathrm{d}}{\mathrm{d}t})\varphi_g = 0$. Since r has real analytic coefficients, the solution is real analytic, too, and the identity property of real analytic functions ensures local uniqueness of w as in Definition 4.1. This completes the proof.

(iv) This statement follows immediately from the definition and from Theorem 2.1. ∎

Note that the representation of the autonomous behavior in Proposition 4.2(iii) is not uniquely defined; it depends on the factorization (2.1). This holds true already for time-invariant systems; see [24, Rem. 5.2.15]. However, the dimension of this autonomous behavior is unique; this follows from the fact that $r(D)$ is unique up to similarity, and the latter preserves the degree; see Theorem 2.1. For time-invariant systems (1.1), the results of Proposition 4.2 can be found in [24, sect. 5.2]. More importantly, the sum of an autonomous behavior and the controllable behavior is indeed uniquely defined. In the following we generalize this result to time-varying systems.

THEOREM 4.3. *Consider the system* $R(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ *with* $R(D) \in \mathcal{M}[D]^{g\times q}$ *and* $\operatorname{rk} R(D) = g$, *factorizations* (2.1), (2.3), *and define, for all* $t \in \mathbb{R}$,

$$\ker_t^{\mathrm{contr}} R = \{w \in \ker_t R \mid \begin{bmatrix} I_g, & 0_{g\times(q-g)} \end{bmatrix} V(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1} w = 0\},$$

$$\ker_t^{\mathrm{aut}} R = \left\{ w \in \ker_t R \;\middle|\; \begin{bmatrix} I_{g-1} & \\ & r(\frac{\mathrm{d}}{\mathrm{d}t}) \end{bmatrix} V(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1} w = 0 \right\},$$

$$\overline{\ker}_t^{\mathrm{aut}} R = \left\{ w \in \ker_t R \;\middle|\; \begin{bmatrix} I_{g-1} & \\ & \bar{r}(\frac{\mathrm{d}}{\mathrm{d}t}) \end{bmatrix} \bar{V}(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1} w = 0 \right\},$$

where the latter is defined with respect to (2.3). *Then*

$$\ker_t R = \ker_t^{\mathrm{aut}} R \oplus \ker_t^{\mathrm{contr}} R = \overline{\ker}_t^{\mathrm{aut}} R \oplus \ker_t^{\mathrm{contr}} R \quad \textit{for almost all } t \in \mathbb{R}. \tag{4.1}$$

Proof. Let $\mathbb{T}$ denote the union of all zeros and poles of the meromorphic coefficients in all entries of $U(D)$, $U(D)^{-1}$, $V(D)$, $V(D)^{-1}$, $r(D)$ and $\bar{U}(D)$, $\bar{U}(D)^{-1}$, $\bar{V}(D)$, $\bar{V}(D)^{-1}$, $\bar{r}(D)$. $\mathbb{T}$ is a discrete set. In the following we consider $t \in \mathbb{R} \setminus \mathbb{T}$ and an open interval $\mathbb{I} \subset \mathbb{T}$ with $t \in \mathbb{I}$. We proceed in several steps.

Step 1. By Proposition 4.2(i) the sums in (4.1) are direct sums.

Step 2. The inclusion

$$\ker_t R \supset \ker_t^{\mathrm{aut}} R \oplus \ker_t^{\mathrm{contr}} R$$

follows from the definition of $\ker_t^{\mathrm{aut}} R$ and $\ker_t^{\mathrm{contr}} R$.

Step 3. We show

$$\ker_t R \subset \ker_t^{\mathrm{aut}} R \oplus \ker_t^{\mathrm{contr}} R. \tag{4.2}$$

Let $w \in \ker_t R$ and set

$$[\varphi_1, \dots, \varphi_q]^T := V(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1} w \in \mathcal{C}^\infty(\mathbb{I}; \mathbb{R}^q).$$

Then

$$\begin{bmatrix} I_{g-1} & \\ & r(\frac{\mathrm{d}}{\mathrm{d}t}) \end{bmatrix} V(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1} w = 0,$$

and hence

$$[\varphi_1, \dots, \varphi_q]^T = [0, \dots, 0, \varphi_g, 0, \dots, 0]^T \quad \text{with } r(\tfrac{\mathrm{d}}{\mathrm{d}t})\varphi_g = 0.$$

Finally

$$w_1 := V(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1} [0, \dots, 0, \varphi_g, 0, \dots, 0]^T \in \ker_t^{\mathrm{aut}} R,$$
$$w_2 := V(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1} [0, \dots, 0, \varphi_{g+1}, \dots, \varphi_q]^T \in \ker_t^{\mathrm{contr}} R$$

yields $w_1 + w_2 = w$, whence (4.2).

Step 4. We show

$$\ker_t^{\mathrm{aut}} R \oplus \ker_t^{\mathrm{contr}} R \subset \overline{\ker}_t^{\mathrm{aut}} R \oplus \ker_t^{\mathrm{contr}} R.$$

Let $w_1 \in \ker_t^{\mathrm{aut}} R$ and $w_2 \in \ker_t^{\mathrm{contr}} R$. Then

$$[0, \dots, 0, \varphi_g, 0, \dots, 0]^T := V(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1} w_1 \in \mathcal{C}^\infty(\mathbb{I}; \mathbb{R}^q) \quad \text{with } r(\tfrac{\mathrm{d}}{\mathrm{d}t})\varphi_g = 0,$$
$$[0, \dots, 0, \varphi_{g+1}, \dots, \varphi_q]^T := V(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1} w_2 \in \mathcal{C}^\infty(\mathbb{I}; \mathbb{R}^q).$$

Since $w := w_1 + w_2 \in \ker_t R$, it follows from (2.3) that

$$\bar{V}(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1} w = [0, \dots, 0, \bar{\varphi}_g, \dots, \bar{\varphi}_q]^T \in \mathcal{C}^\infty(\mathbb{I}; \mathbb{R}^q) \quad \text{with } r(\tfrac{\mathrm{d}}{\mathrm{d}t})\bar{\varphi}_g = 0.$$

Finally, setting

$$\bar{w}_1 := \bar{V}(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1} [0, \dots, 0, \bar{\varphi}_g, 0, \dots, 0]^T \in \overline{\ker}^{\mathrm{aut}} R,$$
$$\bar{w}_2 := \bar{V}(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1} [0, \dots, 0, \bar{\varphi}_{g+1}, \dots, \bar{\varphi}_q]^T \in \overline{\mathfrak{B}}_R^{\mathrm{contr}}$$

shows $w = \bar{w}_1 + \bar{w}_2 \in \overline{\ker}_t^{\mathrm{aut}} R \oplus \ker_t^{\mathrm{contr}} R$.

Step 5. The inclusion

$$\ker_t^{\mathrm{aut}} R \oplus \ker_t^{\mathrm{contr}} R \supset \overline{\ker}_t^{\mathrm{aut}} R \oplus \ker_t^{\mathrm{contr}} R$$

follows by symmetry as in Step 4. This completes the proof of the theorem. ∎

5. Observability. In this section, we study how one behavior can be observed from another. Essential for this are the concepts of adjoints of matrices over $\mathcal{M}[D]$ and the adjoint of a kernel representation $\ker R$.

Definition 5.1. *The* adjoint *for matrices over $\mathcal{M}[D]$ is defined as*

$$.^{\mathrm{ad}} : \mathcal{M}^{n\times m}[D] \to \mathcal{M}^{m\times n}[D], \qquad \sum_{i=0}^{k} P_i D^i \mapsto \left(\sum_{i=0}^{k} P_i D^i\right)^{\mathrm{ad}} := \sum_{i=0}^{k} (-1)^i D^i P_i^T.$$

Proposition 5.2. *The adjoint is an anti-isomorphism; i.e., it is surjective, injective, and satisfies, for arbitrary matrices $P(D), Q(D)$ over $\mathcal{M}[D]$ with appropriate formats,*

$$[P(D) + Q(D)]^{\mathrm{ad}} = P(D)^{\mathrm{ad}} + Q(D)^{\mathrm{ad}}, \tag{5.1}$$
$$[P(D) \cdot Q(D)]^{\mathrm{ad}} = Q(D)^{\mathrm{ad}} \cdot P(D)^{\mathrm{ad}}. \tag{5.2}$$

Proof. Surjectivity, injectivity, and addition are straightforward. It remains to prove the antimultiplication rule (5.2). This is well known in the scalar case; see, for example, [21, p. 25]. To prove the matrix case, denote the entries of $P(D) \in \mathcal{M}^{n\times m}[D]$, $Q(D) \in \mathcal{M}^{m\times l}[D]$ by $p_{ij}(D)$, $q_{ij}(D)$, respectively. Then

$$P(D)^{\mathrm{ad}} = \big(p_{ji}(D)^{\mathrm{ad}}\big)_{1\le i\le n,\, 1\le j\le m}, \qquad Q(D)^{\mathrm{ad}} = \big(q_{ji}(D)^{\mathrm{ad}}\big)_{1\le i\le m,\, 1\le j\le l}$$

and applying this to

$$\big(P(D)\cdot Q(D)\big)_{ij} = \sum_{\lambda=1}^{k} p_{i\lambda}(D) q_{\lambda j}(D)$$

and using the antimultiplication rule (5.2) for scalar polynomials yield the result. This completes the proof. ∎

Definition 5.3. *Let $R(D) \in \mathcal{M}[D]^{g\times q}$ and let $t \in \mathbb{R}$. The* local adjoint *of the kernel representation $\ker_t R$ is the image representation $\mathrm{im}_t R^{\mathrm{ad}}$, i.e., $(\ker_t R)^{\mathrm{ad}} = \mathrm{im}_t R^{\mathrm{ad}}$.*

Certainly, the projection onto the first q components of the kernel representation

$$\left\{ (\tilde{w}, l) \in \mathcal{C}_t^\infty(\mathbb{R}^q) \times \mathcal{C}_t^\infty(\mathbb{R}^g) \;\middle|\; \text{for all } \tau \in \mathrm{dom}\,\tilde{w} \cap \mathrm{dom}\, l : \left[I_q, R(\tfrac{\mathrm{d}}{\mathrm{d}\tau})^{\mathrm{ad}} \right] \begin{bmatrix} \tilde{w}(\tau) \\ l(\tau) \end{bmatrix} = 0 \right\}$$

yields the image representation $\operatorname{im}_t R^{\mathrm{ad}}$.

The following definition is a straightforward generalization of observability for time-invariant systems in the behavioral setup; see [24, Def. 5.3.2].

DEFINITION 5.4. *Let* $[R_1(D), R_2(D)] \in \mathcal{M}[D]^{g\times(q_1+q_2)}$ *and let* $t \in \mathbb{R}$. *Then* $w_2 \in \mathcal{C}_t^\infty(\mathbb{R}^{q_2})$ *is called* locally observable at $t \in \mathbb{R}$ from $w_1 \in \mathcal{C}_t^\infty(\mathbb{R}^{q_1})$ *for* $t \in \mathbb{R}$ *if and only if*

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}, \begin{bmatrix} w_1 \\ \tilde{w}_2 \end{bmatrix} \in \ker_t [R_1, R_2]$$

implies that

$$\text{for all } \tau \in \operatorname{dom} w_2 \cap \operatorname{dom} \tilde{w}_2 : w_2(\tau) = \tilde{w}_2(\tau).$$

An algebraic characterization of observability is given in the following theorem.

THEOREM 5.5. *Let* $[R_1(D), R_2(D)] \in \mathcal{M}[D]^{g\times(q_1+q_2)}$. *Then* w_2 *is locally observable almost everywhere from* w_1 *if and only if* $R_2(D)$ *is left invertible.*

Proof. First note that in view of the linearity of the system, it remains to show that for almost all $t \in \mathbb{R}$ we have

$$\big[w_2 \in \ker_t R_2 \implies w_2 = 0\big] \iff R_2(D) \text{ is left invertible.}$$

"$\Leftarrow$" is immediate.

"$\Rightarrow$": Let $\mathbb{T}$ denote the discrete set of the union of all zeros and poles of the meromorphic coefficients in all entries of $U_2(D), U_2(D)^{-1}, V_2(D), V_2(D)^{-1}, r_2(D)$ which take $R_2(D)$ into a normal form (2.1).

Seeking a contradiction, suppose $R_2(D)$ is not left invertible and let $t \in \mathbb{T}$. Now either $\operatorname{rk}_{\mathcal{M}[D]} R_2(D) < q_2$ (in which case the normal form (2.1) applied to $R_2(D)$ yields the existence of some $w_2 \in \ker_t R_2$ with $w_2 \neq 0$) or, again by Theorem 2.1, there exist $r_2(D) \in \mathcal{M}[D]$ with $\deg r_2(D) \geq 1$ and unimodular $U_2(D) \in \mathcal{M}[D]^{g\times g}$, $V_2(D) \in \mathcal{M}[D]^{q_2\times q_2}$ such that

$$U_2(D)^{-1} R_2(D) V_2(D)^{-1} = \begin{bmatrix} I_{q_2-1} & 0_{(q_2-1)\times 1} \\ 0_{1\times(q_2-1)} & r_2(D) \\ 0_{(g-q_2)\times(q_2-1)} & 0 \end{bmatrix}. \tag{5.3}$$

By $\deg r_2(D) \geq 1$ there exists $\varphi \in \mathcal{C}_t^\infty(\mathbb{R})\backslash\{0\}$ such that $r_2(\frac{\mathrm{d}}{\mathrm{d}t})\varphi = 0$. Therefore $w_2 := (0, \ldots, 0, \varphi)^T \in \ker_t R_2$, which is a contradiction. This completes the proof. ∎

The following theorem relates the concepts of controllability and observability.

THEOREM 5.6. *For* $R(D) \in \mathcal{M}[D]^{g\times q}$ *the following two statements are equivalent:*

(i) *The behavior* $\ker R$ *is locally controllable almost everywhere.*

(ii) *The variable* l *is locally observable almost everywhere from* w *with respect to the behavior induced by*

$$\big[I_q, R^{\mathrm{ad}}\big] \begin{pmatrix} w \\ l \end{pmatrix} = 0.$$

Proof. By Theorem 3.2, statement (i) is equivalent to $R(D)$ being right invertible, which, by Proposition 5.2, is equivalent to $R(D)^{\mathrm{ad}}$ being left invertible. The latter is, by invoking Proposition 5.5, equivalent to assertion (ii). This completes the proof of the theorem. ∎

In order to relate the classical concepts of observability known in the literature to observability as introduced above, we have to permute the columns in the presentation (1.5), (1.7) in the following proposition.

PROPOSITION 5.7. *For a time-varying Rosenbrock system of the form* (1.5) *represented in the form*

$$R(D) = [R_1(D), R_2(D)], \quad R_1(D) = \begin{bmatrix} -Q(D), & 0 \\ W(D), & -I_p \end{bmatrix}, \quad R_2(D) = \begin{bmatrix} P(D) \\ V(D) \end{bmatrix},$$

the following conditions are equivalent:

(i) *w_2 is locally observable from w_1 almost everywhere with respect to the behavior induced by*

$$\left[R_1(\tfrac{\mathrm{d}}{\mathrm{d}t}), R_2(\tfrac{\mathrm{d}}{\mathrm{d}t})\right] \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = 0.$$

(ii) *$R_2(D)$ is left invertible.*

(iii) *$[R_1(D), R_2(D)]$ is observable in the sense defined in* [14].

(iv) *If $R(D)$ represents a time-invariant Rosenbrock system, then it is observable in the sense defined in* [11].

(v) *If $R(D)$ represents a state space system* (1.3) *in the form*

$$R_1(D) = \begin{bmatrix} -B & 0 \\ -F & I_p \end{bmatrix}, \qquad R_2(D) = \begin{bmatrix} DI_n - A \\ -C \end{bmatrix},$$

then it is observable in the classical sense; see, for example, [30].

Proof. The equivalence "(i) $\Leftrightarrow$ (ii)" follows from Theorem 5.5. The equivalences "(ii) $\Leftrightarrow$ (iii)" and "(ii) $\Leftrightarrow$ (iv)" follow from [14, Thm. 6.5] and [11, Cor. 7.6], respectively. They all can be shown directly, but only for state space systems we prove "(i) $\Leftrightarrow$ (v)" directly; it shows how observability in the classical sense and in the behavioral setup are related. Note that in the case of time-varying state space systems and time-invariant Rosenbrock systems the set of critical points $\mathbb{T}$ is empty, and the system is defined on the whole time axis.

Complete observability for time-varying state space systems of the form (1.3) means (see [30, Def. 9.7]) that for any open interval $\mathbb{I} \subset \mathbb{R}$ we have

$$\begin{bmatrix} \frac{\mathrm{d}}{\mathrm{d}t} I_n - A(t) \\ -C(t) \end{bmatrix} z(t) = 0 \quad \text{for all } t \in \mathbb{I} \qquad \Longrightarrow \qquad z(t) = 0 \quad \text{for all } t \in \mathbb{I}. \tag{5.4}$$

(5.4) is equivalent to $R_2(D)$ being left invertible, and hence "(i) $\Leftrightarrow$ (v)" follows from Theorem 5.5. This completes the proof of the theorem. ∎

Example 2. Revisiting Example 1, see also (1.9), with

$$C = \begin{bmatrix} 0 & 0 & I_2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \tag{5.5}$$

corresponding to measuring the positions, we see that the resulting matrix

$$\begin{bmatrix} E(t)D - A(t) \\ C \end{bmatrix}$$

is left invertible if and only if the matrix

$$\begin{bmatrix} D & -1 & 0 \\ -K_{12}(t) & M_{12}(t)D & -F_1 \\ -K_{22}(t) & M_{22}(t)D & 0 \end{bmatrix}$$

is left invertible over the ring $\mathcal{M}[D]$, which holds if and only if $K_{22}(t)$ is nonzero. The latter is typically the case in practice, since the stiffness matrix $K_0(t)$ is symmetric and positive definite. An application of Theorem 5.5 yields the following: x is locally observable from (u, y) at t with respect to the system (1.9), (5.5) if and only if $K_{22}(t)$ is nonzero.

6. Latent variables and elimination. In [24, sect. 6.2], full and manifest behavior is considered for time-invariant systems. We do not repeat these definitions for time-varying systems but show a time-varying version of the crucial Theorem 6.2.6 in [24].

THEOREM 6.1. *Let* $[R(D), S(D)] \in \mathcal{M}[D]^{g\times(q+s)}$. *Then there exists* $R'(D) \in \mathcal{M}[D]^{g'\times q}$ *such that, for almost all* $t \in \mathbb{R}$,

$$\ker_t R' = \{w \in \mathcal{C}^\infty_t(\mathbb{R}^q) \mid \exists\, l \in \mathcal{C}^\infty_t(\mathbb{R}^q) \text{ for all } \tau \in \operatorname{dom} w \cap \operatorname{dom} l : R(\tfrac{\mathrm{d}}{\mathrm{d}\tau})w(\tau) = S(\tfrac{\mathrm{d}}{\mathrm{d}\tau})l(\tau)\}. \tag{6.1}$$

Proof. By Theorem 2.1, there exists some unimodular $U(D) \in \mathcal{M}[D]^{g\times g}$ such that

$$U(D)R(D) = \begin{bmatrix} R'(D) \\ R''(D) \end{bmatrix}, \qquad U(D)S(D) = \begin{bmatrix} 0 \\ S''(D) \end{bmatrix},$$

where $R'(D) \in \mathcal{M}[D]^{g'\times q}$, $R''(D) \in \mathcal{M}[D]^{g''\times q}$, $S''(D) \in \mathcal{M}[D]^{g''\times s}$, and $\operatorname{rk}_{\mathcal{M}[D]} S''(D) = g''$.

Applying Theorem 2.1 again, there exist $\mathcal{M}[D]$-unimodular matrices $U(D)$ and $V(D)$ of sizes g'' and s, and $r(d) \in \mathcal{M}[D]$ such that

$$S''(D) = U(D)^{-1} \left[\begin{array}{cc|c} I_{g''-1} & 0 & \\ 0 & r(D) & 0_{g''\times(q-g'')} \end{array}\right] V(D)^{-1}.$$

Choose $\mathbb{T}$ as the discrete set of the union of all zeros and poles of the meromorphic coefficients in all entries of $U(D), V(D), U(D)^{-1}, V(D)^{-1}, r(D)$. Let $\mathbb{I}$ be an open interval with $\mathbb{I} \subset \mathbb{R} \setminus \mathbb{T}$ and $t \in \mathbb{I}$.

Then, for all $\tau \in \mathbb{I}$,

$$R(\tfrac{\mathrm{d}}{\mathrm{d}\tau})w(\tau) = S(\tfrac{\mathrm{d}}{\mathrm{d}\tau})l(\tau) \iff \begin{bmatrix} R'(\frac{\mathrm{d}}{\mathrm{d}\tau}) & 0 \\ R''(\frac{\mathrm{d}}{\mathrm{d}\tau}) & S''(\frac{\mathrm{d}}{\mathrm{d}\tau}) \end{bmatrix} \begin{bmatrix} w(\tau) \\ l(\tau) \end{bmatrix}.$$

Hence the inclusion "$\supset$" in (6.1) is obvious. To show "$\subset$" in (6.1), let $w \in \ker_t R'$ for $t \in \mathbb{I}$. Let $\tilde{l}_{g''} \in \mathcal{C}^\infty(\mathbb{I}, \mathbb{R})$ denote the solution of

$$r\big(\tfrac{\mathrm{d}}{\mathrm{d}\tau}\big)\tilde{l}_{g''}(\tau) = \big(U\big(\tfrac{\mathrm{d}}{\mathrm{d}\tau}\big)S''\big(\tfrac{\mathrm{d}}{\mathrm{d}\tau}\big)w(\tau)\big)_{g''} \qquad \text{on } \mathbb{I}.$$

This solution exists; see, for example, [34, Chap. IV]. Setting

$$l := V[0, \ldots, 0, \tilde{l}_{g''}]^T$$

yields

$$U(\tfrac{\mathrm{d}}{\mathrm{d}\tau})R''(\tfrac{\mathrm{d}}{\mathrm{d}\tau})w(\tfrac{\mathrm{d}}{\mathrm{d}\tau}) = \left[\begin{array}{cc|c} I_{g''-1} & 0 & \\ 0 & r(D) & 0_{g''\times(q-g'')} \end{array}\right] V(\tfrac{\mathrm{d}}{\mathrm{d}t})^{-1} l(\tau),$$

which is equivalent to $R(\frac{\mathrm{d}}{\mathrm{d}\tau})w(\tau) = S(\frac{\mathrm{d}}{\mathrm{d}\tau})l(\tau)$. This completes the proof of the theorem. ∎

As an "inverse" to Proposition 3.4, we show that any image representation of a behavior may be written as a kernel representation.

COROLLARY 6.2. *Let $M(D) \in \mathcal{M}[D]^{q\times m}$. Then there exist $g \in \mathbb{N}$ and $R'(D) \in \mathcal{M}[D]^{g\times q}$ such that*

$$\mathrm{im}_t\, M = \ker_t R' \qquad \textit{for almost all} \quad t \in \mathbb{R}.$$

Proof. Apply Theorem 6.1 to $[R(D), S(D)] = [I_q, M(D)]$. ∎

Acknowledgment. We are indebted to Jan C. Willems and anonymous referees for their constructive criticisms of an earlier version of this paper.

While the results of the present paper were reviewed, Eve Zerz [42] wrote, based on our findings, a much more elegant algebraic approach where she partially achieves the present results with shorter proofs and also characterizes behaviors included in each other.

REFERENCES

[1] K. E. BRENAN, S. L. CAMPBELL, AND L. R. PETZOLD, *Numerical Solution of Initial-Value Problems in Differential Algebraic Equations*, Elsevier Science, North-Holland, New York, 1996.

[2] R. BYERS, P. KUNKEL, AND V. MEHRMANN, *Regularization of linear descriptor systems with variable coefficients*, SIAM J. Control Optim., 35 (1997), pp. 117–133.

[3] K. ÇAMLIBEL, M. N. BELUR, A. J. SASANE, AND J. C. WILLEMS, *On a class of time varying behaviors*, in Proceedings of the 15th Symposium on the Mathematical Theory of Networks and Systems, Notre Dame, IN, 2002, Session FA5.

[4] S. L. CAMPBELL, *Linearization of DAE's along trajectories*, Z. Angew. Math. Phys., 46 (1995), pp. 70–84.

[5] S. L. CAMPBELL, N. K. NICHOLS, AND W. J. TERRELL, *Duality, observability, and controllability for linear time-varying descriptor systems*, Circuits Systems Signal Process., 10 (1991), pp. 455–470.

[6] P. M. COHN, *Free Rings and Their Relations*, Academic Press, London, New York, 1971.

[7] M. FLIESS, *Some basic structural properties of generalized linear systems*, Systems Control Lett., 15 (1990), pp. 391–396.

[8] S. FRÖHLER AND U. OBERST, *Continuous time-varying linear systems*, Systems Control Lett., 35 (1998), pp. 97–110.

[9] E. GRIEPENTROG AND R. MÄRZ, *Differential-Algebraic Equations and Their Numerical Treatment*, Teubner Verlag, Leipzig, 1986.

[10] E. HAIRER AND G. WANNER, *Solving Ordinary Differential Equations* II: *Stiff and Differential-Algebraic Problems*, Springer-Verlag, Berlin, 1996.

[11] D. HINRICHSEN AND D. PRÄTZEL-WOLTERS, *Solution modules and system equivalence*, Internat. J. Control, 32 (1980), pp. 777–802.

[12] M. HOU AND P. C. MÜLLER, *LQ and Tracking Control of Descriptor Systems with Application to Constrained Manipulator*, Technical report, Sicherheitstechnische Regelungs-und Meßtechnik, Universität Wuppertal, Germany, 1994.

[13] A. ILCHMANN, Y. KUANG, M. KUIJPER, AND C. ZHANG, *Continuous time-varying scalar systems—a behavioural approach*, in Proceedings of the 3rd Third Asian Control Conference, Shanghai, 2000, pp. 429–433.

[14] A. ILCHMANN, I. NÜRNBERGER, AND W. SCHMALE, *Time-varying polynomial matrix systems*, Internat. J. Control, 40 (1984), pp. 329–362.

[15] E. W. KAMEN, *Representation and realization of operational differential equations with time-varying coefficients*, J. Franklin Inst., 301 (1976), pp. 559–570.

[16] P. KUNKEL AND V. MEHRMANN, *A new look at pencils of matrix valued functions*, Linear Algebra Appl., 212/213 (1993), pp. 215–248.

[17] P. KUNKEL AND V. MEHRMANN, *Local and global invariants of linear differential-algebraic equations and their relation*, Electron. Trans. Numer. Anal., 4 (1996), pp. 138–157.

[18] P. Kunkel and V. Mehrmann, *Analysis of over- and underdetermined nonlinear differential-algebraic systems with application to nonlinear control problems*, Math. Control Signals Systems, 14 (2001), pp. 233–256.
[19] P. Kunkel and V. Mehrmann, *Differential-Algebraic Equations—Analysis and Numerical Solution*, EMS Publishing House, Zürich, Switzerland, 2006, to appear.
[20] P. Kunkel, V. Mehrmann, and W. Rath, *Analysis and numerical solution of control problems in descriptor form*, Math. Control Signals Systems, 14 (2001), pp. 29–61.
[21] K. S. Miller, *Linear Differential Equations*, Routledge & Kegan Paul, London, 1964.
[22] U. Oberst, *Multidimensional constant linear systems*, Acta Appl. Math., 20 (1990), pp. 1–175.
[23] O. Ore, *Theory of non-commutative polynomials*, Ann. of Math. (2), 34 (1933), pp. 480–508.
[24] J. W. Polderman and J. C. Willems, *Introduction to Mathematical Systems Theory: A Behavioral Approach*, Springer-Verlag, New York, 1998.
[25] P. J. Rabier and W. C. Rheinboldt, *Nonholonomic Motion of Rigid Mechanical Systems from a DAE Viewpoint*, SIAM, Philadelphia, 2000.
[26] W. Rath, *Feedback Design and Regularization for Linear Descriptor Systems with Variable Coefficients*, Ph.D. thesis, TU Chemnitz-Zwickau, 1996; Shaker Verlag, Aachen, 1997.
[27] H. H. Rosenbrock, *State-Space and Multivariable Theory*, John Wiley, New York, 1970.
[28] H. H. Rosenbrock and C. Storey, *Mathematics of Dynamical Systems*, Nelson, London, 1970.
[29] J. Rudolph, *Duality in time-varying linear systems: A module theoretic approach*, Linear Algebra Appl., 245 (1996), pp. 83–106.
[30] W. J. Rugh, *Linear Systems Theory*, Prentice–Hall, Upper Saddle River, NJ, 1996.
[31] M. Sato, *Theory of hyperfunctions* I, J. Fac. Sci. Univ. Tokyo Sect. I, 8 (1959), pp. 139–193.
[32] M. Sato, *Theory of hyperfunctions* II, J. Fac. Sci. Univ. Tokyo Sect. I, 8 (1960), pp. 387–436.
[33] E. D. Sontag, *Mathematical Control Theory*, 2nd ed., Springer-Verlag, New York, 1998.
[34] W. Walter, *Ordinary Differential Equations*, Springer-Verlag, New York, 1998.
[35] J. C. Willems, *System theoretic models for the analysis of physical systems*, Ricerche Automat., 10 (1981), pp. 71–106.
[36] J. C. Willems, *From time series to linear system,* I: *Finite dimensional linear time invariant systems*, Automatica J. IFAC, 22 (1986), pp. 561–580.
[37] J. C. Willems, *From time series to linear system,* II: *Exact modelling*, Automatica J. IFAC, 22 (1986), pp. 675–694.
[38] J. C. Willems, *From time series to linear system,* III: *Approximate modelling*, Automatica J. IFAC, 23 (1987), pp. 87–115.
[39] W. A. Wolovich, *Linear Multivariable Systems*, Springer-Verlag, New York, 1974.
[40] R. Ylinen, *On the Algebraic Theory of Linear Differential and Difference Systems with Time-Varying or Operator Coefficients*, Systems Theory Laboratory Report, B23, Helsinki University, Helsinki, 1975.
[41] R. Ylinen, *An algebraic theory for analysis and synthesis of time-varying linear systems*, Acta Polytech. Scand. Math. Comput. Sci. Ser., 32 (1980), pp. 1–61.
[42] E. Zerz, *An algebraic analysis approach to linear time-varying systems*, IMA J. Math. Control Inform., to appear.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 5, pp. 1748–1765

A BEHAVIORAL APPROACH TO TIME-VARYING LINEAR SYSTEMS. PART 2: DESCRIPTOR SYSTEMS*

ACHIM ILCHMANN† AND VOLKER MEHRMANN‡

Abstract. In the sequel to [A. Ilchmann and V. Mehrmann, *SIAM J. Control Optim.*, 44 (2005), pp. 1725–1747], we discuss a behavioral approach for linear, time-varying, differential algebraic (descriptor) systems with real analytic coefficients. The analysis is "almost global" in the sense that the analysis is not restricted to an interval $\mathbb{I} \subset \mathbb{R}$ but is allowed for the "time axis" $\mathbb{R}\setminus\mathbb{T}$, where $\mathbb{T}$ is a discrete set of critical points, at which the solution may exhibit a finite escape time. Controllable, observable, and autonomous behavior for linear time-varying descriptor systems is characterized.

Key words. time-varying linear systems, descriptor systems, behavioral approach, controllability, observability, autonomous system

AMS subject classifications. 93B11, 93B40, 93B36

DOI. 10.1137/040609021

1. Introduction. In [12], a behavioral approach was developed for linear time-varying systems with real analytic coefficients. In this paper, this approach will be studied for the specific case of linear time-varying descriptor systems described by differential-algebraic equations of the form

$$\begin{array}{rcl} E(t)\,\dot{x}(t) &=& A(t)x(t) \;+\; B(t)u(t), \\ y(t) &=& C(t)x(t) \;+\; F(t)u(t), \end{array} \tag{1.1}$$

with real analytic matrices $A \in \mathcal{A}^{l\times n}$, $B \in \mathcal{A}^{l\times m}$, $C \in \mathcal{A}^{p\times n}$, $F \in \mathcal{A}^{p\times m}$, where $E \in \mathcal{A}^{l\times n}$ is allowed to be singular in the sense that rk $E(t) < \min\{l, n\}$ for some $t \in \mathbb{R}$. Throughout this paper, the nomenclature as introduced and listed in [12] will be used.

As in [12], we make use of the skew-polynomial rings $\mathcal{A}[D]$ and $\mathcal{M}[D]$ (see [6, 13]) of differential polynomials with coefficients in $\mathcal{A}$, $\mathcal{M}$, respectively, and indeterminate D representing the differential operator $\frac{\mathrm{d}}{\mathrm{d}t}$, and the multiplication rule $Df = fD + \dot{f}$. The algebraic object

$$R(D) = \sum_{i=0}^{n} R_i D^i \in \mathcal{M}[D]^{g\times q} \cong \mathcal{M}^{g\times q}[D]$$

acts on $\mathcal{C}^\infty$-functions w via

$$R(\tfrac{\mathrm{d}}{\mathrm{d}t})w(t) = \sum_{i=0}^{n} R_i(t) w^{(i)}(t)\,.$$

In this notation, time-varying descriptor systems (1.1) may be rewritten as

$$R(\tfrac{\mathrm{d}}{\mathrm{d}t})w = 0, \tag{1.2}$$

*Received by the editors May 26, 2004; accepted for publication (in revised form) March 8, 2005; published electronically November 23, 2005.
http://www.siam.org/journals/sicon/44-5/60902.html
†Institut für Mathematik, Technische Universität Ilmenau, Weimarer Straße 25, D-98693 Ilmenau, Germany (ilchmann@mathematik.tu-ilmenau.de).
‡Institut für Mathematik, MA 4-5, Technische Universität Berlin, Straße des 17. Juni 136, D-10623 Berlin, Germany (mehrmann@math.tu-berlin.de). This research was supported by DFG Research Center MATHEON Mathematics for key technologies in Berlin.

where

$$R(D) = \begin{bmatrix} ED - A & -B & 0 \\ -C & -F & I_p \end{bmatrix}, \quad \text{and} \quad w = \left[x^T, u^T, y^T\right]^T.$$

Systems of differential algebraic equations (often called descriptor systems) play an important role in modelling and control of multibody systems, electric circuits, or coupled systems of partial differential equations; see [1, 9].

The analysis of the behavior of (1.1) has to cope with three essential difficulties. First, the solutions of time-varying systems may exhibit critical points, i.e., a finite escape time. Second, descriptor systems behave quite differently from classical state space systems (i.e., $E = I_n$ in (1.1)). For state space systems, the function $u(\cdot)$ can be considered as an input function free to choose, and initial conditions can be arbitrary. This is in general not true for descriptor systems (1.1), since descriptor systems may contain algebraic constraints, which restrict the solutions, the set of possible inputs, and also the initial values to some manifold. Third, some of the constraints that arise (the hidden constraints) are not explicit and thus it is not clear how to choose the underlying spaces for the descriptor variables x, u, y. Finally, the analytic property of the solution or behavior is local, which is in contrast to the global algebraic properties of $R(D)$. These difficulties are illustrated by the following example.

Example 1.

(i) The scalar differential equations $t\dot{x} = -x$, $t^2\dot{x} = -x$, $t\dot{x} = x$, have local solutions $t \mapsto t^{-1}, e^{1/t}, t$, respectively. Hence at $t = 0$ the solution might be rational with a pole, not even analytic, or does not have any pole, respectively.

(ii) The variables $x_1, \ldots, x_4, u_1, u_2$ of the descriptor system (1.1) with

$$E = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \; A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \; B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix},$$
$$C = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}, \; F = 0_{1\times 2}$$

satisfy the equivalent description

$$u_2 = 0, \; \dot{x}_2 = x_1, \; y = x_4, \; \dot{x}_3 = x_2 + u_1.$$

Thus, u_2 is constrained to be 0 and cannot be freely chosen, as it could in the case of state space systems. The variables x_1 and x_4 can be viewed as input or state variables; the system description does not determine this.

Note also that if we choose the input u_1 as a step function, then either x_3 is chosen as input as well to compensate the delta distribution in u_1 (since $x_1 = \dot{x}_2 = \ddot{x}_3 - \dot{u}_1$), or we may have to enlarge the solution space to allow that x_1 is a delta distribution. But even if we do so, then we have the problem that x_1 is not observable from the output y, which means that internally the system has impulsive parts of the solution, which are not observed. For many types of practical systems, such as, for example, mechanical systems, this would be a disaster: impulses in the solution cannot be tolerated. □

Example 1 indicates that the behavioral viewpoint, where state-, output-, and input-variables are not distinguished, seems the appropriate concept for the analysis of descriptor systems. The behavioral approach has been introduced by Willems [25, 26, 27, 28]; see also the textbook [21] and [12] for a general presentation.

Motivated by Example 1 and as introduced in [12], we study, for $R(D) \in \mathcal{M}[D]^{g\times q}$, *local solutions* of $R(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ belonging to

$$\mathcal{C}_t^\infty(\mathbb{R}^q) \;:=\; \left\{w \in \mathcal{C}^\infty(\mathbb{I},\mathbb{R}^q)\,\middle|\, \mathbb{I} \subset \mathbb{R} \text{ an open interval with } t \in \mathbb{I}\right\}, \quad t \in \mathbb{R},$$

as the *almost global behavior* given by the kernel representation

$$\ker R \;=\; \left\{w \in \mathcal{C}_{\mathrm{pw}}^\infty(\mathbb{R}^q)\,\middle|\, R(\tfrac{\mathrm{d}}{\mathrm{d}\tau})w(\tau) = 0 \text{ for almost all } \tau \in \mathbb{R}\right\}.$$

The *local behavior*

$$\ker_t R \;=\; \left\{w \in \mathcal{C}_t^\infty(\mathbb{R}^q)\,\middle|\, R(\tfrac{\mathrm{d}}{\mathrm{d}\tau})w(\tau) = 0 \text{ for all } \tau \in \text{ dom } w\right\}, \qquad t \in \mathbb{R},$$

becomes a *real vector space* if endowed, for $w_1, w_2 \in \ker_t R$, with addition

$$(w_1 + w_2)(\tau) \;:=\; w_1(\tau) + w_2(\tau) \qquad \forall \tau \in \operatorname{dom} w_1 \cap \operatorname{dom} w_2\,,$$

and obvious scalar multiplication.

We also have to consider those points of the real axis, where the local solution is no longer extendable.

Definition 1.1. *Consider the descriptor system* (1.2). *The set of* critical points, *where the solution is not defined, is given by*

$$\mathbb{T}_R^{\mathrm{crit}} := \left\{ t' \in \mathbb{R} \;\middle|\; \begin{array}{l} \text{there exists, for some } \varepsilon > 0, \text{ a } \mathcal{C}^\infty \text{ function} \\ w : (t'-\varepsilon, t') \to \mathbb{R}^q \text{ or } w : (t', t'+\varepsilon) \to \mathbb{R}^q \\ \text{which solves (1.2) and cannot be extended to} \\ (t'-\varepsilon, t'] \text{ or } [t', t'+\varepsilon)\,, \text{ respectively.} \end{array} \right\} \tag{1.3}$$

Note that for the three differential equations in Example 1(i), the sets of critical points are $\{0\}$, $\{0\}$, $\emptyset$, respectively.

Since E in (1.1) is real analytic, it follows that for almost all $\hat{t} \in \mathbb{R}$, the rank of the matrix $E(\hat{t}) \in \mathbb{R}^{l\times m}$ is equal to $\mathrm{rk}_{\mathcal{A}} E$, and the set of critical points is a discrete set. It is an open problem to characterize the set of critical points. However, we will determine discrete sets which include all critical points.

We define the appropriate behavior, i.e., the solution space, of (1.2) on the time-axis $\mathbb{R}\backslash\mathbb{T}$, where $\mathbb{T}$ is discrete and includes the set of critical points of (1.2). Controllability and observability are defined in terms of trajectories (descriptor variables), which is a conceptual generalization of controllability and observability for state space systems. For these systems in [5] controllability and observability have been studied in terms of derivative arrays. In [4] a first behavior-like approach for analytic coefficients has been discussed. A more general approach that allows for larger classes of coefficients and that can also be implemented numerically has been introduced in [16]. In [11] a first approach in the spirit of the present paper was presented for scalar systems. A completely different approach results from the study of differential-algebraic equations; see [1, 8, 17]. A general solvability theory for nonsquare linear time-varying systems was first given in [15] and analyzed for control problems in a behavioral context in [4, 18, 22]; see also [16] for the general nonlinear case. In these papers, however, mainly the concept of regularization has been discussed, i.e., the problem of finding appropriate feedback that makes the system regular and also decreases the index. Here we consider controllability and observability in the behavioral context.

This paper is organized as follows. In section 2, we define critical points and follow the concepts of [15, 22] by deriving condensed forms for time-varying descriptor systems (1.2) to determine sets covering the critical points. In section 3, controllability is defined, algebraically characterized, and related to the well-known concepts of controllability. In section 4, we apply results from [12] and briefly discuss autonomous behavior and observability for descriptor systems.

2. Condensed forms. In this section, condensed forms with respect to state and input transformations are studied for time-varying descriptor systems (1.2). The condensed form allows us to classify the solution sets and to identify the constraint manifolds for the variables. These forms are akin to the forms derived in [4, 17, 18].

The construction of the condensed forms is based on the computation of analytic singular value decompositions that were introduced in [2] for analytic matrices and that are also valid for real analytic matrices. This result states that for a matrix function $A \in \mathcal{A}^{l\times n}$ there exist real orthogonal matrix functions $U \in \mathcal{A}^{l\times l}, V \in \mathcal{A}^{n\times n}$ and a diagonal matrix $\Sigma \in \mathcal{A}^{r\times r}$, where $r = \mathrm{rk}A$, such that

$$U^T A V \;=\; \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}.$$

It should be noted, though, that, in contrast to the usual singular value decomposition for matrices, the diagonal elements of $\Sigma(t)$, in general, cannot be chosen positive or in descending order. In this way, however, the analytic singular value decomposition is not uniquely defined. Essentially, there is freedom to perform orthogonal transformations in the spaces associated with multiple singular values. This freedom can be removed by choosing minimal variation curves or by always choosing the analytic singular value decomposition to be closest to a reference point [3, 20].

THEOREM 2.1. *Consider a time-varying descriptor system of the form* (1.2) *with*

$$R(D) = \begin{bmatrix} ED - A & -B & 0 \\ -C & -F & I_p \end{bmatrix} \in \mathcal{A}[D]^{(l+p)\times(n+m+p)}.$$

(i) *There exist orthogonal matrices $U_1 \in \mathcal{A}^{l\times l}, V_1 \in \mathcal{A}^{n\times n}$ so that*

$$\begin{bmatrix} U_1 & 0 \\ 0 & I_p \end{bmatrix} R(D) \begin{bmatrix} V_1 & 0 \\ 0 & I_{m+p} \end{bmatrix} \tag{2.1}$$

corresponds to the descriptor system

$$\begin{array}{rcl} \Sigma_d \dot{x}_1 &=& A_{11} x_1 + A_{12} x_2 + A_{13} x_3 + B_1 u, \\ 0 &=& A_{21} x_1 + \Sigma_a x_2 + B_2 u, \\ 0 &=& A_{31} x_1 + B_3 u, \\ 0 &=& A_{41} x_1, \\ 0 &=& 0_{l-\nu}, \\ y &=& C_1 x_1 + C_2 x_2 + C_3 x_3 + F u, \end{array} \tag{2.2}$$

where $\Sigma_d \in \mathcal{A}^{d\times d}, \Sigma_a \in \mathcal{A}^{a\times a}$ are diagonal and invertible over $\mathcal{M}$ with $d = \mathrm{rk}\, E$, and $B_3 \in \mathcal{A}^{\gamma\times m}$, $A_{41} \in \mathcal{A}^{f\times d}$ with full row rank; i.e., $\gamma = \mathrm{rk}\, B_3$, $f = \mathrm{rk}\, A_{41}$, and $\nu = d + a + \gamma + f$. All matrices are real analytic and of conforming formats.

(ii) *There exist orthogonal matrices* $U_2 \in \mathcal{A}^{l\times l}, V_2 \in \mathcal{A}^{n\times n}, W \in \mathcal{A}^{p\times p}, Z \in \mathcal{A}^{m\times m}$ *so that*

$$\begin{bmatrix} U_2 & 0 \\ 0 & W \end{bmatrix} R(D) \begin{bmatrix} V_2 & 0 & 0 \\ 0 & Z & 0 \\ 0 & 0 & I_p \end{bmatrix} \tag{2.3}$$

corresponds to the following descriptor system in condensed form:

$$\begin{aligned}
\Sigma_d \dot{x}_1 &= A_{11}x_1 + A_{12}x_2 + A_{13}x_3 + A_{14}x_4 + A_{15}x_5 + B_{11}u_1 + B_{12}u_2,\\
0 &= A_{21}x_1 + \Sigma_a x_2 + \qquad\qquad\quad B_{21}u_1 + B_{22}u_2,\\
0 &= A_{31}x_1 + \qquad\qquad\qquad\quad \Sigma_\gamma u_1,\\
0 &= \qquad\qquad\quad \Sigma_f x_5,\\
0 &= 0_{l-\nu},\\
y_1 &= C_{11}x_1 + C_{12}x_2 + \Sigma_\omega x_3 + \qquad C_{15}x_5 + F_{11}u_1 + F_{12}u_2,\\
y_2 &= C_{21}x_1 + C_{22}x_2 + \qquad\qquad C_{25}x_5 + F_{21}u_1 + F_{22}u_2,
\end{aligned} \tag{2.4}$$

where $\Sigma_d, \Sigma_a, \Sigma_\gamma, \Sigma_f, \Sigma_\omega$ *are diagonal matrices that are invertible over* $\mathcal{M}$ *and have sizes* d, a, γ, f, ω*, respectively. Furthermore,* $\nu = d + a + \gamma + f$ *and all matrices are real analytic and of conforming formats.*

(iii) *There exist matrices* $U \in \mathcal{A}^{(l-p)\times(l-p)}$, $V \in \mathcal{M}^{n\times n}$ *invertible over* $\mathcal{M}$, $X \in \mathcal{M}^{p\times(l-p)}$, $W \in \mathcal{A}^{p\times p}$ *orthogonal,* $Z \in \mathcal{A}^{m\times m}$ *orthogonal, a scalar function* $\sigma \in \mathcal{A}$*, and a permutation matrix* $P \in \mathcal{A}^{(n+m)\times(n+m)}$ *so that*

$$\tilde{R}(D) := \begin{bmatrix} U & 0 \\ X & W \end{bmatrix} R(D) \begin{bmatrix} P & 0 \\ 0 & I_p \end{bmatrix} \tag{2.5}$$

$$= \left[\begin{array}{c|ccc|ccc|cc}
\sigma D I_d - \tilde{A}_{11} & -\tilde{A}_{13} & -\tilde{A}_{14} & -\tilde{B}_{12} & 0 & 0 & 0 & 0 & 0 \\ \hline
0 & 0 & 0 & \Sigma_a^{-1}\tilde{B}_{22} & I_a & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & I_f & 0 & 0 & 0 \\
\Sigma_\gamma^{-1}\tilde{A}_{31} & 0 & 0 & 0 & 0 & 0 & I_\gamma & 0 & 0 \\ \hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline
0 & -\Sigma_\omega & 0 & -\sigma^{-1}\tilde{F}_{12} & 0 & 0 & 0 & I_\omega & 0 \\
-\sigma^{-1}\tilde{C}_{21} & 0 & 0 & -\sigma^{-1}\tilde{F}_{22} & 0 & 0 & 0 & 0 & I_p
\end{array}\right]$$

corresponds to the meromorphic descriptor system in standard condensed

form

$$
\begin{aligned}
\sigma I_d \dot{x}_1 &= \tilde{A}_{11} x_1 + [\tilde{A}_{13},\ \tilde{A}_{14},\ \tilde{B}_{12}] \begin{bmatrix} x_3 \\ x_4 \\ u_2 \end{bmatrix}, \\
\begin{bmatrix} x_2 \\ x_5 \\ u_1 \end{bmatrix} &= \begin{bmatrix} 0 & 0 & 0 & -\Sigma_a^{-1}\tilde{B}_{22} \\ 0 & 0 & 0 & 0 \\ -\Sigma_\gamma^{-1}\tilde{A}_{31} & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_3 \\ x_4 \\ u_2 \end{bmatrix}, \qquad (2.6) \\
\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} &= \begin{bmatrix} 0 \\ \sigma^{-1}\tilde{C}_{21} \end{bmatrix} x_1 + \begin{bmatrix} \Sigma_\omega & 0 & \sigma^{-1}\tilde{F}_{12} \\ 0 & 0 & \sigma^{-1}\tilde{F}_{22} \end{bmatrix} \begin{bmatrix} x_3 \\ x_4 \\ u_2 \end{bmatrix},
\end{aligned}
$$

where

$$
\sigma(t) := \det \Sigma_d(t)\, \det \Sigma_a(t)\, \det \Sigma_\gamma(t)\, \det \Sigma_f(t) \quad \text{for all } t \in \mathbb{R}\,, \qquad (2.7)
$$

and all matrices are real analytic and of conforming formats. The integers d, a, γ, ω, f *are invariants of* (1.2).

Proof. The proof is constructive using a sequence of real analytic singular value decompositions. When multiplying with D, we will always use the product rule without saying so.

(i) Consider the first equation of (1.1) and choose orthogonal matrices $\tilde{U} \in \mathcal{A}^{l\times l}$, $\tilde{V} \in \mathcal{A}^{n\times n}$ so that

$$
[\tilde{R}(D), -\tilde{B}] = \tilde{U}[ED - A, -B] \begin{bmatrix} \tilde{V} & 0 \\ 0 & I_m \end{bmatrix} = \left[\begin{bmatrix} \Sigma_d & 0 \\ 0 & 0 \end{bmatrix} D - \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ \tilde{A}_{21} & \tilde{A}_{22} \end{bmatrix}, \begin{bmatrix} \tilde{B}_1 \\ \tilde{B}_2 \end{bmatrix} \right],
$$

where $\Sigma_d \in \mathcal{A}^{d\times d}$ with $d = \mathrm{rk} E$ is diagonal.

Next, choose orthogonal matrices $\bar{U} \in \mathcal{A}^{(l-d)\times(l-d)}$, $\bar{V} \in \mathcal{A}^{(n-d)\times(n-d)}$ so that

$$
\begin{aligned}
[\bar{R}(D),\ -\bar{B}] &= \begin{bmatrix} I_d & 0 \\ 0 & \bar{U} \end{bmatrix} [\tilde{R}(D), -\tilde{B}] \begin{bmatrix} I_d & 0 & 0 \\ 0 & \bar{V} & 0 \\ 0 & 0 & I_m \end{bmatrix} \\
&= \left[\begin{bmatrix} \Sigma_d & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} D - \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} & \bar{A}_{13} \\ \bar{A}_{21} & \Sigma_a & 0 \\ \bar{A}_{31} & 0 & 0 \end{bmatrix}, \begin{bmatrix} \bar{B}_1 \\ \bar{B}_2 \\ \bar{B}_3 \end{bmatrix} \right],
\end{aligned}
$$

where $\Sigma_a \in \mathcal{A}^{a\times a}$ is diagonal and invertible over $\mathcal{M}$. Finally, choose an orthogonal $\hat{U} \in \mathcal{A}^{(l-d-a)\times(l-d-a)}$, so that $\left[\begin{smallmatrix} I_{d+a} & 0 \\ 0 & \hat{U} \end{smallmatrix}\right][\bar{R}(D), -\bar{B}]$ has the form (2.2) with $B_3 \in \mathcal{A}^{\gamma\times m}$, $\gamma = \mathrm{rk}\, B_3 = \mathrm{rk}\, \bar{B}_3$, and $A_{41} \in \mathcal{A}^{f\times d}$, $f = \mathrm{rk}\, A_{41}$. Performing all the transformations also on C and partitioning analogously shows (2.2).

(ii) We apply the so-called *index reduction process* as introduced in [18] to (2.4): Fix f variables of x_1, corresponding to some f linearly independent columns of A_{41}, i.e., choose a unitary matrix $Q \in \mathcal{A}^{d\times d}$ such that $A_{41}Q = [A_{41}^\alpha, A_{41}^\beta]$ with $A_{41}^\alpha \in \mathcal{A}^{f\times f}$ is invertible over $\mathcal{M}$. Then

$$
0 = A_{41}x_1 = A_{41}^\alpha x_1^\alpha + A_{41}^\beta x_1^\beta, \quad \begin{bmatrix} x_1^\alpha \\ x_1^\beta \end{bmatrix} := Q x_1,
$$

and so

$$\dot{x}_1^\alpha = -(A_{41}^\alpha)^{-1} A_{41}^\beta \dot{x}_1^\beta - \frac{d}{dt}\left((A_{41}^\alpha)^{-1} A_{41}^\beta\right) x_1^\beta.$$

Inserting $\dot{x}_1^\alpha$ into the differential equation of (2.2) leaves $d-f$ differential equations. Note that we may have introduced meromorphic functions by the inverse of A_{41}^α and its derivative. A multiplication from the left with a real analytic function yields a description in the form (1.1); however, the d differential equations have been reduced to $d-f$ differential equations and we may apply part (i) again. This index reduction process stops after finitely many iterations, and we arrive at the following condensed form:

$$\begin{array}{rcccccccccc}
\Sigma_d \dot{x}_1 & = & \hat{A}_{11}x_1 & + & \hat{A}_{12}x_2 & + & \hat{A}_{13}x_3 & + & \hat{A}_{14}x_4 & + & \hat{B}_1 u, \\
0 & = & \hat{A}_{21}x_1 & + & \Sigma_a x_2 & & & & & + & \hat{B}_2 u, \\
0 & = & \hat{A}_{31}x_1 & & & & & & & + & \hat{B}_3 u, \\
0 & = & & & & & & + & \Sigma_f x_4, & & \\
0 & = & 0_{l-\nu}, & & & & & & & & \\
y & = & \hat{C}_1 x_1 & + & \hat{C}_2 x_2 & + & \hat{C}_3 x_3 & + & \hat{C}_4 x_4 & + & F u,
\end{array} \tag{2.8}$$

where $\Sigma_d, \Sigma_a, \Sigma_f$ are diagonal matrices, invertible over $\mathcal{M}$, and of sizes d, a, f, respectively, and $\hat{B}_3 \in \mathcal{A}^{\gamma\times m}$ has full row rank over $\mathcal{A}$.

As a final step we perform an analytic singular value decomposition of $\hat{C}_3, \hat{B}_3$, respectively, and derive (2.4).

(iii) Using the fact that the fourth equation in (2.4) implies that $x_5 \equiv 0$, which can be extended even at points where Σ_f is singular, we can eliminate all terms invoking x_5 from all the other equations. This corresponds to multiplying (2.4) from the left first by

$$\begin{bmatrix}
I_d & -A_{12}\Sigma_a^{-1} & -[B_{11} - A_{12}\Sigma_a^{-1}B_{21}]\Sigma_\gamma^{-1} & 0 & 0 & 0 & 0 \\
0 & I_a & -B_{21}\Sigma_\gamma^{-1} & 0 & 0 & 0 & 0 \\
0 & 0 & I_\gamma & -A_{15}\Sigma_f^{-1} & 0 & 0 & 0 \\
0 & 0 & 0 & I_f & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & I_{l-\nu} & 0 & 0 \\
0 & -C_{12}\Sigma_a^{-1} & -F_{11}\Sigma_\gamma^{-1} & 0 & 0 & I_\omega & 0 \\
0 & -C_{22}\Sigma_a^{-1} & -F_{21}\Sigma_\gamma^{-1} & 0 & 0 & 0 & I_{p-\omega}
\end{bmatrix}$$

and then by

$$\begin{bmatrix} \sigma\Sigma_d^{-1} & 0 \\ 0 & I_{l-d+p} \end{bmatrix}$$

and from the right by

$$\begin{bmatrix}
I_d & 0 & 0 & 0 \\
-[A_{21} - B_{21}\Sigma_\gamma^{-1}A_{31}]\Sigma_a^{-1} & I_a & 0 & 0 \\
-C_{11}\Sigma_\omega^{-1} & 0 & I_\gamma & 0 \\
0 & 0 & 0 & I_{(n-\nu+f+m+p)\times(n-\nu+f+m+p)}
\end{bmatrix},$$

yielding the transformed system

$$
\begin{aligned}
\sigma\,\dot{x}_1 &= \tilde{A}_{11}x_1 + \quad \tilde{A}_{13}x_3 + \tilde{A}_{14}x_4 + \quad \tilde{B}_{12}u_2,\\
0 &= \quad \Sigma_a x_2 + \quad \sigma^{-1}\tilde{B}_{22}u_2,\\
0 &= \tilde{A}_{31}x_1 + \quad \Sigma_\gamma u_1,\\
0 &= \quad \Sigma_f x_5,\\
0 &= 0_{l-v},\\
y_1 &= \quad \Sigma_\omega x_3 + \quad \sigma^{-1}\tilde{F}_{12}u_2,\\
y_2 &= \sigma^{-1}\tilde{C}_{21}x_1 + \quad \sigma^{-1}\tilde{F}_{22}u_2,
\end{aligned}
\tag{2.9}
$$

where all matrices are real analytic. This proves (2.6). □

Remark 1.

(i) If the descriptor system (1.2) is time-invariant, then all transformations in Theorem 2.1 may be chosen as constant matrices and $\sigma = 1$. In this case, the condensed forms in Theorem 2.1 are well known; see, for example [3].

(ii) To derive (2.2), only an orthogonal transformation on the variables x in (2.1) has been applied. To derive (2.4), the transformations on the variables x and u have not been mixed.
To derive (2.6), we have used nonsingular transformations on x and orthogonal transformations on u. If we allow further linear combinations (which for classical systems where y, x, and u are fixed a priori as outputs, states, and controls, respectively, correspond to state feedback or output feedback), then we can simplify (2.6) further by removing blocks such as $\tilde{A}_{31}$ or by introducing almost everywhere invertible diagonal blocks in diagonal positions of the transformed matrices E or A. Note that the transformation of derivative feedback is not an equivalence transformation, because under derivative feedback the characteristic quantities d, a, γ, f, w are not invariants and hence the properties of the system may be altered by this transformation completely; see [18].

(iii) The description (2.6) is not quite of the form (1.1), since the coefficients of x_1 and u_2 in y_1 and y_2 may have poles at the zeros of σ.

(iv) An immediate consequence of (2.6) is that the variables in x_1 represent couplings between algebraic equations and differential equations that are not influenced by u_1. Systems where such couplings between differential equations and algebraic equations occur are typically called *high index systems*. For a detailed discussion of different index concepts see [1, 8, 17]

(v) The transformation leading to (2.6) does not invoke any differentiation of u. Hence, if the variables denoted by u are classified as inputs a priori, then no extra differentiability conditions for these variables arise; see [4, 18].

(vi) The condensed forms (2.1), (2.4), and (2.6) allow us to detect candidates for critical points, given by

$$
\mathbb{T}_R^{\mathrm{crit}} \subset \mathbb{T}_R := \{t' \in \mathbb{R} \,|\sigma(t') = 0\}\,. \tag{2.10}
$$

As can be seen from the first system considered in Example 1(i), the set $\mathbb{T}_R^{\mathrm{crit}} = \emptyset$ can be a strict subset of $\mathbb{T}_R = \{0\}$.

(vii) The reader may wonder why we display equations of the form $0 = 0$ in the condensed form. These arise typically when automatic modelling systems are used and describe redundant equations in the system. □

To characterize controllability we will need the following staircase form which generalizes the staircase form of Van Dooren [24] to systems with analytic coefficient matrices.

LEMMA 2.2. *For real analytic matrices $A \in \mathcal{A}^{n\times n}, B \in \mathcal{A}^{n\times m}$ there exist orthogonal matrices $P \in \mathcal{A}^{n\times n}$ and $Q \in \mathcal{A}^{m\times m}$ so that*

$$P \left[\ D I_n - A, -B\ \right] \begin{bmatrix} P^T & 0 \\ 0 & Q \end{bmatrix} \tag{2.11}$$

$$= \left[\begin{array}{cccc|cc} DI_{n_1} - A_{11} & -A_{12} & \cdots & -A_{1,s-1} & -A_{1,s} & -B_1 \quad 0 \\ -[\hat{A}_{21}, 0] & \ddots & & \vdots & \vdots & 0 \quad 0 \\ & \ddots & \ddots & -A_{s-2,s-1} & \vdots & \vdots \quad \vdots \\ & & -[\hat{A}_{s-1,s-2}, 0] & DI_{n_{s-1}} - A_{s-1,s-1} & -A_{s-1,s} & 0 \quad 0 \\ \hline 0 & \cdots & 0 & 0 & DI_{n_s} - A_{s,s} & 0 \quad 0 \end{array}\right],$$

where $n_1 \geq n_2 \geq \cdots \geq n_{s-1} \geq n_s \geq 0$, $n_{s-1} > 0$, and $B_1 \in \mathcal{A}^{n_1\times n_1}$ and $\hat{A}_{i,i-1} \in \mathcal{A}^{n_i\times n_i}$ are invertible over $\mathcal{M}$ for $i = 1, \ldots, s-1$.

Proof. A constructive proof is given by the following generalization of the so-called *Staircase Algorithm* to systems with real analytic coefficients. Whenever we use Σ in the following, it denotes a diagonal matrix.

Step 0. Choose orthogonal $U_B \in \mathcal{A}^{n\times n}$, $V_B \in \mathcal{A}^{m\times m}$ so that

$$B = U_B^T \begin{bmatrix} \Sigma_B & 0 \\ 0 & 0 \end{bmatrix} V_B \in \mathcal{A}^{n\times m} \quad \text{with invertible } \Sigma_B \in \mathcal{A}^{n_1\times n_1},$$

and set

$$\begin{aligned} A_0 &:= U_B A U_B^T + \dot{U}_B U_B^T = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \quad \text{with } A_{21} \in \mathcal{A}^{(n-n_1)\times n_1}, \\ B_0 &:= U_B B V_B = \begin{bmatrix} \Sigma_B & 0 \\ 0 & 0 \end{bmatrix}. \end{aligned}$$

Then, using the product rule, we have

$$U_B \left[\ D I_n - A, -B\ \right] \begin{bmatrix} U_B^T & 0 \\ 0 & V \end{bmatrix} = \left[\ D I_n - A_0, -B_0\ \right].$$

Step 1. If $n_1 < n$ and $A_{21} \neq 0$, then choose orthogonal $U_{21} \in \mathcal{A}^{(n-n_1)\times(n-n_1)}$, $V_{21} \in \mathcal{A}^{n_1\times n_1}$ so that

$$A_{21} = U_{21} \begin{bmatrix} \Sigma_{21} & 0 \\ 0 & 0 \end{bmatrix} V_{21}^T \in \mathcal{A}^{(n-n_1)\times n_1} \quad \text{with invertible } \Sigma_{21} \in \mathcal{A}^{n_2\times n_2},$$

and set

$$\begin{aligned} P_1 &:= \begin{bmatrix} V_{21}^T & 0 \\ 0 & U_{21}^T \end{bmatrix}, \\ A_1 &:= P_1 A_0 P_1^T + \dot{P}_1 P_1^T = \left[\begin{array}{cc|c} * & * & * \\ \Sigma_{21} & 0 & * \\ 0 & 0 & * \end{array}\right] + \left[\begin{array}{c|c} \dot{V}_{21}^T V_{21} & 0 \\ \hline 0 & \dot{U}_{21}^T U_{21} \end{array}\right], \\ B_1 &:= V_{21}^T \Sigma_B, \\ \widetilde{B}_1 &:= \begin{bmatrix} B_1 & 0 \\ 0 & 0 \end{bmatrix} \in \mathcal{A}^{n\times n}. \end{aligned}$$

Using the product rule, for some $\tilde{A}_{32} \in \mathcal{A}^{(n-n_1-n_2)\times n_2}$ this gives

$$P_1 \left[\ DI_n - A_0, -B_0\ \right] \begin{bmatrix} P_1^T & 0 \\ 0 & I_m \end{bmatrix} = \left[\ DI_n - A_1, -\widetilde{B}_1\ \right]$$

$$= \left[\begin{array}{c|cc} DI_n - \left[\begin{array}{c|cc} * & * & * \\ \hline [\Sigma_{21}, 0] & * & * \\ 0 & \tilde{A}_{32} & * \end{array}\right], & - \left[\begin{array}{c|c} B_1 & 0 \\ \hline 0 & 0 \\ 0 & 0 \end{array}\right] \end{array}\right].$$

Step 2. If $n_1 + n_2 < n$ and $\tilde{A}_{32} \neq 0$, then choose orthogonal matrices $U_{32} \in \mathcal{A}^{(n-n_1-n_2)\times(n-n_1-n_2)}$, $V_{32}^T \in \mathcal{A}^{n_2\times n_2}$ so that

$$\tilde{A}_{32} = U_{32} \begin{bmatrix} \Sigma_{32} & 0 \\ 0 & 0 \end{bmatrix} V_{32}^T \in \mathcal{A}^{(n-n_1-n_2)\times n_2} \quad \text{with invertible } \Sigma_{32} \in \mathcal{A}^{n_3\times n_3},$$

and set

$$\begin{aligned} P_2 &:= \operatorname{diag}\left\{I_{n_1}, V_{32}^T, U_{32}^T\right\}, \\ \hat{A}_{21} &:= V_{32}^T \Sigma_{21}, \\ A_2 &:= P_2 A_1 P_2^T + \dot{P}_2 P_2^T \\ &= \left[\begin{array}{c|cc} * & * & * \\ \hline V_{32}^T[\Sigma_{21}, 0] & * & * \\ 0 & U_{32}^T \tilde{A}_{32} V_{32} & * \end{array}\right] + \left[\begin{array}{c|cc} 0 & 0 & 0 \\ \hline 0 & \dot{V}_{32}^T V_{32} & 0 \\ 0 & 0 & \dot{U}_{32}^T U_{32} \end{array}\right] \\ &= \left[\begin{array}{cc|cc|c} * & * & * & * & * \\ \hline \hat{A}_{21} & 0 & * & * & * \\ \hline 0 & 0 & \Sigma_{32} & 0 & * \\ 0 & 0 & 0 & 0 & * \end{array}\right]. \end{aligned}$$

Then, for some $\tilde{A}_{43} \in \mathcal{A}^{(n-n_1-n_2-n_3)\times n_3}$,

$$\begin{aligned} P_2 P_1 &\left[\ DI_n - A_0, -B_0\ \right] \begin{bmatrix} P_1^T P_2^T & 0 \\ 0 & I_m \end{bmatrix} \\ &= \left[\ DI_n - A_2, -\widetilde{B}_1\ \right] \\ &= \left[\begin{array}{c|cc} DI_n - \left[\begin{array}{cc|cc|cc} * & * & * & * & * & * \\ \hline \hat{A}_{21} & 0 & * & * & * & * \\ \hline 0 & 0 & \Sigma_{32} & 0 & * & * \\ 0 & 0 & 0 & 0 & \tilde{A}_{43} & * \end{array}\right], & - \left[\begin{array}{c|c} B_1 & 0 \\ \hline 0 & 0 \\ \hline 0 & 0 \\ 0 & 0 \end{array}\right] \end{array}\right]. \end{aligned}$$

Step 3. In the remainder of the proof we proceed analogously as in Step 2 and terminate after finitely many steps with the form (2.11). This completes the proof. □

Example 2. As an example consider the model of a two-dimensional, three-link constrained mobile manipulator studied in [10]; see also [12]. This model leads, after linearization along a trajectory, to a system of the form

$$\begin{aligned} M_0(t)\,\ddot{z}(t) + D_0(t)\,\dot{z}(t) + K_0(t)\,z(t) &= S_0\,u(t) + F_0^T\,\mu(t), \\ F_0\,z(t) &= 0\,, \end{aligned} \tag{2.12}$$

where $M_0, D_0, K_0 \in \mathcal{C}^\omega(\mathbb{I}, \mathbb{R}^{3\times 3})$ and $S_0, F_0^T \in \mathbb{R}^{3\times 2}$ with S_0 having full rank. Introducing the eight-dimensional variable $x(t) = [z(t)^T, \dot{z}(t)^T, \mu(t)^T]^T$ results in the equivalent descriptor system description (1.1) with $F \equiv 0$,

$$E(t) = \begin{bmatrix} I_3 & 0 & 0 \\ 0 & M_0(t) & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A(t) = \begin{bmatrix} 0 & I_3 & 0 \\ -K_0(t) & -D_0(t) & F_0^T \\ F_0 & 0 & 0 \end{bmatrix}, \quad B \equiv \begin{bmatrix} 0 \\ S_0 \\ 0 \end{bmatrix}, \tag{2.13}$$

and the specification of C is left open for the time being.

The critical points of (2.13) include those values of t where the mass matrix $M_0(t)$ changes rank. This happens, for example, when two arms of the manipulator are in one straight line.

Without loss of generality (by using an appropriate permutation of the basis), we may assume that the coordinate system for the Lagrange multipliers is such that $F_0 = [F_1\ 0]$ with nonsingular $F_1 \in \mathbb{R}^{2\times 2}$ and if we partition

$$\begin{aligned} -K_0 &= \begin{bmatrix} K_{11}(t) & K_{12}(t) \\ K_{21}(t) & K_{22}(t) \end{bmatrix}, \ M_0 = \begin{bmatrix} M_{11}(t) & M_{12}(t) \\ M_{21}(t) & M_{22}(t) \end{bmatrix}, \\ -D_0 &= \begin{bmatrix} D_{11}(t) & D_{12}(t) \\ D_{21}(t) & D_{22}(t) \end{bmatrix}, \ S_0 = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix}, \end{aligned}$$

with $K_{11}(t), M_{11}(t), D_{11}(t), S_1 \in \mathbb{R}^{2\times 2}$ and all other formats accordingly, then system (2.13) may be written as

$$\begin{bmatrix} I_2 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & M_{11}(t) & M_{12}(t) & 0 \\ 0 & 0 & M_{21}(t) & M_{22}(t) & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \\ \dot{x}_5 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 & I_2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ K_{11}(t) & K_{12}(t) & D_{11}(t) & D_{12}(t) & F_1^T \\ K_{21}(t) & K_{22}(t) & D_{21}(t) & D_{22}(t) & 0 \\ F_1 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ S_1 \\ S_2 \\ 0 \end{bmatrix} u.$$

Since F_1 is constant and nonsingular, we obtain $x_1 = 0$ and $\dot{x}_1 = 0$. Inserting this

and changing the order of equations and blocks leads to

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & M_{11}(t) & M_{12}(t) & 0 & 0 \\ 0 & M_{21}(t) & M_{22}(t) & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \\ \dot{x}_5 \\ \dot{x}_1 \end{bmatrix}$$
$$= \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ K_{12}(t) & D_{11}(t) & D_{12}(t) & F_1^T & 0 \\ K_{22}(t) & D_{21}(t) & D_{22}(t) & 0 & 0 \\ 0 & 0 & 0 & 0 & F_1 \\ 0 & I_2 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_1 \end{bmatrix} + \begin{bmatrix} 0 \\ S_1 \\ S_2 \\ 0 \\ 0 \end{bmatrix} u.$$

We can repeat the reduction process once more by using that $x_3 = 0$, and hence $\dot{x}_3 = 0$, which gives a system

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & M_{22}(t) & 0 & 0 & 0 \\ 0 & M_{12}(t) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_2 \\ \dot{x}_4 \\ \dot{x}_3 \\ \dot{x}_5 \\ \dot{x}_1 \end{bmatrix}$$
$$= \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ K_{22}(t) & D_{22}(t) & 0 & 0 & 0 \\ K_{12}(t) & D_{12}(t) & 0 & F_1^T & 0 \\ 0 & 0 & 0 & 0 & F_1 \\ 0 & 0 & I_2 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_2 \\ x_4 \\ x_3 \\ x_5 \\ x_1 \end{bmatrix} + \begin{bmatrix} 0 \\ S_2 \\ S_1 \\ 0 \\ 0 \end{bmatrix} u.$$

Since the mass matrix M_0 is positive definite almost everywhere, we can eliminate the block M_{12} and obtain the system

$$\left[\begin{array}{cc|ccc} 1 & 0 & 0 & 0 & 0 \\ 0 & M_{22}(t) & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right] \left[\begin{array}{c} \dot{x}_2 \\ \dot{x}_4 \\ \hline \dot{x}_3 \\ \dot{x}_5 \\ \dot{x}_1 \end{array}\right]$$
$$= \left[\begin{array}{cc|ccc} 0 & 1 & 0 & 0 & 0 \\ K_{22}(t) & D_{22}(t) & 0 & 0 & 0 \\ \hline \tilde{K}_{12}(t) & \tilde{D}_{12}(t) & 0 & F_1^T & 0 \\ 0 & 0 & 0 & 0 & F_1 \\ 0 & 0 & I_2 & 0 & 0 \end{array}\right] \left[\begin{array}{c} x_2 \\ x_4 \\ \hline x_3 \\ x_5 \\ x_1 \end{array}\right] + \left[\begin{array}{c} 0 \\ S_2 \\ \hline \tilde{S}_1 \\ 0 \\ 0 \end{array}\right] u. \tag{2.14}$$

This system is essentially (apart from diagonal matrices Σ) in the condensed form (2.2), with

$$\Sigma_d = \begin{bmatrix} 1 & 0 \\ 0 & M_{22}(t) \end{bmatrix}, \quad \Sigma_a = \begin{bmatrix} 0 & F_1^T & 0 \\ 0 & 0 & F_1 \\ I_2 & 0 & 0 \end{bmatrix}, \quad B_2 = \begin{bmatrix} \tilde{S}_1 \\ 0 \\ 0 \end{bmatrix}.$$

It is then obvious how the more refined forms (2.4) and (2.6) can be determined. ∎

3. Controllability. In this section we discuss the concept of controllability for descriptor systems of the form (1.2). Recall that (local) controllability for general systems of the form $R(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$, where $R(D) \in \mathcal{M}[D]^{g\times q}$, is introduced in [12, Def. 3.1] and discussed in [12, Rem. 3.2].

Remark 2. For descriptor systems with constant coefficients, several different controllability concepts have been introduced; see [3, 7, 19].

(i) System (1.1) with constant coefficients is called

R-controllable	iff	$\operatorname{rk}[\lambda E - A, B]$ is full for all $\lambda \in \mathbb{C}$,
I-controllable	iff	$\operatorname{rk}[E, AS_\infty, B]$ is full, where S_∞ spans the kernel of E,
strongly controllable	iff	the system is R-controllable and I-controllable.

We stress that these algebraic characterizations are sometimes misleading in the literature, since it is sometimes assumed that the rank of $[E, B]$ is full and sometimes not.

It follows that if system (1.1) is square and time-invariant (thus, in particular, $l = n$), then system (1.1) is *I-controllable* if and only if $n-(d+a+\gamma+f) = 0$. The constants a, d, f, γ are defined in Theorem 2.1(ii). I-controllability is related to regularization and index reduction; i.e., in particular it is needed to avoid impulsive solutions in the case of nondifferentiable input functions. In our framework this concept is not relevant.

(ii) If the descriptor system (1.2) is time-varying, then [12, Def. 3.1] is new; see [5, 22, 18] for a discussion of different controllability concepts for time-varying descriptor systems.

(iii) For time-invariant state-space systems, i.e. (1.1) with $E = I_n$, the algebraic conditions can be checked numerically via the Staircase Algorithm of [24]. In a similar fashion Lemma 2.2 may be used to check controllability for time-varying systems.

(iv) For time-invariant systems (1.2), [12, Def. 3.1] corresponds to the concept of R-controllability. This follows from Theorem 3.1 below. □

Theorem 2.1 and Lemma 2.2 put us in a position to characterize controllability of time-varying descriptor systems (1.2).

THEOREM 3.1. *Consider a time-varying descriptor system* (1.2) *and assume that* $R(D)$ *has full row rank over* $\mathcal{M}[D]$. *Consider the condensed form* (2.6) *and* σ *as defined in* (2.7). *Set, for notational convenience,*

$$G(t) := \tilde{A}_{11}(t), \quad S(t) := [\tilde{A}_{13}(t),\ \tilde{A}_{14}(t),\ \tilde{B}_{12}(t)], \quad v(t) := [x_3(t)^T,\ x_4(t)^T,\ u_2(t)^T]^T.$$

Then the following conditions are equivalent:

(i) (1.2) *is locally controllable almost everywhere.*
(ii) $R(D)$ *is right invertible over* $\mathcal{M}[D]$.
(iii) (2.3), *respectively* (2.4), *is locally controllable almost everywhere.*
(iv) $\hat{R}(D) := [\sigma DI_d - G,\ S]$ *is right invertible over* $\mathcal{M}[D]$.
(v) *In the staircase form* (2.11) *of the pair* $[DI_d - G,\ S]$, *the lower block is not present; i.e.,* $n_s = 0$.
(vi) *There exists a discrete set* $\mathbb{T} \subset \mathbb{R}$ *such that for every*

$$\begin{bmatrix} x_1^0 \\ v^0 \end{bmatrix}, \begin{bmatrix} x_1^1 \\ v^1 \end{bmatrix} \in \ ker_t \hat{R}$$

and for every open interval $\mathbb{I} \subset \mathbb{R}\setminus\mathbb{T}$ *and all* $t_0 \in \mathbb{I}$, *there exists* $t_1 > t_0,\ t_1 \in \mathbb{I}$, *and* $[x_1^T, v^T]^T \in \ ker_t\, \hat{R}$, *such that*

$$\begin{bmatrix} x_1(t) \\ v(t) \end{bmatrix} = \begin{cases} \begin{bmatrix} x_1^0(t) \\ v^0(t) \end{bmatrix} & \text{if } t \in (-\infty, t_0] \cap \mathbb{R} \setminus \mathbb{T}, \\ \begin{bmatrix} x_1^1(t) \\ v^1(t) \end{bmatrix} & \text{if } t \in [t_1, \infty) \cap \mathbb{R} \setminus \mathbb{T}. \end{cases}$$

Proof.

(i) $\Leftrightarrow$ (ii): This is proved in [12, Prop. 3.6].

(ii) $\Leftrightarrow$ (iii): The equivalence of local controllability almost everywhere of (1.2) and (2.4), respectively (1.1) and (2.6), follows from (2.3) by invoking orthogonality of U_2, V_2, W, Z.

(ii) $\Leftrightarrow$ (iv): By (2.5), there exist invertible matrices $\tilde{U} \in \mathcal{M}^{(l+p)\times(l+p)}$, $\tilde{V} \in \mathcal{M}^{(n+m+p)\times(n+m+p)}$ so that (1.1) is related to (2.6) in the form (1.2) by the transformation

$$\tilde{U} \begin{bmatrix} E & D-A & -B & 0 \\ & -C & -F & I_p \end{bmatrix} \tilde{V}$$

$$= \begin{bmatrix} \sigma DI_d - \tilde{A}_{11} & 0 & -\tilde{A}_{13} & -\tilde{A}_{14} & 0 & 0 & -\tilde{B}_{12} & 0 & 0 \\ 0 & -\Sigma_a & 0 & 0 & 0 & 0 & -\tilde{B}_{22} & 0 & 0 \\ -\tilde{A}_{31} & 0 & 0 & 0 & 0 & -\Sigma_\gamma & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\Sigma_f & 0 & 0 & 0 & 0 \\ 0_{(l-\nu)\times d} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\sigma\Sigma_\omega & 0 & 0 & 0 & -\tilde{F}_{12} & \sigma I_\omega & 0 \\ -\tilde{C}_{21} & 0 & 0 & 0 & 0 & 0 & -\tilde{F}_{22} & 0 & \sigma I_{p-\omega} \end{bmatrix}. \tag{3.1}$$

The right-hand side is right invertible if and only if $l - \nu = l - d - a - \gamma - f = 0$ (which is a consequence of the full row rank assumption) and $[\sigma DI_d - G,\, S]$ is right invertible over $\mathcal{M}[D]$.

(iv) $\Leftrightarrow$ (v): By Lemma 2.2 there exist orthogonal matrices P and Q so that $P[\sigma DI_d - G, S]\left[\begin{smallmatrix} P^T & 0 \\ 0 & Q \end{smallmatrix}\right]$ is of the staircase form (2.11). Note that σ does not affect the staircase form. Now the equivalence (iv) $\Leftrightarrow$ (v) follows immediately since B_1 and $\hat{A}_{i,i-1}$ are invertible over $\mathcal{M}$ for $i = 1, \ldots, s-1$.

(ii) $\Leftrightarrow$ (vi): This equivalence follows readily from [12, Def. 3.1] and from (3.1), since the set of zeros and poles of the coefficients of $\tilde{U}$ and $\tilde{V}$ is a discrete set. ∎

Note that the assumption that (1.2) has full row rank over $\mathcal{M}[D]$ is equivalent to $l - d - a - \gamma - f = 0$ in (2.6).

Note further that the characterization in Theorem 3.1(ii) does not require a reinterpretation of variables as is done in [4]. Moreover, in contrast to the case of controllability of state space systems, here $u_1(\cdot)$ in (1.1) is not a "free input" variable.

For standard time-invariant state space systems (i.e., $E = I_n$), the right invertibility of $R(D)$ in Theorem 3.1 is derived differently in [21, Thm. 5.2.10].

Remark 3. For time-varying systems (1.1) with $E = I_n$, i.e., state space systems, it is well known that controllability of the system yields that it can be controlled in an arbitrary short time. The interval $\mathbb{I}$ in [12, Def. 3.1] can be replaced by any arbitrary short open interval $\widehat{\mathbb{I}} \subset \mathbb{I}$. This also holds true for descriptor systems (1.2), since $\hat{R}(D)$ in Theorem 3.1(iv) can be viewed locally as a state space system, namely, at those $t \in \mathbb{R}$ where $\sigma(t) \neq 0$; note that the zeros of σ are a discrete set. An alternative and constructive proof is given in [12, Thm. 3.3] for general systems of the form $R(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$. ∎

Example 3.

(i) $R(D) = [t^2D + 1, 1]$ has right inverse $[0, 1]^T$, and hence, by Theorem 3.1, $R(\frac{\mathrm{d}}{\mathrm{d}t})w = 0$ is controllable.

(ii) Revisit the linearized model (2.13) of the three-link constrained mobile manipulator. In Example 2 it is shown that (2.13) is equivalent to (2.11). Rewriting (2.11) in the form (1.2) and invoking that M_{22} is invertible over $\mathcal{M}$ and F_1 is nonsingular, it is easy to see that the corresponding $R(D)$ is right invertible. Therefore, by Theorem 3.1, the linearized model (2.13) is controllable. ∎

4. Observability and autonomous behavior. In [12, sect. 4], the concept of autonomous behavior $\ker_t^{\mathrm{aut}} R$ also has been introduced, and it has been shown that the behavior of a system (1.2) (and hence also of (1.1)) can be decomposed into the direct sum of a controllable and an autonomous behavior. It also immediately follows from the results in [12] that an autonomous behavior $\ker_t^{\mathrm{aut}} R$ of the system (1.2) is invariant under all transformations (2.1), (2.2), (2.4), (2.10). Loosely speaking, an autonomous behavior consists of those solutions which are uniquely determined if they are known on an arbitrarily small open interval. For systems (1.2) we have to cope with the problem of finite escape time.

Example 4. Consider a time-varying state space system (1.2) with $E = I_n$. By [14] there exists $T \in \mathcal{A}^{n\times n}$ invertible over $\mathcal{A}$ so that the coordinate transformation $z := T^{-1}x$ converts (1.1) into

$$\begin{array}{rcl} \frac{\mathrm{d}}{\mathrm{d}t}z_1(t) &=& A_{11}(t)z_1(t) + A_{12}(t)z_2(t) + B_1(t)u(t), \\ \frac{\mathrm{d}}{\mathrm{d}t}z_2(t) &=& A_{22}(t)z_2(t), \\ y(t) &=& C_1(t)z_1(t) + C_2(t)z_2(t) + F(t)u(t), \end{array} \tag{4.1}$$

with all matrices real analytic of conforming formats, and controllable subsystem $\frac{\mathrm{d}}{\mathrm{d}t}z_1(t) = A_{11}(t)z_1(t) + B_1(t)u(t)$. Since (4.1) is a state space system, finite escape time does not occur and the controllable and autonomous subspaces can be described globally. Set

$$\hat{R}(D) := \begin{bmatrix} DI - A_{11} & -A_{12} & -B_1 & 0 \\ 0 & DI - A_{22} & 0 & 0 \\ -C_1 & -C_2 & -F & -I_p \end{bmatrix}.$$

Then, for all $t \in \mathbb{R}$,

$$\ker_t^{\mathrm{contr}} \hat{R} = \left\{ w = [z_1^T, z_2^T, u^T, y^T]^T \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^{(n+m+p)}) \;\middle|\; \hat{R}(\tfrac{\mathrm{d}}{\mathrm{d}t})w = 0 \ \wedge \ z_2 = 0 \right\}$$

and

$$\ker_t^{\mathrm{aut}} \hat{R} = \left\{ w = [z_1^T, z_2^T, u^T, y^T]^T \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^{(n+m+p)}) \;\middle|\; \begin{array}{l} \hat{R}(\frac{\mathrm{d}}{\mathrm{d}t})w = 0,\ z_1 = 0, \\ u = 0,\ \dot{z}_2 = A_{22}\, z_2 \end{array} \right\}$$

is an autonomous behavior, and, hence, in the original coordinates, we have

$$\ker_t^{\mathrm{contr}} R = \begin{bmatrix} T(t) & 0 \\ 0 & I_{m+p} \end{bmatrix} \ker_t^{\mathrm{aut}} R \oplus \begin{bmatrix} T(t) & 0 \\ 0 & I_{m+p} \end{bmatrix} \ker_t^{\mathrm{contr}} \hat{R} \qquad \forall\, t \in \mathbb{R}.$$

Remark 4. Consider a time-varying descriptor system (1.1) in the condensed form (2.6). If (2.6) were controllable, then $\ker^{\mathrm{aut}} R = \{0\}$ would be the only autonomous

behavior of (2.6). To see this, note that x_3, x_4, u_2 are free to choose and hence cannot be a nonzero component of an autonomous behavior. Furthermore, since $[\sigma DI_d - G, S]$ is controllable by Theorem 3.1(iii), it follows that x_1 is uniquely (modulo initial condition) determined by x_3, x_4, u_2, and hence also not a nontrivial component of an autonomous behavior. Finally, (2.6) yields that the remaining components x_2, x_5, u_1, y_1, y_2 are uniquely determined by x_3, x_4, u_2, x_1. This shows $\ker^{\mathrm{aut}} R = \{0\}$.

If (2.6) is not controllable but has a nontrivial uncontrollable subspace, then there exists $\ker^{\mathrm{aut}} R \neq \{0\}$ which is determined by the uncontrollable subspace as for state space systems; see Example 4. □

In [12] it has also been discussed how one behavior can be observed from another. We refer to this paper for the definition of adjoints and observable behavior which generalize well-known concepts of observability, such as for time-varying state space systems (see, for example, [23]) and time-varying Rosenbrock systems (see [13]). It has also been shown that local observability and local controllability are dual concepts.

An application of [12, Thms. 5.5 and 5.6] to descriptor systems (1.2) yields the following result.

THEOREM 4.1. *Consider a descriptor system* (1.2) *with* $R(D) = [R_1(D), R_2(D)]$ *partitioned as*

$$R_1(D) = \begin{bmatrix} E\,D - A \\ -C \end{bmatrix}, \qquad R_2(D) = \begin{bmatrix} -B & 0 \\ -F & I_p \end{bmatrix}.$$

Then the following are equivalent:

(i) *The trajectory x is locally observable from (u, y) almost everywhere.*
(ii) *$R_1(D)$ is left invertible over $\mathcal{M}[D]$.*
(iii) *The matrix*

$$\begin{bmatrix} \sigma\, DI_d - \tilde{A}_{11} & -\tilde{A}_{13} & -\tilde{A}_{14} \\ -\tilde{A}_{31} & 0 & 0 \\ 0 & -\Sigma_\omega & 0 \\ -\sigma^{-1}\tilde{C}_{21} & 0 & 0 \end{bmatrix} \tag{4.2}$$

is left invertible over $\mathcal{M}[D]$, where the matrices in (4.2) *are from the condensed form* (2.4).

Proof. The equivalence (i)↔ (ii) follows from [12, Thms. 5.5 and 5.6]. To see (ii)↔ (iii), note that left invertibility of $R_2(D)$ is equivalent to

$$\begin{bmatrix} U & 0 \\ X & W \end{bmatrix} R_2(D)V = \begin{bmatrix} \sigma\, DI_d - \tilde{A}_{11} & 0 & -\tilde{A}_{13} & -\tilde{A}_{14} & 0 \\ 0 & -\Sigma_a & 0 & 0 & 0 \\ -\tilde{A}_{31} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\Sigma_f \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\Sigma_\omega & 0 & 0 \\ -\sigma^{-1}\tilde{C}_{21} & 0 & 0 & 0 & 0 \end{bmatrix}$$

being left invertible, where U, X, V, W are specified in Theorem 2.1(iii). Since $\left[\begin{smallmatrix} U & 0 \\ X & W \end{smallmatrix}\right]$ and V are invertible over $\mathcal{M}$, the latter holds true if and only if (4.2) is left invertible. This completes the proof. □

Example 5. Consider again the linearized model (2.13) of the three-link constrained mobile manipulator. Suppose that the positions can be measured, corre-

sponding to the additional equation

$$y = \begin{bmatrix} 0 & 0 & I_2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix}. \tag{4.3}$$

In Example 2 we have shown that $x_1 = 0$ and $x_3 = 0$ and thus $\dot{x}_1 = 0$ and $\dot{x}_3 = 0$, and permuting the variables accordingly to (2.14), we obtain

$$y = \left[\begin{array}{cc|ccc} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{array}\right] \begin{bmatrix} x_2 \\ x_4 \\ \hline x_3 \\ x_5 \\ x_1 \end{bmatrix}. \tag{4.4}$$

Hence by Theorem 4.1, x is observable from (u, y) with respect to the system (2.13), (4.3) or, equivalently, system (2.11), (4.4) if and only if

$$\left[\begin{array}{cc|ccc} D-1 & 0 & 0 & 0 & 0 \\ -K_{22} & M_{22}D - D_{22} & 0 & 0 & 0 \\ \hline -\tilde{K}_{12} & -\tilde{D}_{12} & 0 & -F_1^T & 0 \\ 0 & 0 & 0 & 0 & F_1 \\ 0 & 0 & -I_2 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right] \tag{4.5}$$

is left invertible over $\mathcal{M}[D]$. Since F_1 is invertible over $\mathcal{M}$, (4.5) is left invertible if and only if $\left[\begin{smallmatrix} D-1 & 0 \\ -K_{22} & M_{22}D - D_{22} \end{smallmatrix}\right]$ is invertible over $\mathcal{M}[D]$. Summarizing, x is observable from (u, v) almost everywhere if and only if K_{22} is invertible over $\mathcal{M}$. ∎

5. Conclusion. We have introduced a general behavioral approach to linear descriptor systems with real analytic coefficients. We have characterized autonomous, controllable, and observable behavior and have generalized results on time-varying ordinary differential equations and on time-invariant linear algebraic-differential equations. The results have been illustrated by several examples, which demonstrates that the approach also helps in understanding practical problems such as constrained multibody systems.

REFERENCES

[1] K. E. Brenan, S. L. Campbell, and L. R. Petzold, *Numerical Solution of Initial-Value Problems in Differential Algebraic Equations*, 2nd ed., Elsevier–North Holland, New York, 1996.

[2] A. Bunse-Gerstner, R. Byers, V. Mehrmann, and N. K. Nichols, *Numerical computation of an analytic singular value decomposition of a matrix valued function*, Numer. Math., 60 (1991), pp. 1–40.

[3] A. Bunse-Gerstner, R. Byers, V. Mehrmann, and N. K. Nichols, *Feedback design for regularizing descriptor systems*, Linear Algebra Appl., 299 (1999), pp. 119–151.

[4] R. Byers, P. Kunkel, and V. Mehrmann, *Regularization of linear descriptor systems with variable coefficients*, SIAM J. Control Optim., 35 (1997), pp. 117–133.

[5] S. L. Campbell, N. K. Nichols, and W. J. Terrell, *Duality, observability, and controllability for linear time-varying descriptor systems*, Circuits Systems Signal Process., 10 (1991), pp. 455–470.
[6] P. M. Cohn, *Free Rings and Their Relations*, London Math. Soc. Monogr. (NS) 2, Academic Press, London, New York, 1971.
[7] L. Dai, *Singular Control Systems*, Springer-Verlag, Berlin, 1989.
[8] E. Griepentrog and R. März, *Differential-Algebraic Equations and Their Numerical Treatment*, Teubner-Texte Math. 88, Teubner Verlag, Leipzig, Germany, 1986.
[9] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations*. II: *Stiff and Differential-Algebraic Problems*, Springer Ser. Comput. Math. 14, 2nd ed., Springer-Verlag, Berlin, 1996.
[10] M. Hou and P. C. Müller, *LQ and Tracking Control of Descriptor Systems with Application to Constrained Manipulator*, Tech. report, Sicherheitstechnische Regelungs- und Meßtechnik, Universität Wuppertal, Wuppertal, Germany, 1994.
[11] A. Ilchmann, Y. Kuang, M. Kuijper, and C. Zhang, *Continuous time-varying scalar systems—A behavioral approach*, in Proceedings of the Third Asian Control Conference, Shanghai, China, 2000, pp. 429–433.
[12] A. Ilchmann and V. Mehrmann, *A behavioral approach to time-varying linear systems. Part 1: General theory*, SIAM J. Control Optim., 44 (2005), pp. 1725–1747.
[13] A. Ilchmann, I. Nürnberger, and W. Schmale, *Time-varying polynomial matrix systems*, Internat. J. Control, 40 (1984), pp. 329–362.
[14] R. E. Kalman, *Canonical structure of linear dynamical systems*, Proc. Natl. Acad. Sci. USA, 48 (1962), pp. 596–600.
[15] P. Kunkel and V. Mehrmann, *A new look at pencils of matrix valued functions*, Linear Algebra Appl., 212/213 (1994), pp. 215–248.
[16] P. Kunkel and V. Mehrmann, *Analysis of over- and underdetermined nonlinear differential-algebraic systems with application to nonlinear control problems*, Math. Control Signals Systems, 14 (2001), pp. 233–256.
[17] P. Kunkel and V. Mehrmann, *Differential-Algebraic Equations—Analysis and Numerical Solution*, EMS Publishing House, Zürich, Switzerland, to appear.
[18] P. Kunkel, V. Mehrmann, and W. Rath, *Analysis and numerical solution of control problems in descriptor form*, Math. Control Signals Systems, 14 (2001), pp. 29–61.
[19] V. Mehrmann, *The Autonomous Linear Quadratic Control Problem: Theory and Numerical Algorithms*, Lecture Notes in Control and Inform. Sci. 163, Springer-Verlag, Berlin, 1991.
[20] V. Mehrmann and W. Rath, *Numerical methods for the computation of analytic singular value decompositions*, Electron Trans. Numer. Anal., 1 (1993), pp. 72–88.
[21] J. W. Polderman and J. C. Willems, *Introduction to Mathematical Systems Theory: A Behavioral Approach*, Texts Appl. Math. 26, Springer-Verlag, New York, 1998.
[22] W. Rath, *Feedback Design and Regularization for Linear Descriptor Systems with Variable Coefficients*, Ph.D. thesis, TU Chemnitz-Zwickau, Shaker Verlag, Aachen, Germany, 1997.
[23] W. J. Rugh, *Linear System Theory*, Prentice–Hall, Englewood Cliffs, NJ, 1993.
[24] P. Van Dooren, *The computation of Kronecker's canonical form of a singular pencil*, Linear Algebra Appl., 27 (1979), pp. 103–121.
[25] J. C. Willems, *System theoretic models for the analysis of physical systems*, Ricerche Automat., 10 (1979), pp. 71–106.
[26] J. C. Willems, *From time series to linear system*. I. *Finite-dimensional linear time invariant systems*, Automatica J. IFAC, 22 (1986), pp. 561–580.
[27] J. C. Willems, *From time series to linear system*. II. *Exact modelling*, Automatica J. IFAC, 22 (1986), pp. 675–694.
[28] J. C. Willems, *From time series to linear system*. III. *Approximate modelling*, Automatica J. IFAC, 23 (1987), pp. 87–115.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 5, pp. 1766–1786

INTERIOR POINT METHODS IN FUNCTION SPACE*

MARTIN WEISER†

Abstract. A primal-dual interior point method for optimal control problems is considered. The algorithm is directly applied to the infinite-dimensional problem. Existence and convergence of the central path are analyzed, and linear convergence of a short-step path-following method is established.

Key words. interior point methods in function space, optimal control, complementarity functions

AMS subject classifications. 49M15, 90C48, 90C51

DOI. 10.1137/S0363012903437277

1. Introduction.

Numerical methods for solving optimal control problems governed by ODEs fall into two categories, the indirect methods [2, 3, 4, 6, 14, 15, 31] relying on Pontryagin's maximum principle, and the direct methods [7, 17, 21, 30, 37] based on the Karush–Kuhn–Tucker necessary conditions. Direct methods can be characterized by several features. Among them are the following:

(i) Position of discretization: Discretize-then-optimize approaches use an a priori parameterization of the control and possibly the state variables to reduce the optimal control problem to a finite-dimensional nonlinear program. These large nonlinear programs can then be solved by standard NLP solvers. Adaptive mesh refinement can be performed after the finite-dimensional optimum has been reached. On the other hand, optimize-then-discretize approaches formulate the optimization algorithms directly in the infinite-dimensional function space, employing discretization only for solving linear operator equations. Adaptive mesh refinement is used to meet the accuracy requirements imposed on the solution of the linear equations by the optimization algorithm.
Somewhere in between are function space sequential quadratic programming (SQP) methods where linear-quadratic programs are discretized.

(ii) Type of optimization algorithm: Among the most popular algorithms employed for solving the optimization problems arising in optimal control are SQP and interior point methods. A recent alternative are semismooth Newton methods [5, 34].

Discretize-then-optimize methods are covered by a vast amount of published literature using almost any available algorithm for solving the finite-dimensional NLPs. Solutions on consecutive mesh refinement levels or in consecutive SQP steps often exhibit pronounced similarities. This redundancy can be directly exploited by active set–type methods. In contrast, interior point methods are considered to benefit less from this redundancy [20, 40]. Nevertheless, interior point methods are reported to be very efficient for solving optimal control problems—a fact that is not well explained by straightforward application of finite-dimensional interior point convergence theory to

*Received by the editors November 5, 2003; accepted for publication (in revised form) April 14, 2005; published electronically November 23, 2005. This work was supported by the DFG Research Center MATHEON in Berlin. A preprint of this paper appeared as ZIB Report 03-35.
http://www.siam.org/journals/sicon/44-5/43727.html
†Zuse Institute Berlin, Takustr. 7, 14195 Berlin, Germany (weiser@zib.de).

the discretized problems. The best currently known convergence rates of $1-\text{const}/\sqrt{n}$ would instead predict a pronounced mesh dependence of the convergence.

Among the optimize-then-discretize approaches, the SQP methods dominate the published material [1, 17, 22, 23, 27, 32, 33]. Here, Robinson's theory of generalized equations [29] can be used to analyze the function space methods, which leaves, however, the question of how to solve the infinite-dimensional linear-quadratic programs. This is implicitly addressed by infinite-dimensional interior point methods, which have nevertheless attracted less attention [35, 36, 24].

The present paper presents an infinite-dimensional interior point method directly applied to optimal control problems in function space in section 2. Existence and convergence of the central path are analyzed in section 3. Finally, linear convergence of a theoretical short-step path-following algorithm with classical predictor is shown in section 4. In particular, the rate of convergence does not depend on the size of any discretization.

Notation. The Lebesgue spaces and Sobolev spaces of functions with values in $\mathbb{R}^n$ are denoted by L_p^n and $(W_p^m)^n$, respectively. $S(x,\rho)$ is the open ball around x with radius ρ.

Some variables and operators are constructed such that they have a natural block partitioning corresponding to the components u and y of x. The individual blocks are denoted by the corresponding component as a superscript, e.g.,

$$g(x) = \begin{bmatrix} g^u(u) \\ g^y(y) \end{bmatrix} \quad \text{and} \quad \Psi(g(x),\eta) = \begin{bmatrix} \Psi^u(g^u(u),\eta^u) \\ \Psi^y(g^y(y),\eta^y) \end{bmatrix}.$$

2. Problem setting. On the time interval $\Omega = [0,1]$ we consider the optimal control problem

$$\min J(x) \quad \text{subject to} \quad \begin{aligned} c(x) &= 0 \quad \text{a.e.}, \\ r(x) &= 0, \\ g(x) &\geq 0 \quad \text{a.e.} \end{aligned} \tag{2.1}$$

with a partitioning of the variable $x = (u,y) \in X = L_\infty^{n_u}(\Omega) \times (W_\infty^1)^{n_y}(\Omega)$ into controls and states, a Lagrange-type cost functional

$$J(x) = \int_0^1 \tilde{f}(u(t),y(t))\,dt,$$

ordinary differential equations with boundary conditions

$$c(x) = \begin{bmatrix} \bar{c}(x) \\ y(0) - y_0 \end{bmatrix}, \quad \bar{c}(x)(t) = \tilde{c}(x(t)) - \dot{y}(t), \tag{2.2}$$

$$r(x) = \tilde{r}(y(1)) \tag{2.3}$$

as equality constraints, and pointwise state and control constraints

$$g(x)(t) = \begin{bmatrix} \tilde{g}^u(u(t)) \\ \tilde{g}^y(y(t)) \end{bmatrix}.$$

For the whole paper, we will restrict the discussion to the fixed time interval Ω and, hence, simplify the notation by omitting it from the function spaces. We assume all the functions $\tilde{f} : \mathbb{R}^{n_u} \times \mathbb{R}^{n_y} \to \mathbb{R}$, $\tilde{c} : \mathbb{R}^{n_u} \times \mathbb{R}^{n_y} \to \mathbb{R}^{n_y}$, $\tilde{r} : \mathbb{R}^{n_y} \to \mathbb{R}^{n_r}$, $\tilde{g}^u : \mathbb{R}^{n_u} \to \mathbb{R}^{n_{\eta_u}}$, and $\tilde{g}^y : \mathbb{R}^{n_y} \to \mathbb{R}^{n_{\eta_y}}$ to be twice Lipschitz-continuously differentiable.

For convenience, we give here a theorem on Nemyckii operators in L_∞, the straightforward proof of which can be found in [38].

THEOREM 2.1. *If $f : \mathbb{R}^n \to \mathbb{R}^m$ is k times differentiable and its kth derivative satisfies the Lipschitz condition*

$$|f^{(k)}(x) - f^{(k)}(y)| \le \kappa |x - y|, \tag{2.4}$$

the corresponding Nemyckii operator $\mathbf{f}$ defined by $\mathbf{f}(u)(t) = f(u(t))$ maps L^n_∞ into L^m_∞ and is k times Fréchet differentiable. For $1 \le p \le \infty$ its kth derivative can be continuously extended to an operator $\mathbf{f}^{(k)}(u) : (\prod_{j=1}^k L^n_{pk}) \to L^m_p$ that inherits boundedness and Lipschitz continuity from $f^{(k)}$:

$$\left\|\mathbf{f}^{(k)}(u)\right\|_{(\prod_{j=1}^k L^n_{pk}) \to L^m_p} \le \sup_{|x| \le \|u\|_{L^n_\infty}} |f^{(k)}(x)|, \tag{2.5}$$

$$\left\|\mathbf{f}^{(k)}(u+\delta u) - \mathbf{f}^{(k)}(u)\right\|_{(\prod_{j=1}^k L^n_{pk}) \to L^m_p} \le \kappa \left\|\delta u\right\|_{L^n_\infty}. \tag{2.6}$$

If in addition f is $k+1$ times differentiable and its $k+1$st derivative satisfies the Lipschitz condition

$$|f^{(k+1)}(x) - f^{(k+1)}(y)| \le \kappa |x - y|,$$

then $\mathbf{f}$ maps $(W^1_\infty)^n$ into $(W^1_\infty)^m$ and is k times differentiable. For $p \ge 1$ its kth derivative can be continuously extended to an operator $\mathbf{f}^{(k)}(u) : (\prod_{j=1}^k (W^1_{pk})^n) \to (W^1_p)^m$ that inherits boundedness and Lipschitz continuity from $f^{(k)}$ and $f^{(k+1)}$:

$$\begin{aligned} \left\|\mathbf{f}^{(k)}(u)\right\|_{(\prod_{j=1}^k (W^1_{pk})^n) \to (W^1_p)^m} &\le \sup_{|x| \le \|u\|_{L^n_\infty}} (k+1)|f^{(k)}(x)| + |f^{(k+1)}(x)|, \\ \left\|\mathbf{f}^{(k)}(u+\delta u) - \mathbf{f}^{(k)}(u)\right\|_{(\prod_{j=1}^k (W^1_{pk})^n) \to (W^1_p)^m} &\le (k+2)\kappa \left\|\delta u\right\|_{(W^1_\infty)^n}. \end{aligned} \tag{2.7}$$

If the derivatives of $f : \mathbb{R}^n \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}^n$ commute, then so do the derivatives of the corresponding Nemyckii operators $\mathbf{f}'$ and $\mathbf{g}'$.

With Theorem 2.1 earlier, we conclude that

$$\begin{aligned} J &: L^{n_u}_\infty \times (W^1_\infty)^{n_y} \to \mathbb{R}, \\ c &: L^{n_u}_\infty \times (W^1_\infty)^{n_y} \to L^{n_y}_\infty, \quad \text{and} \\ g &: L^{n_u}_\infty \times (W^1_\infty)^{n_y} \to L^{n_{\eta u}}_\infty \times L^{n_{\eta y}}_\infty \end{aligned}$$

are twice Lipschitz-continuously differentiable operators.

The aim of the interior point method discussed here is to approximate Kuhn–Tucker points x_*. These are feasible points characterized by the existence of Lagrange multipliers $\lambda_c \in \mathbb{R}^{n_y} \times (L^{n_y}_\infty)^*$, $\lambda_r \in \mathbb{R}^{n_r}$, and $\eta \in (L^{n_{\eta u}}_\infty)^* \times ((W^1_\infty)^{n_{\eta y}})^*$ such that the following conditions are satisfied:

$$\begin{gathered} J'(x_*) - c'(x_*)^*\lambda_c - r'(x_*)^*\lambda_r - g'(x_*)^*\eta = 0, \\ c(x_*) = 0, \quad r(x_*) = 0, \\ g(x_*) \ge 0, \quad \eta \ge 0, \quad \langle \eta, g(x_*) \rangle = 0. \end{gathered} \tag{2.8}$$

Under certain assumptions (see, e.g., [26, 28]) these conditions are necessary for x_* to be a local solution of (2.1). Thus, Kuhn–Tucker points are promising candidates for solutions.

Unfortunately, the unwieldy complementarity condition (2.8) is difficult to handle numerically. The idea of primal-dual interior point methods is to relax the complementarity condition by

$$\eta \cdot g(x) = \mu, \quad \eta \geq 0, \quad g(x) \geq 0 \tag{2.9}$$

and to consider the homotopy $\mu \to 0$. Alternatively, complementarity functions $\psi(a, b; \mu) : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ can be used to construct Nemyckii operators Ψ such that

$$\Psi(g(x), \eta; \mu) = 0$$

is more or less equivalent to the classical interior point relaxation (2.9).

These relaxations, however, are only well defined if $\eta \in L_1$, and are continuously differentiable only in case $\eta \in L_\infty$. Note that this is required to hold only during the homotopy for $\mu > 0$, not at the Kuhn–Tucker point itself. We will prove in Theorem 3.4 that the homotopy can indeed be performed in the more regular setting of $\eta \in L_\infty^{n_{\eta u}} \times L_\infty^{n_{\eta y}} \subset (L_\infty^{n_{\eta u}})^* \times ((W_\infty^1)^{n_{\eta y}})^*$ for $\mu > 0$.

Define the Lagrangian

$$L(x, \lambda_c, \lambda_r, \eta) = J(x) - \langle \lambda_c, c(x) \rangle - \langle \lambda_r, r(x) \rangle - \langle \eta, g(x) \rangle.$$

Let

$$F(x, \lambda_c, \lambda_r, \eta; \mu) = \begin{bmatrix} \partial_x L(x, \lambda_c, \lambda_r, \eta) \\ -c(x) \\ -r(x) \\ \Psi(\eta, g(x); \mu) \end{bmatrix}. \tag{2.10}$$

As will be shown in Theorem 3.2 later, F maps

$$V \times \mathbb{R}_+ = (L_\infty^{n_u} \times (W_\infty^1)^{n_y}) \times (\mathbb{R}^{n_y} \times L_\infty^{n_y}) \times \mathbb{R}^{n_r} \times (L_\infty^{n_{\eta u}} \times L_\infty^{n_{\eta y}}) \times \mathbb{R}_+ \tag{2.11}$$

into

$$Z = (L_\infty^{n_u} \times (W_1^1)^{n_y *}) \times (\mathbb{R}^{n_y} \times L_\infty^{n_y}) \times \mathbb{R}^{n_r} \times (L_\infty^{n_{\eta u}} \times L_\infty^{n_{\eta y}}).$$

3. The central path. The main object of analytical interest is the *central path* defined by the homotopy (2.9) in μ. First we consider its actual existence in the regular setting given by (2.11) before discussing convergence.

Throughout the paper, we will use the Fischer–Burmeister function [18]

$$\psi(a, b; \mu) = a + b - \sqrt{a^2 + b^2 + 2\mu} \tag{3.1}$$

as an example from a large class of different complementarity functions (see [11, 12, 13, 25]).

3.1. Existence. We begin with establishing some bounds on derivatives of the complementarity function and their inverses.

LEMMA 3.1. *The complementarity function* Ψ *defined via* (3.1) *maps* $L^n_\infty \times L^n_\infty \times \mathbb{R}$ *continuously into* L^n_∞. *Its derivative* $\partial_g \Psi(g,\eta;\mu)$ *is symmetric positive semidefinite, bounded by*

$$\|\partial_g \Psi\|_{L_\infty \to L_\infty} \le 2, \tag{3.2}$$

$$\left\|(\partial_g \Psi)^{-1}\right\|_{L_\infty \to L_\infty} \le \max\left(3, \frac{2}{\mu}\|g\|^2_{L_\infty}\right), \tag{3.3}$$

and Lipschitz continuous with a Lipschitz constant of $\mu^{-1/2}$. *The corresponding holds for* $\partial_\eta \Psi(g,\eta;\mu)$. *Furthermore, the derivatives commute.*

Proof. The claimed properties of the Nemyckii operator Ψ are directly inherited from ψ due to Theorem 2.1. From $(1+\phi)^{-1/2} \le \max(1-\phi/4, 2/3)$ for $\phi > 0$ we infer

$$\begin{aligned}
\min\left(\frac{\mu}{2a^2}, \frac{1}{3}\right) &= 1 - \max\left(1 - \frac{\mu}{2a^2}, \frac{2}{3}\right) \le 1 - \frac{1}{\sqrt{1+\frac{2\mu}{a^2}}} \\
&\le 1 - \frac{1}{\sqrt{1+\frac{b^2}{a^2}+\frac{2\mu}{a^2}}} = 1 - \frac{|a|}{\sqrt{a^2+b^2+2\mu}} \\
&\le \partial_a \psi(a,b;\mu) \\
&\le 1 + \frac{|a|}{\sqrt{a^2+b^2+2\mu}} \le 2.
\end{aligned} \tag{3.4}$$

Thus, $\partial_a \psi$ is uniformly positive definite. Due to Theorem 2.1, the derivative $\partial_g \Psi(g,\eta;\mu)$ of the Nemyckii operator Ψ is bounded by (3.2) and has an inverse that is bounded by (3.3).

As for the Lipschitz continuity, we estimate

$$|\partial_a^2 \psi| = \left| \frac{\sqrt{a^2+b^2+2\mu} - \frac{a^2}{\sqrt{a^2+b^2+2\mu}}}{a^2+b^2+2\mu} \right| \le \frac{1 - \frac{a^2}{a^2+b^2+2\mu}}{\sqrt{a^2+b^2+2\mu}} \le \frac{1}{\sqrt{2\mu}}$$

and

$$|\partial_{ab} \psi| = \left| \frac{ab}{(a^2+b^2+2\mu)^{3/2}} \right| \le \frac{|ab|}{(2|ab|+2\mu)^{3/2}} \le \frac{2}{3\sqrt{6\mu}}$$

such that $\|\psi''\| \le \mu^{-1/2}$. This Lipschitz constant for $\partial_a \psi$ is inherited by $\partial_g \Psi$. Because of symmetry, the same holds for $\partial_\eta \Psi$, which commutes with $\partial_g \Psi$. ∎

THEOREM 3.2. *The complementarity formulation* (2.10) *is a continuously differentiable mapping from* $V \times \mathbb{R}_+$ *to* Z. *Moreover, for any bounded set* $D \subset V$ *there is a constant* $c(D)$ *such that the derivative* $\partial_v F$ *satisfies the Lipschitz condition*

$$\|\partial_v F(v+\delta v;\mu) - \partial_v F(v;\mu)\|_{V \to Z} \le c(1+\mu^{-1/2})\,\|\delta v\|_V \tag{3.5}$$

on D.

Proof. The image spaces and differentiability of the second to fourth component of F have already been established in section 2 and Lemma 3.1. Only the adjoint expression

$$J'(x) - c'(x)^* \lambda_c - r'(x)^* \lambda_r - g'(x)^* \eta$$

remains to be discussed. We consider the terms separately.

First we write $J(x) = \langle \mathbf{1}, \tilde{\mathbf{f}}(x) \rangle$ with $\tilde{\mathbf{f}}'(x) \in \mathcal{L}(L_1^{n_u} \times (W_1^1)^{n_y}, L_1)$ due to Theorem 2.1 and thus obtain

$$J'(x) = \tilde{\mathbf{f}}'(x)^* \mathbf{1} \in \left(L_1^{n_u} \times (W_1^1)^{n_y}\right)^*. \tag{3.6}$$

With δ_0 denoting the point evaluation of the y component at $t = 0$, we have

$$c'(x) = \begin{bmatrix} \tilde{\mathbf{c}}'(x) - \partial_t \\ \delta_0 \end{bmatrix} \in \mathcal{L}\left(L_1^{n_u} \times (W_1^1)^{n_y} \to L_1^{n_y} \times \mathbb{R}^{n_r}\right)$$

again by Theorem 2.1 such that

$$c'(x)^* \lambda_c \in \left(L_1^{n_u} \times (W_1^1)^{n_y}\right)^*. \tag{3.7}$$

Similarly, we obtain

$$r'(x)^* \lambda_r \in \left(L_1^{n_u} \times (W_1^1)^{n_y}\right)^* \quad \text{and} \quad g'(x)^* \eta \in \left(L_1^{n_u} \times (W_1^1)^{n_y}\right)^*. \tag{3.8}$$

Collecting (3.6)–(3.8), $F(v; \mu) \in Z$ is verified. Continuous differentiability is inherited from J, c, g, and ψ.

As for the Lipschitz continuity of the derivative, we have to estimate the differences of

$$\partial_v F(v; \mu) = \begin{bmatrix} \partial_x^2 L(v) & -c'(x)^* & -r'(x)^* & -g'(x)^* \\ -c'(x) & & & \\ -r'(x) & & & \\ \partial_g \Psi(g(x), \eta; \mu) g'(x) & & & \partial_\eta \Psi(g(x), \eta; \mu) \end{bmatrix}$$

for arguments v_1 and v_2. We cover the blocks separately. First we see that

$$c'(x_1) - c'(x_2) = \tilde{\mathbf{c}}'(x_1) - \tilde{\mathbf{c}}'(x_2).$$

Since x_1 and x_2 are bounded in terms of D, the derivative of the Nemyckii operator $\tilde{\mathbf{c}}$ inherits the Lipschitz constant $\kappa_c(D)$ of $\tilde{c}'$ due to (2.6) of Theorem 2.1 with $p = \infty$. Thus, we conclude

$$\|c'(x_1) - c'(x_2)\|_{X \to L_\infty^{n_y} \times \mathbb{R}^{n_y}} \le \kappa_c(D) \|x_1 - x_2\|_X.$$

Analogously, we obtain

$$\|g'(x_1) - g'(x_2)\|_{X \to L_\infty^{n_\eta}} \le \kappa_g(D) \|x_1 - x_2\|_X.$$

Concerning the dual operators $c'(x)^*$ and $g'(x)^*$, we apply Theorem 2.1 with $p = 1$ in (2.6) and obtain

$$\|c'(x_1)^* - c'(x_2)^*\|_{L_\infty^{n_y} \times \mathbb{R}^{n_y} \to L_\infty^{n_u} \times ((W_1^1)^{n_y})^*} \le \kappa_c(D) \|x_1 - x_2\|_X$$

and

$$\|g'(x_1)^* - g'(x_2)^*\|_{L_\infty^{n_\eta} \to L_\infty^{n_u} \times ((W_1^1)^{n_y})^*} \le \kappa_g(D) \|x_1 - x_2\|_X.$$

Similar estimates for $r'(x)$ and $r'(x)^*$ are straightforward. As for $\partial_x^2 L(v)$, we estimate

$$\begin{aligned} \|J''(x_1) - J''(x_2)\|_{X \to L_\infty^{n_u} \times ((W_1^1)^{n_y})^*} &\le \kappa_f(D) \|x_1 - x_2\|_X, \\ \|c''(x_1)^* - c''(x_2)^*\|_{X \times L_\infty^{n_y} \times \mathbb{R}^{n_y} \to L_\infty^{n_u} \times ((W_1^1)^{n_y})^*} &\le \kappa_c(D) \|x_1 - x_2\|_X, \\ \|g''(x_1)^* - g''(x_2)^*\|_{X \times L_\infty^{n_\eta} \to L_\infty^{n_u} \times ((W_1^1)^{n_y})^*} &\le \kappa_g(D) \|x_1 - x_2\|_X \end{aligned}$$

as before. In view of

$$c''(x_1)^*\lambda_{c1} - c''(x_2)^*\lambda_{c2} = c''(x_1)^*(\lambda_{c1} - \lambda_{c2}) + (c''(x_1)^* - c''(x_2)^*)\lambda_{c2}$$

and the boundedness of $c''(x_1)^*$ due to (2.5) of Theorem 2.1, we derive a constant $\kappa(D)$ for

$$\|c''(x_1)^*\lambda_{c1} - c''(x_2)^*\lambda_{c2}\|_{X\to L_\infty^{n_u}\times((W_1^1)^{n_y})^*} \le \bar\kappa_c(D)\|v_1 - v_2\|_X.$$

Treating $r''(x)^*\lambda_r$ and $g''(x)^*\eta$ similarly, we obtain the desired estimate

$$\|\partial_x^2 L(v_1) - \partial_x^2 L(v_2)\|_{X\to L_\infty^{n_u}\times((W_1^1)^{n_y})^*} \le \kappa_L(D)\|v_1 - v_2\|_X.$$

Up to now, the Lipschitz constants have been completely independent of μ. For the blocks $\partial_g\Psi(v)g'(x)$ and $\partial_\eta\Psi(v)$ we obtain a Lipschitz constant of $\kappa_\Psi \le \text{const}(1 + \mu^{-1/2})$. Combining the Lipschitz constants of the individual blocks finally verifies (3.5). ∎

In order to prove the existence of the central path via an implicit function theorem, we first have to establish bounds on the inverse of $\partial_v F$.

THEOREM 3.3. *Suppose there exist an open bounded set $D \subset V$ and constants $\beta > 0$ and $\alpha > 0$ such that the following conditions hold uniformly for all $v \in D$ and $\mu > 0$:*

1. *The state equation satisfies the following inf-sup condition:*

$$\inf_{\xi\in\mathbb{R}^{n_r}} \sup_{\delta u\in L_2^{n_u}} \frac{\xi^T\partial_y r(x)\partial_y c(x)^{-1}\partial_u c(x)\delta u}{|\xi|\,\|\delta u\|_{L_2^{n_u}}} \ge \beta.$$

(*The linearized state equation is controllable.*)

2. *A strengthened Legendre–Clebsch-type condition holds:*

$$\xi^T M_u(t)\xi \ge \alpha|\xi|^2$$

for all $\xi \in \mathbb{R}^{n_u}$ and almost all $t \in \Omega$. Here,

$$\begin{aligned} M_u(t) := {}& \partial_u^2\tilde f(x(t)) - \partial_u^2\tilde c(x(t))^T\lambda_c(t) - (\tilde g^u)''(u(t))^T\eta^u(t) \\ & + (\tilde g^u)'(u(t))^T\partial_\eta\psi(\tilde g^u(u(t)),\eta^u(t);\mu)^{-1}\partial_g\psi(\tilde g^u(u(t)),\eta^u(t);\mu)(\tilde g^u)'(u(t)). \end{aligned}$$

3. *The augmented second derivative of the Lagrangian is uniformly positive definite on the nullspace of the state equation:*

$$\langle\xi, (\partial_x^2 L(v) + g'(x)^*\partial_\eta\Psi(g(x),\eta)^{-1}\partial_g\Psi(g(x),\eta)g'(x))\xi\rangle \ge \alpha\|\xi\|^2_{L_2^{n_u}\times(W_2^1)^{n_y}}$$

for all $\xi \in \ker c'(x)$.

Then $\partial_v F(v;\mu)$ has an inverse which is bounded by

$$\left\|\partial_v F(v;\mu)^{-1}\right\|_{Z\to V} \le \text{const}(1 + \mu^{-3}) \tag{3.9}$$

uniformly for $v \in D$.

Proof. We show that there is a unique solution of $\partial_v F(v;\mu)\Delta v = z$ with $\|\Delta v\|_V \le \text{const}(1 + \mu^{-3})\,\|z\|_Z$.

In order to simplify the notation, let $C = -c'(x)$, $C_u = -\partial_u c(x)$, $C_y = -\partial_y c(x)$, and analogously G, G_u, G_y, R, and R_y. Define $\Psi_\eta = \partial_\eta\Psi(g(x),\eta)$, $\Psi_g = \partial_g\Psi(g(x),\eta)$,

$\Psi^u_\eta = \partial_{\eta^u}\Psi^u(g^u(u),\eta^u)$, $\Psi^u_g = \partial_{g^u}\Psi^u(g^u(u),\eta^u)$, and analogously Ψ^y_η and Ψ^y_g. Moreover, let $M_u = \partial^2_u L(v) + G^*_u(\Psi^u_\eta)^{-1}\Psi^u_g G_u$, and analogously M_y. Finally, let $M_{uy} = \partial_{uy}L(v)$ and $M_{yu} = \partial_{yu}L(v)$.

The state derivative C_y represents the linearization of the initial value problem (2.2) and has a bounded solution for any right-hand side. Thus, C_y has a bounded inverse. More precisely, for any $p \geq 1$,

$$C_y^{-1} : L_p^{n_y} \to (W_p^1)^{n_y} \text{ is bounded uniformly for } v \in D. \tag{3.10}$$

Therefore, we can define the solution operator $S = C_y^{-1}C_u$.

In the following, we will refrain from writing the number of components of the function spaces, which should be clear from context.

In a first step, we reduce the system

$$\partial_v F(v;\mu)(\Delta x, \Delta\lambda_c, \Delta\lambda_r, \Delta\eta)^{\mathrm{T}} = [z_a, z_c, z_r, z_p]^{\mathrm{T}}$$

to a simple saddle point problem. Elimination of the inequality constraints' multipliers $\Delta\eta = \Psi_\eta^{-1}(z_p - \Psi_g G\Delta x)$ by Lemma 3.1 yields the equivalent system

$$\begin{bmatrix} M_u & M_{uy} & C_u^* & \\ M_{yu} & M_y & C_y^* & R_y^* \\ C_u & C_y & & \\ & R_y & & \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta y \\ \Delta\lambda_c \\ \Delta\lambda_r \end{bmatrix} = \begin{bmatrix} \bar z_a^u \\ \bar z_a^y \\ z_c \\ z_r \end{bmatrix},$$

where $(\bar z_a^u, \bar z_a^y)^{\mathrm{T}} = \bar z_a = z_a - G^*\Psi_\eta^{-1}z_p$. Then, $\Delta y = C_y^{-1}z_c - S\Delta u$ and $\Delta\lambda_c = C_y^{-*}(\bar z_a^y - M_y C_y^{-1}z_c - (M_{yu} - M_y S)\Delta u - R_y^*\Delta\lambda_r)$ can be eliminated, which yields

$$\begin{bmatrix} M_u + S^*M_yS - (M_{uy}S + S^*M_{yu}) & -S^*R_y^* \\ -R_yS & \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta\lambda_r \end{bmatrix} = \begin{bmatrix} \hat z_a^u \\ \hat z_r \end{bmatrix}. \tag{3.11}$$

Here we set $\hat z_a^u = \bar z_a^u - M_{uy}C_y^{-1}z_c - S^*(\bar z_a^y - M_y C_y^{-1}z_c)$ and $\hat z_r = z_r - R_y C_y^{-1}z_c$.

In the second step, we establish the existence of a bounded solution of (3.11), first in $L_2^{n_u}\times\mathbb{R}^{n_r}$ and then in $L_\infty^{n_u}\times\mathbb{R}^{n_r}$. Due to Theorem 2.1 and the observation (3.10), M_u, S^*M_yS, $M_{uy}S$, and S^*M_{yu} can all be continuously extended to L_2. Then, $M_u + S^*M_yS - (M_{uy}S + S^*M_{yu}) : L_2^{n_u} \to L_2^{n_u}$ is positive definite due to assumption 3. Moreover, R_yS satisfies the inf-sup-condition of assumption 1. Therefore, Brezzi's splitting theorem [10, 8] guarantees the existence of a solution $(\Delta u, \Delta\lambda_r) \in L_2^{n_u}\times\mathbb{R}^{n_r}$ of (3.11) with

$$\begin{aligned} \|\Delta u\|_{L_2} &\leq \text{const}\left(\|\hat z_a^u\|_{L_2} + \kappa|\hat z_r|\right) \quad \text{and} \\ |\Delta\lambda_r| &\leq \text{const}\left(\kappa\|\hat z_a^u\|_{L_2} + \kappa^2|\hat z_r|\right), \end{aligned} \tag{3.12}$$

where

$$\kappa = 1 + \|M_u + S^*M_yS - (M_{uy}S + S^*M_{yu})\|_{L_2\to L_2},$$

and the constants depend on α and β. Using Lemma 3.1 and, again, the extension of Nemyckii operators to L_2 provided by Theorem 2.1, we obtain the following dependencies on μ:

$$\begin{aligned} \|M_u\|_{L_2\to L_2} &= \|\partial_u^2 L\|_{L_2\to L_2} + \|G_u^*\left(\Psi_\eta^u\right)^{-1}\Psi_g^u G_u\|_{L_2\to L_2} \\ &\leq \text{const} + \|G_u^*\|_{L_2\to L_2}\|\left(\Psi_\eta^u\right)^{-1}\Psi_g^u\|_{L_2\to L_2}\|G_u\|_{L_2\to L_2} \\ &\leq \text{const}(1 + \|\left(\left(\Psi_\eta^u\right)^{-1}\Psi_g^u\|_{L_2\to L_2}\right) \\ &\leq \text{const}(1+\mu^{-1}), \end{aligned}$$

$$\begin{aligned}\|M_y\|_{W_2^1\to(W_2^1)^*} &= \|\partial_y^2 L\|_{W_2^1\to(W_2^1)^*} + \|G_y^*\big(\Psi_\eta^y\big)^{-1}\Psi_g^y G_y\|_{W_2^1\to(W_2^1)^*}\\ &\le \text{const} + \|G_y^*\|_{L_2\to(W_2^1)^*}\|\Psi\tbinom{y}{\eta}^{-1}\Psi_g^y\|_{L_2\to L_2}\|G_y\|_{W_2^1\to L_2}\\ &\le \text{const}\left(1+\|\big(\Psi_\eta^y\big)^{-1}\Psi_g^y\|_{L_2\to L_2}\right)\\ &\le \text{const}(1+\mu^{-1}), \qquad (3.13)\end{aligned}$$

$$\begin{aligned}\kappa &\le 1 + \|M_u\|_{L_2\to L_2} + \text{const}\,\|M_y\|_{W_2^1\to(W_2^1)^*} + \text{const}\\ &\le \text{const}(1+\mu^{-1}).\end{aligned}$$

As for Δu and $\Delta\lambda_r$, we first observe

$$\|\tilde z_a^u\|_{L_2} \le \|z_a\|_{L_2} + \|G_u^*\big(\Psi_\eta^u\big)^{-1} z_p^u\|_{L_2} \le \text{const}(1+\mu^{-1})\|z\|_Z,$$

$$\begin{aligned}\|S^*M_y C_y^{-1} z_c\|_{L_2} &\le \|S^*\|_{(W_2^1)^*\to L_2}\|M_y\|_{W_2^1\to(W_2^1)^*}\|C_y^{-1}z_c\|_{W_2^1}\\ &\le \text{const}(1+\mu^{-1})\|z_c\|_{L_2} \le \text{const}(1+\mu^{-1})\|z\|_Z,\end{aligned}$$

and hence

$$\|\hat z_a^u\|_{L_2} \le \text{const}(1+\mu^{-1})\|z\|_Z. \qquad (3.14)$$

From this we conclude that

$$\|\Delta u\|_{L_2} \le \text{const}(1+\mu^{-1})\|z\|_Z \quad\text{and}\quad |\Delta\lambda_r| \le \text{const}(1+\mu^{-2}).$$

Moreover, $|\hat z_r| \le \text{const}\,\|z\|_Z$ is evident from (3.10). Observing that $S : L_2^{n_u} \to (W_2^1)^{n_y}$ and $S^* : (W_1^1)^{n_y *} \to L_\infty^{n_u}$ due to (3.10), and additionally $R_y^* : \mathbb{R}^{n_r} \to (W_1^1)^{n_y *}$, we infer

$$(S^*M_yS - M_{uy}S - S^*M_{yu}) : L_2^{n_u} \to L_\infty^{n_u} \quad\text{and}\quad S^*R_y^* : \mathbb{R}^{n_r} \to L_\infty^{n_u}$$

such that (3.11) implies

$$M_u\Delta u = \hat z_a^u - (S^*M_yS - M_{uy}S - S^*M_{yu})\Delta u + S^*R_y^*\Delta\lambda_r \in L_\infty^{n_u}.$$

Using assumption 2, the desired regularity $\Delta u \in L_\infty^{n_u}$ is readily established

$$\|\Delta u\|_{L_\infty} \le \text{const}\,\|\hat z_a^u - (S^*M_yS - M_{uy}S - S^*M_{yu})\Delta u + S^*R_y^*\Delta\lambda_r\|_{L_\infty}. \qquad (3.15)$$

In order to estimate the right-hand side of (3.15), we first note that since $\dot y$ appears linearly in c, M_y is a Nemyckii operator. We thus infer

$$\begin{aligned}\|M_y\|_{L_\infty\to L_\infty} &\le \|\partial_y^2 L\|_{L_\infty\to L_\infty} + \|G_y^*\|_{L_\infty\to L_\infty}\|(\Psi_\eta^y)^{-1}\Psi_g^y\|_{L_\infty\to L_\infty}\|G_y\|_{L_\infty\to L_\infty}\\ &\le \text{const}(1+\mu^{-1}),\end{aligned}$$

where we used Theorem 2.1 to obtain $G_y \in \mathcal{L}(L_1, L_1)$, which implies $G_y^* \in \mathcal{L}(L_\infty, L_\infty)$. Then we derive upper bounds for the individual terms in (3.15) as follows:

$$\begin{aligned}&\|S^*M_yS - M_{uy}S - S^*M_{yu}\|_{L_2\to L_\infty}\|\Delta u\|_{L_2}\\ &\quad\le \|S^*\|_{L_\infty\to L_\infty}\|M_y\|_{L_\infty\to L_\infty}\|S\|_{L_2\to L_\infty}\,\text{const}(1+\mu^{-1})\|z\|_Z\\ &\quad\le \text{const}(1+\mu^{-2})\|z\|_Z,\end{aligned}$$

$$\|S^*R_y^*\|_{\mathbb{R}^{n_r}\to L_\infty}|\Delta\lambda_r| \le \text{const}(1+\mu^{-2})\|z\|,$$

and $\|\hat{z}_a^u\|_{L_\infty} \le \text{const}(1+\mu^{-1})$ analogously to (3.14). Thus, we conclude

$$\|\Delta u\|_{L_\infty} \le \text{const}(1+\mu^{-2})\|z\|_Z. \tag{3.16}$$

In the final step of the proof, we will now trace back the elimination chain from the beginning. First we get

(3.17)

$$\begin{aligned}
\|\Delta\lambda_c\|_{\mathbb{R}^{n_r}\times L_\infty} &= \|C_y^{-*}\Big(\bar{z}_a^y - M_y C_y^{-1} z_c - (M_{yu} - M_y S)\Delta u - R_y^*\Delta\lambda_r\Big)\|_{\mathbb{R}^{n_r}\times L_\infty} \\
&\le \text{const}\,\|\bar{z}_a^y - M_y C_y^{-1} z_c - (M_{yu} - M_y S)\Delta u - R_y^*\Delta\lambda_r\|_{(W_1^1)^*} \\
&\le \text{const}\,\Big(\|\bar{z}_a^y\|_{(W_1^1)^*} + \|M_y\|_{L_\infty\to L_\infty}\|C_y^{-1}z_c\|_{W_1^1} \\
&\qquad + \|M_{yu} - M_y S\|_{L_\infty\to L_\infty}\|\Delta u\|_{L_\infty} \\
&\qquad + \|R_y^*\|_{\mathbb{R}^{n_r}\to(W_1^1)^*}|\Delta\lambda_r|\Big) \\
&\le \text{const}\,\Big(\|z_a^y - G_y^*(\Psi_\eta^y)^{-1}\big(z_p^y - \Psi_w^y z_s^y\big)\|_{(W_1^1)^*} + (1+\mu^{-1})\|z\|_Z \\
&\qquad + (1+\mu^{-1})\|\Delta u\|_{L_\infty} + |\Delta\lambda_r|\Big) \\
&\le \text{const}\,\Big(\|z\|_Z + \|G_y^*\|_{L_\infty\to(W_1^1)^*}\|\big(\Psi_\eta^y\big)^{-1}\|_{L_\infty\to L_\infty}\|z_p^y - \Psi_w^y z_s^y\|_{L_\infty} \\
&\qquad + (1+\mu^{-3})\|z\|_Z\Big) \\
&\le \text{const}(1+\mu^{-3})\|z\|_Z.
\end{aligned}$$

The state Δy is bounded by

$$\|\Delta y\|_{W_\infty^1} \le \|C_y^{-1}z_c\|_{W_\infty^1} + \|S\|_{L_\infty\to W_\infty^1}\|\Delta u\|_{L_\infty} \le \text{const}(1+\mu^{-2})\|z\|_Z. \tag{3.18}$$

Finally, we obtain for the Lagrange multiplier $\Delta\eta$ the estimate

$$\begin{aligned}
\|\Delta\eta\|_{L_\infty} &\le \|\Psi_\eta^{-1}\|_{L_\infty\to L_\infty}\big(\|z_p\|_{L_\infty} + \|\Psi_g G\Delta x\|_{L_\infty}\big) \\
&\le \text{const}(1+\mu^{-3})\|z\|_Z.
\end{aligned} \tag{3.19}$$

Collecting (3.12) and (3.16)–(3.19) we obtain the claim (3.9). ∎

Now we are ready to prove that the central path exists locally, and that it can be continued up to $\mu = 0$ unless it leaves its bounded set of definition.

COROLLARY 3.4. *Suppose the assumptions of Theorem* 3.3 *are satisfied. If there are $v_0 \in D$ and $\mu_0 > 0$ with $F(v_0;\mu_0) = 0$, then there exists a maximal open interval $I_\mu \subset R_+$ around μ_0 and a continuously differentiable central path $v : I_\mu \to D$ with the following properties:*

1. *$v(\mu_0) = v_0$.*
2. *$F(v(\mu);\mu) = 0$ for all $\mu \in I_\mu$.*
3. *Either $\text{dist}(v(I_\mu), \partial D) = 0$ or $\inf I_\mu = 0$ holds.*

Proof. Due to Theorems 3.2 and 3.3 there is an open neighborhood of (v_0,μ_0) on which F and $\partial_v F$ are continuous and $\partial_v F$ is bijective. The implicit function theorem (cf. [41, Thm. 4.B]) guarantees the existence of a continuously differentiable central path $v(\mu)$ with $F(v(\mu),\mu) = 0$ on an open interval around μ_0. A closer inspection of the proof of the implicit function theorem and using the bounds derived in Theorems 3.2 and 3.3 shows that there is a constant $\epsilon = \epsilon(\text{dist}(v_0, D))$ independent of μ such that $v(\mu)$ exists on the open interval $]\mu_0 - \epsilon\mu^{-4}, \mu_0 + \epsilon\mu^{-4}[$.

Let $I_\mu \subset \mathbb{R}_+$ be a maximal open interval around μ_0, such that property 2 holds. Now assume that property 3 does not hold, i.e., $\operatorname{dist}(v(I_\mu), \partial D) \geq \varepsilon > 0$ and $\delta = \inf I_\mu > 0$. We consider $\mu = \delta + \epsilon\delta^{-4}/2$ with $\epsilon = \epsilon(\varepsilon)$. Again, due to the implicit function theorem, there is an open interval $J_\mu =]\mu - \epsilon\mu^{-4}, \mu + \epsilon\mu^{-4}[$ such that property 2 holds on J_μ and hence on $J_\mu \cup I_\mu$. Since $\mu - \epsilon\mu^{-4} < \delta$, this consequence contradicts the maximality of I_μ, and property 3 must be true. ∎

3.2. Convergence. Corollary 3.4 does not guarantee the existence of the central path for all $\mu > 0$, since the path may reach the boundary of D for some $\mu_{\lim} > 0$. Moreover, the upper bound for $\|\partial_v F(v;\mu)^{-1}\|$ which has been established in Theorem 3.3 is useless for proving convergence of the path towards a Kuhn–Tucker limit point. The two reasons are the possible occurrence of Dirac parts in the state constraints' multipliers at the beginning or end of constrained arcs, and the naive block elimination of the multipliers $\Delta\eta$ in the proof of Corollary 3.4.

Under more restrictive assumptions, in particular, the restriction to purely control constrained problems, a splitting into nearly active and nearly inactive constraints can be used to show both boundedness of the central path and independence of $\|\partial_v F(v;\mu)^{-1}\|$ with respect to μ.

DEFINITION 3.5. *For some $\rho > 0$ and functions $u \in L_\infty^{n_u}$ and $\eta \in L_\infty^{n_\eta^u}(\Omega)$, define the characteristic function $\chi^A = \chi^A(t;u,\eta,\mu)$ of the nearly active set vector Ω^A componentwise as*

$$\chi_i^A(t) = \begin{cases} 1, & \tilde{g}_i^u(u_i(t)) \leq \rho\eta_i^u(t), \\ 0 & \text{otherwise.} \end{cases}$$

The corresponding characteristic function χ^I of the nearly inactive set vector Ω^I is defined as $\mathbf{1} - \chi^A$, where $\mathbf{1} \in L_\infty^{n_\eta^u}$ is the constant function with value 1.

Note that pointwise multiplication with χ^A defines an orthogonal projector onto the corresponding L_∞ space over the nearly active set vector Ω^A.

First we address the issue of the central path leaving a bounded domain of definition. Assuming a suitable constraint qualification for nearly active constraints of points on the central path, we establish a priori bounds for the central path.

THEOREM 3.6. *Suppose $n_\eta^y = 0$; i.e., there are no state constraints. Assume that the following conditions are satisfied:*

(i) *The feasible region $D_u := \{u \in L_\infty^{n_u} : g(u) \geq 0\}$ is bounded.*

(ii) *The state contribution function in the state equation is linearly bounded:*

$$|\tilde{c}(u,y)| \leq \text{const}(1 + |y|) \quad \text{for all } y \in \mathbb{R}^{n_y} \text{ and } u \in D_u.$$

Then there is a bounded set $D_y \subset (W_\infty^1)^{n_y}$ such that for all $\mu > 0$ every solution v of $F(v;\mu) = 0$ satisfies $u \in D_u$ and $y \in D_y$.

If, in addition, there is a constant $\beta > 0$ such that the equality constraints and nearly active control constraints satisfy the inf-sup condition

$$\inf_{h \in \mathbb{R}^{n_r}, \xi \in L_\infty^{n_\eta^u}} \sup_{\delta u \in L_1^{n_u}} \frac{h^T \partial_y r(x) \partial_y c(x)^{-1} \partial_u c(x) \delta u + \langle \chi^A \xi, g'(u)\delta u \rangle}{(|h| + \|\chi^A \xi\|_{L_\infty^{n_\eta^u}}) \, \|\delta u\|_{L_1^{n_u}}} \geq \beta \tag{3.20}$$

uniformly for central path solutions v with $x \in D_u \times D_y$, then there is a bounded set $D_0 \subset V$ such that $v \in D_0$.

Proof. Suppose $v = (u, y, \lambda_c, \lambda_r, \eta)$ is a central path solution of $F(v;\mu) = 0$ for some $\mu > 0$. Since $\Psi(g(u), \eta) = 0$ implies $g(u) \geq 0$, we have $u \in D_u$ by assumption (i).

Assumption (ii) then guarantees the existence of a constant $\gamma_y < \infty$ such that $y \in S(0, \gamma_y) =: D_y$.

Now consider the state part of the adjoint equation

$$\partial_y J(x) - \partial_y c(x)^* \lambda_c - \partial_y r(y)^* \lambda_r = 0.$$

Due to the formulation of c as initial value problem, the inverse of $\partial_y c(x) : (W_1^1) \to L_1 \times \mathbb{R}^{n_y}$ is uniformly bounded on $D_u \times D_y$. Thus, we can conclude that

$$\begin{aligned}\|\lambda_c\|_{L_\infty \times \mathbb{R}^{n_y}} &\le \|\partial_y c(x)^{-*}\|_{(W_\infty^1)^* \to L_\infty \times \mathbb{R}^{n_y}} \|\partial_y J(x) - \partial_y r(y)^* \lambda_r\|_{(W_\infty^1)^*} \\ &\le \text{const}\, \|\partial_y J(x) - \partial_y r(y)^* \lambda_r\|_{(W_\infty^1)^*}.\end{aligned}$$

Since $\partial_y \tilde{f}(x)$ is uniformly bounded in $L_\infty^{n_y}$ for $x \in D_u \times D_y$, so is $\|\partial_y J(x)\|_{(W_\infty^1)^*}$, and we obtain

$$\|\lambda_c\|_{L_\infty \times \mathbb{R}^{n_y}} \le \text{const}(1 + |\lambda_r|). \tag{3.21}$$

Inserting $\lambda_c = \partial_y c(x)^{-*}(\partial_y J(x) - \partial_y r(y)^* \lambda_r)$ into the control part of the adjoint equation

$$\partial_u J(x) - \partial_u c(x)^* \lambda_c - g'(u)^* \eta = 0,$$

and splitting the Lagrange multiplier η into nearly active and nearly inactive parts yields

$$\begin{aligned}&\partial_u J(x) - \partial_u c(x)^* \partial_y c(x)^{-*} \partial_y J(x) - g'(u)^* \chi^I \eta \\ &\quad = (\partial_y r(y) \partial_y c(x)^{-1} \partial_u c(x))^* \lambda_r + g'(u)^* \chi^A \eta.\end{aligned}$$

Then the inf-sup condition of assumption (3.20) provides the estimate

$$\begin{aligned}\beta(|\lambda_r| + \|\chi^A \eta\|_{L_\infty}) &\le \sup_{u \in L_1} \frac{\langle (\partial_y r(x) \partial_y c(x)^{-1} \partial_u c(x))^* \lambda_r + g'(u)^* \chi^A \eta, u \rangle}{\|u\|_{L_1}} \\ &\le \|(\partial_y r(x) \partial_y c(x)^{-1} \partial_u c(x))^* \lambda_r + g'(u)^* \chi^A \eta\|_{L_\infty} \\ &= \|\partial_u J(x) - \partial_u c(x)^* \partial_y c(x)^{-*} \partial_y J(x) + g'(u) \chi^I \eta\|_{L_\infty}.\end{aligned}$$

Note that $\|\chi^I \eta\|_{L_\infty}$ is bounded by $\rho^{-1} \|g(u)\|_{L_\infty}$ and $\|\partial_y c(x)^{-*} \partial_y J(x)\|_{L_\infty \times \mathbb{R}^{n_y}}$ is bounded as shown earlier. Similarly, $\|\partial_u J(x)\|_{L_\infty}$ is bounded. $\|g'(u)\|_{L_\infty \to L_\infty}$ and $\|\partial_u c(x)\|_{L_1 \to L_1 \times \mathbb{R}^{n_y}}$ are bounded by Theorem 2.1. Thus, we conclude that

$$|\lambda_r| + \|\chi^A \eta\|_{L_\infty} \le \text{const}\, \beta^{-1}.$$

Combining this with $x \in D_u \times D_y$ verifies the boundedness of v. ∎

The splitting of the domain into nearly active and inactive regions leads also to improved estimates for the dependency of the complementarity function on the homotopy parameter μ.

The reason for the dependence of $\|\partial_v F(v; \mu)^{-1}\|$ on μ in Theorem 3.3 is the increase of $\|\partial_\eta \Psi^{-1}\|$ as $\mu \to 0$. This can be overcome by more sophisticated elimination of variables in the proof. As a preparation, we first prove a refinement of Lemma 3.1.

LEMMA 3.7. *The Fischer–Burmeister complementarity function satisfies the following estimates:*

$$\|\chi^A \partial_g \Psi(g(u), \eta)^{-1}\|_{L_\infty \to L_\infty} \le \left(1 - \frac{\rho}{\sqrt{1 + \rho^2}}\right)^{-1}, \tag{3.22}$$

$$\|\chi^I \partial_\eta \Psi(g(u), \eta)^{-1}\|_{L_\infty \to L_\infty} \le \left(1 - \frac{1}{\sqrt{1 + \rho^2}}\right)^{-1}. \tag{3.23}$$

In particular, both bounds are independent of μ.

Proof. In the relevant inequality (3.4) we now assume that $a \le \rho b$. This leads to

$$\partial_a \psi(a,b;\mu) \ge 1 - \frac{1}{\sqrt{1+\frac{b^2}{a^2}+\frac{2\mu}{a^2}}} \ge 1 - \frac{1}{\sqrt{1+\frac{1}{\rho^2}+\frac{2\mu}{a^2}}} \ge 1 - \frac{1}{\sqrt{1+\frac{1}{\rho^2}}}.$$

On the nearly active region, this assumption holds, such that due to the projection onto the nearly active region the estimate transfers to $\chi^A \partial_g \Psi(g(u),\eta)^{-1}$. Thus, (3.22) is verified. By symmetry, (3.23) is verified using the complementary assumption $a > \rho b$. ∎

THEOREM 3.8. *Assume* $n_\eta^y = 0$*; i.e., only control constraints are present. Suppose there exist a bounded set* $D \subset V$ *and constants* $\beta > 0$ *and* $\alpha > 0$ *such that the following conditions hold uniformly for all central path solutions* $v = v(\mu) \in D$ *with* $F(v(\mu);\mu) = 0$ *and* $\mu > 0$.

1. *State equation and nearly inactive control constraints satisfy the inf-sup condition*

$$\inf_{h\in\mathbb{R}^{n_r},\xi\in L_p^{n_\eta^u}} \sup_{\delta u\in L_q^{n_u}} \frac{h^T \partial_y r(x) \partial_y c(x)^{-1} \partial_u c(x) \delta u + \langle \chi^A \xi, g'(u)\delta u\rangle}{(|h| + \|\chi^A \xi\|_{L_p^{n_\eta^u}})\, \|\delta u\|_{L_q^{n_u}}} \ge \beta$$

for both $(p,q) = (\infty,1)$ *and* $(p,q) = (2,2)$.

2. *The augmented second derivative of the Lagrangian*

$$M = \partial_x^2 L(v) + g'(x)^* \partial_\eta \Psi(g(x),\eta)^{-1} \chi^I \partial_g \Psi(g(x),\eta) g'(x)$$

is positive semidefinite on the nullspace of the linearized state equation:

$$\langle \xi, M\xi\rangle \ge 0 \qquad \text{for } \xi \in \ker c'(x), \tag{3.24}$$

$$\langle \xi, M\xi\rangle \ge \alpha \|\xi\|^2_{L_2^{n_u}\times (W_2^1)^{n_y}} \qquad \text{for } \xi \in \ker c'(x) \cap \ker \chi^A g'(u). \tag{3.25}$$

Then $\partial_v F(v;\mu)$ *has an inverse which is bounded uniformly for* $(v,\mu) \in D \times \mathbb{R}_+$.

Before delving into the proof, let us briefly discuss the assumptions of Theorem 3.8. Mostly, they have counterparts in well-known optimality conditions, but they need to be extended a priori to a neighborhood of the central path in order to be able to show convergence.

Assumption 1 is a direct generalization of the linear independence constraint qualification (LICQ; see, e.g., [19, Def. 2.9]) from nonlinear programming to the infinite-dimensional setting. It is also a reinterpretation of regular points (cf. [28, (2.1)]) in the setting of interior points. It provides uniqueness of the Lagrange multipliers and is therefore necessary for proving invertibility of $\partial_v F$.

Convexity of the Lagrangian on the nullspace of the linearized state equation is generally required for sufficient second order optimality conditions. In particular, requirement (3.25) can be interpreted as an adaptation of the convexity condition given by Maurer [28, Thm. 3.5], whereas (3.24) is only technically necessary for invoking a certain saddle point lemma in the proof. In the control constrained setting, the Legendre–Clebsch condition that has been assumed explicitly in Theorem 3.3 is implied by the earlier convexity assumption.

LEMMA 3.9. *Assumption 2 of Theorem 3.8 implies a strengthened Legendre–Clebsch-type condition for almost all* $t \in \Omega$*:*

$$\begin{aligned} M_u(t) := {} & \partial_u^2 \tilde f(x(t)) - \partial_u^2 \tilde c(x(t))^T \lambda_c(t) - \tilde g''(u(t))^T \eta(t) \\ & + \tilde g'(u(t))^T \partial_\eta \psi(\tilde g(u(t)),\eta(t);\mu)^{-1} \chi^I \partial_g \psi(\tilde g(u(t)),\eta(t);\mu) \tilde g'(u(t)) \end{aligned}$$

satisfies

$$\xi^T M_u(t)\xi \geq 0 \qquad \text{for } \xi \in \mathbb{R}^{n_u}, \tag{3.26}$$

$$\xi^T M_u(t)\xi \geq \alpha|\xi|^2 \qquad \text{for } \xi \in \ker \chi^A(t)\tilde{g}'(u(t)). \tag{3.27}$$

Proof. Let $\xi \in \mathbb{R}^{n_u}$ be arbitrary and define $\delta u = \xi\chi_{[t-\epsilon,t+\epsilon]}$ for arbitrary $t \in \operatorname{int}(\Omega)$ and sufficiently small $\epsilon > 0$. Defining M_{yu}, M_{uy}, and S as in Theorem 3.3, we introduce $\delta y = S\delta u$ such that $(\delta u, \delta y) \in \ker c'(x)$. From standard ODE theory we know that $\|\delta y\|_{L_\infty} \leq \operatorname{const}\|\delta u\|_{L_1} \leq \operatorname{const}\epsilon$. Let $M_u = \partial_u^2 L(v) + g'(u)^*\chi^I\Psi_\eta(g(u),\eta)^{-1}\Psi_g(g(u),\eta)g'(u)$, $M_y = \partial_y^2 L(v)$, and

$$M = \begin{bmatrix} M_u & M_{uy} \\ M_{yu} & M_y \end{bmatrix}.$$

Since M_y, M_{yu}, and M_{uy} are uniformly bounded Nemyckii operators, we have by (3.24)

$$\begin{aligned}\langle \delta u, M_u \delta u\rangle &= \langle(\delta u,\delta y), M(\delta u,\delta y)\rangle - \langle \delta y, M_y\delta y\rangle - \langle \delta y, M_{uy}\delta u\rangle - \langle \delta u, M_{yu}\delta y\rangle \\ &\geq 0 - \operatorname{const}\|\delta y\|_{L_\infty}^2 - 2\operatorname{const}\|\delta y\|_{L_\infty}\|\delta u\|_{L_1} \\ &\geq -\operatorname{const}\epsilon^2\end{aligned}$$

for all t and $\epsilon > 0$, and hence $\xi^T M_u(t)\xi \geq 0$ for all ξ and almost all $t \in \Omega$, which verifies (3.26). Restricting ξ to $\ker\chi^A(t)\tilde{g}'(u(t))$ and using (3.25) instead of (3.24) finally proves (3.27). ∎

Proof of Theorem 3.8. The structure and line of argument is similar to the proof of Theorem 3.3. We, therefore, concentrate on the differences and extensions. Define $C, C_u, C_y, R, R_y, M_{uy}, M_{yu}$, and S as before. Let $G = -g'(u)$. Define $\Psi_g = \partial_g\Psi(g(u),\eta)$ and analogously Ψ_η. Finally, define M_u and M_y as in Lemma 3.9.

As before, the first step consists of eliminating the Lagrange multiplier, but here only the nearly inactive part $\chi^I\eta = \chi^I\Psi_\eta^{-1}(z_p - \Psi_g G\Delta u)$. In order to symmetrize the remaining system, the nearly active part of the complementarity equation is multiplied by Ψ_g^{-1}:

$$\begin{bmatrix} M_u & M_{uy} & C_u^* & & G^*\chi^A \\ M_{yu} & M_y & C_y^* & R_y^* & \\ C_u & C_y & & & \\ & R_y & & & \\ \chi^A G & & & & -\chi^A\Psi_g^{-1}\Psi_\eta \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta y \\ \Delta\lambda_c \\ \Delta\lambda_r \\ \chi^A\Delta\eta \end{bmatrix} = \begin{bmatrix} \bar{z}_a^u \\ z_a^y \\ z_c \\ z_r \\ \chi^A\Psi_g^{-1}z_p \end{bmatrix}$$

with $\bar{z}_a^u = z_a^u - G^*\chi^I\Psi_\eta^{-1}z_p$. Note that χ^A, Ψ_g^{-1}, and Ψ_η commute. Continuing with the elimination of Δy and λ_c as in the proof of Theorem 3.3, we end up with

$$\begin{bmatrix} T & -(R_yS)^* & G^*\chi^A \\ -R_yS & & \\ \chi^A G & & -\chi^A\Psi_g^{-1}\Psi_\eta \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta\lambda_r \\ \chi^A\Delta\eta \end{bmatrix} = \begin{bmatrix} \hat{z}_a^u \\ \hat{z}_r \\ \chi^A\Psi_g^{-1}z_p \end{bmatrix},$$

where $T = M_u + S^*M_yS - (M_{uy}S + S^*M_{yu})$, $\hat{z}_a^u = \bar{z}_a^u - M_{uy}C_y^{-1}z_c - S^*(\bar{z}_a^y - M_yC_y^{-1}z_c)$, and $\hat{z}_r = z_r - R_yC_y^{-1}z_c$. Due to assumption 2, T is positive definite on the nullspace of $\chi^A G$ and positive semidefinite on the whole space. Assumption 1 provides the inf-sup condition for the combined operator

$$\begin{bmatrix} -R_yS \\ \chi^A G \end{bmatrix},$$

and $\chi^A\Psi_g^{-1}\Psi_\eta$ is positive semidefinite. In this situation, the application of Brezzi's splitting theorem is substituted by a theorem of Braess and Blömer [9] on saddle point problems with penalty term. This guarantees the existence of a solution $(\Delta u, \Delta\lambda_r, \chi^A\Delta\eta) \in L_2^{n_u} \times \mathbb{R}^{n_r} \times L_2(\Omega^A)$ with

$$\|\Delta u\|_{L_2} + |\Delta\lambda_r| + \|\Delta\eta^A\|_{L_2} \le \text{const}\,\kappa(\|\hat{z}_a^u\|_{L_2} + |\hat{z}_r| + \|\chi^A\Psi_g^{-1}z_p\|_{L_2}),$$

where $\kappa = \|T\|+\|R_yS\|+\|G_A\|+\|\chi^A\Psi_g^{-1}\Psi_\eta\|+\alpha+\beta$. Note that due to Lemma 3.7 the operators $\chi^A\Psi_g^{-1}\Psi_\eta$ and $\chi^I\Psi_\eta^{-1}\Psi_g$ are bounded independently of μ. This property is inherited by κ and $\|\chi^A\Psi_g^{-1}z_p\|$, such that $\|\Delta u\|_{L_2}$, $|\Delta\lambda_r|$, and $\|\chi^A\Delta\eta\|_{L_2}$ are bounded independently of μ.

Subsequently, the L_∞-regularity of Δu and $\chi^A\Delta\eta$ is established. As in the proof of Theorem 3.3, we have

$$(S^*M_yS - M_{uy}S - S^*M_{yu})\Delta u + S^*R_y^*\Delta\lambda_r \in L_\infty^{n_u}$$

such that for almost all $t \in \Omega$ the finite-dimensional linear equation system

$$\begin{bmatrix} M_u(t) & \tilde{g}'(u(t))^T\chi^A(t) \\ \chi^A(t)\tilde{g}'(u(t)) & -B \end{bmatrix} \begin{bmatrix} \Delta u(t) \\ \chi^A\Delta\eta(t) \end{bmatrix} = \begin{bmatrix} a \\ \chi^A(t)b \end{bmatrix} \tag{3.28}$$

holds, with $B = \chi^A(t)\partial_g\psi(g(u(t)),\eta(t))^{-1}\partial_\eta\psi(g(u(t)),\eta(t))$. Here, a and b denote generic right-hand side vectors the norm of which is bounded by a constant independent of μ. By Lemma 3.9, $M_u(t)$ is positive definite on the nullspace of $\tilde{g}'(u(t))$, such that we can again apply the lemma by Braess and Blömer, now for the finite-dimensional equation (3.28). This yields

$$|\Delta u(t)| + |\chi^A\Delta\eta(t)| \le \text{const}(\|M_u(t)\| + \|\tilde{g}'(u(t))\| + \|B\| + \alpha + \beta)(|a| + |b|) \tag{3.29}$$

for almost all $t \in \Omega$, and hence

$$\|\Delta u\|_{L_\infty} \le \text{const}, \tag{3.30}$$

$$\|\chi^A\Delta\eta^A\|_{L_\infty} \le \text{const} \tag{3.31}$$

independently of μ. Finally, tracing back the elimination stack as in Theorem 3.3 verifies the claim. ∎

As in Corollary 3.4, local existence of the central path can be shown. Moreover, the a priori bound of the solution given by Theorem 3.6 eliminates the possibility of premature termination of the path. Finally, the fact that the inverse of $\partial_v F$ can be bounded independently of μ limits the length of the path and thus ensures convergence.

THEOREM 3.10. *Assume Theorem* 3.6 *holds, providing a bounded set* $D_0 \subset V$ *containing the central path. Define* $D = \bigcup_{v\in D_0} S(v,\epsilon)$ *for some* $\epsilon > 0$. *Suppose the assumptions of Theorem* 3.8 *hold on* D.

If there are $v_0 \in D_0$ *and* $\mu_0 > 0$ *with* $F(v_0;\mu_0) = 0$, *then the central path* $v(\mu)$ *exists for all* $0 < \mu \le \mu_0$ *and converges to a Kuhn–Tucker point* $v(0)$:

$$\|v(\mu) - v(0)\|_V \le \text{const}\,\sqrt{\mu}.$$

Proof. First we notice that due to Theorem 3.2, there is some $\epsilon > 0$ such that $\partial_v F(v;\mu)^{-1}$ is uniformly bounded on the neighborhood

$$U = \bigcup_{(v;\mu) \text{ with } F(v;\mu)=0} S((v,\mu),\epsilon)$$

of the central path solutions $v(\mu)$. As in the proof of Theorem 3.4, the central path exists on a maximal interval I_μ containing μ_0. Since due to Theorem 3.6 this central path is bounded away from ∂D, we have $\inf I_\mu = 0$. Thus, the central path exists for all $0 < \mu \le \mu_0$.

Next we estimate $\partial_\mu F(v(\mu);\mu)$. Since only the complementarity function Ψ depends on μ, this is given by $\partial_\mu \Psi(g(u),\eta;\mu) = -(g(u)^2+\eta^2+2\mu)^{-1/2}$. On the central path, we have $g(u)\cdot\eta = \mu$ a.e. and thus

$$\|\partial_\mu \Psi(g(u),\eta;\mu)\|_{L_\infty} \le (4\mu)^{-1/2}.$$

Now the derivative of the central path is given by

$$v'(\mu) = \partial_v F(v(\mu);\mu)^{-1}\partial_\mu F(v(\mu);\mu).$$

Theorem 3.8 yields

$$\|v'(\mu)\|_V \le \|\partial_v F(v(\mu);\mu)^{-1}\|_{Z\to V}\|\partial_\mu F(v(\mu);\mu)\|_Z \le \text{const}\,\mu^{-1/2}. \tag{3.32}$$

Therefore, the central path is uniformly continuous and converges to some limit point $v(0) \in D$ at a rate of

$$\|v(\mu)-v(0)\|_V \le \int_0^\mu \|v'(s)\|_V\,ds \le \text{const}\int_0^\mu s^{-1/2}\,ds = \text{const}\,\sqrt{\mu}.$$

The continuity of F on $D\times[0,\infty[$ implies that $F(v(0);0)=0$, such that $v(0)$ satisfies the first order necessary conditions (2.8). □

In the remainder of the section, we will apply the preceding theorems to a class of prototypical optimal control problems. We consider

$$\begin{aligned} \min \int_0^1 \left(\tilde f^y(y(t)) + \frac{\alpha}{2}|u(t)|^2\right) dt \\ \text{subject to} \quad \dot y(t) = Ay(t) + Bu(t), \\ y(0) = y_0, \\ a \le u(t) \le b. \end{aligned}$$

THEOREM 3.11. *Suppose that $\tilde f^y$ is convex and twice Lipschitz-continuously differentiable, $\alpha > 0$, $a < b$, $A \in \mathbb{R}^{n_y\times n_y}$, and $B \in \mathbb{R}^{n_y\times n_u}$. Assume there are v_0 and $\mu_0 > 0$ such that $F(v_0;\mu_0) = 0$. Then the central path $v(\mu)$ converges to a Kuhn–Tucker point $v(0) \in D$ at a rate of*

$$\|v(\mu)-v(0)\| \le \text{const}\,\sqrt{\mu}.$$

Proof. We restrict the discussion to a scalar control, i.e., $n_u = 1$. The extension to vector valued controls is straightforward but notationally more involved. We start with Theorem 3.6, choosing

$$\rho < \frac{1}{\mu_0}\left(\frac{b-a}{2}\right)^2 \tag{3.33}$$

for separating nearly active and nearly inactive constraints. Due to the box constraints and the linearity of the state equation, conditions (i) and (ii) are satisfied. Since no terminal boundary conditions are given, the inf-sup condition (3.20) simplifies to

$$\inf_{\xi\in L_p^2}\sup_{\delta u\in L_q^1}\frac{\langle \chi^A\xi, g'(u)\delta u\rangle}{\|\chi^A\xi\|_{L_p^2}\|\delta u\|_{L_q^1}} \ge \beta \quad \text{with } g'(u) = \begin{pmatrix} I \\ -I\end{pmatrix}.$$

Assume that for a central path solution (v, μ) with $\mu \le \mu_0$, the lower constraint $u \ge a$ is nearly active at t, i.e., $\rho\eta^a(t) \ge u(t) - a$. For simplicity, we will omit the argument t in the following. Together with (3.33) and the interior point condition $\eta^a(u-a) = \mu = \eta^b(b-u)$ holding for all central path solutions, this implies

$$\begin{aligned} b - u = b - a - (u - a) &\ge b - a - \sqrt{\rho\eta^a(u-a)} = b - a - \sqrt{\rho\mu} \\ &\ge b - a - \frac{b-a}{2} = \frac{b-a}{2} > \sqrt{\rho\mu} = \sqrt{\rho\eta^b(b-u)}. \end{aligned}$$

Squaring and dividing by $b-u$ finally yields $b-u > \rho\eta^b$, which implies that the upper constraint $u \le b$ is nearly inactive whenever the lower constraint is nearly active. Analogously, the converse can be shown, such that at most one of the two constraints is active. Since in $\chi^A\xi$ at least one component vanishes, we see that

$$\inf_{\xi\in L_p^2} \sup_{\delta u\in L_q^1} \frac{\langle \chi^A\xi, g'(u)\delta u\rangle}{\|\chi^A\xi\|_{L_p^2}\|\delta u\|_{L_q^1}} \ge \inf_{\xi\in L_p^1} \sup_{\delta u\in L_q^1} \frac{\langle \xi, \delta u\rangle}{\|\xi\|_{L_p^1}\|\delta u\|_{L_q^1}} \ge 1 \tag{3.34}$$

for both $(p, q) = (\infty, 1)$ and $(p, q) = (2, 2)$, which confirms the inf-sup condition.

Now we verify the assumptions of Theorem 3.8 on the whole space $D = V$. Assumption 1 is again the inf-sup condition (3.34). The Legendre–Clebsch condition 2 is satisfied due to $\alpha > 0$ and the linearity of the constraints, as is the positive definiteness condition 3 for $\partial_x^2 L(v)$. Since Theorem 3.8 thus holds on V, we can apply Theorem 3.10, which yields the claim. ∎

Remark 3.12. The main conditions to verify are the inf-sup constraint qualification and the convexity. While the latter has been explicitly assumed, the former is a direct consequence of the box constraints. More complex optimization problems require more work to verify the assumptions of Theorem 3.10. Nonlinearity of the state equation needs to be compensated by convexity and an a priori bound on λ as given by Theorem 3.6 in order to obtain convexity of the Lagrangian with respect to x. The inf-sup constraint qualification can be shown for more general constraints, e.g., pointwise convex polyhedric admissible sets for the control. It needs to be verified that at most n_u constraints are nearly active.

Numerical results for a specific problem of this class are given in [39].

4. A short-step path-following method. With the refined estimates from section 3.2, we can show linear convergence of a short-step path-following method. Note that this is a purely theoretical algorithm, since it relies on the exact solution of operator equations in function space and on knowledge of global Lipschitz constants. For an implementable approximation via inexact Newton corrector and inexact tangential predictor, we refer to [39].

We consider the following simple algorithm.

ALGORITHM 4.1.

1 initialize v_0, μ_0 such that $F(v_0; \mu_0) = 0$
2 choose $\sigma < 1$ sufficiently large
3 while $\mu_k > 0$
4 advance $\mu_{k+1} \leftarrow \sigma\mu_k$
5 compute one corrector step $\partial_v F(v_k; \mu_{k+1})\delta v_k = -\partial_\mu F(v_k; \mu_{k+1})$
6 advance $v_{k+1} \leftarrow v_k + \delta v_k$, $k \leftarrow k + 1$

The sequence v_k of iterates converges to the Kuhn–Tucker point $v(0)$.

First, we recall the essentials of an affine covariant Newton–Mysovskikh theorem from [16].

THEOREM 4.2. *Assume $F : X \to Y$ is a differentiable mapping with $F(x^*) = 0$. Assume the derivative $F'(x)$ is invertible on $D = S(x^*, \delta)$ and satisfies*

$$\|F'(x)^{-1}(F'(y) - F'(x))\| \le \omega\|y - x\| \tag{4.1}$$

for $x, y \in D$. Let the ordinary Newton sequence x^k starting at $x^0 \in D$ be defined by $x^{k+1} = x^k - F'(x^k)^{-1}F(x^k)$. Then x^k converges to x^* at a rate of

$$\|x^{k+1} - x^*\| \le \frac{\omega}{2}\|x^k - x^*\|^2.$$

THEOREM 4.3. *Suppose that F satisfies the assumptions of Theorem* 3.10*, providing a bounded set D. Let $v_0 \in D$ and $\mu_0 > 0$ be given such that $F(v_0; \mu_0) = 0$. Then there is a constant $\sigma < 1$ such that the sequence v_k of iterates generated by Algorithm* 4.1 *converges linearly to the limit point $v(0)$ of the central path.*

Proof. To begin with, we verify the assumptions of Theorem 4.2. By Theorems 3.2 and 3.8 there are constants γ_1 and γ_2 independent of $\mu \le \mu_0$, such that $\|\partial_v F(v; \mu) - \partial_v F(v(\mu); \mu)\|_{V\to Z} \le \gamma_1\mu^{-1/2}$ and $\|\partial_v F(v(\mu); \mu)^{-1}\|_{Z\to V} \le \gamma_2$. Omitting the argument μ from F, we use the Banach perturbation lemma to derive

$$\begin{aligned}
&\|\partial_v F(v)^{-1}\|_{Z\to V} \\
&\quad \le \|\partial_v F(v(\mu))^{-1}\|_{Z\to V}\|(I - (\partial_v F(v(\mu)) - \partial_v F(v))\partial_v F(v(\mu))^{-1})^{-1}\|_{Z\to Z} \\
&\quad \le \frac{\gamma_2}{1 - \gamma_1\mu^{-1/2}\|v - v(\mu)\|_V\gamma_2} \le 2\gamma_2
\end{aligned}$$

for $v \in D = S\big(v(\mu), \sqrt{\mu}/(2\gamma_2\gamma_1)\big)$. For $v_1, v_2 \in D$ we thus obtain

$$\begin{aligned}
&\|\partial_v F(v_1)^{-1}(\partial_v F(v_2) - \partial_v F(v_2))\|_{V\to V} \\
&\quad \le \|\partial_v F(v_1)^{-1}\|_{Z\to V}\|(\partial_v F(v_2) - \partial_v F(v_1))\|_{V\to Z} \\
&\quad \le 2\gamma_2\gamma_1\mu^{-1/2}\|v_2 - v_1\|_V,
\end{aligned}$$

which establishes the Lipschitz condition (4.1) with

$$\omega(\mu) \le \frac{2\gamma_2\gamma_1}{\sqrt{\mu}}.$$

As in (3.32) in the proof of Theorem 3.10, we obtain a bound on the derivative of the central path in the form of

$$\|v'(\mu)\|_V \le \frac{\beta}{\sqrt{\mu}}$$

with $\beta < \infty$ independent of μ. Define

$$\delta = (2\gamma_2\gamma_1)^{-1} \quad \text{and} \quad \sigma \ge \left(1 - \frac{\delta}{2(\delta + \beta)}\right)^2. \tag{4.2}$$

Let us assume by induction that $\|v_k - v(\mu_k)\|_V \le \delta\sqrt{\mu_k}/2$. Then we have

$$\begin{aligned}
\|v_k - v(\sigma\mu_k)\|_V &\le \|v_k - v(\mu_k)\|_V + (1 - \sigma)\mu_k \sup_{\mu\in[\sigma\mu_k, \mu_k]} \|v'(\mu)\|_V \\
&\le \frac{\delta\sqrt{\mu_k}}{2} + (1 - \sigma)\mu_k\beta(\sigma\mu_k)^{-1/2} \\
&= \sqrt{\mu_k}\left(\frac{\delta}{2} + \frac{\beta}{\sqrt{\sigma}} + \beta\sqrt{\sigma}\right).
\end{aligned}$$

With σ given by (4.2), some tedious calculation verifies

$$\frac{\delta}{2} + \frac{\beta}{\sqrt{\sigma}} + \beta\sqrt{\sigma} \le \delta\sqrt{\sigma}$$

and hence

$$\|v_k - v(\sigma\mu_k)\|_V \le \delta\sqrt{\mu_k\sigma}.$$

Now the corrector step, which is a Newton step for the problem $F(v;\sigma\mu_k) = 0$, leads to

$$\begin{aligned}\|v_{k+1} - v(\mu_{k+1})\|_V &\le \frac{\omega(\mu)}{2}\|v_k - v(\mu_{k+1})\|_V^2 \le \frac{\omega(\mu)}{2}\delta^2\mu_{k+1} \\ &\le \frac{\delta}{2}\sqrt{\mu_{k+1}},\end{aligned}$$

which completes the induction. As for the convergence of the iterates, we observe that by Theorem 3.10

$$\begin{aligned}\|v_k - v(0)\|_V &\le \|v_k - v(\mu_k)\|_V + \|v(\mu_k) - v(0)\|_V \\ &\le \frac{\delta}{2}\sqrt{\mu_k} + \text{const}\,\sqrt{\mu_k} \\ &\le \text{const}\,\sigma^{k/2}\sqrt{\mu_0},\end{aligned}$$

which proves linear convergence of $v_k \to v(0)$. ∎

Acknowledgment. The author gratefully acknowledges careful reading of the manuscript by A. Schiela.

REFERENCES

[1] W. Alt and K. Malanowski, *The Lagrange–Newton method for state constrained optimal control problems*, Comput. Optim. Appl., 4 (1995), pp. 217–239.

[2] U. Ascher, J. Christiansen, and R. Russell, *Collocation software for boundary-value ODEs*, ACM Trans. Math. Software, 7 (1981), pp. 209–222.

[3] U. Ascher, R. Mattheij, and R. Russell, *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[4] G. Bader and U. Ascher, *A new basis implementation for a mixed order boundary value ODE solver*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 483–500.

[5] M. Bergounioux, M. Haddou, M. Hintermüller, and K. Kunisch, *A comparison of a Moreau–Yosida-based active set strategy and interior point methods for constrained optimal control problems*, SIAM J. Optim., 11 (2000), pp. 495–521.

[6] H. Bock, *Numerische Behandlung von zustandsbeschränkten und Chebychef-Steuerungs-Problemen*, Technical report, Carl-Cranz-Gesellschaft, Oberpfaffenhofen, 1981.

[7] H. Bock and K.-J. Plitt, *A multiple shooting algorithm for direct solution of optimal control problems*, in Proceedings of the 9th IFAC World Congress, Budapest, Pergamon Press, Elmsford, NY, 1984.

[8] D. Braess, *Finite Elements*, 2nd ed., Cambridge University Press, Cambridge, UK, 2001.

[9] D. Braess and C. Blömer, *A multigrid method for a parameter dependent problem in solid mechanics*, Numer. Math., 57 (1990), pp. 747–761.

[10] F. Brezzi, *On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers*, Rev. Franc. Automat. Inform. Rech. Oper., 8 (1974), pp. 129–151.

[11] J. Burke and S. Xu, *The global linear convergence of a noninterior path-following algorithm for linear complementarity problems*, Math. Oper. Res., 23 (1998), pp. 719–734.

[12] B. Chen and N. Xiu, *A global linear and local quadratic noninterior continuation method for nonlinear complementarity problems based on Chen–Mangasarian smoothing functions*, SIAM J. Optim., 9 (1999), pp. 605–623.
[13] C. Chen and O. L. Mangasarian, *A class of smoothing functions for nonlinear and mixed complementarity problems*, Comput. Optim. Appl., 5 (1996), pp. 97–138.
[14] P. Deuflhard, *A modified Newton method for the solution of ill-conditioned systems of nonlinear equations with application to multiple shooting*, Numer. Math., 22 (1974), pp. 289–315.
[15] P. Deuflhard, *A stepsize control for continuation methods and its special application to multiple shooting techniques*, Numer. Math., 33 (1979), pp. 115–146.
[16] P. Deuflhard and F. Potra, *Asymptotic mesh independence of Newton–Galerkin methods via a refined Mysovskii theorem*, SIAM J. Numer. Anal., 29 (1992), pp. 1395–1412.
[17] A. Dontchev, W. Hager, and V. Veliov, *Uniform convergence and mesh independence of Newton's method for discretized variational problems*, SIAM J. Control Optim., 39 (2000), pp. 961–980.
[18] A. Fischer, *A special Newton-type optimization method*, Optimization, 24 (1992), pp. 269–284.
[19] A. Forsgren, P. Gill, and M. Wright, *Interior methods for nonlinear optimization*, SIAM Rev., 44 (2002), pp. 525–597.
[20] N. Gould and P. Toint, *Numerical methods for large-scale non-convex quadratic programming*, in Trends in Industrial and Applied Mathematics, A. Siddiqi and M. Kocvara, eds., Kluwer Academic, Norwell, MA, 2002, pp. 149–179.
[21] W. Hager, *Runge–Kutta methods in optimal control and the transformed adjoint system*, Numer. Math., 87 (2000), pp. 247–282.
[22] M. Heinkenschloss and M. Tröltzsch, *Analysis of the Lagrange–SQP–Newton method for the control of a phase field equation*, Control Cybernet., 28 (1999), pp. 177–211.
[23] M. Hintermüller and M. Hinze, *A SQP-Semi-Smooth Newton-Type Algorithm Applied to Control of the Instationary Navier–Stokes System Subject to Control Constraints*, Technical report TR 03-11, Department of Computational and Applied Mathematics, Rice University, Houston, TX, 2003.
[24] S. Ito, C. Kelley, and E. Sachs, *Inexact primal-dual interior point iteration for linear programs in function spaces*, Comput. Optim. Appl., 4 (1995), pp. 189–201.
[25] C. Kanzow, *Some noninterior continuation methods for linear complementarity problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 851–868.
[26] S. Kurcyusz, *On the existence and nonexistence of Lagrange multipliers in Banach spaces*, J. Optim. Theory Appl., 20 (1976), pp. 81–110.
[27] K. Machielsen, *Numerical Solution of Optimal Control Problems with State Constraints by Sequential Quadratic Programming in Function Space*, CWI Tract 53, Centrum voor Wiskunde en Informatica, Amsterdam, 1988.
[28] H. Maurer, *First and second order sufficient optimality conditions in mathematical programming and optimal control*, Math. Program. Study, 14 (1981), pp. 163–177.
[29] S. Robinson, *Generalized equations*, in Mathematical Programming. The State of the Art, A. Bachem, M. Grötschel, and B. Korte, eds., Springer, Berlin, 1983, pp. 346–367.
[30] V. Schulz, *Solving discretized optimization problems by partially reduced SQP methods*, Comput. Vis. Sci., 1 (1998), pp. 83–96.
[31] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, Springer, Berlin, 1993.
[32] F. Tröltzsch, *An SQP method for the optimal control of a nonlinear heat equation*, Control Cybernet., 23 (1994), pp. 267–288.
[33] F. Tröltzsch, *On the Lagrange–Newton–SQP method for the optimal control of semilinear parabolic equations*, SIAM J. Control Optim., 38 (1999), pp. 294–312.
[34] M. Ulbrich, *Semismooth Newton methods for operator equations in function spaces*, SIAM J. Optim., 13 (2003), pp. 805–842.
[35] M. Ulbrich and S. Ulbrich, *Superlinear convergence of affine-scaling interior-point Newton methods for infinite-dimensional nonlinear problems with pointwise bounds*, SIAM J. Control Optim., 38 (2000), pp. 1938–1984.
[36] M. Ulbrich, S. Ulbrich, and M. Heinkenschloss, *Global convergence of trust-region interior-point algorithms for infinite-dimensional nonconvex minimization subject to pointwise bounds*, SIAM J. Control Optim., 37 (1999), pp. 731–764.
[37] O. von Stryk, *Numerical solution of optimal control problems by direct collocation*, in Optimal Control. Calculus of Variations, Optimal Control Theory and Numerical Methods, Internat. Ser. Numer. Math. 111, R. Bulirsch, ed., Birkhäuser, Boston, 1993, pp. 129–143.
[38] M. Weiser, *Function Space Complementarity Methods for Optimal Control Problems*, Ph.D. thesis, Free University of Berlin, Berlin, 2001.

[39] M. Weiser and P. Deuflhard, *The Central Path towards the Numerical Solution of Optimal Control Problems*, ZIB report 01-12, Zuse Institute Berlin, 2001.
[40] E. Yildirim and S. Wright, *Warm-start strategies in interior-point methods for linear programming*, SIAM J. Optim., 12 (2002), pp. 782–810.
[41] E. Zeidler, *Nonlinear Functional Analysis and Its Applications*, Vol. I, Springer, New York, 1986.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 5, pp. 1787–1810

A SEARCH ALGORITHM FOR A CLASS OF OPTIMAL FINITE-PRECISION CONTROLLER REALIZATION PROBLEMS WITH SADDLE POINTS*

JUN WU†, SHENG CHEN‡, GANG LI§, AND JIAN CHU†

Abstract. With game theory, we review the optimal digital controller realization problems that maximize a finite word length (FWL) closed-loop stability measure. For a large class of these optimal FWL controller realization problems which have saddle points, a minimax-based search algorithm is derived for finding a global optimal solution. The algorithm consists of two stages. In the first stage, the closed form of a transformation set is constructed which contains global optimal solutions. In the second stage, a subgradient approach searches this transformation set to obtain a global optimal solution. This algorithm does not suffer from the usual drawbacks associated with using direct numerical optimization methods to tackle these FWL realization problems. Furthermore, for a small class of optimal FWL controller realization problems which have no saddle point, the proposed algorithm also provides useful information to help solve them.

Key words. closed-loop stability, digital controller, finite word length, game theory, optimization, saddle points

AMS subject classifications. 93D99, 91A80, 15A18, 90C47, 90C90, 90C30

DOI. 10.1137/S0363012903435084

1. Introduction.

There has been a growing awareness that finite-precision controller implementation can have a serious influence on the actual performance of a digital closed-loop control system [1], [2], [3]. Due to the finite word length (FWL) errors, a casual controller implementation may degrade the designed closed-loop performance, or even destabilize the designed stable closed-loop system, if the controller implementation structure is not carefully chosen. The FWL effect has become more critical with the growing popularity of robust controller design methods which focus only on dealing with large plant uncertainty and result in controllers of much higher order and complexity than traditional classical control [2]. There are generally two types of FWL errors in the digital controller implementation. The first one is the rounding errors that occur in arithmetic operations [4], [5], and the second one is the controller parameter representation errors which have critical influence on closed-loop stability [6], [7], [8], [9], [10], [11], [12]. Typically, these two types of errors are investigated separately for the reason of mathematical tractability.

In general, there exist two different strategies, called the direct and indirect strategies, for constructing digital controllers that can tolerate FWL implementation errors. For the indirect strategy, the transfer function of the digital controller has been de-

*Received by the editors September 25, 2003; accepted for publication (in revised form) May 17, 2005; published electronically November 23, 2005.

http://www.siam.org/journals/sicon/44-5/43508.html

†National Key Laboratory of Industrial Control Technology, Institute of Advanced Process Control, Zhejiang University, Hangzhou, 310027, People's Republic of China. These authors were supported by the National Natural Science Foundation of China (grants 60174026, 60374002, and 60421002) and 973 program of China (grant 2002CB312200) (jwu@iipc.zju.edu.cn, chuj@iipc.zju.edu.cn).

‡Corresponding author. School of Electronics and Computer Science, University of Southampton, Highfield, Southampton SO17 1BJ, UK (sqc@ecs.soton.ac.uk). This author was supported by the United Kingdom Royal Academy of Engineering.

§School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore (egli@ntu.edu.sg).

signed by some controller synthesis methods. It is well known that a transfer function can be fulfilled with different realizations, and different realizations possess different degrees of robustness to FWL errors. This property can be utilized to select "optimal" realizations that optimize some FWL performance measures. Various FWL performance measures have been investigated, and these include the averaged roundoff noise gain [5], the complex stability radius measure [6], the transfer function sensitivity measure [7], the l_1-based stability measure [8], the Frobenius-norm pole sensitivity measure [9], and the 1-norm pole sensitivity measure [10], [11]. In the direct strategy, controller design involves explicitly the considerations of FWL implementation. By extending the standard H_∞ control design to include FWL controller parameter perturbations, the work of [12] developed a Riccati inequality approach, which directly obtains optimal controller realizations satisfying both the H_∞ robustness and FWL closed-loop stability requirements. Similarly, by extending the standard linear quadratic Gaussian (LQG) control design to include the effects of FWL roundoff noise, the work of [4] developed a FWL-LQG controller design method. The direct strategy appears to be better than the indirect strategy, since the former does not make specific assumptions on the controller, and in theory it should be a preferred approach. However, except for a few methods, such as H_∞ and LQG, it is very difficult to extend various controller design methods to this direct strategy. But this difficulty does not exist in the indirect strategy, where controller synthesis and controller realization are two separate steps. Various existing controller design methods can be used to attain a transfer function or an initial realization of the controller, which can then be optimized to satisfy FWL implementation requirements.

This paper adopts the indirect strategy with the Frobenius-norm pole sensitivity measure proposed in [9]. Our motivation is as follows. The Frobenius-norm pole sensitivity measure was derived in [9], and the optimal controller realization problem was defined as the maximization of this measure over all the possible controller realizations. An analytical solution to this class of optimal realization problems was attempted in [9]. However, it was pointed out that the conditions presented in [9] are not sufficient to provide an optimal realization [13]. Consequently, the solution expression presented in [9] is in general a suboptimal solution, and numerical optimization methods have to be adopted [14] to find optimal solutions. Since these optimal FWL realization problems are highly complicated nonlinear and nonconvex optimization problems, especially when the order of the controller is large, a direct numerical optimization is computationally very expensive. Moreover, chances of search being trapped at some bad local solutions increase for large-scale problems, and it is impossible to tell whether or not a solution obtained is a global optimum. In this paper, these optimal FWL controller realization problems are reviewed with game theory [15], [16]. They are consequently divided into two types: optimization problems which have saddle points and optimization problems which do not have a saddle point.

For the class of optimal FWL realization problems with saddle points, this paper derives a minimax-based search algorithm for finding global optimal solutions. Our search algorithm is computationally much more efficient than usual numerical optimization for tackling this class of complicated optimization problems. Moreover, when this algorithm attains a solution, it is guaranteed to be a global optimal realization. Comments are made regarding why in practice the class of these optimization problems with saddle points is much larger than the class having no saddle point. It is shown that our proposed search algorithm is also useful in helping to solve the small class of these optimal FWL realization problems which have no saddle point. The

remainder of the paper is organized as follows. Section 2 defines the optimal FWL controller realization problem considered in this study and introduces some necessary mathematical preliminaries. In section 3, the proposed two-stage search algorithm is derived. Section 4 discusses the practical value of this algorithm. Section 5 presents several design examples, and the paper concludes with section 6.

2. Problem definition and preliminaries. For a complex-valued matrix $\mathbf{M} = [m_{ij}]$, $\mathbf{M}^T$ is the transposed matrix of $\mathbf{M}$, $\mathbf{M}^H$ is the Hermitian adjoint matrix of $\mathbf{M}$, $\mathbf{M}^*$ is conjugate to $\mathbf{M}$,

$$\|\mathbf{M}\|_{\max} \triangleq \max_{i,j} |m_{ij}|, \tag{1}$$

and the Frobenius norm is defined as

$$\|\mathbf{M}\|_F \triangleq \left(\sum_{i,j} |m_{ij}|^2 \right)^{1/2}. \tag{2}$$

Let $\mathrm{Vec}(\cdot)$ be the column stacking operator such that $\mathrm{Vec}(\mathbf{M})$ is a vector. For a real-valued positive semidefinite matrix $\mathbf{D} \geq 0$, the matrix $\mathbf{D}^{1/2}$ satisfies $\mathbf{D}^{1/2}(\mathbf{D}^{1/2})^T = \mathbf{D}$. For two real-valued matrices $\mathbf{M} = [m_{ij}]$ and $\mathbf{N} = [n_{ij}]$ of the same dimension, denote

$$\langle \mathbf{M}, \mathbf{N} \rangle = \sum_{i,j} m_{ij} n_{ij}. \tag{3}$$

2.1. Problem definition. Consider the discrete-time closed-loop control system, consisting of a linear time-invariant plant $P(z)$ and a digital controller $C(z)$. The plant model $P(z)$ is assumed to be strictly proper with a state-space description

$$\begin{cases} \mathbf{x}_P(t+1) = \mathbf{A}_P \mathbf{x}_P(t) + \mathbf{B}_P \mathbf{u}(t), \\ \mathbf{z}(t) = \mathbf{C}_P \mathbf{x}_P(t), \end{cases} \tag{4}$$

where $\mathbf{A}_P \in \mathcal{R}^{m \times m}$, $\mathbf{B}_P \in \mathcal{R}^{m \times l}$, and $\mathbf{C}_P \in \mathcal{R}^{q \times m}$. The digital controller $C(z)$ is described by

$$\begin{cases} \mathbf{x}_C(t+1) = \mathbf{A}_C \mathbf{x}_C(t) + \mathbf{B}_C \mathbf{z}(t), \\ \mathbf{u}(t) = \mathbf{C}_C \mathbf{x}_C(t) + \mathbf{D}_C \mathbf{z}(t), \end{cases} \tag{5}$$

with $\mathbf{A}_C \in \mathcal{R}^{n \times n}$, $\mathbf{B}_C \in \mathcal{R}^{n \times q}$, $\mathbf{C}_C \in \mathcal{R}^{l \times n}$, and $\mathbf{D}_C \in \mathcal{R}^{l \times q}$. Denote the *realization* of $C(z)$ as

$$\mathbf{X} \triangleq \begin{bmatrix} \mathbf{D}_C & \mathbf{C}_C \\ \mathbf{B}_C & \mathbf{A}_C \end{bmatrix}. \tag{6}$$

Assume that an initial realization of $C(z)$,

$$\mathbf{X}_0 \triangleq \begin{bmatrix} \mathbf{D}_C^0 & \mathbf{C}_C^0 \\ \mathbf{B}_C^0 & \mathbf{A}_C^0 \end{bmatrix}, \tag{7}$$

has been given by some controller synthesis method. Then all the realizations of $C(z)$ form a set

$$\mathcal{S}_C \triangleq \left\{ \mathbf{X} : \mathbf{X} = \mathbf{X}(\mathbf{T}) = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-1} \end{bmatrix} \mathbf{X}_0 \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T} \end{bmatrix} \right\}, \tag{8}$$

where the transformation $\mathbf{T} \in \mathcal{R}^{n\times n}$ is an arbitrary nonsingular matrix, and $\mathbf{0}$ and $\mathbf{I}$ denote the zero and identity matrices of appropriate dimensions, respectively. $\mathcal{S}_C$ is not a convex set, as

$$\lambda \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_1^{-1} \end{bmatrix} \mathbf{X}_0 \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_1 \end{bmatrix} + (1-\lambda) \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_2^{-1} \end{bmatrix} \mathbf{X}_0 \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_2 \end{bmatrix} \tag{9}$$

may not belong to $\mathcal{S}_C$ for any nonsingular $\mathbf{T}_1, \mathbf{T}_2 \in \mathcal{R}^{n\times n}$ and $0 < \lambda < 1$. The stability of the closed-loop control system depends on the eigenvalues of the closed-loop transition matrix

$$\begin{aligned} \overline{\mathbf{A}}(\mathbf{X}) &= \begin{bmatrix} \mathbf{A}_P + \mathbf{B}_P\mathbf{D}_C\mathbf{C}_P & \mathbf{B}_P\mathbf{C}_C \\ \mathbf{B}_C\mathbf{C}_P & \mathbf{A}_C \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{A}_P & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{B}_P & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \mathbf{X} \begin{bmatrix} \mathbf{C}_P & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \triangleq \mathbf{M}_0 + \mathbf{M}_1\mathbf{X}\mathbf{M}_2. \end{aligned} \tag{10}$$

All the different realizations $\mathbf{X}$ in $\mathcal{S}_C$ have exactly the same set of closed-loop poles if they are implemented with infinite precision. Since the closed-loop system has been designed to be stable, all the eigenvalues $\lambda_k(\overline{\mathbf{A}}(\mathbf{X}))$, $1 \le k \le m+n$, of $\overline{\mathbf{A}}(\mathbf{X})$ are within the unit disk.

When $\mathbf{X}$ is implemented with an FWL digital processor of fixed-point format, it is perturbed to $\mathbf{X} + \Delta\mathbf{X}$. Each element of $\Delta\mathbf{X}$ is bounded by $\pm\varepsilon$; that is, $\|\Delta\mathbf{X}\|_{\max} \le \varepsilon$, where ε is the maximum representation error of the digital processor. With the perturbation $\Delta\mathbf{X}$, $\lambda_k(\overline{\mathbf{A}}(\mathbf{X}))$ is moved to $\lambda_k(\overline{\mathbf{A}}(\mathbf{X}+\Delta\mathbf{X}))$. If an eigenvalue of $\overline{\mathbf{A}}(\mathbf{X}+\Delta\mathbf{X})$ is outside the open unit disk, the closed-loop system, designed to be stable, becomes unstable with the finite-precision implemented $\mathbf{X}$. It is therefore critical to know when the FWL error will cause closed-loop instability. This means that we would like to know the largest open "hypercube" in the perturbation space within which the closed-loop system remains stable. The size of this perturbation hypercube quantifies the FWL characteristics of $\mathbf{X}$ and is therefore a true FWL closed-loop stability measure for $\mathbf{X}$ [17].

Computing the size of this largest stable perturbation hypercube, however, is an unsolved open problem. An approximation to this true FWL closed-loop stability measure is the following Frobenius-norm pole sensitivity measure defined in [9]:

$$f(\mathbf{X}) \triangleq \min_{k\in\{1,\ldots,m+n\}} \frac{1 - |\lambda_k(\overline{\mathbf{A}}(\mathbf{X}))|}{\sqrt{(l+n)(q+n)}\left\|\frac{\partial\lambda_k(\overline{\mathbf{A}}(\mathbf{X}))}{\partial\mathbf{X}}\right\|_F}. \tag{11}$$

Rigorous discussions regarding the rationality of $f(\mathbf{X})$ as an FWL closed-loop stability measure can be found in [9], [11]. Basically, under some mild assumptions and using a first-order approximation, it can be shown that the closed-loop system remains stable if $\|\Delta\mathbf{X}\|_{\max} < f(\mathbf{X})$. It has been argued in [18] that estimates obtained from first-order perturbation theory are often more realistic than rigorous bounds obtained by other means. Thus, the larger $f(\mathbf{X})$ is, the larger an FWL error $\Delta\mathbf{X}$ that the closed-loop system can tolerate. Moreover, $f(\mathbf{X})$ is computationally tractable, as is summarized in the following lemma given by [19].

LEMMA 1. *Let* $\overline{\mathbf{A}}(\mathbf{X}) = \mathbf{M}_0 + \mathbf{M}_1\mathbf{X}\mathbf{M}_2$ *given in* (10) *be diagonalizable. Denote* $\mathbf{p}_k$ *a right eigenvector of* $\overline{\mathbf{A}}(\mathbf{X})$ *corresponding to the eigenvalue* $\lambda_k(\overline{\mathbf{A}}(\mathbf{X}))$. *The*

reciprocal left eigenvector $\mathbf{y}_k$ *related to* $\mathbf{p}_k$ *is obtained from* $[\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{m+n}] = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_{m+n}]^{-H}$. *Then*

$$\frac{\partial \lambda_k(\overline{\mathbf{A}}(\mathbf{X}))}{\partial \mathbf{X}} = \mathbf{M}_1^T \mathbf{y}_k^* \mathbf{p}_k^T \mathbf{M}_2^T \quad \forall k \in \{1, \ldots, m+n\}. \tag{12}$$

As different controller realizations $\mathbf{X}$ result in different values of $f(\mathbf{X})$, it is natural to search for "optimal" controller realizations that maximize the measure defined in (11). This leads to the following optimal FWL realization problem [9]:

$$\upsilon \triangleq \max_{\mathbf{X} \in \mathcal{S}_C} f(\mathbf{X}). \tag{13}$$

Numerical optimization methods have been used to attain solutions of this optimal realization problem (e.g., [14]). In general, the optimization problem (13) is highly nonlinear and nonconvex. Thus, numerical optimization methods do not guarantee attaining a global optimal solution and suffer from high costs, particularly for large-scale systems.

Now, let us define

$$g(\mathbf{X}, k) \triangleq \frac{1 - |\lambda_k(\overline{\mathbf{A}}(\mathbf{X}))|}{\sqrt{(l+n)(q+n)} \left\| \frac{\partial \lambda_k(\overline{\mathbf{A}}(\mathbf{X}))}{\partial \mathbf{X}} \right\|_F}. \tag{14}$$

Obviously, the optimal FWL realization problem (13) can be viewed as

$$\upsilon = \max_{\mathbf{X} \in \mathcal{S}_C} \min_{k \in \{1, \ldots, m+n\}} g(\mathbf{X}, k). \tag{15}$$

2.2. Saddle points and minimax theorem. This subsection introduces without proofs some properties of saddle points and the minimax theorem, which are useful in solving the optimization problem (15). The detailed discussion of this topic can be found in the standard game theory textbooks, such as [15], [16].

DEFINITION 1. $(\mathbf{X}', k') \in \mathcal{S}_C \times \{1, \ldots, m+n\}$ *is said to be a saddle point of* $g(\mathbf{X}, k)$ *if*

$$g(\mathbf{X}, k') \le g(\mathbf{X}', k') \le g(\mathbf{X}', k) \quad \forall \mathbf{X} \in \mathcal{S}_C, \ \forall k \in \{1, \ldots, m+n\}. \tag{16}$$

THEOREM 1. *If both* $(\mathbf{X}', k')$ *and* $(\mathbf{X}'', k'')$ *are saddle points of* $g(\mathbf{X}, k)$, *then*

$$g(\mathbf{X}', k') = g(\mathbf{X}'', k''). \tag{17}$$

The following theorem is the well-known minimax theorem in game theory.

THEOREM 2. *If and only if there exists at least a saddle point* $(\mathbf{X}', k')$ *of* $g(\mathbf{X}, k)$, *then*

$$\max_{\mathbf{X} \in \mathcal{S}_C} \min_{k \in \{1, \ldots, m+n\}} g(\mathbf{X}, k) = \min_{k \in \{1, \ldots, m+n\}} \max_{\mathbf{X} \in \mathcal{S}_C} g(\mathbf{X}, k) = g(\mathbf{X}', k'). \tag{18}$$

A direct corollary of Theorem 2 is stated as follows.

COROLLARY 1. *If* $g(\mathbf{X}, k)$ *has no saddle point, then*

$$\max_{\mathbf{X} \in \mathcal{S}_C} \min_{k \in \{1, \ldots, m+n\}} g(\mathbf{X}, k) < \min_{k \in \{1, \ldots, m+n\}} \max_{\mathbf{X} \in \mathcal{S}_C} g(\mathbf{X}, k). \tag{19}$$

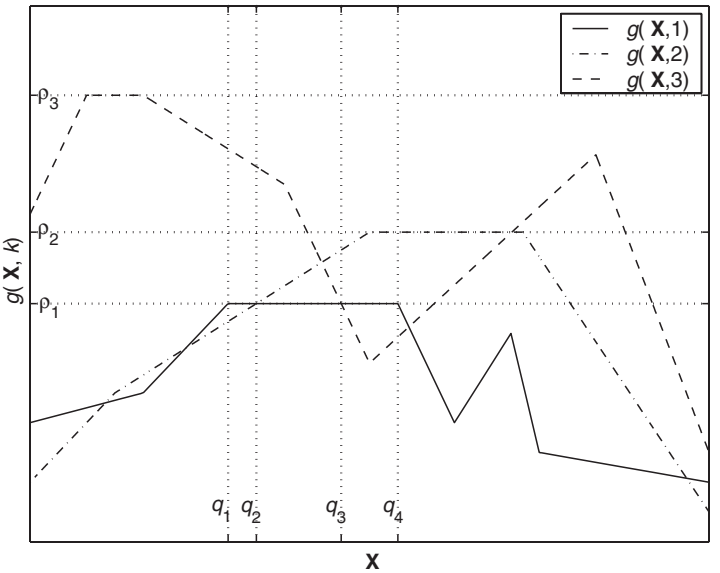


FIG. 1. *A simple illustration of ρ_k, $\mathcal{X}$, and saddle points.*

Theorems 1 and 2 show that for the optimal FWL realization problem (15) which has saddle points, any saddle point of $g(\mathbf{X}, k)$ is a global optimal solution of (15). Define

$$\rho_k \stackrel{\triangle}{=} \max_{\mathbf{X} \in \mathcal{S}_C} g(\mathbf{X}, k) \tag{20}$$

for $k \in \{1, \ldots, m+n\}$ and the index

$$k' = \arg \min_{k \in \{1,\ldots,m+n\}} \rho_k. \tag{21}$$

There exist an infinite number of $\mathbf{X} \in \mathcal{S}_C$ such that $g(\mathbf{X}, k') = \rho_{k'}$. Define

$$\mathcal{X} \stackrel{\triangle}{=} \{\mathbf{X} : g(\mathbf{X}, k') = \rho_{k'}, \mathbf{X} \in \mathcal{S}_C\}. \tag{22}$$

Figure 1 depicts a simple illustration for a case of ρ_k with $k \in \{1, 2, 3\}$. It is easily seen that in this case $\mathcal{X}$ is the segment between q_1 and q_4 on the $\mathbf{X}$-axis. It can also be observed in Figure 1 that the points between q_2 and q_3 (a subset of $\mathcal{X}$) are the realizations corresponding to saddle points. This observation accords with the following theorem, which provides a method for finding a saddle point.

THEOREM 3. *If and only if $\mathbf{X}' \in \mathcal{X}$ satisfies*

$$g(\mathbf{X}', k) \geq \rho_{k'} \quad \forall k \in \{1, \ldots, m+n\} \setminus \{k'\}, \tag{23}$$

then $(\mathbf{X}', k')$ is a saddle point of $g(\mathbf{X}, k)$.

3. Search algorithm. A main objective of this paper is how to find a global optimal solution to the optimal FWL realization problem (15) which has saddle points. In other words, we assume that there exist saddle points for $g(\mathbf{X}, k)$ in the problem (15). What happens if the problem has no saddle point and how to deal with it will be

discussed in section 4. Based on Theorem 3, a two-stage algorithm is developed to find a saddle point of the optimal FWL controller realization problem (15). The first stage focuses the attention on solving the optimization problem (20) for $k \in \{1, \ldots, m+n\}$, and the index k' and the closed-form expression of $\mathcal{X}$ are obtained in this stage. The second stage searches $\mathcal{X}$ for a controller realization $\mathbf{X}_{\text{opt}}$ that meets the condition $g(\mathbf{X}_{\text{opt}}, k) \geq \rho_{k'}\ \forall k \in \{1, \ldots, m+n\} \setminus \{k'\}$. Such an $\mathbf{X}_{\text{opt}}$ is a global optimal solution to the optimal FWL controller realization problem (13). We now discuss this two-stage algorithm in detail.

3.1. Stage 1 of the algorithm. It is known easily from (8) and (10) that

$$\overline{\mathbf{A}}(\mathbf{X}) = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-1} \end{bmatrix} \overline{\mathbf{A}}(\mathbf{X}_0) \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T} \end{bmatrix}. \tag{24}$$

This means that $\forall \mathbf{X} \in \mathcal{S}_C$, $\lambda_k(\overline{\mathbf{A}}(\mathbf{X})) = \lambda_k(\overline{\mathbf{A}}(\mathbf{X}_0))$. Thus, from (14), solving the maximization problem (20) is equivalent to solving the following minimization problem:

$$\eta_k \stackrel{\triangle}{=} \min_{\mathbf{X} \in \mathcal{S}_C} \left\| \frac{\partial \lambda_k(\overline{\mathbf{A}}(\mathbf{X}))}{\partial \mathbf{X}} \right\|_F. \tag{25}$$

Combining Lemma 1 with the definition of $\|\cdot\|_F$, one has

$$\left\| \frac{\partial \lambda_k(\overline{\mathbf{A}}(\mathbf{X}))}{\partial \mathbf{X}} \right\|_F = \|\mathbf{M}_1^T \mathbf{y}_k\|_F \, \|\mathbf{M}_2 \mathbf{p}_k\|_F \,. \tag{26}$$

Let $\mathbf{p}_k$ and $\mathbf{y}_k$ be partitioned into

$$\mathbf{p}_k = \begin{bmatrix} \mathbf{p}_k(1) \\ \mathbf{p}_k(2) \end{bmatrix}, \quad \mathbf{y}_k = \begin{bmatrix} \mathbf{y}_k(1) \\ \mathbf{y}_k(2) \end{bmatrix}, \quad \mathbf{p}_k(1), \mathbf{y}_k(1) \in \mathcal{C}^m, \quad \mathbf{p}_k(2), \mathbf{y}_k(2) \in \mathcal{C}^n. \tag{27}$$

Then it follows from (24) that

$$\begin{bmatrix} \mathbf{p}_k(1) \\ \mathbf{p}_k(2) \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{p}_{0k}(1) \\ \mathbf{p}_{0k}(2) \end{bmatrix}, \quad \begin{bmatrix} \mathbf{y}_k(1) \\ \mathbf{y}_k(2) \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^T \end{bmatrix} \begin{bmatrix} \mathbf{y}_{0k}(1) \\ \mathbf{y}_{0k}(2) \end{bmatrix}, \tag{28}$$

where $\left[\mathbf{p}_{0k}^T(1)\ \mathbf{p}_{0k}^T(2)\right]^T$ and $\left[\mathbf{y}_{0k}^T(1)\ \mathbf{y}_{0k}^T(2)\right]^T$ are the right and reciprocal left eigenvectors of $\overline{\mathbf{A}}(\mathbf{X}_0)$ corresponding to $\lambda_k(\overline{\mathbf{A}}(\mathbf{X}_0))$, respectively. Combining (10) and (26)–(28), we have

$$\begin{aligned} \left\| \frac{\partial \lambda_k(\overline{\mathbf{A}}(\mathbf{X}))}{\partial \mathbf{X}} \right\|_F^2 &= \|\mathbf{T}^{-1}\mathbf{p}_{0k}(2)\|_F^2 \|\mathbf{T}^T\mathbf{y}_{0k}(2)\|_F^2 \\ &\quad + \alpha_k^2 \|\mathbf{T}^T\mathbf{y}_{0k}(2)\|_F^2 + \beta_k^2 \|\mathbf{T}^{-1}\mathbf{p}_{0k}(2)\|_F^2 + \alpha_k^2\beta_k^2, \end{aligned} \tag{29}$$

where the constants $\alpha_k = \|\mathbf{C}_P \mathbf{p}_{0k}(1)\|_F$ and $\beta_k = \|\mathbf{B}_P^T \mathbf{y}_{0k}(1)\|_F$. It is easy to see that, in order to attain ρ_k, we need to minimize the function

$$\xi(\mathbf{T}, \alpha, \beta, \mathbf{p}, \mathbf{y}) \stackrel{\triangle}{=} \|\mathbf{T}^{-1}\mathbf{p}\|_F^2 \|\mathbf{T}^T\mathbf{y}\|_F^2 + \alpha^2 \|\mathbf{T}^T\mathbf{y}\|_F^2 + \beta^2 \|\mathbf{T}^{-1}\mathbf{p}\|_F^2 + \alpha^2\beta^2. \tag{30}$$

There are three different cases on minimizing $\xi(\mathbf{T}, \alpha, \beta, \mathbf{p}, \mathbf{y})$, depending on whether $\mathbf{p}$ and $\mathbf{y}$ are real-valued or complex-valued.

Case 1. $\mathbf{p}, \mathbf{y} \in \mathcal{R}^n$ and $\mathbf{y}^T\mathbf{p} \neq 0$.

Case 2. $\mathbf{p}, \mathbf{y} \in \mathcal{C}^n$ and $\det((\Upsilon(\mathbf{y}))^T\Upsilon(\mathbf{p})) > 0$, where

$$\Upsilon(\mathbf{y}) \triangleq [\mathrm{Re}(\mathbf{y})\ \mathrm{Im}(\mathbf{y})] \tag{31}$$

with $\mathrm{Re}(\mathbf{y})$ and $\mathrm{Im}(\mathbf{y})$ denoting the real and imaginary parts of $\mathbf{y}$, respectively.

Case 3. $\mathbf{p}, \mathbf{y} \in \mathcal{C}^n$ and $\det((\Upsilon(\mathbf{y}))^T\Upsilon(\mathbf{p})) < 0$.

Let $\mathbf{e}_i$ denote the ith coordinate vector, and define

$$\mathbf{r} \triangleq \begin{cases} \mathbf{y} & \text{for Case 2,} \\ \mathbf{y}^* & \text{for Case 3.} \end{cases} \tag{32}$$

The following theorem gives the results on minimizing $\xi(\mathbf{T}, \alpha, \beta, \mathbf{p}, \mathbf{y})$ for Cases 2 and 3. Case 1 is much simpler than Cases 2 and 3, and the result for Case 1 can easily be obtained in a similar way.

THEOREM 4. *Given positive* $\alpha, \beta \in \mathcal{R}$, $\mathbf{p}$ *and* $\mathbf{y}$ *being of Case* 2 *or* 3, *we have*

$$\min_{\substack{\mathbf{T}\in\mathcal{R}^{n\times n} \\ \det \mathbf{T}\neq 0}} \xi(\mathbf{T}, \alpha, \beta, \mathbf{p}, \mathbf{y}) = (|\mathbf{r}^H\mathbf{p}| + \alpha\beta)^2, \tag{33}$$

and $\xi(\mathbf{T}, \alpha, \beta, \mathbf{p}, \mathbf{y})$ *achieves the minimum if and only if*

$$\mathbf{T} = \mathbf{Q}\begin{bmatrix} \mathbf{H}^{1/2} & \mathbf{0} \\ \mathbf{F}(\mathbf{H}^{1/2})^{-T} & \mathbf{\Omega} \end{bmatrix}\mathbf{V}, \tag{34}$$

where $\mathbf{V} \in \mathcal{R}^{n\times n}$ *is an arbitrary orthogonal matrix, and* $\mathbf{\Omega} \in \mathcal{R}^{(n-2)\times(n-2)}$ *is an arbitrary nonsingular matrix; the orthogonal matrix* $\mathbf{Q}$ *can be obtained from the QR factorization of* $\Upsilon(\mathbf{r})$, *that is,*

$$\Upsilon(\mathbf{r}) = \mathbf{Q}\begin{bmatrix} \gamma_{11} & 0 & 0 & \cdots & 0 \\ \gamma_{12} & \gamma_{22} & 0 & \cdots & 0 \end{bmatrix}^T; \tag{35}$$

and the matrices $\mathbf{H}$ *and* $\mathbf{F}$ *are determined by*

$$\mathbf{H} = \frac{\beta}{\alpha}\begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \end{bmatrix}^{-T}(\Upsilon(\mathbf{r}))^T\Upsilon(\mathbf{p})\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}\begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \end{bmatrix}^{-1} \tag{36}$$

and

$$\mathbf{F} = \frac{\beta}{\alpha}\begin{bmatrix} \mathbf{e}_3^T \\ \vdots \\ \mathbf{e}_n^T \end{bmatrix}\mathbf{Q}^T\Upsilon(\mathbf{p})\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}\begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \end{bmatrix}^{-1} \tag{37}$$

with $\theta \in [0, 2\pi)$ *which is solved from*

$$\begin{cases} \tan\theta = \frac{a_{21}-a_{12}}{a_{11}+a_{22}}, \\ a_{11}\cos\theta - a_{12}\sin\theta > 0 \end{cases} \tag{38}$$

and

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \triangleq (\Upsilon(\mathbf{r}))^T\Upsilon(\mathbf{p}). \tag{39}$$

Proof. See the appendix. ∎

Using Theorem 4, the single-pole peak ρ_k for $k \in \{1, \ldots, m+n\}$ can be computed. For example, when $\mathbf{p}_{0k}(2), \mathbf{y}_{0k}(2) \in \mathcal{C}^n$, and $\det((\Upsilon(\mathbf{y}_{0k}(2)))^T \Upsilon(\mathbf{p}_{0k}(2))) > 0$, we have

$$\rho_k = \frac{1 - |\lambda_k(\overline{\mathbf{A}}(\mathbf{X}_0))|}{\sqrt{(l+n)(q+n)}(|\mathbf{y}_{0k}^H(2)\mathbf{p}_{0k}(2)| + \|\mathbf{C}_P\mathbf{p}_{0k}(1)\|_F \|\mathbf{B}_P^T\mathbf{y}_{0k}(1)\|_F)}. \tag{40}$$

Thus, the index k' is readily given from $\rho_{k'} = \min_{k\in\{1,\ldots,m+n\}} \rho_k$. In addition, Theorem 4 with (34)–(39) provides the closed-form transformation set

$$\mathcal{T} \stackrel{\triangle}{=} \left\{\mathbf{T} : g(\mathbf{X}(\mathbf{T}), k') = \rho_{k'}, \mathbf{T} \in \mathcal{R}^{n\times n}, \det \mathbf{T} \neq 0\right\}. \tag{41}$$

Since $\mathbf{X}$ depends on $\mathbf{T}$ as is defined in (8), the realization set $\mathcal{X}$ given in (22) is defined on the transformation set $\mathcal{T}$ as

$$\mathcal{X} = \left\{\mathbf{X} : \mathbf{X} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-1} \end{bmatrix} \mathbf{X}_0 \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T} \end{bmatrix}, \mathbf{T} \in \mathcal{T}\right\}. \tag{42}$$

3.2. Stage 2 of the algorithm. This stage searches in $\mathcal{T}$ for an optimal transformation $\mathbf{T}_{\text{opt}}$ that satisfies $g(\mathbf{X}(\mathbf{T}_{\text{opt}}), k) \geq \rho_{k'}\ \forall k \in \{1, \ldots, m+n\}\backslash\{k'\}$. According to Theorem 3, the corresponding realization $\mathbf{X}_{\text{opt}} = \mathbf{X}(\mathbf{T}_{\text{opt}})$ is a global optimal solution for the optimal realization problem (13). Without any loss of generality, we will assume that $\mathbf{p}_{k'}$ and $\mathbf{y}_{k'}$ is of Case 2. From Theorem 4, the transformation set (41) is specified by

$$\mathcal{T} = \left\{\mathbf{T} : \mathbf{T} = \mathbf{Q} \begin{bmatrix} \mathbf{H}^{1/2} & \mathbf{0} \\ \mathbf{F}(\mathbf{H}^{1/2})^{-T} & \mathbf{\Omega} \end{bmatrix} \mathbf{V}\right\}, \tag{43}$$

where $\mathbf{Q}$, $\mathbf{H}$, and $\mathbf{F}$ are determined in Theorem 4 by setting $\alpha = \|\mathbf{C}_P\mathbf{p}_{0k'}(1)\|_F$, $\beta = \|\mathbf{B}_P^T\mathbf{y}_{0k'}(1)\|_F$, $\mathbf{p} = \mathbf{p}_{0k'}(2)$, and $\mathbf{r} = \mathbf{y} = \mathbf{y}_{0k'}(2)$, $\mathbf{\Omega} \in \mathcal{R}^{(n-2)\times(n-2)}$ is an arbitrary nonsingular matrix, and $\mathbf{V} \in \mathcal{R}^{n\times n}$ is an arbitrary orthogonal matrix. From (14), (29), and the definition of $\|\cdot\|_F$, it can be seen that $g(\mathbf{X}(\mathbf{T}), k) = g(\mathbf{X}(\mathbf{TV}), k)$ for any orthogonal $\mathbf{V} \in \mathcal{R}^{n\times n}$ and nonsingular $\mathbf{T} \in \mathcal{R}^{n\times n}$. This means that $\mathbf{V}$ plays no role in computing the value of $g(\mathbf{X}, k)$, and hence we simply set $\mathbf{V} = \mathbf{I}$. Thus we explore only those

$$\mathbf{T} = \mathbf{T}(\mathbf{\Omega}) = \mathbf{Q} \begin{bmatrix} \mathbf{H}^{1/2} & \mathbf{0} \\ \mathbf{F}(\mathbf{H}^{1/2})^{-T} & \mathbf{\Omega} \end{bmatrix}, \tag{44}$$

and the objective becomes to search for a nonsingular $\mathbf{\Omega}_{\text{opt}} \in \mathcal{R}^{(n-2)\times(n-2)}$ such that $g(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_{\text{opt}})), k) \geq \rho_{k'}\ \forall k \in \{1, \ldots, m+n\} \setminus \{k'\}$. The detailed search procedure is as follows.

Initialization: Arbitrarily select a nonsingular $\mathbf{\Omega} \in \mathcal{R}^{(n-2)\times(n-2)}$ to obtain an initial point $\mathbf{X}(\mathbf{T}(\mathbf{\Omega}))$, let N be a large enough integer and τ a small positive number, and set $N_t = 1$.

Step 1: Find out

$$e = \arg \min_{k\in\{1,\ldots,m+n\}} g(\mathbf{X}, k).$$

If $g(\mathbf{X}, e) = \rho_{k'}$, which means that (23) holds, then $\mathbf{\Omega}_{\text{opt}} = \mathbf{\Omega}$ and terminate the routine. If $g(\mathbf{X}, e) > \rho_{k'}$ but $N_t \geq N$, which means that no saddle point is found after a large number of iterations, then the routine is also terminated for practical consideration.

Step 2: $\mathbf{\Omega} = \mathbf{\Omega} + \tau \frac{\partial g(\mathbf{X},e)}{\partial \mathbf{\Omega}} \| \frac{\partial g(\mathbf{X},e)}{\partial \mathbf{\Omega}} \|_F^{-1}$, $N_t = N_t + 1$, and go to *Step* 1.

For calculating $\frac{\partial g(\mathbf{X}(\mathbf{T}(\mathbf{\Omega})),e)}{\partial \mathbf{\Omega}}$, let $\mathbf{e}_i$ denote the ith coordinate vector. The following well-known fact is useful: given any element y_{ij} in a nonsingular $\mathbf{Y} \in \mathcal{R}^{n\times n}$ with $i \in \{1,\ldots,n\}$ and $j \in \{1,\ldots,n\}$,

$$\frac{\partial \mathbf{Y}}{\partial y_{ij}} = \mathbf{e}_i \mathbf{e}_j^T \quad \text{and} \quad \frac{\partial \mathbf{Y}^{-1}}{\partial y_{ij}} = -\mathbf{Y}^{-1}\mathbf{e}_i \mathbf{e}_j^T \mathbf{Y}^{-1}. \tag{45}$$

From (10), (14), (28), and Lemma 1, we know that

$$g(\mathbf{X}(\mathbf{T}(\mathbf{\Omega})),e) = \frac{(1-|\lambda_e|)/\sqrt{(l+n)(q+n)}}{\left\| \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^T(\mathbf{\Omega}) \end{bmatrix} \mathbf{M}_1^T \mathbf{y}_{0e}^* \mathbf{p}_{0e}^T \mathbf{M}_2^T \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-T}(\mathbf{\Omega}) \end{bmatrix} \right\|_F}. \tag{46}$$

From (44), we have

$$\left\| \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^T(\mathbf{\Omega}) \end{bmatrix} \mathbf{M}_1^T \mathbf{y}_{0e}^* \mathbf{p}_{0e}^T \mathbf{M}_2^T \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^{-T}(\mathbf{\Omega}) \end{bmatrix} \right\|_F = \left\| \mathbf{U}_1^T \mathbf{\Phi}_e \mathbf{U}_2^{-T} \right\|_F, \tag{47}$$

where $\mathbf{U}_1$, $\mathbf{U}_2$, and $\mathbf{\Phi}_e$ are given, respectively, by ($\mathbf{I}$ in $\mathbf{U}_1$ and $\mathbf{U}_2$ have different dimensions)

$$\mathbf{U}_1 = \left[\begin{array}{c|cc} \mathbf{I} & \multicolumn{2}{c}{\mathbf{0}} \\ \hline \mathbf{0} & \mathbf{H}^{1/2} & \mathbf{0} \\ & \mathbf{F}(\mathbf{H}^{1/2})^{-T} & \mathbf{\Omega} \end{array} \right], \tag{48}$$

$$\mathbf{U}_2 = \left[\begin{array}{c|cc} \mathbf{I} & \multicolumn{2}{c}{\mathbf{0}} \\ \hline \mathbf{0} & \mathbf{H}^{1/2} & \mathbf{0} \\ & \mathbf{F}(\mathbf{H}^{1/2})^{-T} & \mathbf{\Omega} \end{array} \right], \tag{49}$$

$$\mathbf{\Phi}_e = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}^T \end{bmatrix} \mathbf{M}_1^T \mathbf{y}_{0e}^* \mathbf{p}_{0e}^T \mathbf{M}_2^T \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}^{-T} \end{bmatrix}. \tag{50}$$

For any element ψ_{ts} in $\mathbf{\Psi}_e = \mathbf{U}_1^T \mathbf{\Phi}_e \mathbf{U}_2^{-T}$, where $t \in \{1,\ldots,l+n\}$ and $s \in \{1,\ldots,q+n\}$, and any ω_{ij} in $\mathbf{\Omega}$, where $i \in \{1,\ldots,n-2\}$ and $j \in \{1,\ldots,n-2\}$,

$$\begin{aligned} \frac{\partial \psi_{ts}}{\partial \omega_{ij}} &= \mathbf{e}_t^T \frac{\partial \mathbf{U}_1^T}{\partial \omega_{ij}} \mathbf{\Phi}_e \mathbf{U}_2^{-T} \mathbf{e}_s + \mathbf{e}_t^T \mathbf{U}_1^T \mathbf{\Phi}_e \frac{\partial \mathbf{U}_2^{-T}}{\partial \omega_{ij}} \mathbf{e}_s \\ &= \mathbf{e}_t^T \mathbf{e}_{l+2+j} \mathbf{e}_{l+2+i}^T \mathbf{\Phi}_e \mathbf{U}_2^{-T} \mathbf{e}_s - \mathbf{e}_t^T \mathbf{U}_1^T \mathbf{\Phi}_e \mathbf{U}_2^{-T} \mathbf{e}_{q+2+j} \mathbf{e}_{q+2+i}^T \mathbf{U}_2^{-T} \mathbf{e}_s \\ &= \mathbf{e}_t^T \mathbf{e}_{l+2+j} \mathbf{e}_{l+2+i}^T \mathbf{\Phi}_e \mathbf{U}_2^{-T} \mathbf{e}_s - \mathbf{e}_t^T \mathbf{\Psi}_e \mathbf{e}_{q+2+j} \mathbf{e}_{q+2+i}^T \mathbf{U}_2^{-T} \mathbf{e}_s. \end{aligned} \tag{51}$$

That is,

$$\begin{aligned} \frac{\partial \psi_{ts}}{\partial \mathbf{\Omega}} = &\begin{bmatrix} \mathbf{e}_t^T & & \\ & \ddots & \\ & & \mathbf{e}_t^T \end{bmatrix} \left(\begin{bmatrix} \mathbf{e}_{l+3}\mathbf{e}_{l+3}^T \mathbf{\Phi}_e & \cdots & \mathbf{e}_{l+n}\mathbf{e}_{l+3}^T \mathbf{\Phi}_e \\ \vdots & \cdots & \vdots \\ \mathbf{e}_{l+3}\mathbf{e}_{l+n}^T \mathbf{\Phi}_e & \cdots & \mathbf{e}_{l+n}\mathbf{e}_{l+n}^T \mathbf{\Phi}_e \end{bmatrix} \right. \\ &\left. - \begin{bmatrix} \mathbf{\Psi}_e \mathbf{e}_{q+3}\mathbf{e}_{q+3}^T & \cdots & \mathbf{\Psi}_e \mathbf{e}_{q+n}\mathbf{e}_{q+3}^T \\ \vdots & \cdots & \vdots \\ \mathbf{\Psi}_e \mathbf{e}_{q+3}\mathbf{e}_{q+n}^T & \cdots & \mathbf{\Psi}_e \mathbf{e}_{q+n}\mathbf{e}_{q+n}^T \end{bmatrix} \right) \begin{bmatrix} \mathbf{U}_2^{-T}\mathbf{e}_s & & \\ & \ddots & \\ & & \mathbf{U}_2^{-T}\mathbf{e}_s \end{bmatrix}. \end{aligned} \tag{52}$$

Since

$$g(\mathbf{X}(\mathbf{T}(\mathbf{\Omega})),e)=\frac{(1-|\lambda_e|)/\sqrt{(l+n)(q+n)}}{\sqrt{\sum_{t=1}^{l+n}\sum_{s=1}^{q+n}\psi_{ts}^{*}\psi_{ts}}}, \tag{53}$$

we can readily calculate

$$\frac{\partial g(\mathbf{X}(\mathbf{T}(\mathbf{\Omega})),e)}{\partial \mathbf{\Omega}}=-\frac{1-|\lambda_e|}{\sqrt{(l+n)(q+n)}\,\|\mathbf{\Psi}_e\|_F^3}\mathrm{Re}\left[\sum_{t=1}^{l+n}\sum_{s=1}^{q+n}\psi_{ts}^{*}\frac{\partial \psi_{ts}}{\partial \mathbf{\Omega}}\right]. \tag{54}$$

Comment 1. In a way, the above search procedure solves

$$\min_{\mathbf{\Omega}\in\mathcal{R}^{(n-2)\times(n-2)}}\ \max_{k\in\{1,\ldots,m+n\}}\left(-g(\mathbf{X}(\mathbf{T}(\mathbf{\Omega})),k)\right). \tag{55}$$

The function $h(\mathbf{\Omega})=\max_{k\in\{1,\ldots,m+n\}}\left(-g(\mathbf{X}(\mathbf{T}(\mathbf{\Omega})),k)\right)$ to be minimized has corners where differentiability fails, although $g(\mathbf{X}(\mathbf{T}(\mathbf{\Omega})),k)$ is differentiable for any $k\in\{1,\ldots,m+n\}$. In fact, the problem (55) is a classical optimization problem which requires nondifferentiable optimization approaches, such as subgradient methods [22]. Subdifferentiation of h at $\mathbf{\Omega}$ is defined as

$$\aleph h(\mathbf{\Omega})=Conv\left\{\mathbf{J}\in\mathcal{R}^{(n-2)\times(n-2)}\ \middle|\ \begin{array}{c}\mathbf{J}=\lim\frac{\partial h(\mathbf{\Omega}_i)}{\partial \mathbf{\Omega}_i},\mathbf{\Omega}_i\to\mathbf{\Omega},\\ \frac{\partial h(\mathbf{\Omega}_i)}{\partial \mathbf{\Omega}_i}\ \text{exists},\frac{\partial h(\mathbf{\Omega}_i)}{\partial \mathbf{\Omega}_i}\ \text{converges}\end{array}\right\}, \tag{56}$$

where *Conv* denotes the convex hull. The elements of $\aleph h(\mathbf{\Omega})$ are called subgradients. Denote the directional derivative

$$h^{\circ}(\mathbf{\Omega},\mathbf{\Gamma})=\lim_{\substack{t\to 0\\ t>0}}\frac{h(\mathbf{\Omega}+t\mathbf{\Gamma})-h(\mathbf{\Omega})}{t} \tag{57}$$

in every direction $\mathbf{\Gamma}\in\mathcal{R}^{(n-2)\times(n-2)}$. A relationship between subgradients and the directional derivative is given in [22], which is restated in the following lemma.

LEMMA 2. $h^{\circ}(\mathbf{\Omega},\mathbf{\Gamma})=\max_{\mathbf{J}\in\aleph h(\mathbf{\Omega})}\langle\mathbf{J},\mathbf{\Gamma}\rangle$.

It is seen that $-\frac{\partial g(\mathbf{X},e)}{\partial \mathbf{\Omega}}$ is a subgradient of $h(\mathbf{\Omega})$ and our method is a subgradient algorithm. Since $h(\mathbf{\Omega})$ is differentiable almost everywhere when $\mathbf{\Omega}$ is not a local optimal point, there exists a neighborhood $\mathcal{B}_r=\left\{\mathbf{\Theta}\in\mathcal{R}^{(n-2)\times(n-2)}\ |\ \|\mathbf{\Theta}-\mathbf{\Omega}\|_F<r\right\}$ such that

$$h^{\circ}(\mathbf{\Omega},\mathbf{\Xi}-\mathbf{\Omega})<0 \tag{58}$$

and

$$\mathbf{\Xi}=\min_{\mathbf{\Theta}\in\mathcal{B}_r}h(\mathbf{\Omega}). \tag{59}$$

Then we have the following theorem.

THEOREM 5. *There exists* $\tau_m>0$ *such that for Step* 2 *of the above search algorithm*

$$\left\|\mathbf{\Omega}+\tau\frac{\partial g(\mathbf{X},e)}{\partial \mathbf{\Omega}}\left\|\frac{\partial g(\mathbf{X},e)}{\partial \mathbf{\Omega}}\right\|_F^{-1}-\mathbf{\Xi}\right\|_F<\|\mathbf{\Omega}-\mathbf{\Xi}\|_F \tag{60}$$

$\forall\tau\in(0,\tau_m)$.

Proof. By the definition of Frobenius norm,

$$\begin{aligned}&\left\|\boldsymbol{\Xi}-\boldsymbol{\Omega}-\tau\frac{\partial g(\mathbf{X},e)}{\partial\boldsymbol{\Omega}}\left\|\frac{\partial g(\mathbf{X},e)}{\partial\boldsymbol{\Omega}}\right\|_F^{-1}\right\|_F^2\\&=\|\boldsymbol{\Xi}-\boldsymbol{\Omega}\|_F^2+2\tau\left\langle-\frac{\partial g(\mathbf{X},e)}{\partial\boldsymbol{\Omega}}\left\|\frac{\partial g(\mathbf{X},e)}{\partial\boldsymbol{\Omega}}\right\|_F^{-1},\boldsymbol{\Xi}-\boldsymbol{\Omega}\right\rangle+\tau^2.\end{aligned}\tag{61}$$

Since $-\frac{\partial g(\mathbf{X},e)}{\partial\boldsymbol{\Omega}}$ is a subgradient, from Lemma 2 and (58), one has

$$\left\langle-\frac{\partial g(\mathbf{X},e)}{\partial\boldsymbol{\Omega}}\left\|\frac{\partial g(\mathbf{X},e)}{\partial\boldsymbol{\Omega}}\right\|_F^{-1},\boldsymbol{\Xi}-\boldsymbol{\Omega}\right\rangle\le h^\circ(\boldsymbol{\Omega},\boldsymbol{\Xi}-\boldsymbol{\Omega})\left\|\frac{\partial g(\mathbf{X},e)}{\partial\boldsymbol{\Omega}}\right\|_F^{-1}<0.\tag{62}$$

Thus, for $0<\tau<\tau_m=2\big\langle\frac{\partial g(\mathbf{X},e)}{\partial\boldsymbol{\Omega}}\big\|\frac{\partial g(\mathbf{X},e)}{\partial\boldsymbol{\Omega}}\big\|_F^{-1},\ \boldsymbol{\Xi}-\boldsymbol{\Omega}\big\rangle$,

$$2\tau\left\langle-\frac{\partial g(\mathbf{X},e)}{\partial\boldsymbol{\Omega}}\left\|\frac{\partial g(\mathbf{X},e)}{\partial\boldsymbol{\Omega}}\right\|_F^{-1},\boldsymbol{\Xi}-\ \boldsymbol{\Omega}\right\rangle+\tau^2<0.\tag{63}$$

This together with (61) proves the assertion. ∎

The above result shows that, for sufficiently small $\tau>0$, $\frac{\partial g(\mathbf{X},e)}{\partial\boldsymbol{\Omega}}$ is a good direction along which to update $\boldsymbol{\Omega}$ so that it becomes closer to $\boldsymbol{\Xi}$, although occasionally the updated $h(\boldsymbol{\Omega})$ may be worse. Therefore, $h(\boldsymbol{\Omega})$ will be improved significantly after some iterations. Our numerical examples listed in section 5 show that this simplest subgradient optimization algorithm behaves satisfactorily in practice, provided that τ is chosen appropriately. Of course, if this simplest subgradient algorithm fails in some cases, various enhanced subgradient algorithms [22], [23], [24] can be adopted to tackle the problem.

Comment 2. The termination at $N_t\ge N$ does not mean that the problem (55) has no saddle point. As $h(\boldsymbol{\Omega})$ may be nonconvex, our subgradient search sequence may possibly oscillate around a local optimum which is worse than $\rho_{k'}$. Regardless of whether or not the problem (55) has saddle points, when the routine does not find a saddle point, we can further increase the value of $\min_{k\in\{1,\ldots,m+n\}}g(\mathbf{X},k)$ by a direct numerical optimization. This is further discussed in the next section.

4. Discussions. The function $g(\mathbf{X},k)$ having saddle points is the main assumption in this paper. Here we explain heuristically that for many practical control systems this assumption is valid. First, from section 3.1, it is known that k', $\rho_{k'}$, and $\mathcal{X}$ exist regardless of whether or not $g(\mathbf{X},k)$ has saddle points. Second, Theorem 3 shows that if and only if there exist $\mathbf{T}\in\mathcal{T}$ satisfying

$$g(\mathbf{X}(\mathbf{T}),k)\ge\rho_{k'}\quad\forall k\in\{1,\ldots,m+n\}\setminus\{k'\},\tag{64}$$

the saddle points of $g(\mathbf{X},k)$ exist. From the definition of $g(\mathbf{X},k)$ in (14), $g(\mathbf{X},k)$ is proportional to the single-pole stability margin $1-|\lambda_k(\overline{\mathbf{A}}(\mathbf{X}))|$, which is a fixed value, and inverse proportional to its eigenvalue sensitivity, which depends on $\mathbf{X}$. For practical digital closed-loop control systems, there exist usually only a few dominant poles which are near the unit circle and/or have relatively high eigenvalue sensitivities,

compared with all the other nondominant poles. For this reason, the index k' defined in (21) is usually the index of a dominant pole, and the values of $g(\mathbf{X}, k)$ for those nondominant poles at $\mathbf{X}(\mathbf{T})$ are larger than $\rho_{k'}$ for most $\mathbf{T} \in \mathcal{T}$. Therefore, to satisfy condition (64), one needs only to consider the few dominant poles whose indices are not k'. It should be observed that $\mathbf{T}$ in $\mathcal{T}$ has a fairly large degree of freedom. Specifically, the free parameter $\mathbf{\Omega}$ in (44) can be any nonsingular matrix in $\mathcal{R}^{(n-2)\times(n-2)}$. This large degree of freedom, together with the fact that there are typically just a few dominant poles to consider, means that most likely there exist $\mathbf{T} \in \mathcal{T}$ satisfying (64). Thus $g(\mathbf{X}, k)$ has saddle points for many practical problems. We conjecture without a rigorous proof that the class of optimal FWL controller realization problems (15) which have saddle points is much larger than the class having no saddle point. Empirically, we have tested a total of six FWL controller design examples that we found in the FWL controller design literature. Only one example, which is given in [14], was shown to possibly have no saddle point.

The routine presented in section 3.2 is computationally much more attractive than a direct numerical optimization of (13). Actually, all that is needed is to find a $\mathbf{T} \in \mathcal{T}$ such that $g(\mathbf{X}(\mathbf{T}), k) \geq \rho_{k'}$ for $k \in \{1, \ldots, m+n\} \setminus \{k'\}$, rather than to directly maximize $f(\mathbf{X}(\mathbf{T}))$ over $\mathcal{R}^{n\times n}$ (and, of course, $\det \mathbf{T} \neq 0$). The former objective can be attained often easily even for large-scale problems. In addition, the number of saddle points is infinite when $g(\mathbf{X}, k)$ has saddle points. Hence our algorithm can find global optimal solutions for most practical problems which have saddle points even though we do not strictly prove the convergence of the subgradient routine. An additional advantage of the algorithm presented, which is particularly important in practical applications, is that when the algorithm attains a solution the user knows for sure that it is a global optimal solution to the optimal realization problem (13). This should be compared with direct numerical optimization of (13) where even when it converges to a solution, there is no way to tell whether or not the solution is a global optimal one.

It should be pointed out that our algorithm, presented for the problems having saddle points, is also useful in helping to solve those optimal FWL realization problems which do not have a saddle point. Actually, the algorithm given in section 3 can be executed even for the problems which do not have a saddle point. Using the results of section 3.1, k' and $\rho_{k'}$ can be computed, and $\mathcal{X}$ is obtained in closed form. Corollary 1 tells us that $\rho_{k'}$ is an upper bound of the optimal value of the realization problem having no saddle point. After executing N iterations of the routine given in section 3.2, the resulting realization $\mathbf{X}_t$ obviously does not satisfy (64). But through these N iterations, $\min_{k\in\{1,\ldots,m+n\}} g(\mathbf{X}, k)$ has been increased to as close to $\rho_{k'}$ as possible under $\mathbf{X} \in \mathcal{X}$. Therefore, the value of $f(\mathbf{X}_t)$ is not much less than $\rho_{k'}$. This provides a small region $[f(\mathbf{X}_t), \rho_{k'}]$ within which the optimal value of the FWL controller realization problem lies. Of course, this also provides an excellent guess from which a direct numerical optimization approach can be used to find a (local) optimal solution for those optimization problems having no saddle point.

Obviously, the same idea is equally applicable to the problems whose saddle points are not found after N iterations of the search routine. In fact, when the subgradient routine is terminated after N iterations but the condition (64) is not met, one cannot answer the question of whether or not the problem (55) has any saddle point. However, one knows the small region within which the global optimal value lies, and the solution obtained after N iterations provides an excellent initial guess for a direct numerical optimization.

5. Design examples. Six examples are used to illustrate the effectiveness of the proposed design algorithm.

Example 1. The example in [25] is discretized with a sampling frequency of 5 Hz to obtain the discrete-time plant model

$$\mathbf{A}_P = \begin{bmatrix} 3.2439e-1 & -4.5451e+0 & -4.0535e+0 & -2.7003e-3 & 0 \\ 1.4518e-1 & 4.9477e-1 & -4.6945e-1 & -3.1274e-4 & 0 \\ 1.6814e-2 & 1.6491e-1 & 9.6681e-1 & -2.2114e-5 & 0 \\ 1.1889e-3 & 1.8209e-2 & 1.9829e-1 & 1.0000e+0 & 0 \\ 6.1301e-5 & 1.2609e-3 & 1.9930e-2 & 2.0000e-1 & 1.0000e+0 \end{bmatrix},$$

$$\mathbf{B}_P = [\,1.4518e-1 \quad 1.6814e-2 \quad 1.1889e-3 \quad 6.1301e-5 \quad 2.4979e-6\,]^T,$$

$$\mathbf{C}_P = [\,0 \quad 0 \quad 1.6188e+0 \quad -1.5750e-1 \quad -4.3943e+1\,]$$

and the initially designed digital controller

$$\mathbf{A}_C^0 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & -4.7086e-1 \\ 1 & 0 & 0 & 0 & 0 & 2.6885e+0 \\ 0 & 1 & 0 & 0 & 0 & -6.6649e+0 \\ 0 & 0 & 1 & 0 & 0 & 9.4410e+0 \\ 0 & 0 & 0 & 1 & 0 & -8.2537e+0 \\ 0 & 0 & 0 & 0 & 1 & 4.2600e+0 \end{bmatrix}, \quad \mathbf{B}_C^0 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{D}_C^0 = [4.6000e-2],$$

$$\mathbf{C}_C^0 = [\,2.1187e-1 \quad 9.4498e-2 \quad 1.0887e-2 \quad -4.4171e-2 \quad -7.6000e-2 \quad -8.8562e-2\,].$$

The corresponding closed-loop transition matrix $\overline{\mathbf{A}}(\mathbf{X}_0)$ is then formed using (10), from which the eigenvalues and the eigenvectors of the ideal closed-loop system are computed. These 11 eigenvalues and their absolute values are

$$\begin{bmatrix} \lambda_{1,2} \\ \lambda_{3,4} \\ \lambda_{5,6} \\ \lambda_{7,8} \\ \lambda_{9,10} \\ \lambda_{11} \end{bmatrix} = \begin{bmatrix} 4.8368e-1 \pm j8.5569e-1 \\ 4.8135e-1 \pm j8.5363e-1 \\ 9.9993e-1 \pm j3.7887e-4 \\ 8.3967e-1 \pm j1.6514e-1 \\ 8.0884e-1 \pm j1.2026e-1 \\ 8.1905e-1 \end{bmatrix}, \quad \begin{bmatrix} |\lambda_{1,2}| \\ |\lambda_{3,4}| \\ |\lambda_{5,6}| \\ |\lambda_{7,8}| \\ |\lambda_{9,10}| \\ |\lambda_{11}| \end{bmatrix} \begin{bmatrix} 9.8293e-1 \\ 9.7999e-1 \\ 9.9993e-1 \\ 8.5575e-1 \\ 8.1774e-1 \\ 8.1905e-1 \end{bmatrix}.$$

This closed-loop system has five pairs of conjugate complex-valued eigenvalues and one real-valued eigenvalue. Using the method developed in section 3.1, the single-pole peak for each eigenvalue is computed, and they are

$$\begin{bmatrix} \rho_{1,2} \\ \rho_{3,4} \\ \rho_{5,6} \\ \rho_{7,8} \\ \rho_{9,10} \\ \rho_{11} \end{bmatrix} = \begin{bmatrix} 2.5072e-3 \\ 2.1295e-3 \\ 6.7344e-6 \\ 2.8586e-3 \\ 3.0832e-3 \\ 4.3181e-3 \end{bmatrix}.$$

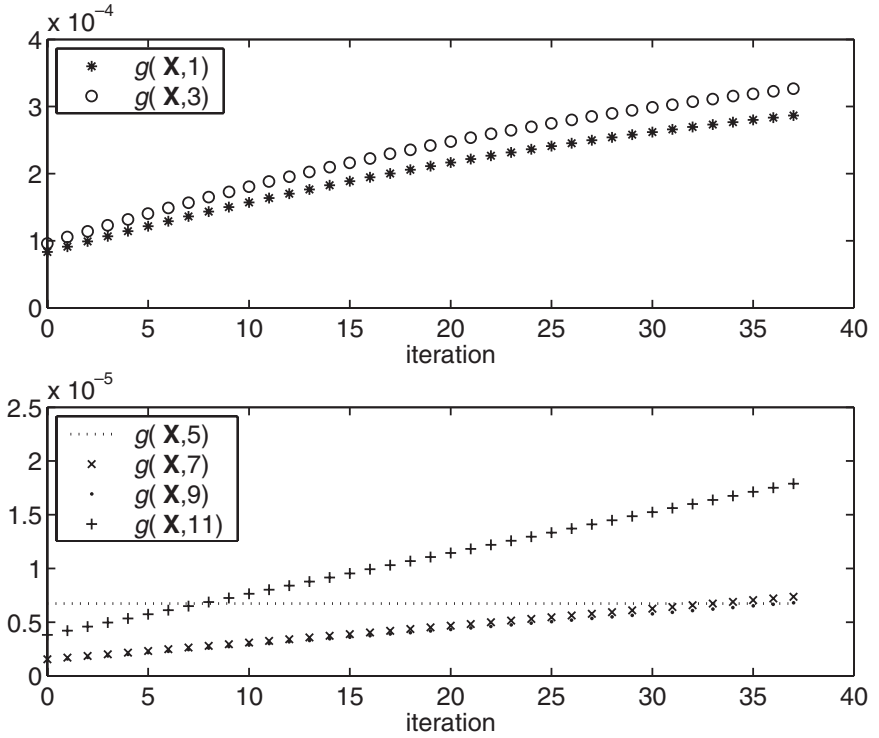


FIG. 2. *The values of* $g(\mathbf{X}, k)$ *in each iteration of the algorithm for Example* 1.

Obviously, the minimum value of all the ρ_k's is ρ_5 (or ρ_6). Therefore, $k' = 5$ and the corresponding matrices $\mathbf{Q}$, $\mathbf{H}$, and $\mathbf{F}$ in the set (44) are given by

$$\mathbf{Q} = \begin{bmatrix} -6.6011e-2 & -8.4915e-2 & -4.3670e-1 \\ -3.7006e-1 & -4.3518e-1 & -4.9156e-1 \\ -5.0566e-1 & -3.8025e-1 & 7.1063e-1 \\ -5.2127e-1 & -8.6900e-2 & -2.2452e-1 \\ -4.5786e-1 & 3.1775e-1 & -1.0190e-1 \\ -3.4878e-1 & 7.4183e-1 & 4.3249e-2 \end{bmatrix}$$

$$\begin{bmatrix} -5.1206e-1 & -5.2972e-1 & -5.0490e-1 \\ -2.2314e-1 & 1.7033e-1 & 5.9434e-1 \\ -2.5387e-1 & -1.6560e-1 & -5.3367e-2 \\ 7.4814e-1 & -2.4759e-1 & -2.2204e-1 \\ -2.0850e-1 & 6.8322e-1 & -4.1079e-1 \\ -1.4270e-1 & -3.6725e-1 & 4.1345e-1 \end{bmatrix},$$

$$\mathbf{H} = \begin{bmatrix} 2.6322e+0 & -3.9258e+2 \\ -3.9258e+2 & 6.9856e+6 \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} 4.8432e+4 & -8.8104e+8 \\ -5.2079e+4 & 9.4682e+8 \\ 2.4998e+4 & -4.5374e+8 \\ -2.4644e+4 & 4.4816e+8 \end{bmatrix}.$$

Set $\tau = 0.1$ and the initial $\mathbf{\Omega} = \mathbf{I}$. Figure 2 illustrates the changes of $g(\mathbf{X}, k)$ in each iteration. From Figure 2, it can be seen that at the 37th iteration, the optimal

controller realization is found, since at this iteration the conditions of Theorem 3 are met and the algorithm terminates. The resulting matrix $\mathbf{\Omega}_{\text{opt}}$ is

$$\mathbf{\Omega}_{\text{opt}} = \begin{bmatrix} 2.3184e+0 & -1.6411e+0 & 5.5681e-1 & -7.6953e-1 \\ -1.6411e+0 & 2.4047e+0 & -8.2094e-1 & 7.0079e-1 \\ 5.5680e-1 & -8.2095e-1 & 1.2097e+0 & -3.7643e-1 \\ -7.6954e-1 & 7.0078e-1 & -3.7643e-1 & 1.3454e+0 \end{bmatrix}$$

and the corresponding global optimal transformation matrix is

$$\mathbf{T}_{\text{opt}} = \begin{bmatrix} -6.9470e+1 & -3.2765e+4 & -7.8507e-2 & & \\ 3.0977e+2 & 1.5431e+5 & -1.1360e+0 & & \\ -6.0267e+2 & -3.0945e+5 & 2.0130e+0 & & \\ 6.5537e+2 & 3.4747e+5 & -1.7153e+0 & & \\ -4.1530e+2 & -2.2683e+5 & 8.0247e-1 & & \\ 1.1931e+2 & 6.9580e+4 & -1.8821e-1 & & \end{bmatrix}$$
$$\begin{bmatrix} -4.3363e-1 & -2.7354e-1 & -5.0267e-1 \\ 5.4680e-1 & -1.0820e-1 & 9.5739e-1 \\ -1.6781e+0 & 4.2386e-1 & -7.3423e-1 \\ 2.2151e+0 & -9.5513e-1 & 4.9153e-1 \\ -1.1829e+0 & 1.0956e+0 & -8.7755e-1 \\ 1.7712e-1 & -4.5868e-1 & 5.6121e-1 \end{bmatrix}.$$

Example 2. The second example is taken from [14]. In this example, $m = 4$, $n = 10$, $l = 2$, and $q = 2$, and hence it is a closed-loop system of order 14. The initial controller realization $\mathbf{X}_0$ of $C(z)$ has been given [20]. The corresponding closed-loop transition matrix $\overline{\mathbf{A}}(\mathbf{X}_0)$ is formed using (10), from which the eigenvalues and the eigenvectors of the ideal closed-loop system are computed. This closed-loop system has six pairs of conjugate complex-valued eigenvalues and two real-valued eigenvalues given by

$$\begin{bmatrix} \lambda_{1,2} \\ \lambda_{3,4} \\ \lambda_5 \\ \lambda_{6,7} \\ \lambda_{8,9} \\ \lambda_{10,11} \\ \lambda_{12} \\ \lambda_{13,14} \end{bmatrix} = \begin{bmatrix} -8.4482e-1 \pm j7.8204e-2 \\ -3.7557e-1 \pm j3.3602e-1 \\ 2.1624e-1 \\ 7.1567e-1 \pm j9.6631e-3 \\ 9.2895e-1 \pm j1.2923e-1 \\ 9.8506e-1 \pm j7.5831e-2 \\ 8.3133e-1 \\ 8.8267e-1 \pm j3.7235e-2 \end{bmatrix}.$$

Using the method developed in section 3.1, the single-pole peaks for every eigenvalues are computed, and they are

$$\begin{bmatrix} \rho_{1,2} \\ \rho_{3,4} \\ \rho_5 \\ \rho_{6,7} \\ \rho_{8,9} \\ \rho_{10,11} \\ \rho_{12} \\ \rho_{13,14} \end{bmatrix} = \begin{bmatrix} 8.3118e-3 \\ 4.0562e-2 \\ 6.2954e-2 \\ 8.0984e-3 \\ 3.7768e-3 \\ 5.4246e-3 \\ 5.8442e-3 \\ 8.0773e-3 \end{bmatrix}.$$

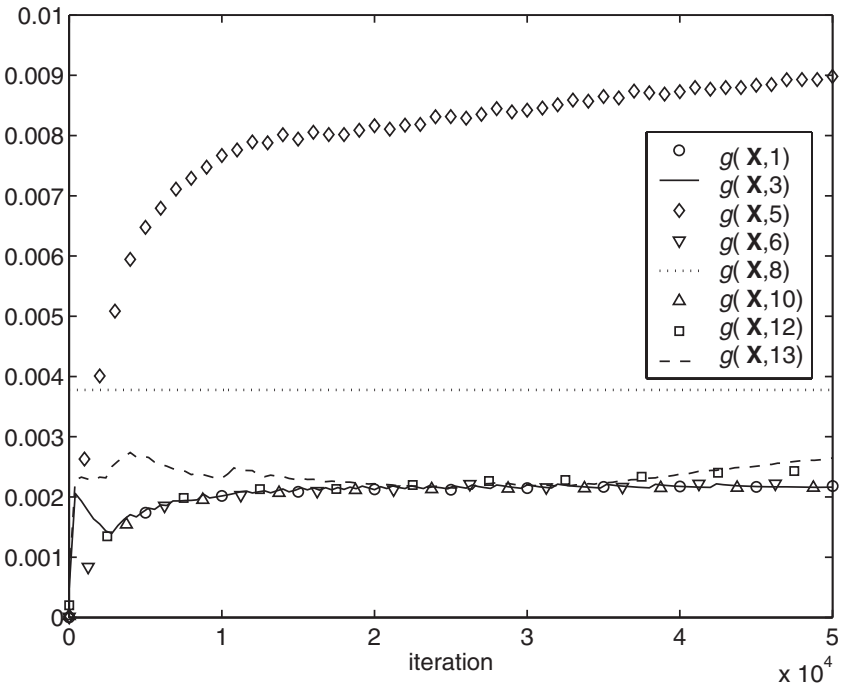


FIG. 3. *The values of* $g(\mathbf{X}, k)$ *in each iteration of the algorithm for Example* 2.

Obviously, the minimum value of all the ρ_k's is ρ_8 (or ρ_9). Therefore, $k' = 8$ and the corresponding matrices $\mathbf{Q}$, $\mathbf{H}$, and $\mathbf{F}$ in (44) are computed according to Theorem 4. With $\mathbf{T}$ in (44), the second stage of our algorithm can be executed. Figure 3 illustrates the changes of $g(\mathbf{X}, k)$ in each iteration of the second stage. From Figure 3, it can be seen that after the $N = 50000$ iteration, we still cannot find a realization satisfying (64). This suggests that this example most likely has no saddle point (although one cannot be sure). So we terminate the algorithm at the 50000 iteration and obtain a realization $\mathbf{X}_t$. Although this $\mathbf{X}_t$ is not an optimal realization, it is much better than $\mathbf{X}_0$, since $f(\mathbf{X}_t) = 2.1539e{-}3$ while $f(\mathbf{X}_0) = 1.1734e{-}4$. In particular, we notice that $\mathbf{X}_t$ is also better than the "optimal" realization given in [14], which was found by a direct numerical optimization search using the simulated annealing algorithm and has a FWL measure value of $1.5844e - 3$ [14]. At this stage, we are sure that the optimal solution given in [14] is not a global optimal one at all. Using the realization $\mathbf{X}_t$ obtained by our search algorithm as the initial point, we then use a direct numerical optimization method to solve for the optimization problem (13) and obtain a new optimal realization whose FWL measure value is $3.1929e - 3$. This optimal value is more than double the one given in [14]. Obviously, we cannot tell whether or not this new optimal realization is a global optimal one. However, we know that the optimal value of the FWL realization problem for this example lies in the range of $[3.1929e - 3,\ 3.7768e - 3]$. For this example, no other design has found a controller realization whose FWL closed-loop stability measure $f(\mathbf{X})$ is larger than $3e - 3$. Our algorithm is the first one to achieve a $f(\mathbf{X}) > 3e - 3$.

The saddle points (or the global optimal solutions) of the following four examples are found successfully by our proposed method.

Example 3. This example is a fluid power speed controller given in [8], where $m = 4$, $n = 4$, $l = 1$, and $q = 1$.

Example 4. This example is a discretized version of an H_∞ robust controller given in [26] with a sampling frequency of 250 Hz, where $m = 2$, $n = 3$, $l = 1$, and $q = 1$.

Example 5. This example is taken from [6], where $m = 3$, $n = 2$, $l = 1$, and $q = 1$.

Example 6. This example is a steel rolling mill proportional-integral-derivative controller given in [8], where $m = 3$, $n = 2$, $l = 1$, and $q = 1$.

As mentioned previously, the realizations of $C(z)$ are not unique. For instance, in Example 1, the initially designed controller $(\mathbf{A}_C^0, \mathbf{B}_C^0, \mathbf{C}_C^0, \mathbf{D}_C^0)$ is the controllable companion-form realization for

$$C(z) = \frac{0.046z^6 + 0.0159z^5 - 0.4284z^4 + 0.9227z^3 - 1.0043z^2 + 0.5983z - 0.1503}{z^6 - 4.26z^5 + 8.2537z^4 - 9.441z^3 + 6.6649z^2 - 2.6885z + 0.4709}.$$

Apart from the controllable companion form, denoted as $\mathbf{X}_c$, a controller is also often implemented in the parallel or series form in practice. Denote these two realizations of $C(z)$ as

$$\mathbf{X}_p = \begin{bmatrix} \mathbf{D}_C^p & \mathbf{C}_C^p \\ \mathbf{B}_C^p & \mathbf{A}_C^p \end{bmatrix} \tag{65}$$

and

$$\mathbf{X}_s = \begin{bmatrix} \mathbf{D}_C^s & \mathbf{C}_C^s \\ \mathbf{B}_C^s & \mathbf{A}_C^s \end{bmatrix}, \tag{66}$$

respectively. The parallel-form realization of $C(z)$ for Example 1 is given by

$$C(z) = 0.046 + \frac{1.8921e-7}{z-1} + \frac{-0.0024z + 0.0013}{z^2 - 0.9670z + 0.9589} + \frac{0.1056z - 0.1487}{z^2 - 1.6016z + 0.7103} + \frac{0.1087}{z - 0.6913}$$

with

$$\mathbf{A}_C^p = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & -9.5886e-1 & 9.6700e-1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -7.1030e-1 & 1.6016e0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 6.9134e-1 \end{bmatrix},$$

$$\mathbf{B}_C^p = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}^T, \quad \mathbf{D}_C^p = \begin{bmatrix} 4.6000e-2 \end{bmatrix},$$

$$\mathbf{C}_C^p = \begin{bmatrix} 1.8921e-7 & 1.2816e-3 & -2.3654e-3 & -1.4868e-1 & 1.0555e-1 \\ 1.0869e-1 \end{bmatrix},$$

while the series-form realization is

$$C(z) = 0.046\left(\frac{0.1812}{z-1}+1\right)\left(\frac{0.6344z+0.2556}{z^2-1.6016z+0.7103}+1\right) \times \left(\frac{4.8231}{z-0.6913}+1\right)\left(\frac{-1.0329z+0.0410}{z^2-0.9670z+0.9589}+1\right)$$

with

$$\mathbf{A}_C^s = \begin{bmatrix} 1 & 0 & 1.8120e-1 & 1.8120e-1 & 0 & 1.8120e-1 \\ 0 & 0 & -7.1030e-1 & 2.5562e-1 & 0 & 2.5562e-1 \\ 0 & 1 & 1.6016e0 & 6.3442e-1 & 0 & 6.3442e-1 \\ 0 & 0 & 0 & 6.9134e-1 & 0 & 4.8231e0 \\ 0 & 0 & 0 & 0 & 0 & -9.5886e-1 \\ 0 & 0 & 0 & 0 & 1 & 9.6700e-1 \end{bmatrix},$$

$$\mathbf{B}_C^s = \begin{bmatrix} 1.8120e-1 & 2.5562e-1 & 6.3442e-1 & 4.8231e0 & 4.1007e-2 \\ -1.0329e0 \end{bmatrix}^T,$$

$$\mathbf{C}_C^s = \begin{bmatrix} 4.6000e-2 & 0 & 4.6000e-2 & 4.6000e-2 & 0 & 4.6000e-2 \end{bmatrix},$$

$$\mathbf{D}_C^s = \begin{bmatrix} 4.6000e-2 \end{bmatrix}.$$

The above three realizations, $\mathbf{X}_c$, $\mathbf{X}_p$, and $\mathbf{X}_s$, are sparse because they contain many trivial parameters (0, 1, or -1). For Example 1, $\mathbf{X}_0$ has 13 nontrivial parameters, while $\mathbf{X}_p$ and $\mathbf{X}_s$ have only 12 nontrivial parameters (the repeated values, such as $1.8120e-1$ in $\mathbf{X}_s$, are counted as one nontrivial parameter). Clearly, a trivial parameter requires no arithmetic operation in a fixed-point implementation and does not cause any computational error. A sparse controller realization has computational advantages in practical implementations. An FWL closed-loop stability measure, which is similar to the one defined in (11) but takes into account the sparsity of controller realization, is defined in [9] as

$$f_{sp}(\mathbf{X}) \triangleq \min_{k\in\{1,\ldots,m+n\}} \frac{1-|\lambda_k(\overline{\mathbf{A}}(\mathbf{X}))|}{\sqrt{N_s \sum_{i,j} \delta(x_{ij}) \left|\frac{\partial \lambda_k(\overline{\mathbf{A}}(\mathbf{X}))}{\partial x_{ij}}\right|^2}}, \tag{67}$$

where

$$\delta(x_{ij}) = \begin{cases} 1, & x_{ij} \text{ is nontrivial,} \\ 0, & x_{ij} \text{ is trivial,} \end{cases} \tag{68}$$

and N_s is the number of nontrivial parameters in $\mathbf{X}$. Comparing the definitions of $f_{sp}(\mathbf{X})$ and $f(\mathbf{X})$, it follows that

$$f_{sp}(\mathbf{X}) \geq f(\mathbf{X}). \tag{69}$$

Table 1 lists the values of $f(\mathbf{X})$, $f_{sp}(\mathbf{X})$, and N_s for $\mathbf{X}_{\text{opt}}$, $\mathbf{X}_p$, $\mathbf{X}_s$, and $\mathbf{X}_c$ of every example except for Example 2. Example 2 is a multiple-input multiple-output system for which no parallel-form or series-form realization is defined. It can be seen that the optimal realization $\mathbf{X}_{\text{opt}}$ found by the proposed method has the best FWL

TABLE 1
Comparison of performance measures for different realizations.

		$\mathbf{X}_c$	$\mathbf{X}_p$	$\mathbf{X}_s$	$\mathbf{X}_{\text{opt}}$
Example 1	$f(\mathbf{X})$	$3.1797e-11$	$8.0156e-9$	$2.8727e-9$	$6.7344e-6$
	$f_{sp}(\mathbf{X})$	$7.4944e-11$	$1.8464e-8$	$7.1095e-9$	$6.7344e-6$
	N_s	13	12	12	49
Example 3	$f(\mathbf{X})$	$5.0963e-10$	$1.5234e-5$	$3.0949e-6$	$2.7321e-4$
	$f_{sp}(\mathbf{X})$	$8.5965e-10$	$2.7908e-5$	$5.4711e-6$	$2.7321e-4$
	N_s	9	8	8	25
Example 4	$f(\mathbf{X})$	$1.6555e-10$	$8.3351e-10$	$1.4611e-7$	$5.0786e-5$
	$f_{sp}(\mathbf{X})$	$6.1068e-10$	$1.5627e-7$	$3.0905e-7$	$5.0786e-5$
	N_s	7	7	7	16
Example 5	$f(\mathbf{X})$	$1.6699e-4$	$5.4326e-4$	$4.8802e-4$	$3.2716e-3$
	$f_{sp}(\mathbf{X})$	$2.5956e-4$	$2.4426e-3$	$7.3417e-4$	$3.2716e-3$
	N_s	5	4	4	9
Example 6	$f(\mathbf{X})$	$6.7163e-4$	$1.0775e-3$	$1.0774e-3$	$4.8968e-3$
	$f_{sp}(\mathbf{X})$	$9.5044e-4$	$3.5239e-3$	$1.6347e-3$	$4.8968e-3$
	N_s	5	4	4	9

closed-loop stability robustness as measured by either $f(\mathbf{X})$ or $f_{sp}(\mathbf{X})$, compared with the other three realizations. It can also be seen that the optimal realization obtained by the proposed search algorithm is a fully parameterized nonsparse one. The other three sparse realizations have similar numbers of nontrivial parameters, and thus have the same lighter computational load than that of the optimal one given here. However, it is worth pointing out that $\mathbf{X}_{\text{opt}}$ is not unique since $\mathbf{V}$ in (43) is an arbitrary orthogonal matrix. By choosing $\mathbf{V}$ in an appropriate way, one can obtain a sparse optimal realization $\mathbf{X}_{\text{opt}}$. The topic of how to make $\mathbf{X}_{\text{opt}}$ sparse is beyond the scope of this paper, and interested readers are referred to the work [1] for details.

6. Conclusions. We have developed an efficient search algorithm for solving the class of optimal FWL controller realization problems based on the Frobenius-norm pole sensitivity measure, which have saddle points. Our approach first constructs the closed form of a transformation matrix set which contains global optimal solutions and then searches this set with a subgradient routine to find a global optimal solution. The proposed algorithm has considerable advantages over using direct numerical optimization methods to tackle this class of optimal FWL realization problems. In particular, when the subgradient routine converges to a solution, it is guaranteed to be a global optimal solution. It has been conjectured with some empirical support that for many practical control systems the assumption of having saddle points is a valid one and the cases of optimal FWL controller realization problems which do not have saddle points are less common. It has been demonstrated that for this smaller class of optimal FWL realization problems without saddle points our algorithm also provides useful information to help solve them.

Appendix. Proof of Theorem 4. We present the proof for Case 2. The proof for Case 3 is similar and hence is omitted.

LEMMA 3 (see [21]). *Let real-valued matrices* $\mathbf{M}_{22}$, $\mathbf{M}_{21}$, *and* $\mathbf{M}_{11} > 0$ *be given with appropriate dimensions. Then*

$$\begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{21}^T \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix} > 0 \tag{70}$$

if and only if $\mathbf{M}_{22} - \mathbf{M}_{21}\mathbf{M}_{11}^{-1}\mathbf{M}_{21}^T > 0$.

LEMMA 4. *Given positive* $\alpha, \beta \in \mathcal{R}$, $\mathbf{p}, \mathbf{y} \in \mathcal{C}^n$, *and for any nonsingular* $\mathbf{T} \in \mathcal{R}^{n\times n}$, *we have*

$$\xi(\mathbf{T}, \alpha, \beta, \mathbf{p}, \mathbf{y}) \geq (|\mathbf{y}^H \mathbf{p}| + \alpha\beta)^2. \tag{71}$$

The equality occurs if and only if there exist $\mathbf{W} \in \mathcal{R}^{n\times n}$, $\mathbf{W} > 0$, *and* $\theta \in [0, 2\pi)$ *satisfying*

$$\mathbf{W}\Upsilon(\mathbf{y}) = \frac{\beta}{\alpha}\Upsilon(\mathbf{p}) \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}. \tag{72}$$

When (72) *has solutions, the equality in* (71) *occurs only at the transformation matrix* $\mathbf{T} = \mathbf{W}^{1/2}\mathbf{V}$, *where* $\mathbf{V} \in \mathcal{R}^{n\times n}$ *is an arbitrary orthogonal matrix.*

Proof. First,

$$\begin{aligned} &\|\mathbf{T}^{-1}\mathbf{p}\|_F^2 \|\mathbf{T}^T\mathbf{y}\|_F^2 + \alpha^2 \|\mathbf{T}^T\mathbf{y}\|_F^2 + \beta^2 \|\mathbf{T}^{-1}\mathbf{p}\|_F^2 + \alpha^2\beta^2 \\ &\geq (\|\mathbf{T}^{-1}\mathbf{p}\|_F \|\mathbf{T}^T\mathbf{y}\|_F + \alpha\beta)^2. \end{aligned} \tag{73}$$

The equality holds if and only if

$$\alpha\|\mathbf{T}^T\mathbf{y}\|_F = \beta\|\mathbf{T}^{-1}\mathbf{p}\|_F. \tag{74}$$

Using the Cauchy–Schwarz inequality, we have

$$(\|\mathbf{T}^{-1}\mathbf{p}\|_F \|\mathbf{T}^T\mathbf{y}\|_F + \alpha\beta)^2 \geq (\|(\mathbf{T}^T\mathbf{y})^H \mathbf{T}^{-1}\mathbf{p}\|_F + \alpha\beta)^2 \geq (|\mathbf{y}^H\mathbf{p}| + \alpha\beta)^2. \tag{75}$$

The equality holds if and only if

$$\mathbf{T}^T\mathbf{y} = c\mathbf{T}^{-1}\mathbf{p} \tag{76}$$

for some complex number c.

To achieve (73) and (75) with equality, one needs to satisfy both of the conditions (74) and (76). This implies that $c = (\cos\theta + j\sin\theta)\frac{\beta}{\alpha}$ and $\theta \in [0, 2\pi)$. Thus,

$$\mathbf{T}^T\mathbf{y} = (\cos\theta + j\sin\theta)\frac{\beta}{\alpha}\mathbf{T}^{-1}\mathbf{p}. \tag{77}$$

As $\mathbf{T}$ is nonsingular, equality (77) is equivalent to

$$\mathbf{W}\mathbf{y} = (\cos\theta + j\sin\theta)\frac{\beta}{\alpha}\mathbf{p} \tag{78}$$

with $\mathbf{W} > 0$ and $\mathbf{T} = \mathbf{W}^{1/2}\mathbf{V}$. Noticing the map Υ defined in (31), condition (78) can be viewed as

$$\mathbf{W}\Upsilon(\mathbf{y}) = \frac{\beta}{\alpha}\Upsilon(\mathbf{p}) \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}. \tag{79}$$

This completes the proof. □

LEMMA 5. *Given positive* $\alpha, \beta \in \mathcal{R}$, $\mathbf{p}, \mathbf{y} \in \mathcal{C}^n$, *and* $\operatorname{rank}(\Upsilon(\mathbf{y})) = 2$, (79) *has solutions if and only if* $\det((\Upsilon(\mathbf{y}))^T\Upsilon(\mathbf{p})) > 0$. *Moreover, any solution to* (79) *can be expressed as*

$$\left.\begin{aligned} &\tan\theta = \frac{a_{21} - a_{12}}{a_{11} + a_{22}} \\ &a_{11}\cos\theta - a_{12}\sin\theta > 0 \\ &\mathbf{W} = \mathbf{Q}\begin{bmatrix} \mathbf{H} & \mathbf{F}^T \\ \mathbf{F} & \mathbf{G} \end{bmatrix}\mathbf{Q}^T \end{aligned}\right\}, \tag{80}$$

where

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = (\Upsilon(\mathbf{y}))^T \Upsilon(\mathbf{p}); \tag{81}$$

the orthogonal matrix $\mathbf{Q}$ *can be obtained from the QR factorization of* $\Upsilon(\mathbf{y})$*, that is,*

$$\Upsilon(\mathbf{y}) = \mathbf{Q} \begin{bmatrix} \gamma_{11} & 0 & 0 & \cdots & 0 \\ \gamma_{12} & \gamma_{22} & 0 & \cdots & 0 \end{bmatrix}^T; \tag{82}$$

$\mathbf{H}$ *and* $\mathbf{F}$ *are determined by*

$$\mathbf{H} = \frac{\beta}{\alpha} \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \end{bmatrix}^{-T} (\Upsilon(\mathbf{y}))^T \Upsilon(\mathbf{p}) \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \end{bmatrix}^{-1}, \tag{83}$$

$$\mathbf{F} = \frac{\beta}{\alpha} \begin{bmatrix} \mathbf{e}_3^T \\ \vdots \\ \mathbf{e}_n^T \end{bmatrix} \mathbf{Q}^T \Upsilon(\mathbf{p}) \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \end{bmatrix}^{-1}; \tag{84}$$

and $\mathbf{G}$ *is given as*

$$\mathbf{G} = \mathbf{F}\mathbf{H}^{-1}\mathbf{F}^T + \mathbf{U} \tag{85}$$

with $\mathbf{U} \in \mathcal{R}^{(n-2)\times(n-2)}$ *being an arbitrary positive definite matrix.*

Proof. If $\det((\Upsilon(\mathbf{y}))^T \Upsilon(\mathbf{p})) > 0$, it is easy to verify that $\mathbf{W}$ and θ given by (80)–(85) are a solution to (79). If, on the other hand, (79) has a solution $\mathbf{W}$ and θ, it can be seen that

$$(\Upsilon(\mathbf{y}))^T \mathbf{W} \Upsilon(\mathbf{y}) = \frac{\beta}{\alpha} (\Upsilon(\mathbf{y}))^T \Upsilon(\mathbf{p}) \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}. \tag{86}$$

On account of $(\Upsilon(\mathbf{y}))^T \mathbf{W} \Upsilon(\mathbf{y}) > 0$, we have

$$(\Upsilon(\mathbf{y}))^T \Upsilon(\mathbf{p}) \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} > 0. \tag{87}$$

A necessary condition to satisfy (87) is that

$$\det\left((\Upsilon(\mathbf{y}))^T \Upsilon(\mathbf{p}) \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \right) > 0. \tag{88}$$

Since the left side of the above inequality is equal to $\det((\Upsilon(\mathbf{y}))^T \Upsilon(\mathbf{p}))$, the condition (88) becomes $\det((\Upsilon(\mathbf{y}))^T \Upsilon(\mathbf{p})) > 0$. This completes the proof of the first part of Lemma 5.

Now, when (81) is given, (87) holds if and only if all of the following three conditions are satisfied:

$$\left.\begin{array}{l} a_{21}\cos\theta - a_{22}\sin\theta = a_{11}\sin\theta + a_{12}\cos\theta \\ a_{11}\cos\theta - a_{12}\sin\theta > 0 \\ \det((\Upsilon(\mathbf{y}))^T \Upsilon(\mathbf{p})) > 0 \end{array}\right\}. \tag{89}$$

From the first line of (89), we directly obtain $\tan\theta = \frac{a_{21}-a_{12}}{a_{11}+a_{22}}$. Denote

$$\mathbf{S} = \mathbf{Q}^T \mathbf{W} \mathbf{Q}. \tag{90}$$

Then, from (79), (82), and (90), one has

$$\begin{aligned}\mathbf{S}\,[\,\mathbf{e}_1 \quad \mathbf{e}_2\,]\begin{bmatrix}\gamma_{11} & \gamma_{12}\\ 0 & \gamma_{22}\end{bmatrix} &= \mathbf{S}\begin{bmatrix}\gamma_{11} & 0 & 0 & \cdots & 0\\ \gamma_{12} & \gamma_{22} & 0 & \cdots & 0\end{bmatrix}^T\\ &= \frac{\beta}{\alpha}\mathbf{Q}^T\Upsilon(\mathbf{p})\begin{bmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{bmatrix}.\end{aligned} \tag{91}$$

Partition $\mathbf{S}$ into

$$\mathbf{S} = \begin{bmatrix}\mathbf{H} & \mathbf{F}^T\\ \mathbf{F} & \mathbf{G}\end{bmatrix}, \tag{92}$$

where $\mathbf{H} \in \mathcal{R}^{2\times 2}$, $\mathbf{F} \in \mathcal{R}^{(n-2)\times 2}$, and $\mathbf{G} \in \mathcal{R}^{(n-2)\times(n-2)}$. Then from (91) and noticing

$$(\Upsilon(\mathbf{y}))^T = \begin{bmatrix}\gamma_{11} & \gamma_{12}\\ 0 & \gamma_{22}\end{bmatrix}^T [\,\mathbf{e}_1 \quad \mathbf{e}_2\,]^T\,\mathbf{Q}^T, \tag{93}$$

we have

$$\begin{aligned}\mathbf{H} &= \begin{bmatrix}\mathbf{e}_1^T\\ \mathbf{e}_2^T\end{bmatrix}\mathbf{S}\,[\,\mathbf{e}_1 \quad \mathbf{e}_2\,]\\ &= \frac{\beta}{\alpha}\begin{bmatrix}\gamma_{11} & \gamma_{12}\\ 0 & \gamma_{22}\end{bmatrix}^{-T}(\Upsilon(\mathbf{y}))^T\Upsilon(\mathbf{p})\begin{bmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{bmatrix}\begin{bmatrix}\gamma_{11} & \gamma_{12}\\ 0 & \gamma_{22}\end{bmatrix}^{-1},\end{aligned} \tag{94}$$

$$\mathbf{F} = \begin{bmatrix}\mathbf{e}_3^T\\ \vdots\\ \mathbf{e}_n^T\end{bmatrix}\mathbf{S}\,[\,\mathbf{e}_1 \quad \mathbf{e}_2\,] = \frac{\beta}{\alpha}\begin{bmatrix}\mathbf{e}_3^T\\ \vdots\\ \mathbf{e}_n^T\end{bmatrix}\mathbf{Q}^T\Upsilon(\mathbf{p})\begin{bmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{bmatrix}\begin{bmatrix}\gamma_{11} & \gamma_{12}\\ 0 & \gamma_{22}\end{bmatrix}^{-1}. \tag{95}$$

From Lemma 3 and $\mathbf{S} > 0$, it is known that $\mathbf{G} = \mathbf{F}\mathbf{H}^{-1}\mathbf{F}^T + \mathbf{U}$, where $\mathbf{U} \in \mathcal{R}^{(n-2)\times(n-2)}$ is an arbitrary positive definite matrix. □

Combining Lemmas 4 and 5 leads to Theorem 4 for Case 2.

REFERENCES

[1] M. Gevers and G. Li, *Parameterizations in Control, Estimation and Filtering Problems: Accuracy Aspects*, Springer-Verlag, London, 1993.

[2] R. S. H. Istepanian and J. F. Whidborne, eds., *Digital Controller Implementation and Fragility: A Modern Perspective*, Springer-Verlag, London, 2001.

[3] G. F. Franklin, J. D. Powell, and M. L. Workman, *Digital Control of Dynamic Systems*, 3rd ed., Addison–Wesley, Reading, MA, 1998.

[4] K. Liu, R. E. Skelton, and K. Grigoriadis, *Optimal controllers for finite wordlength implementation*, IEEE Trans. Automat. Control, 37 (1992), pp. 1294–1304.

[5] G. Li, J. Wu, S. Chen, and K. Y. Zhao, *Optimum structures of digital controllers in sampled-data systems*: *A roundoff noise analysis*, IEE Proc. Control Theory and Applications, 149 (2002), pp. 247–255.

[6] I. J. Fialho and T. T. Georgiou, *Computational algorithms for sparse optimal digital controller realizations*, in Digital Controller Implementation and Fragility: A Modern Perspective, R. S. H. Istepanian and J. F. Whidborne, eds., Springer-Verlag, London, 2001, pp. 105–121.

[7] A. G. Madievski, B. D. O. Anderson, and M. Gevers, *Optimum realizations of sampled-data controllers for FWL sensitivity minimization*, Automatica, 31 (1995), pp. 367–379.

[8] J. F. Whidborne, J. Wu, and R. S. H. Istepanian, *Finite word length stability issues in an l_1 framework*, Internat J. Control, 73 (2000), pp. 166–176.

[9] G. Li, *On the structure of digital controllers with finite word length consideration*, IEEE Trans. Automat. Control, 43 (1998), pp. 689–693.
[10] P. E. Mantey, *Eigenvalue sensitivity and state-variable selection*, IEEE Trans. Automat. Control, 13 (1968), pp. 263–269.
[11] J. Wu, S. Chen, G. Li, and J. Chu, *Optimal finite-precision state-estimate feedback controller realizations of discrete-time systems*, IEEE Trans. Automat. Control, 45 (2000), pp. 1550–1554.
[12] G.-H. Yang, J. L. Wang, and C. Lin, *H_∞ control for linear systems with additive controller gain variations*, Internat J. Control, 73 (2000), pp. 1500–1506.
[13] J. F. Whidborne, J. Wu, R. S. H. Istepanian, and J. Chu, *Comments on "On the structure of digital controllers with finite word length consideration"*, IEEE Trans. Automat. Control, 45 (2000), p. 344.
[14] R. S. H. Istepanian, J. Wu, and J. F. Whidborne, *Controller realizations of a teleoperated dual-wrist assembly system with finite word length considerations*, IEEE Trans. Control Systems Technology, 9 (2001), pp. 624–628.
[15] G. Owen, *Game Theory*, 3rd ed., Academic Press, San Diego, CA, 1995.
[16] J. Szép and F. Forgó, *Introduction to the Theory of Games*, Akadémiai Kiadó, Budapest, 1985.
[17] I. J. Fialho and T. T. Georgiou, *On stability and performance of sampled-data systems subject to wordlength constraint*, IEEE Trans. Automat. Control, 39 (1994), pp. 2476–2481.
[18] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, UK, 1988.
[19] S. Chen, J. Wu, R. H. Istepanian, J. Chu, and J. F. Whidborne, *Optimising stability bounds of finite-precision controller structures for sampled-data systems in the δ-operator domain*, IEE Proc. Control Theory and Applications, 146 (1999), pp. 517–526.
[20] J. Yan and S. E. Salcudean, *Teleoperation controller design using H_∞-optimization with application to motion-scaling*, IEEE Trans. Control Systems Technology, 4 (1996), pp. 244–258.
[21] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in Systems and Control Theory*, Stud. Appl. Math. 15, SIAM, Philadelphia, 1994.
[22] J. Zowe, *Nondifferentiable optimization*, in Computational Mathematical Programming, K. Schittkowski, ed., Springer-Verlag, Berlin, 1985, pp. 323–356.
[23] N. Z. Shor, *Minimization Methods for Non-Differentiable Functions*, Springer-Verlag, Berlin, 1985.
[24] K. C. Kiwiel, *Methods of Descent for Nondifferentiable Optimization*, Springer-Verlag, Berlin, 1985.
[25] T. Chen and B. A. Francis, *Input-output stability of sampled-data systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 50–58.
[26] L. H. Keel and S. P. Bhattacharryya, *Robust, fragile, or optimal*, IEEE Trans. Automat. Control, 42 (1997), pp. 1098–1105.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 5, pp. 1811–1843

RISK SENSITIVE PORTFOLIO MANAGEMENT WITH COX–INGERSOLL–ROSS INTEREST RATES: THE HJB EQUATION*

TOMASZ R. BIELECKI†, STANLEY R. PLISKA‡, AND SHUENN-JYI SHEU§

Abstract. This paper presents an application of risk sensitive control theory in financial decision making. The investor has an infinite horizon objective that can be interpreted as maximizing the portfolio's risk adjusted exponential growth rate. There are two assets, a stock and a bank account, and two underlying Brownian motions, so this model is incomplete. The novel feature here is that the interest rate for the bank account is governed by Cox–Ingersoll–Ross dynamics. This is significant for risk sensitive portfolio management because the factor process, unlike in the Gaussian and all other cases treated in the literature, cannot be negative (under appropriate parameterization).

Key words. risk sensitive control, optimal portfolios, Cox–Ingersoll–Ross interest rates, incomplete model

AMS subject classifications. 60J20, 90A09, 90C40, 93E20

DOI. 10.1137/S0363012903437952

1. Introduction.

Beginning with the pioneering work by Merton [22], [23], [24] and continuing through the recent books by Karatzas and Shreve [18] and Korn [20], some very sophisticated stochastic control methods have been developed for portfolio management. Virtually all of these studies make use of an expected utility criterion. But recently a new criterion has emerged from the control theory literature. Called the *risk sensitive* criterion, this was originally used (see, for example, Whittle [27]) for a decision maker seeking to maximize some (random) cash reward (or minimize some cash payment) while simultaneously being concerned about the risk or uncertainty in the size of the reward. Essentially, this criterion equals the expected value of the reward minus a penalty term that is proportional to the variance of the reward. The constant of proportionality is a parameter whose value can be chosen to achieve for the decision maker an appropriate trade-off between the expectation of the reward and its variance.

Recognizing its relevance to portfolio management, Bielecki and Pliska [5] applied the risk sensitive idea to a version of Merton's [23] intertemporal capital asset pricing model. The result was an infinite horizon criterion that they called the *risk adjusted growth rate* and viewed as being analogous to the classical Markowitz single-period approach except that instead of trading off single-period criteria the investor is trading off the portfolio's long run growth rate versus its average volatility (see Bielecki and Pliska [9] for a detailed study of various economic and mathematical properties of this criterion). Bielecki and Pliska also showed in [5] and subsequent work (see [2], [3], [4], [6], [7], [8], and [10]) that the resulting models usually have the virtue of being more

*Received by the editors November 25, 2003; accepted for publication (in revised form) February 19, 2005; published electronically December 6, 2005.
http://www.siam.org/journals/sicon/44-5/43795.html

†Applied Mathematics Department, Illinois Institute of Technology, E1 Building, 10 W. 32nd Street, Chicago, IL 60616 (bielecki@iit.edu). The research of this author was partially supported by NSF grants DMS-9971307 and DMS-0202851.

‡Department of Finance, University of Illinois at Chicago, 601 S. Morgan Street, Chicago, IL 60607-7124 (srpliska@uic.edu).

§Institute of Mathematics, Academia Sinica, Nankang, Taipei, Taiwan 11529, Republic of China (sheusj@math.sinica.edu.tw). The research of this author was partially supported by NSC 92-2115-M-001-035.

tractable than corresponding models which use traditional expected utility criteria. Other studies of the risk sensitive criterion for portfolio management include Bagchi and Kumar [1], Fleming and Sheu [14], [15], [16], Kuroda and Nagai [21], Nagai [25], and Nagai and Peng [26]. Kaise and Sheu [17] discuss the solution of a general equation (in R^n) that is related to the Hamilton–Jacobi–Bellman (HJB) equation in this paper.

Throughout all this work on risk sensitive portfolio management the underlying factor process, if any, was taken to be Gaussian or, at least (see Nagai [25]), a process whose domain is all of some Euclidean space. The aim of this paper is to provide some initial results on risk sensitive portfolio management for a case where this kind of condition does not hold. Since interest rate processes are commonly taken as factor processes and since the so-called Cox–Ingersoll–Ross [12] interest rate process (a popular one in the finance literature) cannot be negative, this model of the factor process was chosen for our object of study.

The result is a risk sensitive portfolio optimization model having a factor process whose domain is the nonnegative portion of the real line. Since this is a model of interest rates, it is more realistic than, say, Gaussian models, but it comes with a price: the resulting analysis is exceptionally lengthy, complex, and technical. This is true even though our model is rather simple, having just this scalar-valued factor process, two assets (the usual bank account and a risky stock), and two underlying Brownian motions. Consequently, this paper will study only the associated HJB equation, saving the verification of optimality and related issues for a future, separate paper.

After formulation of our model in section 2, the main results are presented in section 3. Chief among these is Theorem 3.1, which asserts the HJB equation has a unique solution. Needed for its proof and of separate interest are some results pertaining to a related, "truncated" problem: for some fixed number M the investor is required to keep all of his or her money in the bank account whenever the interest rate exceeds M. Existence of a unique solution to the HJB equation for this truncated problem is established by Theorem 3.2. Intuitively, one should expect the solution of the truncated HJB equation to converge to the solution of the original one as $M \to \infty$; this is indeed the case, as stated in Theorem 3.3. The rest of the paper is devoted to the proofs of these three theorems. Theorem 3.2 is proved in section 4, while the other two are proved in section 5. Various technical results are relegated to an appendix.

2. Formulation of the optimal risk sensitive asset management problem. In this section we formulate an optimal dynamic asset management problem featuring a risk sensitive optimality criterion. Let $(\Omega, \{\mathcal{F}_t\}_{t\geq 0}, \mathcal{F}, \mathbf{P})$ be the underlying probability space. The securities market involves a single factor, namely, an interest rate r that is subject to the so-called Cox–Ingersoll–Ross [12] dynamics

$$dr_t = -c(r_t - \bar{r})dt + \lambda\sqrt{r_t}dW_t, \tag{1}$$

where $c, \bar{r}$, and λ are three specified positive scalar parameters. In order to ensure that the interest rate process is always strictly positive (not only is absorption at zero unrealistic, but with the interest rate fixed at zero our investment problem becomes trivial: continuously rebalance to some fixed proportion), we make the following assumption (see Feller [13]).

Assumption 2.1. $2c\bar{r} > \lambda^2$.

There are two assets. One is the customary bank account:

$$\frac{dS_0(t)}{S_0(t)} = r_t dt; \tag{2}$$

here $S_0(t)$ represents the time-t amount of money in the bank account assuming none is added or withdrawn after time 0. The other asset is a stock (or stock index) whose price process satisfies

$$\frac{dS_1(t)}{S_1(t)} = \mu(r_t)dt + \sigma dW_t + \rho d\bar{W}_t. \tag{3}$$

Here W_t and $\bar{W}_t$ are two independent Brownian motions, σ and ρ are two specified scalar parameters, and

$$\mu(r) := \mu_1 + \mu_2 r, \tag{4}$$

where μ_1 and μ_2 are two specified scalar parameters. Note that with $\mu_2 \neq 0$ we can allow the level of interest rates to affect the return properties of the stock, and with $\sigma \neq 0$ the residuals of the interest rate process will be correlated with the residuals of the stock's return process. For instance, with suitable values of σ and ρ this correlation is negative.

Trading strategies will be adapted real-valued stochastic processes that are denoted h. We shall interpret h_t as the proportion of the investor's time-t wealth that is invested in the stock. In general, for each time t we allow h_t to be any real number; that is, we do not impose any constraints on h_t, such as short selling restrictions.

Remark 2.1. For a well posed optimization problem one needs additional assumptions about admissible trading strategies to the effect that the SDE (5) below admits a unique, strong solution. In particular, we need to require that for any admissible strategy the measure transformations used throughout the paper are well defined. It follows from the results of Bielecki, Pliska, and Yong [11] that a strategy that is a feedback strategy in the state variable r_t and grows linearly in r_t satisfies the above requirements of admissibility under appropriate parameterization. Other choices are also possible. Since the main concern of this paper is analysis of the HJB equation, and since derivation of the equation is done by a formal argument, then the choice of a class of admissible trading strategies is not an issue here. Of course, when one turns to studying the optimality of a strategy, the choice of the class of admissible trading strategies will be an important and, frequently, a quite delicate issue. This issue will be addressed in a future work.

The investor's time-t wealth will be denoted V_t. Under the trading strategy h, the corresponding wealth process V will satisfy

$$\frac{dV_t}{V_t} = [(1-h_t)r_t + h_t\mu(r_t)]dt + h_t(\sigma dW_t + \rho d\bar{W}_t). \tag{5}$$

By standard results, there exists a unique, strong, and almost surely positive solution to this equation; it is given by

$$V_t = V_0 exp\left(\int_0^t h_t\sigma dW_t + \int_0^t h_t\rho d\bar{W}_t + \int_0^t \left[-\frac{1}{2}(\sigma^2+\rho^2)h_t^2 + (1-h_t)r_t + h_t\mu(r_t)\right]dt\right). \tag{6}$$

In this paper we consider the following family of risk sensitized optimal investment problems, labeled as $\mathcal{P}_\theta$:

$$\begin{gathered} \textit{for } \theta \in (0,\infty), \textit{ maximize the risk sensitized expected growth rate} \\ J_\theta(v,r;h) := \liminf_{t\to\infty}(-2/\theta)t^{-1} ln\ \mathbf{E}^h\ [e^{-(\theta/2)lnV_t}|V_0 = v, r_0 = r] \\ \textit{over the class of all admissible investment processes } h, \end{gathered} \tag{7}$$

where $\mathbf{E}^h$ is the expectation with respect to $\mathbf{P}$. The notation $\mathbf{E}^h$ emphasizes that the expectation is evaluated for the wealth process V corresponding to the investment strategy h.

The parameter θ here is interpreted as the measure of the investor's attitude toward risk; the bigger the value of θ, the more risk averse the investor. This is because the criterion can be interpreted, at least approximately, as the portfolio's exponential growth rate minus a penalty term which equals $\theta/4$ times the portfolio's asymptotic variance. A comprehensive interpretation of this risk sensitive objective for portfolio management can be found in Bielecki and Pliska [9].

We note that the techniques used in this paper can also be used to study problems $\mathcal{P}_\theta$ for negative values of θ, corresponding to risk seeking investors. The risk null case, for $\theta = 0$, can be studied independently or as the limit of the risk averse situation when the risk sensitivity parameter θ goes to zero. However, in this paper we shall not consider cases where $\theta \leq 0$.

For much of what follows we find it convenient to introduce the scalar parameter

$$\gamma := -\theta/2. \tag{8}$$

Since θ is always strictly positive, the parameter γ should always be regarded as strictly negative. Moreover, the reader should keep in mind that corresponding to any appearance of the parameter γ is $\theta = -2\gamma$.

3. Analysis of the Hamilton–Jacobi–Bellman equation. In this section we formulate our model and present our main results concerning the HJB equation corresponding to the investor's portfolio optimization problem $\mathcal{P}_\theta$. We not only establish existence and uniqueness of a solution, but also establish some important properties of this solution. This analysis is rather involved, and so the balance of this paper is devoted to the proof of the results in this section.

In view of our risk sensitive objective, we are interested in computing the expectation of quantities like V_t^γ for some $\gamma < 0$. Since by (6)

$$V_t^\gamma = V_0^\gamma exp\Bigg(\gamma \int_0^t h_t \sigma dW_t + \gamma \int_0^t h_t \rho d\bar{W}_t + \int_0^t \gamma \Bigg[-\frac{1}{2}(\sigma^2 + \rho^2) h_t^2 + (1 - h_t) r_t + h_t \mu(r_t) \Bigg] dt \Bigg), \tag{9}$$

we recognize that it is convenient to make a Girsanov-type change of probability measure. In particular, it is straightforward to show for each trading strategy h and $T > 0$ that

$$E[V_T^\gamma] = \tilde{E}\left[V_0^\gamma exp\left(\gamma \int_0^T L(r_t, h_t) dt \right) \right], \tag{10}$$

where we have introduced the notation $\tilde{E}$ for expectation under the new probability measure and the additional functions

$$L(r, u) := -\frac{1}{2}(1 - \gamma)(\sigma^2 + \rho^2) u^2 + \bar{\mu}(r) u + r, \tag{11}$$

and

$$\bar{\mu}(r) := \mu(r) - r. \tag{12}$$

Moreover, under this new probability measure the dynamics for the interest rate process r are given by

$$dr_t = (-c(r_t - \bar{r}) + \gamma\sigma\lambda\sqrt{r_t}h_t)dt + \lambda\sqrt{r_t}d\tilde{W}_t, \tag{13}$$

where $\tilde{W}$ denotes a (scalar-valued) Brownian motion under this new probability measure.

Using standard methods of risk sensitive control theory (see, for example, [8], [15], and [21]), it is now straightforward to specify the HJB dynamic programming equation:

$$\begin{aligned}\Lambda = {} & \frac{1}{2}\lambda^2 r\frac{d^2\Phi}{dr^2} - c(r-\bar{r})\frac{d\Phi}{dr} + \frac{1}{2}\lambda^2 r\left(\frac{d\Phi}{dr}\right)^2 \\ & + \inf_{\{u\in\mathbf{R}\}}\left[\gamma\sigma\lambda\sqrt{r}u\frac{d\Phi}{dr} + \gamma L(r,u)\right], \quad r > 0.\end{aligned} \tag{14}$$

We seek a solution in terms of the scalar Λ and the *bias function* Φ such that Λ is the optimal risk adjusted growth rate in problem $\mathcal{P}_\theta$ and such that the minimal selector identifies an optimal (or, at least, an ϵ-optimal) trading strategy.

It is convenient to transform this equation into a simpler form. Since the stock proportion h_t is unrestricted, we see that the minimizing value of u in the HJB equation must satisfy the first order condition $\gamma\sigma\lambda\sqrt{r}\Phi' + \gamma[-(1-\gamma)(\sigma^2+\rho^2)u + \bar{\mu}(r)] = 0$. In other words, our candidate h^* for the optimal trading strategy will satisfy the expression $h_t^* = u^*(r_t)$, where

$$u^*(r) := \frac{1}{1-\gamma}\frac{1}{\sigma^2+\rho^2}\left(\bar{\mu}(r) + \sigma\lambda\sqrt{r}\frac{d\Phi}{dr}\right). \tag{15}$$

Remark 3.1. In view of Theorem 3.1 below, the candidate optimal strategy has a linear growth in the state variable r_t. Thus it is an admissible strategy (cf. Remark 2.1).

Substituting the preceding expression for u in the HJB equation, introducing the function

$$g := \frac{d\Phi}{dr},$$

and doing a little algebra enables one to see that the original HJB equation is equivalent to

$$\Lambda = \frac{1}{2}\lambda^2 r\frac{dg}{dr} + \frac{1}{2}\lambda^2 r\left(1 + \frac{\gamma}{1-\gamma}\frac{\sigma^2}{\sigma^2+\rho^2}\right)g^2 + b(r)g + d(r), \quad r > 0, \tag{16}$$

where we have introduced for convenience the functions

$$b(r) := -c(r-\bar{r}) + \frac{\gamma}{1-\gamma}\frac{\sigma\lambda}{\sigma^2+\rho^2}\sqrt{r}\bar{\mu}(r) \tag{17}$$

and

$$d(r) := \frac{1}{2}\frac{\gamma}{1-\gamma}\frac{1}{\sigma^2+\rho^2}[\bar{\mu}(r)]^2 + \gamma r. \tag{18}$$

The following theorem is our main result about the HJB equation.

THEOREM 3.1. *The HJB equation* (16) *has a unique solution* (Λ^*, g^*) *satisfying the following two properties:*

$$\lim_{r\to 0} g^*(r) = \frac{1}{c\bar{r}}\left[\Lambda^* - \frac{1}{2}\frac{\gamma}{1-\gamma}\frac{\mu_1^2}{\sigma^2+\rho^2}\right] \tag{19}$$

and either

$$\lim_{r\to\infty} \frac{g^*(r)}{\sqrt{r}} = -\left(1+\frac{\gamma}{1-\gamma}\frac{\sigma^2}{\sigma^2+\rho^2}\right)^{-1}\left[\frac{|\mu_2-1|}{\lambda}\sqrt{\frac{-\gamma}{1-\gamma}\frac{1}{\sigma^2+\rho^2}} + \frac{\gamma}{1-\gamma}\frac{\sigma}{\sigma^2+\rho^2}\frac{\mu_2-1}{\lambda}\right] \tag{20}$$

for $\mu_2 \neq 1$ *or*

$$\lim_{r\to\infty} g^*(r) = -\left(1+\frac{\gamma}{1-\gamma}\frac{\sigma^2}{\sigma^2+\rho^2}\right)^{-1}\left(\frac{c}{\lambda^2} - \left(\frac{c^2}{\lambda^4} - 2\frac{\gamma}{\lambda^2}\left(1+\frac{\gamma}{1-\gamma}\frac{\sigma^2}{\sigma^2+\rho^2}\right)\right)^{\frac{1}{2}}\right) \tag{21}$$

for $\mu_2 = 1$. *Moreover,* Λ^* *is characterized as the smallest* Λ *such that the HJB equation has a solution defined for all* r.

In order to study problem $\mathcal{P}_\theta$, as well as to investigate a related problem of separate interest, consider exactly the same problem except that now, for some arbitrary positive number M, we impose the trading strategy constraint that $h_t = 0$ if $r_t > M$. Analogous to the unconstrained problem, the dynamic programming equation for this constrained, truncated problem is

$$\Lambda_M = \frac{1}{2}\lambda^2 r\frac{dg}{dr} + \frac{1}{2}\lambda^2 r\left(1+\frac{\gamma}{1-\gamma}\frac{\sigma^2}{\sigma^2+\rho^2}\right)g^2 + b(r)g + d(r), \quad 0 < r \leq M, \tag{22}$$

$$\Lambda_M = \frac{1}{2}\lambda^2 r\frac{dg}{dr} + \frac{1}{2}\lambda^2 r g^2 - c(r-\bar{r})g + \gamma r, \quad r > M. \tag{23}$$

In addition, our candidate for the optimal trading strategy is now given by

$$u_M^*(r) := \frac{1}{1-\gamma}\frac{1}{\sigma^2+\rho^2}\Big(\bar{\mu}(r) + \sigma\lambda\sqrt{r}g(r)\Big), \qquad r \leq M, \tag{24}$$

$$u_M^*(r) := 0, \qquad r > M.$$

Moreover, for this constrained problem we have the following important result.

THEOREM 3.2. *The HJB equation* (22), (23) *for the constrained problem has a unique solution* (Λ_M^*, g_M^*) *satisfying the following two properties:*

$$\lim_{r\to 0} g_M^*(r) = \frac{1}{c\bar{r}}\left[\Lambda_M^* - \frac{1}{2}\frac{\gamma}{1-\gamma}\frac{\mu_1^2}{\sigma^2+\rho^2}\right] \tag{25}$$

and

$$\lim_{r\to\infty} g_M^*(r) = \frac{c}{\lambda^2} - \sqrt{\frac{c^2}{\lambda^4} - \frac{2\gamma}{\lambda^2}}. \tag{26}$$

As our next main result indicates, the solutions of the two kinds of investment problems are related in an intuitive manner.

THEOREM 3.3. *The following hold:*

$$\lim_{M\to\infty} g_M^*(r) = g^*(r) \quad \forall r > 0 \tag{27}$$

and

$$\lim_{M\to\infty} \Lambda_M^* = \Lambda^*. \tag{28}$$

Remark 3.2. For (16), there is a smallest Λ such that (16) has a smooth solution g. This follows from the argument in [17]. We can show that Λ^* in Theorem 3.1 is this smallest value of Λ; see Theorem 5.1. The argument in [17] is applicable to equations in multidimensional spaces, and therefore it can be applied to a model with several assets and multiple factor processes. However, it is difficult to fully understand and analyze the solutions in such multidimensional cases.

These three theorems are proved in the following two sections. We now conclude this section by suggesting a procedure for computing the solution (Λ^*, g^*) of (16). Suppose that for a given number Λ we can solve our equation (16) for a function g satisfying (19). Since Λ^* is characterized as the smallest Λ such that this solution g is finite for all $r > 0$, if some value Λ gives a finite solution, then $\Lambda^* \le \Lambda$. On the other hand, if some value Λ does not correspond to a finite g, then $\Lambda < \Lambda^*$. Hence a suitable iterative procedure should converge to (Λ^*, g^*). To be more precise, we first choose $\Lambda_1 \le \Lambda^* \le \Lambda_2$ by some rough estimate and follow this procedure:

1. $\Lambda = (\Lambda_1 + \Lambda_2)/2$ and solve (16) and (19). If the solution g exists for all $r > 0$, go to step 2. Otherwise, go to step 3.
2. Save the solution (Λ, g), redefine $\Lambda_1 := \Lambda_1,\ \Lambda_2 := \Lambda$, and go to step 1.
3. Redefine $\Lambda_1 := \Lambda,\ \Lambda_2 := \Lambda_2$, and go to step 1.

The solution (Λ, g) in step 2 will give a good approximation of (Λ^*, g^*) after a sufficient number of iterations.

4. Proof of Theorem 3.2. We begin by transforming the constrained HJB equation (22) to a version that will be more convenient for some proofs. Denoting

$$A := 1 + \frac{\gamma}{1-\gamma}\frac{\sigma^2}{\sigma^2+\rho^2},$$

$$\bar{\Lambda} := A\Lambda,$$

$$\bar{d}(r) := Ad(r),$$

and

$$\bar{g} := Ag, \tag{29}$$

we see by simple substitution that (22) is equivalent to

$$\bar{\Lambda} = \frac{1}{2}\lambda^2 r\frac{d\bar{g}}{dr} + \frac{1}{2}\lambda^2 r\bar{g}^2 + b(r)\bar{g} + \bar{d}(r), \quad 0 < r \le M. \tag{30}$$

We would like to know that this equation has a (possibly unique) solution $\bar{g}$ for an arbitrary $\bar{\Lambda}$, but establishing this is not so easy because the second term on the right-hand side is nonlinear and the coefficient of the first derivative term is degenerate at

$r = 0$ (so one cannot proceed with the analysis by considering the solution for all $r \geq 0$ and specifying the value of $\bar{g}(0)$). Our approach will be to address these issues by studying the function

$$\tilde{g}(r) := \bar{g}(r)e(r),$$

where

$$e(r) := r^{\frac{2c\bar{r}}{\lambda^2}} \exp\left(-\frac{2c}{\lambda^2}r + \frac{2\gamma\sigma}{(1-\gamma)\lambda(\sigma^2+\rho^2)}\int_0^r \frac{\bar{\mu}(s)}{\sqrt{s}}ds\right).$$

This is because $\bar{g}$ satisfies (30) if and only if $\tilde{g}$ satisfies

$$\frac{d\tilde{g}}{dr} + \frac{1}{e(r)}\tilde{g}^2 = \frac{2}{\lambda^2 r}e(r)[\bar{\Lambda} - \bar{d}(r)], \tag{31}$$

and this latter differential equation will be easier to analyze. Also, note for future arguments that (1) $e(r) > 0$ for all $r > 0$ and (2) equation (31) is equivalent to

$$\tilde{g}(r) = -\int_0^r \frac{1}{e(s)}\tilde{g}^2(s)ds + \int_0^r \frac{2}{\lambda^2 s}e(s)[\bar{\Lambda} - \bar{d}(s)]ds. \tag{32}$$

Using the preceding transformations we can prove (see Appendix A) the following initial key result in the proof of Theorem 3.2.

PROPOSITION 4.1. *If (Λ_M, g) is a solution of (22) defined on $(0, r_0]$ for some $r_0 > 0$, then either g satisfies (25), with g_M^* replaced with g, or*

$$\lim_{r\to 0} rg(r) = \frac{1}{A}\left(-\frac{2c\bar{r}}{\lambda^2} + 1\right). \tag{33}$$

Note that there is a trading strategy given by (15) associated with each solution g, where $g = d\Phi/dr$. Moreover, note that the asymptotic behavior stated in (33) implies that a solution satisfying (33) might not be compatible with our portfolio optimization problem. So from now on we shall focus on solutions of (22) that satisfy (25) rather than (33). The reason will become apparent below. In particular, see Corollary 4.1, which gives special properties of the solution satisfying (25).

Our next key result shows that there exists a unique solution g of (22) satisfying (25), at least a solution in some neighborhood of $r = 0$.

PROPOSITION 4.2. *Fix $\Lambda = \Lambda_M$. For small enough $r_0 > 0$ there exists a unique g satisfying (22) and (25) for all $r \in (0, r_0]$. It also satisfies*

$$|g(r)e(r)| \leq c_1 r^{\frac{2c\bar{r}}{\lambda^2}}, \qquad r \leq r_0, \tag{34}$$

for some positive number c_1. In addition, we have

$$g(r)e(r) \leq \int_0^r \frac{2}{\lambda^2 s}e(s)[\Lambda - d(s)]ds, \qquad r \leq r_0. \tag{35}$$

Proof. Since there is a correspondence between solutions of (22) and solutions of (32), it suffices to focus on the latter. For some suitable positive numbers r_0 and c_1 (to be decided later), consider the operator T defined for $f \in \mathbf{F}_{c_1}$, where

$$Tf(r) := -\int_0^r f^2(s)\frac{1}{e(s)}ds + \int_0^r \frac{2}{\lambda^2 s}e(s)[\bar{\Lambda} - \bar{d}(s)]ds$$

and

$$\mathbf{F}_{c_1} := \{f : |f(r)| \le c_1 r^{\frac{2c\bar{r}}{\lambda^2}}, 0 \le r \le r_0\}.$$

In order to show that $Tf \in \mathbf{F}_{c_1}$ we need to estimate $|Tf(r)|$. The first term in the definition of T is bounded by

$$c_1^2 \bar{c}_2 \int_0^r s^{\bar{g}} ds = c_1^2 c_2 r^{\delta+1},$$

where $\delta := \frac{2c\bar{r}}{\lambda^2}$ and $1/e(s) \le \bar{c}_2 s^{-\delta}$, $c_2 = \bar{c}_2/(1+\delta)$. The second term is bounded by

$$c_1(\bar{\Lambda})\bar{c}_3 \int_0^r s^{\delta-1} ds = c_1(\bar{\Lambda}) c_3 r^{\delta},$$

where $c_3 = \bar{c}_3/\delta$ and $e(r) \le \bar{c}_3 r^{\delta}$ for r small, and where

$$c_1(\bar{\Lambda}) := \max_{0<r\le 1} \left| \frac{2}{\lambda^2}[\bar{\Lambda} - \bar{d}(r)] \right|.$$

Therefore

$$|Tf(r)| \le [c_1^2 c_2 r + c_1(\bar{\Lambda})c_3] r^{\delta} \le [c_1^2 c_2 r_0 + c_1(\bar{\Lambda})c_3] r^{\delta}$$

if $r \le r_0$. It now follows by taking $c_1 = 2c_1(\bar{\Lambda})c_3$ and $r_0 = 1/[4c_2 c_3 c_1(\bar{\Lambda})]$ that

$$|Tf(r)| \le c_1 r^{\frac{2c\bar{r}}{\lambda^2}},$$

and so $T : \mathbf{F}_{c_1} \to \mathbf{F}_{c_1}$.

On the other hand, for $r \le r_0$

$$|Tf_1(r) - Tf_2(r)| \le \int_0^r |f_1(s) + f_2(s)||f_1(s) - f_2(s)| \frac{1}{e(s)} ds$$
$$\le ||f_1 - f_2|| 2c_1 \int_0^{r_0} s^{\frac{2c\bar{r}}{\lambda^2}} \frac{1}{e(s)} ds \le 2c_1 \bar{c}_2 r_0 ||f_1 - f_2||, \quad r \le r_0,$$

where $||\cdot||$ denotes the supnorm on $[0, r_0]$ and

$$\bar{c}_2 := \max_{r\le 1} \frac{1}{e(r)} r^{\frac{2c\bar{r}}{\lambda^2}}.$$

Hence by taking r_0 small enough so that $2c_1\bar{c}_2 r_0 < 1$, we see that T will be a contraction mapping from $\mathbf{F}_{c_1}$ into $\mathbf{F}_{c_1}$. Hence T has a unique fixed point, say $\tilde{g}$, which means that $\tilde{g}$ satisfies (32).

If we define $g(r) = \tilde{g}(r)/(Ae(r))$, then g is a solution of (22) defined on $(0, r_0]$. Therefore, in view of Proposition 4.1, either one of (33) or (25) holds. Since $\tilde{g}$ is in $\mathbf{F}_{c_1}$, it follows that we have (34). Then (33) cannot be true. Finally, (35) is a consequence of (32). This completes the proof of the proposition. □

The next main step is to show that one can choose Λ so that the value of r_0 in Proposition 4.2 can be taken to be M, that is, so that a solution of (22) will exist on all of $(0, M]$. The following result will be used in this step.

LEMMA 4.1. *Let g_1 and g_2 be the two solutions of* (22) *and* (25) *corresponding to values of Λ_M equal to Λ_1 and Λ_2, respectively. If $\Lambda_1 < \Lambda_2$, then $g_1 < g_2$.*

Proof. Because of the correspondence between solutions of (22) and solutions of (31), it suffices to study the latter. Since $\tilde{g}_1$ and $\tilde{g}_2$ both satisfy (31) (with their respective values of $\bar{\Lambda}$) we can subtract one equation from the other to obtain

$$\frac{d}{dr}(\tilde{g}_1 - \tilde{g}_2) + \frac{1}{e(r)}[\tilde{g}_1 + \tilde{g}_2][\tilde{g}_1 - \tilde{g}_2] = \frac{2}{\lambda^2 r} e(r)[\bar{\Lambda}_1 - \bar{\Lambda}_2].$$

We thus have

$$\tilde{g}_1(r) - \tilde{g}_2(r) = \int_0^r \frac{2e(s)}{\lambda^2 s}[\bar{\Lambda}_1 - \bar{\Lambda}_2] \exp\left(-\int_s^r \frac{\tilde{g}_1(u) + \tilde{g}_2(u)}{e(u)} du\right) ds.$$

This is strictly negative if $\bar{\Lambda}_1 < \bar{\Lambda}_2$, so Lemma 4.1 is established. □

COROLLARY 4.1. *For fixed Λ, suppose g defined on $(0, r_0]$ is the solution of (22) satisfying (25). Let $y < g(r_0)$, and suppose g_0 is a solution of (22) such that $g_0(r_0) = y$. Then $g_0(r)$ exists for $r \in (0, r_0]$ and g_0 satisfies (33).*

Remark 4.1. We omit a proof of this corollary since the argument is very similar to what was used for the proof of Lemma 4.1. While this corollary does not play a crucial role in the proof of Theorem 3.2, it sheds some light on why we are concerned with the boundary condition (33).

We note that if $g(r_0)$ is finite, then g is well defined for $r > r_0$, up to a (possibly infinite) point denoted $r(\bar{\Lambda})$, where $\lim_{r \to r(\bar{\Lambda})} g(r) = -\infty$ (if g explodes, then by (35) it explodes in the negative direction). For each $M > 0$ we now define

$$\bar{\Lambda}_*(M) := \inf\{\Lambda : \text{the corresponding solution } g \text{ of (22) satisfying (25) is finite for all } r \le M\}.$$

We then have the following proposition.

PROPOSITION 4.3. *Fix arbitrary $0 < M < \infty$. Then $\bar{\Lambda}_*(M) < \infty$ and, for each $\Lambda \ge \bar{\Lambda}_*(M)$, the corresponding solution g of (22) and (25) is finite for all $r \le M$. Moreover, $g(M) \to \infty$ as $\Lambda \to \infty$, and $g(M) \to -\infty$ as $\Lambda \to -\infty$.*

Remark 4.2. Note that in Lemma 5.2 below we prove that $\bar{\Lambda}_*(M) > -\infty$.

Proof. We prove $\bar{\Lambda}_*(M) < \infty$. The rest is a consequence of either Lemma 4.1 or a similar argument.

We first take an r_0 small enough and a finite $\bar{\Lambda}_0$, and

$$\tilde{g}_0(r) = \int_0^r \frac{2}{\lambda^2 s} e(s)(\bar{\Lambda}_0 - \bar{d}(s)) ds - \int_0^r \frac{1}{e(s)} \tilde{g}_0^2(s) ds, \quad r \le r_0.$$

We know that $\tilde{g}_0(r)$ is finite for $r \in [0, r_0]$. Now for $\theta > 0$ we consider

$$\tilde{g}_\theta(r) = \int_0^r \frac{2}{\lambda^2 s} e(s)(\bar{\Lambda}_0 + \theta - \bar{d}(s)) ds - \int_0^r \frac{1}{e(s)} \tilde{g}_\theta^2(s) ds, \quad r \le r_0.$$

This has solution $\tilde{g}_\theta(\cdot)$ in a neighborhood of 0 as given in Lemma 4.2 with $\bar{\Lambda} = \bar{\Lambda}_0 + \theta$. We know that $\tilde{g}_\theta(r)$ is finite for $r \in [0, r_0]$ and that

$$\tilde{g}_\theta(r) \ge \tilde{g}_0(r), \quad r \le r_0.$$

Let us fix $0 < r_1 < r_0$, a large $K > 0, K > \|\tilde{g}_0\|_{[r_1, r_0]}$, where $\|\tilde{g}_0\|_{[r_1, r_0]}$ is the maximum of $|\tilde{g}_0(r)|, r \in [r_1, r_0]$. There is a θ_0 such that for $\theta > \theta_0$, $\tilde{g}_\theta(\cdot)$ is increasing

for $r_1 \le r \le r_0$ if $|\tilde{g}_\theta(r)| \le K$. This is due to the following calculation:

$$\frac{d}{dr}\tilde{g}_\theta(r) = \frac{2}{\lambda^2 r}e(r)(\bar{\Lambda}_0 + \theta - \bar{d}(r)) - \frac{1}{e(r)}\tilde{g}_\theta^2(r)$$
$$\ge \frac{2}{\lambda^2 r_0}\inf_{[r_1,r_0]}\{e(r)(\bar{\Lambda}_0 + \theta_0 - \|\bar{d}\|_{[r_1,r_0]})\} - \left\|\frac{1}{e(\cdot)}\right\|_{[r_1,r_0]} K^2 > 0,$$

where the last inequality holds if θ_0 is large enough.

From this, for a fixed K, we must have $\tilde{g}_\theta(r_0) > K$ if θ is large enough. Suppose not. Then using the fact that $\tilde{g}_\theta(r) > \tilde{g}_0(r),\ r_1 \le r \le r_0$, and the above monotonicity result, we can conclude that

$$|\tilde{g}_\theta(r)| < K, \quad r_1 \le r \le r_0.$$

Thus

$$\frac{d}{dr}\tilde{g}_\theta(r) \ge \left(\frac{2}{\lambda^2 r_0}\inf_{[r_1,r_0]}\{e(r)(\bar{\Lambda}_0 + \theta - \|\bar{d}\|_{[r_1,r_0]})\} - \left\|\frac{1}{e(r)}\right\|_{[r_1,r_0]} K^2\right)$$

is larger than a given number (say L) for $r_1 \le r \le r_0$ if θ is large enough. For such θ,

$$\tilde{g}_\theta(r_0) = \tilde{g}_\theta(r_1) + \int_{r_1}^{r_0}\frac{d}{dr}\tilde{g}_\theta(r)dr \ge \tilde{g}_0(r_1) + L(r_0 - r_1),$$

and this is larger than K if L is large enough. This gives a contradiction.

Next, for a fixed $K > 0$, if θ is large enough, then $\tilde{g}_\theta(r_0) > K$ implies $\tilde{g}_\theta(r) > K,\ r_0 \le r \le M$. This follows by using the properties that

$$\inf_{r_0 \le r \le M}\frac{1}{r}e(r) > 0, \qquad \sup_{r_0 \le r \le M}\frac{1}{e(r)} < \infty$$

and the estimate

$$\frac{d}{dr}\tilde{g}_\theta(r) \ge \frac{2}{\lambda^2}\inf_{r_0 \le r \le M}\left\{\frac{1}{r}e(r)(\bar{\Lambda}_0 + \theta_0 - \|\bar{d}\|_{[r_0,M]})\right\} - \left\|\frac{1}{e(r)}\right\|_{[r_0,M]} K^2 > 0$$

for an $r_0 \le r \le M$ satisfying $\tilde{g}_\theta(r) = K$ if θ is large enough.

We conclude from the above analysis that for a $K > \|\tilde{g}_0\|_{[r_1,r_0]}$, there is a θ sufficiently large such that $\tilde{g}_\theta(M) > K$. This implies $\bar{\Lambda}_*(M) \le \bar{\Lambda}_0 + \theta < \infty$. As a consequence of this argument, we also have that $\tilde{g}_\theta(M)$ tends to ∞ as θ tends to ∞. This ends the proof of this proposition. □

We now know there exists a solution of (22) and (25) on all of $(0, M]$, so we turn to the study of the solution of the constrained dynamic programming equation for $r > M$, that is, (23). But the solution of this differential equation must satisfy the boundary condition at $r = M$ that has $g(M)$ taking the value that comes from the solution of (22) and (25) for $r \le M$. Hence the solution of (23) is well defined for $r > M$, at least in some neighborhood of M.

LEMMA 4.2. *Given a specified value of $g(M)$, (23) has a unique solution g on $[M, r_1)$, where $r_1 := \sup\{r > M : g(r) > -\infty\}$. Also, there exists some $K < \infty$, which does not depend on r_1 (but may depend on $g(M)$), such that $g(r) \le K$ on $[M, r_1)$.*

Proof. It is sufficient to prove the existence of K. The existence of r_1 follows from the theory of ordinary differential equations.

First note that (23) can be rewritten as

$$\frac{dg}{dr} - \frac{2c}{\lambda^2}\left[1 - \frac{\bar{r}}{r}\right] g + g^2 = \frac{2}{\lambda^2 r}[\Lambda_M - \gamma r], \quad r \geq M. \tag{36}$$

It suffices to show that if a solution is such that $g(r_0) < c_2$, then $g(r) < c_2$ for all $r > r_0$, where r_0 and c_2 here are large. Suppose, on the contrary, that there is some $r_1 > r_0$ such that $g(r) < c_2$ for $r_0 \leq r < r_1$ and $g(r_1) = c_2$. Then by differential equation (36) we must have $\frac{dg}{dr}(r_1) < 0$. But this is a contradiction, so Lemma 4.2 is established. □

For fixed M and any $\Lambda > \bar{\Lambda}_*(M)$ we know by Proposition 4.3 that (22) with $\Lambda_M = \Lambda$ has a solution g on $(0, M]$ with $g(M)$ finite. So corresponding to each such Λ we can, as in the following lemma, consider the solution g of (23) on $[M, r_1)$ that takes this corresponding value of $g(M)$ at $r = M$. In other words, for each $\Lambda > \bar{\Lambda}_*(M)$ we have a solution of (22) and (23) that is continuous on $(0, r_1)$ for some $r_1 > M$.

LEMMA 4.3. *Fix M and let g_1 and g_2 be two solutions of* (22), (23), *and* (25) *corresponding to values of Λ_M equal to Λ_1 and Λ_2, respectively, where $\Lambda_1, \Lambda_2 > \bar{\Lambda}_*(M)$. Then $\Lambda_1 < \Lambda_2$ implies $g_1(r) < g_2(r)$ if g_1 is defined at r.*

Proof. First consider two differential equations (36), one satisfied by $(g_1, \Lambda_M = \Lambda_1)$ and the other by $(g_2, \Lambda_M = \Lambda_2)$. Subtracting one from the other gives

$$\frac{d}{dr}(g_2 - g_1) + \left[-\frac{2c}{\lambda^2}(1 - \bar{r}/r) + g_2 + g_1\right](g_2 - g_1) = \frac{2}{\lambda^2 r}[\Lambda_2 - \Lambda_1],$$

in which case

$$g_2(r) - g_1(r) = \exp\left(-\int_M^r \left(-\frac{2c}{\lambda^2}(1 - \bar{r}/s) + g_1(s) + g_2(s)\right) ds\right)[g_2(M) - g_1(M)]$$

$$+ \int_M^r \frac{2}{\lambda^2 s}(\Lambda_2 - \Lambda_1) \exp\left(-\int_s^r \left(-\frac{2c}{\lambda^2}(1 - \bar{r}/u) + g_1(u) + g_2(u)\right) du\right) ds.$$

Since $g_2(M) - g_1(M) > 0$ by Lemma 4.1, Lemma 4.3 follows from this. □

For each $M > 0$ we now define

$$\Lambda_M^* := \inf\{\Lambda_M : \text{the corresponding solution } g \text{ of (22), (23) satisfying (25) is finite for all } r > 0\},$$

and we observe that $\Lambda_M^* \geq \bar{\Lambda}_*(M)$. We now have the following key result.

PROPOSITION 4.4. *For each fixed number $M < \infty$ we have $\Lambda_M^* < \infty$.*

Proof. We need to prove the existence of a Λ_M such that $g(r)$ is finite for all $r > 0$, where g is the solution for (22), (23), and (25). By (36), if $g(M) > 0$ and $\Lambda_M > 0$, then $g(r) > 0$ for all $r > M$. We can show by the argument in the proof of Lemma 4.3 that $g(M) > 0$ if Λ_M is sufficiently large. This completes the proof. □

By Lemma 4.3 we know for $\Lambda_M \geq \Lambda_M^*$ that there exists a solution of (22), (23), and (25) on all of $(0, \infty)$. In order to identify the "correct" solution we now investigate the limiting behavior of these solutions g as $r \to \infty$.

PROPOSITION 4.5. *Fix $M < \infty$ and arbitrary $\Lambda_M \geq \Lambda_M^*$, and consider the solution g of (22), (23) satisfying (25). Then exactly one of the following two conditions will hold; that is, either*

$$\lim_{r\to\infty} g(r) = \frac{c}{\lambda^2} - \sqrt{\frac{c^2}{\lambda^4} - \frac{2\gamma}{\lambda^2}} \tag{37}$$

or

$$\lim_{r\to\infty} g(r) = \frac{c}{\lambda^2} + \sqrt{\frac{c^2}{\lambda^4} - \frac{2\gamma}{\lambda^2}}. \tag{38}$$

Proof. Denote

$$\alpha := \frac{c}{\lambda^2} - \sqrt{\frac{c^2}{\lambda^4} - \frac{2\gamma}{\lambda^2}},$$

and note that α is negative and satisfies

$$\alpha^2 - \frac{2c}{\lambda^2}\alpha + \frac{2\gamma}{\lambda^2} = 0.$$

Next, define

$$\hat{g}(r) = g(r) - \alpha,$$

and note that, in view of (23), we must have

$$\frac{d\hat{g}}{dr} + \left(-2\sqrt{\frac{c^2}{\lambda^4} - \frac{2\gamma}{\lambda^2}} + \frac{2c\bar{r}}{\lambda^2}\frac{1}{r}\right)\hat{g} + \hat{g}^2 = \left(\frac{2\Lambda_M}{\lambda^2} - \frac{2c\bar{r}}{\lambda^2}\alpha\right)\frac{1}{r}, \quad r > M. \tag{39}$$

We now claim there is c_1 large enough such that

$$\hat{g}(r) > -c_1/r, \quad r \geq M. \tag{40}$$

Denote $f(r) = \hat{g}(r) + c_1/r$. Then (40) is equivalent to

$$f(r) > 0, \quad r \geq M.$$

To prove this, we have the following observation. By (39), we easily see that

$$\frac{df(r)}{dr} < 0$$

if $f(r) = 0$. This implies that if $f(r_0) \leq 0$ for some $r_0 > 0$, then

$$f(r) < 0, \quad r > r_0.$$

We shall show that this cannot be true. Otherwise, using (39), we have

$$\frac{d\hat{g}}{dr} + \frac{1}{2}\hat{g}^2 < 0, \quad r \geq r_0.$$

This, in turn, implies

$$-\frac{1}{\hat{g}(r)} + \frac{1}{\hat{g}(r_0)} + \frac{1}{2}(r - r_0) < 0, \quad r \geq r_0.$$

But this cannot be true for all $r \geq r_0$, a contradiction that implies (40).

We now consider two cases, depending upon whether or not

$$\frac{2\Lambda_M}{\lambda^2} - \frac{2c\bar{r}}{\lambda^2}\alpha > 0. \tag{41}$$

If (41) is true and $\hat{g}(r_0) > 0$ for some $r_0 > M$, then $\hat{g}(r) > 0$ for all $r \geq r_0$. From this, we conclude that one of the following two possibilities holds: either there is $r_0 > M$ such that

$$\hat{g}(r) > 0, \quad r \geq r_0, \tag{42}$$

or there is $r_0 > M$ such that

$$\hat{g}(r) < 0, \quad r \geq r_0. \tag{43}$$

We have the same conclusion if the opposite of (41) holds, so it suffices to consider (42) and (43) separately. First we assume (43). This together with (40) implies

$$\lim_{r\to\infty} \hat{g}(r) = 0.$$

In other words,

$$\lim_{r\to\infty} g(r) = \alpha, \tag{44}$$

which is equivalent to (37) in this case.

For the rest of this proof we shall assume (42). We consider (41) only (the analysis for other case is similar). We rewrite (39) as follows:

$$\frac{d\hat{g}}{dr} = \left(\left(2\sqrt{\frac{c^2}{\lambda^4} - \frac{2\gamma}{\lambda^2}} - \frac{2c\bar{r}}{\lambda^2}\frac{1}{r}\right) - \hat{g}\right)\hat{g} + \left(\frac{2\Lambda_M}{\lambda^2} - \frac{2c\bar{r}}{\lambda^2}\alpha\right)\frac{1}{r}, \quad r > M. \tag{45}$$

Denote

$$\bar{f}(r) = \hat{g}(r) + \left(-2\sqrt{\frac{c^2}{\lambda^4} - \frac{2\gamma}{\lambda^2}} + \frac{2c\bar{r}}{\lambda^2}\frac{1}{r}\right).$$

Assuming (42), by using (45) it is easy to see that

$$\frac{d\bar{f}(r)}{dr} > 0$$

if $r \geq r_0$ and $\bar{f}(r) < 0$. Moreover, if $\bar{f}(r) < 0$, then

$$\frac{d\hat{g}(r)}{dr} > \left(\frac{2\Lambda_M}{\lambda^2} - \frac{2c\bar{r}}{\lambda^2}\alpha\right)\frac{1}{r}.$$

These two observations imply that we cannot have $\bar{f}(r) < 0$ for all $r \geq r_0$, so there must exist some $r_1 > r_0$ such that

$$\hat{g}(r) + \left(-2\sqrt{\frac{c^2}{\lambda^4} - \frac{2\gamma}{\lambda^2}} + \frac{2c\bar{r}}{\lambda^2}\frac{1}{r}\right) > 0, \quad r \geq r_1. \tag{46}$$

Using the same argument applied to

$$\hat{g}(r) + \left(-2\sqrt{\frac{c^2}{\lambda^4} - \frac{2\gamma}{\lambda^2}} + \frac{2c\bar{r}}{\lambda^2}\frac{1}{r} \right) - \frac{c_2}{r}$$

for a large c_2, we can prove the existence of $r_2 > r_1$ such that

$$\hat{g}(r) + \left(-2\sqrt{\frac{c^2}{\lambda^4} - \frac{2\gamma}{\lambda^2}} + \frac{2c\bar{r}}{\lambda^2}\frac{1}{r} \right) - \frac{c_2}{r} < 0, \quad r \geq r_2. \tag{47}$$

From (46) and (47), we have

$$\lim_{r\to\infty} \hat{g}(r) = 2\sqrt{\frac{c^2}{\lambda^4} - \frac{2\gamma}{\lambda^2}}.$$

This is equivalent to (38). This together with (44) completes the proof of Proposition 4.5. □

Remark 4.3. We shall pay special attention to the limiting behavior (37); (38) will now be ignored. There are reasons for doing this. It will be shown below that the solution satisfying (37) is unique and that it characterizes the solution associated with Λ_M^*. These interesting properties may also be considered as evidence that the minimality of Λ_M^* will play some special role for the portfolio optimization problem. A general study of multidimensional problems related to this observation can be found in [17].

We are now interested in the smallest Λ_M such that (22), (23), and (25) have a solution for all r. The following lemma states that there is at most one value of Λ_M giving a solution of (22), (23), and (25) that also satisfies (37).

LEMMA 4.4. *Suppose g_1 and g_2 are two solutions of* (22), (23), *and* (25) *corresponding to $\Lambda_M = \Lambda_1, \Lambda_2$, respectively. If both g_1 and g_2 satisfy* (37), *then $g_1 = g_2$ and $\Lambda_1 = \Lambda_2$.*

Proof. Subtracting the equation for g_2 from the equation for g_1 gives

$$\frac{d}{dr}(g_2 - g_1) + \left(-\frac{2c}{\lambda^2} + \frac{2c\bar{r}}{\lambda^2}\frac{1}{r} + g_1 + g_2 \right)(g_2 - g_1) = \frac{2}{\lambda^2 r}(\Lambda_2 - \Lambda_1).$$

By (37) we then have

$$-(g_2 - g_1)(r)\bar{e}(r) = \int_r^\infty \frac{2}{\lambda^2 s}(\Lambda_2 - \Lambda_1)\bar{e}(s)ds, \tag{48}$$

where we have introduced the function

$$\bar{e}(r) := \exp\left\{ \int_{r_0}^r \left(-\frac{2c}{\lambda^2} + \frac{2c\bar{r}}{\lambda^2}\frac{1}{s} + g_1(s) + g_2(s) \right) ds \right\}.$$

Here $r_0 > M$ is fixed. The integral on the right-hand side of (48) is finite by (37). Moreover, (48) implies $g_2 - g_1 > 0$ if $\Lambda_2 - \Lambda_1 < 0$. But we also have $g_2 - g_1 < 0$ if $\Lambda_2 - \Lambda_1 < 0$. Therefore $\Lambda_1 = \Lambda_2$ and $g_1 = g_2$; that is, the proof of Lemma 4.4 is completed. □

It remains to prove the solution g^* corresponding to $\Lambda_M = \Lambda_M^*$ satisfies (37). In other words, with Lemma 4.4 establishing uniqueness, it remains to establish existence. This is a consequence of the following lemma, because if for $\Lambda_M = \bar{\Lambda}$ the

corresponding limit is (38), then for all $\Lambda_M < \bar{\Lambda}$ in some neighborhood of $\bar{\Lambda}$ the corresponding limits also satisfy (38). Hence if there exists a solution for $\Lambda_M = \Lambda_M^*$ (the infimum of Λ_M for which there exists a solution; Λ_M^* is finite by Proposition 4.3 above and Lemma 5.2 below, and the infimum is attained by the same kind of argument used below in the proof of Theorem 5.1), this solution must satisfy the other limit, namely, (37).

LEMMA 4.5. *If $\Lambda_M = \hat{\Lambda}$ is such that the corresponding solution of* (22), (23) *satisfies* (25) *and* (38), *then there exists some $\delta > 0$ such that for any $\Lambda_M > \hat{\Lambda} - \delta$ the solution of* (22), (23) *exists for all r.*

Proof. Let $\hat{g}$ be the solution corresponding to $\hat{\Lambda}$; thus

$$\frac{d\hat{g}}{dr} - \frac{2c}{\lambda^2}\left(1 - \frac{\bar{r}}{r}\right)\hat{g} + \hat{g}^2 = \frac{2}{\lambda^2 r}(\hat{\Lambda} - \gamma r), \quad r > M.$$

With g being the solution of (22) and (23) corresponding to Λ, write

$$\bar{g} := g - \hat{g},$$

and so

$$\frac{d\bar{g}}{dr} + \frac{d\hat{g}}{dr} - \frac{2c}{\lambda^2}\left(1 - \frac{\bar{r}}{r}\right)(\bar{g} + \hat{g}) + (\bar{g} + \hat{g})^2 = \frac{2}{\lambda^2 r}(\Lambda - \gamma r).$$

This implies

$$\frac{d\bar{g}}{dr} + \left(\frac{2c}{\lambda^2}\left(1 - \frac{\bar{r}}{r}\right) + 2\hat{g}\right)\bar{g} + \bar{g}^2 = \frac{2}{\lambda^2 r}(\Lambda - \hat{\Lambda}). \tag{49}$$

We now seek the solution of (49) such that $||\bar{g}|| \le \delta_1$, where

$$||\bar{g}|| := \sup_{r \ge M} |\bar{g}(r)|.$$

Here δ_1 will be chosen later in a manner which depends on δ, where $|\Lambda - \hat{\Lambda}| < \delta$. Note that (49) can be rewritten as

$$\bar{g}(r) = \bar{g}(M)\frac{1}{\bar{e}(r)} - \int_M^r \frac{\bar{e}(s)}{\bar{e}(r)}\bar{g}^2(s)ds + \frac{2}{\lambda^2}(\Lambda - \hat{\Lambda})\int_M^r \frac{\bar{e}(s)}{\bar{e}(r)}\frac{1}{s}ds,$$

where we introduce the function

$$\bar{e}(r) := \exp\left\{\int_M^r \left(-\frac{2c}{\lambda^2}(1 - \bar{r}/s) + 2\hat{g}(s)\right)ds\right\}.$$

We use again the fixed-point argument to get a solution g. Denote

$$\mathbf{F} := \{f : [M, \infty) \to \mathbf{R}, ||f|| \le \delta_1, f(M) = \bar{g}(M)\},$$

where $\bar{g}(M) = g(M) - \hat{g}(M)$, and where g is the solution of (22) corresponding to $\Lambda_M = \Lambda$. We denote for $f \in \mathbf{F}$

$$Tf(r) := \bar{g}(M)\frac{1}{\bar{e}(r)} - \int_M^r \frac{\bar{e}(s)}{\bar{e}(r)}f^2(s)ds + \frac{2}{\lambda^2}(\Lambda - \hat{\Lambda})\int_M^r \frac{\bar{e}(s)}{\bar{e}(r)}\frac{1}{s}ds.$$

We know $\bar{g}(M) \to 0$ if $\Lambda \to \hat{\Lambda}$. We consider

$$|\Lambda - \hat{\Lambda}| < \delta, \quad \bar{\delta} = |\bar{g}(M)| \max_{r \le M} \frac{1}{\bar{e}(r)},$$

where δ is small. Then take $\delta_1 > 0$ satisfying

$$\delta_1^2 \sup_{r \ge M} \int_M^r \frac{\bar{e}(s)}{\bar{e}(r)} ds + \bar{\delta} + \frac{2\delta}{\lambda^2} \sup_{r \ge M} \left\{ \int_M^r \frac{\bar{e}(s)}{\bar{e}(r)} \frac{1}{s} ds \right\} < \delta_1.$$

Note that for δ small enough we can take

$$\delta_1 = 2 \left(\bar{\delta} + \frac{2\delta}{\lambda^2} \sup_{r \ge M} \left\{ \int_M^r \frac{\bar{e}(s)}{\bar{e}(r)} \frac{1}{s} ds \right\} \right).$$

Then it is not difficult to show that the operator $T : \mathbf{F} \to \mathbf{F}$. Moreover, for arbitrary $f_1, f_2 \in \mathbf{F}$ we have

$$||Tf_1 - Tf_2|| \le 2\delta_1 \sup_{r \ge M} \frac{1}{\bar{e}(r)} \int_M^r \bar{e}(s) ds ||f_1 - f_2|| = K||f_1 - f_2||.$$

By taking δ_1 small enough one has the number $K < 1$. Then T is a contraction with a unique fixed point in $\mathbf{F}$, which is the unique solution of (23). This completes the proof of Lemma 4.5. □

5. Proofs of Theorems 3.1 and 3.3. The proofs of Theorems 3.1 and 3.3 are accomplished by the three propositions in this section. The first of these shows that the solutions of Theorem 3.2 converge as $M \to \infty$ to a solution of the HJB equation (16) that also satisfies conditions (19) and either (20) (if $\mu_2 \ne 1$) or (21) (if $\mu_2 = 1$). Later in this section we will show uniqueness, thereby completing the proofs of both Theorems 3.1 and 3.3.

PROPOSITION 5.1. *Let Λ_M^* and $g_M^*(r)$ be as in Theorem* 3.2. *Then $\Lambda_M^* \to \Lambda$ and $g_M^*(r) \to g(r)$ as $M \to \infty$, where Λ and $g(r)$ satisfy* (16) *and $g(r)$ also satisfies* (19) *and either* (20) *(in the case $\mu_2 \ne 1$) or* (21) *(in the case $\mu_2 = 1$).*

To prove Proposition 5.1 we need the following four lemmas. The first two of these are based upon the following equation:

$$(50)\quad \Lambda = \frac{1}{2}\lambda^2 r \frac{dg}{dr} + \frac{1}{2}\lambda^2 r \left(1 + \frac{\gamma}{1-\gamma} \frac{\sigma^2}{\sigma^2 + \rho^2}\right) g^2 + b(r)g + d(r), \quad r \le R_0 + 1.$$

Here $R_0 > 0$ is a specified constant, and it is noteworthy that this equation is essentially the same as (22), which is part of the dynamic programming equation in Theorem 3.2.

LEMMA 5.1. *Let $r_0 < R_0$ be arbitrary. Then there is $K > 0$, depending on r_0, R_0, and Λ, such that if* (50) *has a solution g, then*

$$|g(r)| \le K, \quad r_0 \le r \le R_0.$$

Moreover, K can be chosen to be increasing in Λ. Therefore, for r_0, R_0, Λ fixed, the set $\{|g(r)| : r_0 \le r \le R_0; g$ satisfies (50)$\}$ *is bounded.*

Remark 5.1. If the value of the ODE solution g is specified at $r_0 < R_0$, say, then the values of g for all r will be determined. The most interesting part of this lemma is the conclusion that regardless of the initial value we choose for g at r_0, if $g(r)$ is

finite in $(0, R_0+1]$, then $|g(r)| \le K$ for all $r_0 \le r \le R_0$, where K is independent of g, although it may depend on R_0. Consequently, in order to have a solution of (50) we cannot arbitrarily assign a value of g at r_0 (that is, $g(r_0)$ needs to satisfy $|g(r_0)| \le K$). Regarding the dependence of K on Λ, an expression for K is provided at the end of the proof that follows. This dependence will be used in the proof of Proposition 5.1.

Proof. Let g satisfy (50). Take $\phi : [0,\infty) \to [0,1]$ smooth such that

$$\begin{aligned} \phi(r) &= 1, \quad r_0 \le r \le R_0, \\ &= 0, \quad 0 \le r \le \tfrac{r_0}{2},\ R_0+1 < r. \end{aligned} \tag{51}$$

Without loss of generality, we can take ϕ such that it satisfies the following property:

$$\left| \frac{1}{\sqrt{\phi(r)}} \frac{d\phi(r)}{dr} \right| \le K_1.$$

Here K_1 is some number that may depend on r_0 and R_0. To see this, we denote $h(r)$ by

$$h(r) := \frac{1}{\sqrt{\phi(r)}} \frac{d\phi(r)}{dr}, \quad r \ge R_0,$$

and so

$$\sqrt{\phi(r)} = 1 + 2\int_{R_0}^{r} h(u)du, \quad r > R_0.$$

We then choose $h(\cdot)$ such that $h(\cdot)$ is bounded and

$$2\int_{R_0}^{R_0+1} h(u)du = -1.$$

Moreover, we choose $h(r) = 0, r \ge R_0+1$. The derivatives of h of any order at R_0 and R_0+1 are 0. Thus ϕ satisfies the required property on $[R_0,\infty)$. We can apply a similar argument for $r \in (0, r_0]$.

Consider

$$f(r) = \frac{1}{2}\phi(r)g^2(r).$$

Then $f(r)$ takes a maximum at some r_1 satisfying $r_0/2 \le r_1 \le R_0+1$. Denote

$$X^2 = 2f(r_1) = 2\max f(r).$$

Then

$$\frac{df}{dr}(r_1) = 0;$$

that is,

$$\phi(r_1)g(r_1)\frac{dg}{dr}(r_1) = -\frac{1}{2}g^2(r_1)\frac{d\phi}{dr}(r_1). \tag{52}$$

We multiply (50) at $r = r_1$ by $\phi(r_1)g(r_1)$ and use (52) to get

$$\begin{aligned} \Lambda\phi(r_1)g(r_1) = &-\frac{1}{4}\lambda^2 r_1 g^2(r_1)\frac{d\phi}{dr}(r_1) + \frac{1}{2}\lambda^2 r_1\left(1 + \frac{\gamma}{1-\gamma}\frac{\sigma^2}{\sigma^2+\rho^2}\right)\phi(r_1)g^3(r_1) \\ &+ b(r_1)\phi(r_1)g^2(r_1) + d(r_1)\phi(r_1)g(r_1). \end{aligned}$$

Assume $g(r_1) \neq 0$. Then divide the above relation by $g(r_1)$ to obtain

$$\Lambda\phi(r_1) = -\frac{1}{4}\lambda^2 r_1 g(r_1)\frac{d\phi}{dr}(r_1) + \frac{1}{2}\lambda^2 r_1\left(1 + \frac{\gamma}{1-\gamma}\frac{\sigma^2}{\sigma^2+\rho^2}\right)\phi(r_1)g^2(r_1) + b(r_1)\phi(r_1)g(r_1) + d(r_1)\phi(r_1).$$

Then

$$X^2 + 2\alpha X = \beta, \tag{53}$$

where

$$\alpha = \alpha(r_1) = \frac{1}{\lambda^2 r_1(1 + \frac{\gamma}{1-\gamma}\frac{\sigma^2}{\sigma^2+\rho^2})}\left(b(r_1)\sqrt{\phi(r_1)} - \frac{1}{4}\lambda^2 r_1 \frac{1}{\sqrt{\phi(r_1)}}\frac{d\phi}{dr}(r_1)\right),$$

$$\beta = \beta(r_1) = \frac{2}{\lambda^2 r_1(1 + \frac{\gamma}{1-\gamma}\frac{\sigma^2}{\sigma^2+\rho^2})}(-d(r_1)\phi(r_1) + \Lambda\phi(r_1)).$$

From (53),

$$X = -\alpha \pm \sqrt{\alpha^2 + \beta},$$

in which case

$$|X| \leq |\alpha| + \sqrt{\alpha^2 + \beta}.$$

We see $|X| \leq K$, where

$$K = \max\left\{|\alpha(r)| + \sqrt{\alpha(r)^2 + \beta(r)}, \frac{r_0}{2} \leq r \leq R_0 + 1\right\}$$

so that K depends on r_0, R_0, and Λ. Since

$$\max_{r_0 \leq r \leq R_0} |g(r)|^2 \leq X^2,$$

the result follows. □

The next lemma says that the set of all Λ such that (50) has a solution is bounded below.

LEMMA 5.2. *For a fixed R_0, there is a $\Lambda(R_0)$ such that if* (50) *has a solution g, then $\Lambda \geq \Lambda(R_0)$.*

Proof. We take $0 < r_0 < R_0$ and a smooth function ϕ as in (51). We can define $\hat{b}(r)$ for all $r > 0$ such that

$$\hat{b}(r) = b(r), \quad r \leq R_0$$

and such that the diffusion process defined by

$$d\hat{r}(t) = \hat{b}(\hat{r}(t))dt + \lambda\sqrt{\hat{r}(t)}dB(t)$$

has an invariant density that we denote by $\hat{p}(r)$. See [19].

Denote

$$\Phi(r) = \exp\left(\left(1 + \frac{\gamma}{1-\gamma}\frac{\sigma^2}{\sigma^2+\rho^2}\right) W(r)\right),$$

where

$$W(r) = \int_{r_0}^{r} g(u)du.$$

Then

$$\hat{L}\Phi(r) = \hat{\Lambda}\Phi(r) - \hat{d}(r)\Phi(r), \quad r \le R_0,$$

where

$$\hat{\Lambda} = \left(1 + \frac{\gamma}{1-\gamma}\frac{\sigma^2}{\sigma^2+\rho^2}\right)\Lambda,$$

$$\hat{d}(r) = \left(1 + \frac{\gamma}{1-\gamma}\frac{\sigma^2}{\sigma^2+\rho^2}\right) d(r),$$

$$\hat{L}f(r) = \frac{1}{2}\lambda^2 r\frac{d^2 f}{dr^2}(r) + \hat{b}(r)\frac{df}{dr}(r).$$

We have

$$\int \hat{L}\Phi(r)\phi(r)\hat{p}(r)dr = \int \hat{\Lambda}\Phi(r)\phi(r)\hat{p}(r)dr - \int \hat{d}(r)\Phi(r)\phi(r)\hat{p}(r)dr.$$

The equation for the invariant density $\hat{p}$ is

$$\frac{1}{2}\frac{d^2}{dr^2}(\lambda^2 r\hat{p}(r)) - \frac{d}{dr}(\hat{b}(r)\hat{p}(r)) = 0.$$

In one dimension, we have

$$\frac{1}{2}\frac{d}{dr}(\lambda^2 r\hat{p}(r)) - \hat{b}(r)\hat{p}(r) = 0.$$

From this and the integration by parts formula we then have

$$\int \hat{L}\Phi(r)\phi(r)\hat{p}(r)dr = -\frac{1}{2}\int \lambda^2 r\frac{d\Phi}{dr}(r)\frac{d\phi}{dr}(r)\hat{p}(r)dr.$$

Consequently,

$$\hat{\Lambda}\int \Phi(r)\phi(r)\hat{p}(r)dr = \int \hat{d}(r)\Phi(r)\phi(r)\hat{p}(r)dr - \frac{1}{2}\int \lambda^2 r\frac{d\Phi}{dr}(r)\frac{d\phi}{dr}(r)\hat{p}(r)dr. \tag{54}$$

By Lemma 5.1 there is some number K, which depends on Λ, such that

$$\frac{1}{K} \le \Phi(r) \le K,$$

$$\left|\frac{d\Phi}{dr}(r)\right| \le K$$

for $r_0 \le r \le R_0$. From this and (54) we have

$$|\hat{\Lambda}| \int \Phi(r)\phi(r)\hat{p}(r)dr \le K\left(\|\hat{d}\| + \lambda^2 R_0 \left\|\frac{d\phi}{dr}\right\|\right).$$

The left-hand side is larger than

$$|\hat{\Lambda}|\frac{1}{K}\int \phi(r)\hat{p}(r)dr.$$

From these inequalities, $|\hat{\Lambda}|$ has an upper bound depending only on R_0. This completes the proof. □

For the following lemma and subsequent use we shall make use of a quantity that was defined in section 4, namely,

$$\Lambda_M^* := \inf\{\Lambda_M : \quad (22) \text{ and } (23) \text{ has a solution satisfying } (25)\}.$$

LEMMA 5.3. *For each $M_0 > 0$, the set $\{\Lambda_M^*; M \ge M_0\}$ is bounded above.*

Proof. It is enough to show that there is Λ such that (22)–(23) has a solution $g = g_M$ satisfying (25) such that $g(r) > 0$ for all r, for $\Lambda_M = \Lambda$ with $M \ge M_0$, since this will imply $\Lambda_M^* \le \Lambda$ by the definition of Λ_M^*.

We take Λ large enough such that (22) has a solution g satisfying

$$\lim_{r\to 0} g(r) = \frac{1}{c\bar{r}}\left(\Lambda - \frac{\gamma}{2(1-\gamma)}\frac{\mu_1^2}{\sigma^2+\rho^2}\right) > 0.$$

By (22), it is easy to see that $g(r) > 0$ for $0 < r \le M$, since in $0 < r \le M$, g is increasing at the zeros of g. This argument also applies to $M \le r$. That is, $g(r)$ cannot be $-\infty$ for finite r. Therefore, we get a unique solution of (22)–(23) satisfying (25). □

LEMMA 5.4. *Let $g = g_M$ be the solution of* (22) *and* (23) *satisfying* (25) *with $\Lambda = \Lambda_M^*$. Then there is some $M_0 > 0$ such that for $M \ge M_0$,*

$$g(r) < 0, \ r > M_0.$$

Proof. By (37), $g(r)$ will be negative if r is large enough. From (23) and the fact that Λ is bounded below (see Lemma 5.2), it is easy to see that $g(r) < 0$ for $r > M$ if M is large enough, since $g(r)$ for $r > M$ is increasing at zeros of g. This argument also applies to $M_0 \le r \le M$. Therefore, $g(r) < 0$ for $r > M_0$ if M is large enough. This completes the proof. □

Armed with these lemmas, we can now prove Proposition 5.1.

Proof of Proposition 5.1. By Lemmas 5.2 and 5.3, for a fixed $M_0 > 0$, $\{\Lambda_M^*, M \ge M_0\}$ is bounded above and below. We can take a sequence $M_n \to \infty$ as $n \to \infty$ such that $\Lambda_{M_n}^*$ converges to some Λ. Boundedness of $\{\Lambda_{M_n}^*\}$ also implies the uniform boundedness of $\{|g_{M_n}^*(r)|\}$ on compact sets by Lemma 5.1. This further implies the uniform boundedness of $\{|\frac{dg_{M_n}^*}{dr}(r)|\}$ on compact sets, by using (22) and (23). Therefore, we can take a subsequence of $\{M_n\}$ (still denoted by $\{M_n\}$), such that $g_{M_n}^*(r)$ converges to $g(r)$ uniformly on compact sets.

We know Λ, g satisfy (16) and g satisfies (19). In fact, we need only rule out the possibility that g satisfies (33). But since the $g_{M_n}^*(r)$ satisfy (34) for c_1 and r_0 independent of n (see the proof of Proposition 4.2), it follows that (33) cannot hold for g.

It remains to prove that (20) or (21) (depending on the case) holds for g, because then $(\Lambda^*, g^*) = (\Lambda, g)$ satisfies the properties in Theorem 3.1 (see also Theorem 5.1 below). From this it follows that the limit of (Λ_M^*, g_M^*) as $M \to \infty$ is unique, and so Proposition 5.1 will be proved.

We now prove that (20) holds for g when $\mu_2 \neq 1$. By Lemma 5.4, there is M_0 such that

$$g(r) < 0, \ r \geq M_0. \tag{55}$$

We need to know the behaviors of the solutions of (16) as $r \to \infty$. This will be given in Proposition 5.2 below. Now g given above is a solution of (16). Define $\bar{g} = Ag$, $A = 1 + \gamma\sigma^2/(1-\gamma)(\sigma^2+\rho^2)$. According to this theorem, either (56) or (57) holds. From (61), we can conclude the following. If (56) holds, then $g(r) < 0$ for r large. If (57) holds, then $g(r) > 0$ for r large. Since (55) holds, we must have (56). This in turn implies (20) by a simple calculation. The case $\mu_2 = 1$ is treated in a similar manner. This completes the proof. □

PROPOSITION 5.2. *Let (Λ, g) be a solution of* (16) *for $0 < r < \infty$. Then exactly one of the following relations holds:*
either

$$\lim_{r\to\infty} r(\bar{g}(r) - \bar{g}_0(r)) = -\frac{1}{8}\left(1 - \frac{\lambda}{|\lambda|}\left(\frac{-\gamma}{1-\gamma}\right)^{\frac{1}{2}} \frac{\sigma}{\sqrt{\sigma^2+\rho^2}}\right) \tag{56}$$

or

$$\lim_{r\to\infty} \frac{1}{\sqrt{r}}(\bar{g}(r) - \bar{g}_0(r)) = 2\left(-\frac{\gamma}{1-\gamma}\frac{1}{\sigma^2+\rho^2}\right)^{\frac{1}{2}} \frac{|\mu_2 - 1|}{|\lambda|}, \quad \mu_2 \neq 1, \tag{57}$$

$$\lim_{r\to\infty} (\bar{g}(r) - \bar{g}_0(r)) = 2\left(\frac{c^2}{\lambda^4} - 2\frac{\gamma}{\lambda^2}\left(1 + \frac{\gamma}{1-\gamma}\frac{\sigma^2}{\sigma^2+\rho^2}\right)\right)^{\frac{1}{2}}, \quad \mu_2 = 1. \tag{58}$$

Here

$$\bar{g}_0(r) := -\frac{b(r)}{\lambda^2 r} - \left(-\frac{2\left(1 + \frac{\gamma}{1-\gamma}\frac{\sigma^2}{\sigma^2+\rho^2}\right)d(r)}{\lambda^2 r} + \frac{b(r)^2}{\lambda^4 r^2}\right)^{\frac{1}{2}},$$

while $b(\cdot), d(\cdot)$, and $\bar{g}(\cdot)$ are defined by (17), (18), *and* (29), *respectively.*

In order to prove Proposition 5.2 we need three more lemmas. For these we consider a function $g(r)$ that is finite for all r and satisfies (16) and (19). Using this and $\bar{g}_0(r)$ as specified in Proposition 5.2, we then define

$$\hat{g} := \bar{g} - \bar{g}_0.$$

Since $\bar{g}_0$ satisfies

$$\frac{1}{2}\lambda^2 r\bar{g}_0(r)^2 + b(r)\bar{g}_0(r) + \bar{d}(r) = 0,$$

it follows that

$$\frac{1}{2}\lambda^2 r\frac{d}{dr}\hat{g}(r) + \frac{1}{2}\lambda^2 r\hat{g}(r)^2 + \tilde{b}(r)\hat{g}(r) = L(r),$$

where

$$L(r) := \Lambda - \frac{1}{2}\lambda^2 r \frac{d}{dr}\bar{g}_0(r)$$

and

$$\tilde{b}(r) := b(r) + \lambda^2 r \bar{g}_0(r) = -\lambda^2 r \left(-\frac{2\bar{d}(r)}{\lambda^2 r} + \frac{b(r)^2}{\lambda^4 r^2} \right)^{\frac{1}{2}} = -\big(-2\lambda^2 r \bar{d}(r) + b(r)^2 \big)^{\frac{1}{2}}.$$

Notice this equation can be rewritten as

$$\frac{d\hat{g}}{dr} + \hat{g}^2 + \frac{2\tilde{b}(r)}{\lambda^2 r}\hat{g} = \frac{2L(r)}{\lambda^2 r}. \tag{59}$$

In order to investigate the asymptotic properties of (59) we calculate

$$-\frac{2\bar{d}(r)}{\lambda^2 r} + \frac{b(r)^2}{\lambda^4 r^2} = -\frac{2\gamma}{\lambda^2 r}\left(1 + \frac{\gamma}{1-\gamma}\frac{\sigma^2}{\sigma^2+\rho^2}\right)\left(\frac{1}{2}\frac{1}{1-\gamma}\frac{1}{\sigma^2+\rho^2}\bar{\mu}(r)^2 + r\right)$$

$$+ \frac{1}{\lambda^4 r^2}\left(-c(r-\bar{r}) + \frac{\gamma}{1-\gamma}\frac{\sigma\lambda}{\sigma^2+\rho^2}\sqrt{r}\bar{\mu}(r)\right)^2$$

$$= \bar{\mu}(r)^2\left(-\frac{\gamma}{1-\gamma}\frac{1}{\sigma^2+\rho^2}\left(1 + \frac{\gamma}{1-\gamma}\frac{\sigma^2}{\sigma^2+\rho^2}\right)\frac{1}{\lambda^2 r} + \left(\frac{\gamma}{1-\gamma}\right)^2 \frac{\sigma^2}{(\sigma^2+\rho^2)^2}\frac{1}{\lambda^2 r}\right)$$

$$-2\frac{c\sigma\lambda}{\lambda^4}\frac{1}{\sigma^2+\rho^2}\frac{r-\bar{r}}{r}\frac{1}{\sqrt{r}}\bar{\mu}(r) + \frac{c^2}{\lambda^4}\frac{(r-\bar{r})^2}{r^2} - 2\frac{\gamma}{\lambda^2}\left(1 + \frac{\gamma}{1-\gamma}\frac{\sigma^2}{\sigma^2+\rho^2}\right)$$

$$= -\frac{\gamma}{1-\gamma}\frac{1}{\sigma^2+\rho^2}\frac{1}{\lambda^2 r}\bar{\mu}(r)^2 - 2\frac{c\sigma\lambda}{\lambda^4}\frac{1}{\sigma^2+\rho^2}\frac{r-\bar{r}}{r}\frac{1}{\sqrt{r}}\bar{\mu}(r) + \frac{c^2}{\lambda^4}\frac{(r-\bar{r})^2}{r^2}$$
$$- 2\frac{\gamma}{\lambda^2}\left(1 + \frac{\gamma}{1-\gamma}\frac{\sigma^2}{\sigma^2+\rho^2}\right).$$

Since

$$\frac{2\tilde{b}(r)}{\lambda^2 r} = -2\left(-\frac{2\bar{d}(r)}{\lambda^2 r} + \frac{b(r)^2}{\lambda^4 r^2}\right)^{\frac{1}{2}},$$

it follows when $\mu_2 \neq 1$ that

$$\frac{2\tilde{b}(r)}{\lambda^2 r} \cong -2\left(-\frac{\gamma}{1-\gamma}\frac{1}{\sigma^2+\rho^2}\right)^{\frac{1}{2}} \frac{|\mu_2 - 1|}{|\lambda|}\sqrt{r} + O(1) \quad \text{as} \quad r \to \infty. \tag{60}$$

Moreover,

$$\bar{g}_0(r) \cong -\frac{1}{\lambda}\frac{\gamma}{1-\gamma}\frac{\sigma}{\sigma^2+\rho^2}(\mu_2 - 1)\sqrt{r} - \left(\frac{-\gamma}{1-\gamma}\frac{1}{\sigma^2+\rho^2}\frac{1}{\lambda^2}\right)^{\frac{1}{2}}|\mu_2 - 1|\sqrt{r} + O(1) \text{ as } \quad r \to \infty. \tag{61}$$

On the other hand, if $\mu_2 = 1$, then

$$\frac{2\tilde{b}(r)}{\lambda^2 r} \cong -2\left(\frac{c^2}{\lambda^4} - 2\frac{\gamma}{\lambda^2}\left(1 + \frac{\gamma}{1-\gamma}\frac{\sigma^2}{\sigma^2+\rho^2}\right)\right)^{\frac{1}{2}} + O\left(\frac{1}{\sqrt{r}}\right) \quad \text{as} \quad r \to \infty \tag{62}$$

and

$$\bar{g}_0(r) \cong \frac{c}{\lambda^2} - \left(\frac{c^2}{\lambda^4} - 2\frac{\gamma}{\lambda^2}\left(1 + \frac{\gamma}{1-\gamma}\frac{\sigma^2}{\sigma^2+\rho^2}\right)\right)^{\frac{1}{2}} + O\left(\frac{1}{\sqrt{r}}\right) \quad \text{as} \quad r \to \infty.$$

From this we see that if $\mu_2 \neq 1$, then

$$L(r) \cong \frac{|\lambda|}{4}\left(\left(\frac{-\gamma}{1-\gamma}\frac{1}{\sigma^2+\rho^2}\right)^{\frac{1}{2}} + \frac{\lambda}{|\lambda|}\frac{\gamma}{1-\gamma}\frac{\sigma}{\sigma^2+\rho^2}\right)|\mu_2 - 1|\sqrt{r} + O\left(\frac{1}{\sqrt{r}}\right) \quad \text{as} \quad r \to \infty, \tag{63}$$

whereas if $\mu_2 = 1$, then

$$L(r) \cong \Lambda + O\left(\frac{1}{\sqrt{r}}\right) \quad \text{as} \quad r \to \infty.$$

We are now ready for the first of the three lemmas that will be used in the proof of Proposition 5.2.

LEMMA 5.5. *There exist positive numbers c_1 and r_1 such that*

$$\hat{g}(r) > -\frac{c_1}{r} \qquad \forall r \geq r_1. \tag{64}$$

Proof. This proof is by contradiction. Suppose it is false. Then for any $c_1 > 0$ and $r_2 > 0$ there exists some $r_0 > r_2$ such that

$$\hat{g}(r_0) \leq -c_1/r_0.$$

From this we shall prove that

$$\hat{g}(r) \leq -\frac{c_1}{r} \qquad \forall r \geq r_0. \tag{65}$$

But if this is not true, then without loss of generality there is some $r_1 > r_0$ such that $\hat{g}(r_1) = -c_1/r_1$ and

$$\hat{g}(r) < -c_1/r, \quad r_0 < r < r_1.$$

Denoting $f(r) := \hat{g}(r) + c_1/r$, we then see that

$$\frac{df(r_1)}{dr} = -\hat{g}(r_1)^2 + \frac{L(r_1)}{\lambda^2 r_1} - \frac{2\tilde{b}(r_1)}{\lambda^2 r_1}\hat{g}(r_1) - c_1\frac{1}{r_1^2} < 0$$

if we take c_1 large enough. This is a contradiction; (65) must be true if this lemma is false.

By (65) and (59), we have

$$\frac{d\hat{g}}{dr} + \hat{g}^2 < 0.$$

Then

$$\frac{\frac{d\hat{g}}{dr}}{\hat{g}^2} + 1 < 0,$$

which implies

$$\frac{1}{\hat{g}(r_0)} - \frac{1}{\hat{g}(r)} + (r - r_0) < 0$$

for all $r > r_0$. This cannot be true for all $r \geq r_0$, so (65) leads to a contradiction. The proof is complete. □

LEMMA 5.6. *Suppose for some large $r_0 > 0$ that with $r = r_0$ we have*

$$\frac{c_2}{r} < \hat{g}(r) < -\frac{2\tilde{b}(r)}{\lambda^2 r} + \frac{c_2}{r}. \tag{66}$$

If c_2 is large enough, then this inequality also holds for all $r \geq r_0$.

Proof. Suppose $\mu_2 \neq 1$. By (59) and (63), $\hat{g} - c_2/r$ is increasing at $r \geq r_0$ such that $\hat{g} - c_2/r = 0$. Therefore,

$$\frac{c_2}{r} < \hat{g}(r), \quad r \geq r_0.$$

Denote

$$r_1 := \inf\left\{ r > r_0 : \hat{g}(r) \geq -\frac{2\tilde{b}(r)}{\lambda^2 r} + \frac{c_2}{r} \right\}.$$

We have shown $\hat{g}(r) > c_2/r$ for $r_0 < r < r_1$, so it suffices to show we have $r_1 = \infty$.

Assume not. Then $\hat{g}(r_1) = -2\tilde{b}(r_1)/(\lambda^2 r_1) + c_2/r_1$ and

$$\hat{g}(r) < -\frac{2\tilde{b}(r)}{\lambda^2 r} + \frac{c_2}{r}, \quad r_0 \leq r < r_1.$$

We now consider

$$f(r) := \hat{g}(r) + \frac{2\tilde{b}(r)}{\lambda^2 r} - \frac{c_2}{r}$$

and show that $\frac{d}{dr} f(r_1) < 0$, which leads to a contradiction. We have

$$\begin{aligned} \frac{d}{dr} f(r_1) = & \frac{d}{dr}\hat{g}(r_1) + \frac{d}{dr}\left(\frac{2\tilde{b}(r)}{\lambda^2 r}\right)(r_1) + \frac{c_2}{r_1^2} \\ = & \frac{2L(r_1)}{\lambda^2 r_1} - \frac{c_2}{r_1}\left(-\frac{2\tilde{b}(r)}{\lambda^2 r} + \frac{c_2}{r_1}\right) + \frac{d}{dr}\left(\frac{2\tilde{b}(r)}{\lambda^2 r}\right)(r_1) + \frac{c_2}{r_1^2}. \end{aligned}$$

From this, (60), and (63) we can show $\frac{d}{dr} f(r_1) < 0$, thereby completing the proof for the case $\mu_2 \neq 1$. The case $\mu_2 = 1$ is handled in a similar manner. □

LEMMA 5.7. *With r_0 and c_2 as in the preceding lemma such that c_2 is large enough and* (66) *holds, for small $c_1 > 0$ there exists some $r_1 > r_0$ such that*

$$\hat{g}(r) + \frac{2\tilde{b}(r)}{\lambda^2 r} \geq -c_1$$

for all $r \geq r_1$.

Proof. We first show that there is some $r_1 > r_0$ satisfying

$$\hat{g}(r_1) + \frac{2\tilde{b}(r_1)}{\lambda^2 r_1} \geq -c_1. \tag{67}$$

Otherwise,

$$\hat{g}(r) + \frac{2\tilde{b}(r)}{\lambda^2 r} < -c_1 \qquad \forall r > r_0. \tag{68}$$

By (59), we have

$$\frac{d\hat{g}}{dr} \geq c_1 \hat{g} + \frac{2L(r)}{\lambda^2 r} \geq c_1 \hat{g} - \bar{c}\frac{1}{r} \geq \frac{c_1}{2}\hat{g}$$

if c_2 is large enough. Then

$$\hat{g}(r) \geq \exp\left\{\frac{c_1}{2}(r - r_0)\right\} \hat{g}(r_0).$$

But this contradicts (68), so (67) must be true.

Denote

$$f(r) := \hat{g}(r) + \frac{2\tilde{b}(r)}{\lambda^2 r}.$$

Using (59), (60), and (63), we can easily see that

$$\frac{df(r)}{dr} > 0$$

if $f(r) = -c_1$ and c_1 small. From this and (67), $f(r) > -c_1$ for $r > r_1$. This completes the proof.

We are now ready for the proof of Proposition 5.2. Recall that $\hat{g} := \bar{g} - \bar{g}_0$ satisfies (59).

Proof of Proposition 5.2. By Lemmas 5.5–5.7, for positive numbers r_0 and c_1, c_2 (c_1 small, c_2 large) either

$$-\frac{c_2}{r} < -\hat{g}(r) < \frac{c_2}{r}, \quad r \geq r_0 \tag{69}$$

or

$$-c_1 < \hat{g}(r) + \frac{2\tilde{b}(r)}{\lambda^2 r} < \frac{c_2}{r}, \quad r \geq r_0. \tag{70}$$

We first suppose that (69) holds. Denote

$$e(r) := \exp\left(\int_{r_0}^{r} \frac{2\tilde{b}(s)}{\lambda^2 s} ds\right),$$

so that we have for $r \geq r_0$

$$\hat{g}(r) = -\int_r^\infty \frac{L(s)}{\lambda^2 s}\frac{e(s)}{e(r)} ds + \int_r^\infty \hat{g}(s)^2 \frac{e(s)}{e(r)} ds.$$

By (60) and l'Hôpital's rule we then have

$$\lim_{r\to\infty}\frac{r}{e(r)}\int_r^\infty \frac{L(s)}{\lambda^2 s}e(s)ds = \frac{1}{8}\left(1-\frac{\lambda\sigma}{|\lambda|}\left(\frac{-\gamma}{1-\gamma}\right)^{\frac{1}{2}}\left(\frac{1}{\sigma^2+\rho^2}\right)^{\frac{1}{2}}\right)$$

and

$$\lim_{r\to\infty}\frac{r}{e(r)}\int_r^\infty \hat{g}(s)^2 e(s)ds = 0.$$

This implies (56).

On the other hand, suppose (70) holds. Our next step is to show that

$$-c_2\frac{1}{r} < \hat{g}(r) + \frac{2\tilde{b}(r)}{\lambda^2 r} < \frac{c_2}{r}, \quad r \geq r_0. \tag{71}$$

We do this by first showing that there is some number $r_1 > r_0$ satisfying (71) for $r = r_1$. This is true, for if not, then

$$-c_1 \leq \hat{g}(r) + \frac{2\tilde{b}(r)}{\lambda^2 r} \leq -c_2\frac{1}{r}, \quad r \geq r_0,$$

where c_1 is given in (70). Then (59) implies

$$\frac{d\hat{g}}{dr} \geq c_2\frac{1}{r}\hat{g} + \frac{2L(r)}{\lambda^2 r} \geq \frac{c_2}{2}\frac{1}{r}\hat{g}.$$

Integrating this we obtain

$$\hat{g}(r) \geq \hat{g}(r_0)\exp\left(\frac{c_2}{2}\ln\left(\frac{r}{r_0}\right)\right) = \hat{g}(r_0)\left(\frac{r}{r_0}\right)^{c_2/2}.$$

But this contradicts the assertion that $\hat{g}(r) + \frac{2\tilde{b}(r)}{\lambda^2 r} < -\frac{c_2}{r}$ for all $r \geq r_0$, so we know (71) holds for some $r_1 > r_0$.

We now consider $f(r) : \hat{g}(r) + \frac{2\tilde{b}(r)}{\lambda^2 r} + \frac{c_2}{r}$ and use (59) to show that $f(r)$ is increasing at $f(r) = 0$. Therefore, $f(r_1) > 0$ implies $f(r) > 0$ for all $r \geq r_1$. Finally, (57) follows directly from (71), so this proof is completed. □

We now have fully established Proposition 5.1. Thus to complete the proofs of Theorems 3.1 and 3.3 it remains only to establish uniqueness of the solution of the HJB equation. This is accomplished by the following proposition.

PROPOSITION 5.3. *Let g_1 and g_2 be solutions of* (16) *satisfying* (19) *corresponding to Λ_1 and Λ_2, respectively. Let $\hat{g}_1 = \bar{g}_1 - \bar{g}_0$ and $\hat{g}_2 = \bar{g}_2 - \bar{g}_0$ with $\bar{g}_0$ defined as in Proposition* 5.2, *and suppose $\hat{g}_1$ and $\hat{g}_2$ both satisfy limit* (56). *Then $g_1 = g_2$ and $\Lambda_1 = \Lambda_2$.*

Proof. Denote

$$\bar{\Lambda}_1 = \left(1+\frac{\gamma}{1-\gamma}\frac{\sigma^2}{\sigma^2+\rho^2}\right)\Lambda_1, \ \bar{\Lambda}_2 = \left(1+\frac{\gamma}{1-\gamma}\frac{\sigma^2}{\sigma^2+\rho^2}\right)\Lambda_2.$$

We subtract the equation for $\hat{g}_2$ from the equation for $\hat{g}_1$, thereby obtaining

$$\frac{d}{dr}(\hat{g}_2-\hat{g}_1) + \left(\frac{2\tilde{b}(r)}{\lambda^2 r} + \hat{g}_1 + \hat{g}_2\right)(\hat{g}_2-\hat{g}_1) = \frac{\bar{\Lambda}_2 - \bar{\Lambda}_1}{\lambda^2 r}.$$

Denote

$$\tilde{e}(r) := \exp\left(\int_{r_0}^{r}\left(\frac{2\tilde{b}(s)}{\lambda^2 s} + \hat{g}_1(s) + \hat{g}_2(s)\right)ds\right).$$

Then

$$\frac{d}{dr}\Big(\big(\hat{g}_2(r) - \hat{g}_1(r)\big)\tilde{e}(r)\Big) = \frac{\bar{\Lambda}_2 - \bar{\Lambda}_1}{\lambda^2 r}\tilde{e}(r),$$

and so

$$\big(\hat{g}_2(r) - \hat{g}_1(r)\big)\tilde{e}(r) = -\int_r^\infty \frac{\bar{\Lambda}_2 - \bar{\Lambda}_1}{\lambda^2 s}\tilde{e}(s)ds.$$

Without loss of generality, suppose $\Lambda_2 - \Lambda_1 \geq 0$, in which case $\hat{g}_2(r) - \hat{g}_1(r) \leq 0$. But Lemma 4.1 implies $\hat{g}_2(r) - \hat{g}_1(r) \geq 0$. Therefore, $\hat{g}_2(r) = \hat{g}_1(r)$, $\Lambda_2 = \Lambda_1$, and this proof is completed. □

Remark 5.2. The following result, not crucial for the proofs of Theorems 3.1 or 3.3, says that Λ^* is the smallest number such that (16) has a solution defined on $[0, \infty)$. For $\Lambda = \Lambda^*$, (16) has a unique solution. A more general result of this kind is given in the paper by Kaise and Sheu [17].

THEOREM 5.1. *Let* Λ^* *be given in Theorem* 3.1. *Then there is only one solution for* (16) *on* $[0, \infty)$ *with* $\Lambda = \Lambda^*$. *Moreover, if* (16) *has a solution on* $[0, \infty)$ *for some* Λ, *then* $\Lambda \geq \Lambda^*$.

Proof. We consider only the case $\mu_2 \neq 1$ since the argument for the case $\mu_2 = 1$ is similar. Assume $\Lambda = \Lambda^*$ and g is a solution of (16) on $[0, \infty)$. Suppose $g \neq g^*$ with g^* as given in Theorem 3.1. Then (33) holds for g. Since g^* satisfies (25), a simple comparison argument for an ODE shows that $g(r) < g^*(r)$ for all r. But g satisfies either one of (56) or (57). Since g^* satisfies (56), therefore $g < g^*$ implies that g also satisfies (56). Now by Proposition 5.3, we conclude $g = g^*$, a contradiction.

We now consider Λ such that (16) has a solution g_0 defined on $[0, \infty)$. Then (16) must have a solution g defined on $[0, \infty)$ satisfying (25). If g_0 also satisfies (25), then $g = g_0$. See Corollary 4.1. If g_0 satisfies (33), then we have $g(r) > g_0(r)$ for small $r > 0$, and hence for all r. This implies that g is also defined for all $r > 0$. Now if $\Lambda < \Lambda^*$, then $g(r) < g^*(r)$ for all r by Lemma 4.1. By Proposition 5.2, g either satisfies (56) or (57). We know g^* satisfies (56). Together with $g < g^*$, we conclude that g satisfies (56). By Proposition 5.3, we have $g = g^*, \Lambda = \Lambda^*$, a contradiction to our assumption that $\Lambda < \Lambda^*$. This completes the proof. □

Appendix A. Proof of Proposition 4.1. In view of the equivalence between differential equations (22) and (30) it suffices to show that any solution $\bar{g}$ of (30) on $(0, r_0]$ satisfies either

$$\lim_{r\to 0} r\bar{g}(r) = -\frac{2c\bar{r}}{\lambda^2} + 1 \tag{72}$$

or

$$\lim_{r\to 0} \bar{g}(r) = \frac{1}{c\bar{r}}\left(\bar{\Lambda} - \frac{1}{2}A\frac{\gamma}{1-\gamma}\frac{\mu_1^2}{\sigma^2 + \rho^2}\right). \tag{73}$$

To do this we study the solution $\tilde{g}$ of the (equivalent) differential equation (31), from which we have

$$\frac{d\tilde{g}}{dr} \leq \frac{2}{\lambda^2 r}e(r)[\bar{\Lambda} - \bar{d}(r)].$$

So by Lemma A.1 (see the end of this appendix) we have

$$\tilde{g}(r) \leq \int_0^r \frac{2}{\lambda^2 s} e(s)[\bar{\Lambda} - \bar{d}(s)]ds.$$

Substituting for $e(s)$ and so forth, it is apparent that this implies for some number $c_1 > 0$ that

$$\tilde{g}(r) \leq c_1 r^{\frac{2c\bar{r}}{\lambda^2}}, \quad 0 < r < r_0. \tag{74}$$

Take a large c_2 and consider

$$f(r) = \tilde{g}(r) + c_2 r^{\frac{2c\bar{r}}{\lambda^2}-1}.$$

Using (31), it is easily seen that

$$\frac{df(r)}{dr} < 0 \text{ if } f(r) = 0.$$

Therefore, if we take c_2 such that $f(r_0) > 0$, then $f(r) > 0$ for $0 < r \leq r_0$. That is, we have proved

$$\tilde{g}(r) > -c_2 r^{\frac{2c\bar{r}}{\lambda^2}-1}, \quad 0 < r < r_0. \tag{75}$$

By a simple calculation, we have the following two cases:

$$\bar{\Lambda} - \bar{d}(r) > 0 \text{ for } r \text{ in some neighborhood of } 0, \tag{76}$$

or

$$\bar{\Lambda} - \bar{d}(r) < 0 \text{ for } r \text{ in some neighborhood of } 0. \tag{77}$$

First we suppose inequality (76) is true. By (31), $\tilde{g}(r)$ is increasing when r is small and $\tilde{g}(r) = 0$. Using the similar argument for (75), we have the following two possibilities: there is $r_1 > 0$ such that $\tilde{g}(r) < 0,\ r \leq r_1$, or $\tilde{g}(r) > 0,\ r \leq r_1$.

We consider the first possibility. That is, there is r_1 such that $\tilde{g}(r) < 0,\ r \leq r_1$. Conditions (76) and (31) imply

$$\frac{\frac{d\tilde{g}}{dr}}{\tilde{g}^2} \geq -\frac{1}{e(r)},$$

in which case

$$\frac{1}{\tilde{g}(r)} - \frac{1}{\tilde{g}(r_1)} \geq -\int_r^{r_1} \frac{1}{e(s)} ds,$$

that is,

$$-\frac{1}{\tilde{g}(r)} \leq -\frac{1}{\tilde{g}(r_1)} + \int_r^{r_1} \frac{1}{e(s)} ds.$$

It follows that for some constants $c_1, \bar{c}_1$, we have

$$-\tilde{g}(r) > \bar{c}_1 r^{\frac{2c\bar{r}}{\lambda^2}-1}, \tag{78}$$

since

$$\int_r^{r_1} \frac{1}{e(s)} ds \le c_1 r^{-\frac{2c\bar{r}}{\lambda^2}+1}.$$

By (31) again we have

$$\frac{1}{\tilde{g}(r)} - \frac{1}{\tilde{g}(r_1)} = -\int_r^{r_1} \frac{1}{e(s)} ds + \int_r^{r_1} \frac{2}{\lambda^2 s} e(s)[\bar{\Lambda} - \bar{d}(s)] \frac{1}{\tilde{g}^2(s)} ds. \tag{79}$$

We use this to study the limiting behavior of $\tilde{g}(r)$. For the first term on the right-hand side we have by l'Hôpital's rule

$$\lim_{r\to 0} r^{\frac{2c\bar{r}}{\lambda^2}-1} \int_r^{r_0} \frac{1}{e(s)} ds = \frac{1}{\frac{2c\bar{r}}{\lambda^2} - 1}.$$

For the second term on the right-hand side of (79) we have

$$\int_r^{r_0} \frac{2}{\lambda^2 s} e(s)[\bar{\Lambda} - \bar{d}(s)] \frac{1}{\tilde{g}^2(s)} ds \le c \int_r^{r_0} s^{-1+\frac{2c\bar{r}}{\lambda^2}-2\frac{2c\bar{r}}{\lambda^2}+2} ds \le c r^{-\frac{2c\bar{r}}{\lambda^2}+2}.$$

Here we use (78). Thus by (79) we have

$$\lim_{r\to 0} \frac{\tilde{g}(r)}{r^{\frac{2c\bar{r}}{\lambda^2}-1}} = -\frac{2c\bar{r}}{\lambda^2} + 1. \tag{80}$$

Hence for the first possibility (i.e, $\tilde{g}(r_1) < 0$ for some small $r_1 > 0$) and when (76) holds, we have proved (72).

Now we consider the second possibility: there is $r_1 > 0$ such that $\tilde{g}(r) > 0$ for all $r \le r_1$. In view of (32) it follows by l'Hôpital's rule and (74) that

$$\lim_{r\to 0} \frac{\tilde{g}(r)}{r^{\frac{2c\bar{r}}{\lambda^2}}} = \frac{1}{c\bar{r}} \left(\bar{\Lambda} - \frac{1}{2} A \frac{\gamma}{1-\gamma} \frac{\mu_1^2}{\sigma^2 + \rho^2} \right). \tag{81}$$

This, with the relation $\tilde{g}(r) = \bar{g}(r)e(r)$ and the definition of $e(\cdot)$, implies (73).

We summarize what we have shown. Assuming condition (76), we have (80) or (81). They are equivalent to (72) and (73), respectively.

For the remainder of this proof we consider the case that inequality (77) holds. Using the similar argument for (75), we consider

$$\tilde{g}(r) + r^{\frac{2c\bar{r}}{\lambda^2}-\alpha},$$

$1 > \alpha > 1/2$, and deduce there is $r_1 > 0$ such that

$$-\tilde{g}(r) < r^{\frac{2c\bar{r}}{\lambda^2}-\alpha}, \quad r \le r_1, \tag{82}$$

or

$$-\tilde{g}(r) > r^{\frac{2c\bar{r}}{\lambda^2}-\alpha}, \quad r \le r_1. \tag{83}$$

Let $0 < \delta < 1/2$ and $r_1 > 0$ be small, and consider two situations, depending upon whether

$$-\tilde{g}(r) > r^{\frac{2c\bar{r}}{\lambda^2}+\delta-1}, \quad 0 < r \le r_1. \tag{84}$$

For the first situation, assume (84) does not hold. Then by (82), (83) with $\alpha = 1-\delta$, we have

$$-\tilde{g}(r) \le r^{\frac{2c\bar{r}}{\lambda^2}+\delta-1}, \quad r \le r_1.$$

Note that we have already established the property $\tilde{g}(r) < 0$ for r small. From this, together with the property of $e(r)$ defined in section 4 and (31), we have

$$\tilde{g}(r) = -\int_0^r \frac{\tilde{g}^2(s)}{e(s)}ds + \int_0^r \frac{2e(s)}{\lambda^2 s}[\bar{\Lambda} - \bar{d}(s)]ds \ge -c_1 r^{\frac{2c\bar{r}}{\lambda^2}-1+2\delta} - c_2 r^{\frac{2c\bar{r}}{\lambda^2}} \ge -cr^{\frac{2c\bar{r}}{\lambda^2}-1+2\delta},$$

since $\delta < 1/2$. That is,

$$-\tilde{g}(r) \le cr^{\frac{2c\bar{r}}{\lambda^2}-1+2\delta}.$$

Continuing in an iterative fashion one obtains

$$-\tilde{g}(r) \le c_1 r^{\frac{2c\bar{r}}{\lambda^2}-1+2^m\delta}$$

if m is such that $2^m\delta < 1$, where c_1 may depend on m and δ. In particular, this holds for $m = m_0$, $2^{m_0}\delta < 1 \le 2^{m_0+1}\delta$. Apply this same procedure once more to obtain

$$\tilde{g}(r) \ge -c_1 r^{\frac{2c\bar{r}}{\lambda^2}-1+2\bar{\delta}} - c_2 r^{\frac{2c\bar{r}}{\lambda^2}} \ge -cr^{\frac{2c\bar{r}}{\lambda^2}}, \tag{85}$$

where $\bar{\delta} = 2^{m_0}\delta$. The last step is due to $2\bar{\delta} > 1$. We shall now show (81) by the following calculation. In view of (74) and (75) we have

$$\lim_{r\to 0} r^{-\frac{2c\bar{r}}{\lambda^2}} \int_0^r \frac{2e(s)}{\lambda^2 s}[\bar{\Lambda} - \bar{d}(s)]ds = \frac{1}{c\bar{r}}\left(\bar{\Lambda} - \frac{1}{2}A\frac{\gamma}{1-\gamma}\frac{\mu_1^2}{\sigma^2+\rho^2}\right). \tag{86}$$

In addition, by (74) and (85) we have

$$\lim_{r\to 0} r^{-\frac{2c\bar{r}}{\lambda^2}} \int_0^r \frac{1}{e(s)}\tilde{g}(s)^2 ds = 0.$$

This with (86) and (32) implies (81), which is equivalent to (73).

Now we consider the opposite situation, namely, there is $0 < \delta < 1/2$ such that (84) does hold for some r_1. Then

$$\frac{1}{\tilde{g}(r)} - \frac{1}{\tilde{g}(r_1)} = -\int_r^{r_1} \frac{1}{e(s)}ds + \int_r^{r_1} \frac{2}{\lambda^2 s}e(s)[\bar{\Lambda} - \bar{d}(s)]\frac{1}{\tilde{g}^2(s)}ds.$$

Corresponding to the two terms on the right-hand side we have

$$\lim_{r\to 0} r^{\frac{2c\bar{r}}{\lambda^2}-1} \int_r^{r_1} \frac{1}{e(s)}ds = \frac{1}{\frac{2c\bar{r}}{\lambda^2}-1}$$

and

$$\lim_{r\to 0} r^{\frac{2c\bar{r}}{\lambda^2}-1} \int_r^{r_1} \frac{2}{\lambda^2 s}e(s)[\bar{\Lambda} - \bar{d}(s)]\frac{1}{\tilde{g}^2(s)}ds = 0.$$

Hence

$$\lim_{r\to 0} \frac{\tilde{g}(r)}{r^{\frac{2c\bar{r}}{\lambda^2}-1}} = -\frac{2c\bar{r}}{\lambda^s} + 1; \tag{87}$$

thus by the definition of $e(\cdot)$ and the relationship between $\bar{g}$ and $\tilde{g}$ we see that (72) holds. This completes the proof of Proposition 4.1. □

The lemma that was used in the preceding proof follows.

LEMMA A.1. *If the differential equation* (31) *has a solution* $\tilde{g}$ *on* $(0, r_0]$ *for some* $r_0 > 0$, *then* $\tilde{g}(r) \to 0$ *as* $r \to 0$.

Proof. We first prove that for any $c_1 > 0$ there are $r_n, n = 1, 2, \ldots$, which tend to 0 as n tends to infinity and which satisfy $\tilde{g}(r_n) > -c_1$. If not, there is $r_1 > 0$ such that $\tilde{g}(r) \leq -c_1$ for all $0 < r \leq r_1$. Since

$$\frac{\frac{d\tilde{g}}{dr}}{\tilde{g}^2} + \frac{1}{e(r)} = \frac{2}{\lambda^2 r} e(r)[\bar{\Lambda} - \bar{d}(r)] \frac{1}{\tilde{g}^2(r)},$$

we have

$$\frac{1}{\tilde{g}(r)} - \frac{1}{\tilde{g}(r_1)} = -\int_r^{r_1} \frac{1}{e(s)} ds + \int_r^{r_1} \frac{2}{\lambda^2 s} e(s)[\bar{\Lambda} - \bar{d}(s)] \frac{1}{\tilde{g}^2(s)} ds. \tag{88}$$

For small $r > 0$, the first term on the right-hand side is bounded above by $-c_2 r^{-\frac{2c\bar{r}}{\lambda^2}+1}$ for some $c_2 > 0$, and the second term is bounded above by c_3/c_1^2 for some $c_3 > 0$. From this, the right-hand side converges to $-\infty$ as $r \to 0$, in which case $\tilde{g}(r) \to 0$ as $r \to 0$, a contradiction.

By this result we can take a sufficiently small $r_1 > 0$ such that $\tilde{g}(r_1) > -c_1$. Next we show that

$$\tilde{g}(r) > -c_1, \quad r \leq r_1. \tag{89}$$

To see this, suppose this is not true. Then there is some $r_2 < r_1$ such that $\tilde{g}(r_2) = -c_1$ and

$$\tilde{g}(r) > -c_1, \quad r_2 < r < r_1.$$

By (31)

$$\frac{d\tilde{g}}{dr}(r_2) = -\frac{1}{e(r_2)} c_1^2 + \frac{2}{\lambda^2 r_2} e(r_2)[\bar{\Lambda} - \bar{d}(r_2)] < 0.$$

This contradicts the specified property of c_1.

Finally, it suffices to show that for any $c_1 > 0$ there is some $r_3 > 0$ such that

$$\tilde{g}(r) \leq c_1, \quad r \leq r_3, \tag{90}$$

because it is easy to see that our lemma would then follow from (89) and (90). To prove (90), suppose there is $r_4 > 0$ such that $\tilde{g}(r_4) > c_1$. Then

$$\frac{d\tilde{g}}{dr}(r_4) \leq -\frac{1}{e(r_4)} c_1^2 + \frac{2}{\lambda^2 r_4} e(r_4)[\bar{\Lambda} - \bar{d}(r_4)] < 0.$$

Therefore, $\tilde{g}(r)$ is decreasing at r_4. This argument also shows that $\tilde{g}$ is decreasing on the set $\{\tilde{g}(r) > c_1\}$. Then we must have $\tilde{g} > c_1$ on $(0, r_4]$. This leads to a contradiction since using (88) with $r_1 = r_4$ and small r we have the right-hand side tending to $-\infty$ while the left-hand side is bounded. This completes the proof of Lemma A.1. □

REFERENCES

[1] A. Bagchi and K. S. Kumar, *Dynamic asset management: Risk sensitive criterion with non-negative factors constraints*, in Recent Developments in Mathematical Finance, J. Yong, ed., World Scientific, River Edge, NJ, 2002, pp. 1–11.

[2] T. R. Bielecki, A. Harris, Z. Li, and S. R. Pliska, *Risk sensitive asset management: Two empirical examples*, in Proceedings of the Conference on Mathematical Finance (Konstanz, Germany, 2000), M. Kohlmann, ed., Birkhaüser, Basel, Switzerland, 2001, pp. 99–110.

[3] T. R. Bielecki, D. Hernandez-Hernandez, and S. R. Pliska, *Risk sensitive control of finite state Markov chains in discrete time, with applications to portfolio management*, Math. Methods Oper. Res., 50 (1999), pp. 167–188.

[4] T. R. Bielecki, D. Hernandez-Hernandez, and S. R. Pliska, *Risk sensitive asset management with constrained trading strategies*, in Recent Developments in Mathematical Finance, J. Yong, ed., World Scientific, River Edge, NJ, 2002, pp. 127–138.

[5] T. R. Bielecki and S. R. Pliska, *Risk sensitive dynamic asset management*, Appl. Math. Optim., 39 (1999), pp. 337–360.

[6] T. R. Bielecki and S. R. Pliska, *Risk-sensitive dynamic asset management in the presence of transaction costs*, Finance Stoch., 4 (2000), pp. 1–33.

[7] T. R. Bielecki and S. R. Pliska, *Risk sensitive control with applications to fixed income portfolio management*, in Proceedings of the European Congress of Mathematics (Barcelona, 2000), C. Casacuberta et al., eds., Birkhaüser, Basel, Switzerland, 2001, pp. 331–345.

[8] T. R. Bielecki and S. R. Pliska, *Risk sensitive ICAPM with application to fixed-income management*, IEEE Trans. Automat. Control, 49 (2004), pp. 420–432.

[9] T. R. Bielecki and S. R. Pliska, *Economic properties of the risk sensitive criterion for portfolio management*, Rev. of Accounting and Finance, 2 (2003), pp. 3–17.

[10] T. R. Bielecki, S. R. Pliska, and M. Sherris, *Risk sensitive asset allocation*, J. Econom. Dynam. Control, 24 (2000), pp. 1145–1177.

[11] T. R. Bielecki, S. R. Pliska, and J. Yong, *Optimal investment decisions for a portfolio with a rolling horizon bond and a discount bond*, Int. J. Theor. Appl. Finance, to appear.

[12] J. C. Cox, J. E. Ingersoll, and S. A. Ross, *A theory of term structure of interest rates*, Econometrica, 53 (1985), pp. 385–407.

[13] W. Feller, *Two singular diffusion problems*, Ann. Math. (2), 54 (1951), pp. 173–182.

[14] W. H. Fleming and S.-J. Sheu, *Optimal long term growth rate of expected utility of wealth*, Ann. Appl. Probab., 9 (1999), pp. 871–903.

[15] W. H. Fleming and S.-J. Sheu, *Risk sensitive control and an optimal investment model*, Math. Finance, 10 (2000), pp. 197–213.

[16] W. H. Fleming and S.-J. Sheu, *Risk sensitive control and an optimal investment.* II, Ann. Appl. Probab., 12 (2002), pp. 730–767.

[17] H. Kaise and S.-J. Sheu, *On the structure of solutions of ergodic type Bellman equation related to risk-sensitive control*, Ann. Probab., to appear.

[18] I. Karatzas and S. E. Shreve, *Methods of Mathematical Finance*, Appl. Math. 39, Springer-Verlag, New York, 1998.

[19] R. Z. Khasminskii, *Stochastic Stability of Differential Equations*, Sijthoff and Noordhoff, Rockville, MD, 1980.

[20] R. Korn, *Optimal Portfolios. Stochastic Models for Optimal Investment and Risk Management in Continuous Time*, World Scientific, Singapore, 1997.

[21] K. Kuroda and H. Nagai, *Risk-sensitive portfolio optimization on infinite time horizon*, Stochastics Stochastics Rep., 73 (2002), pp. 309–331.

[22] R. C. Merton, *Optimum consumption and portfolio rules in a continuous time model*, J. Econom. Theory, 3 (1971), pp. 373–413.

[23] R. C. Merton, *An intertemporal capital asset pricing model*, Econometrica, 41 (1973), pp. 866–887.

[24] R. C. Merton, *Continuous-Time Finance*, Basil Blackwell, Cambridge, MA, 1990.

[25] H. Nagai, *Optimal strategies for risk-sensitive portfolio optimization problems for general factor models*, SIAM J. Control Optim., 41 (2003), pp. 1779–1880.

[26] H. Nagai and S. Peng, *Risk-sensitive dynamic portfolio optimization with partial information on infinite time horizon*, Ann. Appl. Probab., 12 (2001), pp. 1–23.

[27] P. Whittle, *Risk Sensitive Optimal Control*, John Wiley, New York, 1990.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 5, pp. 1844–1863

A PRIORI ERROR ESTIMATES FOR THE FINITE ELEMENT DISCRETIZATION OF ELLIPTIC PARAMETER IDENTIFICATION PROBLEMS WITH POINTWISE MEASUREMENTS*

R. RANNACHER† AND B. VEXLER‡

Abstract. We develop an a priori error analysis for the finite element Galerkin discretization of parameter identification problems. The state equation is given by an elliptic partial differential equation of second order with a finite number of unknown parameters, which are estimated using pointwise measurements of the state variable.

Key words. parameter identification, finite elements, pointwise measurements, L^∞-error estimates

AMS subject classifications. 65K10, 65N30, 65N21, 49M25, 49K20

DOI. 10.1137/040611100

1. Introduction.

We consider parameter identification problems governed by an elliptic partial differential equation of second order. The finitely many unknown parameters are estimated using the measurements of point values of the *state variable*. Let $\Omega \subset \mathbb{R}^2$ be a convex polygonal domain; $L^2(\Omega)$ the corresponding Lebesgue space with inner product and norm denoted by $(\cdot,\cdot)$ and $\|\cdot\|_2$, respectively; and $H^m(\Omega)$ the Sobolev space of order $m \in \mathbb{N}$. With this notation, we set

$$V := \{v \in H^1(\Omega) \cap C(\overline{\Omega}) \,|\, v = 0 \text{ on } \partial\Omega\}.$$

The state variable $u \in V$ is determined by an elliptic partial differential equation (the *state equation*)

$$\begin{aligned} -\nabla \cdot (A(q)\nabla u) &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{1.1}$$

for given Hölder continuous $f \in C^\alpha(\overline{\Omega})$, $\alpha \in (0,1)$. Here, $Q \subset \mathbb{R}^{n_p}$ denotes the open admissible set of parameters $q \in \mathbb{R}^{n_p}$, for which $A(q) = (A_{ij}(q))$ is a symmetric and positive definite 2×2 matrix with twice continuously differentiable entries $A_{ij} : Q \to C^{1+\alpha}(\overline{\Omega})$. The above conditions guarantee that, for any admissible value of the parameter q, the corresponding solution u of the state equation (1.1) is in $H^2(\Omega)$ (see, e.g., Grisvard [12]). At the corner points of $\partial\Omega$, the second derivatives of the solution may become singular. However, u has Hölder continuous second derivatives, $u \in C^{2+\alpha}(\overline{\Omega}_d)$, for each subdomain $\Omega_d \subset \Omega$ with distance $d > 0$ to the corner points.

The usual weak formulation of (1.1) is

$$a(q)(u,\phi) = (f,\phi) \quad \forall \phi \in V, \tag{1.2}$$

*Received by the editors July 6, 2004; accepted for publication (in revised form) March 1, 2005; published electronically December 6, 2005. This work has been supported by the German Research Foundation (DFG) through the *Sonderforschungsbereich* 359 "Reactive Flows, Diffusion and Transport," and the *Graduiertenkolleg* "Complex Processes: Modeling, Simulation and Optimization" at the Interdisciplinary Center of Scientific Computing (IWR), University of Heidelberg.
http://www.siam.org/journals/sicon/44-5/61110.html

†Institute of Applied Mathematics, University of Heidelberg, Im Neuenheimer Feld 294, D-69120 Heidelberg, Germany (rannacher@iwr.uni-heidelberg.de).

‡Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenberger Strasse 69, A-4040 Linz, Austria (boris.vexler@oeaw.ac.at).

where the bilinear form $a(q)(\cdot,\cdot)$ is defined by

$$a(q)(u,\phi) := (A(q)\nabla u, \nabla\phi). \tag{1.3}$$

Further, the observation operator $C(\cdot)$ describing the mapping of the state variable u to the space of measurements $Z = \mathbb{R}^{n_m}$ is given by

$$C_i(v) = v(\xi_i), \quad i = 1, 2, \dots, n_m, \tag{1.4}$$

where $\{\xi_i\} \subset \Omega$ is a finite set of measurement points. We assume that $n_m \geq n_p$. The Euclidean product and norm on Q and Z are denoted by $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$, respectively, and the same notation is also used for the corresponding natural norms of matrices.

The values of the parameters are estimated from a given set of measurements $\hat{C} \in Z$ using a least squares approach such that we obtain a constrained optimization problem with the cost functional $J : V \to \mathbb{R}$:

$$\text{Minimize} \quad J(u) := \frac{1}{2}\|C(u) - \hat{C}\|^2 \tag{1.5}$$

under the constraint (1.2). Throughout, we assume the existence of a solution $(u, q) \in V \times Q$ of the problem (1.2), (1.5). For an analysis of the existence of solutions for parameter identification problems, see, e.g., Banks and Kunisch [2], Kravaris and Seinfeld [16], and Litvinov [17].

The state equation is discretized by a conforming finite element Galerkin method defined on a family $\{\mathcal{T}_h\}_{h>0}$ of shape regular quasi-uniform meshes $\mathcal{T}_h = \{K\}$ consisting of closed *cells* K which are either triangles or quadrilaterals. The straight parts which make up the boundary ∂K of a cell K are called *faces*. The mesh parameter h is defined as a cellwise constant function by setting $h_{|K} = h_K$, and h_K is the diameter of K. Usually we use the symbol h also for the maximal cell size, i.e.,

$$h = \max_{K\in\mathcal{T}_h} h_K. \tag{1.6}$$

For convenience, we assume that $0 < h < 1$. On the mesh $\mathcal{T}_h$ we define finite element spaces $V_h \subset V$ consisting of linear or bilinear shape functions; see, e.g., Brenner and Scott [5] or Johnson [14]. The corresponding discrete state $u_h \in V_h$ and parameter $q_h \in Q$ are determined by

$$\text{Minimize} \quad J(u_h), \tag{1.7}$$

under the constraint

$$a(q_h)(u_h, \phi_h) = (f, \phi_h) \quad \forall\phi_h \in V_h. \tag{1.8}$$

Since Q is finite dimensional, the parameter q_h is determined in the same space Q.

The main purpose of this paper is to analyze the behavior of the error in parameters $\|q - q_h\|$ for h tending to zero. There are a number of publications in which a priori error estimates are derived for optimal control problems governed by partial differential equations; see, e.g., Falk [8], Arada, Casas, and Tröltzsch [1], Deckelnic and Hinze [6], and Gunzburger and Hou [13]. However, there are only few published results on this topic in the context of parameter identification problems; see Falk [9], Neittaanmäki and Tai [18], and Kärkkäinen [15].

In Vexler [26], an a priori error analysis for the case of V-stable observation operators $C_i(\cdot)$ is developed and optimal-order convergence is shown,

$$\|q - q_h\| = \mathcal{O}(h^2), \tag{1.9}$$

essentially under the assumption that $|C_i(u) - C_i(u_h)| \leq ch^2\|u\|_{H^2}$. However, the generalization of this result to pointwise observations is not straightforward. For this case, we prove in this paper that, under certain regularity conditions,

$$\|q - q_h\| = \mathcal{O}(h^2|\log(h)|^2). \tag{1.10}$$

The proof uses the technique for estimating discrete Green functions developed in Frehse and Rannacher [10]. A complementary result of a posteriori error analysis for parameter identification problems is given in Becker and Vexler [4].

To the authors' knowledge, this is the first a priori error analysis for parameter identification problems with pointwise observations. The consideration of pointwise observations in determining discrete parameters seems very natural in view of practical measurement techniques; see [3] for applications in reactive flow analysis.

The paper is organized as follows. In the next section, we describe an algorithm for solving problem (1.7), (1.8). In section 3, we present a paradigm for a priori error analysis for discretization of a class of optimization problems. Thereafter, in section 4, we derive the announced error estimate using an L^∞-stability result, which is proven in section 5. In section 6, we present a numerical example confirming the asymptotic sharpness of our error estimate. Possible extensions are addressed in the last section.

2. Optimization algorithm. In this section, we reformulate the problem under consideration as an unconstrained optimization problem and describe a solution algorithm for it. Since the coefficient matrix $A(q)$ is assumed to be positive definite for parameters $q \in Q \subset \mathbb{R}^{n_p}$, the relation

$$a(q)(S(q), \phi) = (f, \phi) \quad \forall \phi \in H_0^1(\Omega) \tag{2.1}$$

defines an operator $S : Q \to H_0^1(\Omega)$. By the assumptions on the data of the problem and the Sobolev embedding theorem,

$$S : Q \to H_0^1(\Omega) \cap H^2(\Omega) \subset V. \tag{2.2}$$

The solution operator S can be shown to possess first and second derivatives which are continuous with respect to the norm of V; see Theorem 2.1 below. We recall that the existence of a solution $q \in Q$ of problem (1.5) is assumed. Let $Q_0 \subset Q$ be an open bounded set containing the optimal parameter q on which the coefficient matrix $A(q)$ is uniformly positive definite; i.e., there exists $\gamma \in \mathbb{R}_+$ such that

$$p^* A(q)p \geq \gamma\|p\|^2 \quad \forall p \in Q, \quad \forall q \in Q_0, \tag{2.3}$$

uniformly with respect to $x \in \overline{\Omega}$. We introduce the *reduced observation operator* $c : Q_0 \to Z$ by

$$c(q) := C(S(q)). \tag{2.4}$$

This allows us to reformulate the problem under consideration as an unconstrained optimization problem with the reduced cost functional $j : Q_0 \to \mathbb{R}$:

$$\text{Minimize} \quad j(q) := \frac{1}{2}\|c(q) - \hat{C}\|^2, \quad q \in Q_0. \tag{2.5}$$

Denoting by $G = c'(q) \in \mathbb{R}^{n_p \times n_m}$ the Jacobian matrix of the reduced observation operator $c(\cdot)$, the first-order necessary optimality condition $j'(q) = 0$ for (2.5) reads

$$G^*(c(q) - \hat{C}) = 0, \tag{2.6}$$

where G^* denotes the transpose of G. The positive semidefiniteness of the Hessian matrix $H := \nabla^2 j(q)$ is the second-order necessary optimality condition. A solution q of problem (2.5) is called *stable* if the sufficient optimality condition holds, i.e., if the Hessian H is positive definite. Throughout, we will assume the solution q to be stable. The stability of the solution is given, for instance, if the value of the cost functional $\|C(u) - \hat{C}\|$ is small enough and the matrix G has full rank n_p; see, e.g., [26] for details.

Since by assumption the matrix coefficient $A(\cdot)$ is twice continuously differentiable, there holds

$$\sup_{\xi\in Q_0} |||A(\xi)|||_{1,\infty} + \sup_{\xi\in Q_0} |||A'_{q_j}(\xi)|||_{1,\infty} + \sup_{\xi\in Q_0} |||A''_{q_j q_k}(\xi)|||_{1,\infty} < \infty, \tag{2.7}$$

where $|||B|||_{1,\infty} := \max_{i,j=1,2} \|B_{ij}\|_{1,\infty}$ for a matrix function $B = (B_{ij}) \in C^1(\overline{\Omega})^{2\times 2}$.

In the following propositions, we give representations of the Jacobian G of $c(\cdot)$, the Hessian H of $j(\cdot)$, and the Hessian of $c_i(\cdot)$.

THEOREM 2.1. *Let the reduced observation operator $c(\cdot)$ and the reduced functional $j(\cdot)$ be defined as in* (2.4) *and* (2.5)*, respectively.*

(i) *The elements of the Jacobian of $c(\cdot)$ at some $q \in Q_0$ are given by*

$$G_{ij} = \frac{\partial c_i}{\partial q_j}(q) = C_i(w_j), \quad i = 1, \dots, n_m,\ j = 1, \dots, n_p, \tag{2.8}$$

where $w_j \in V$ are the solutions of the problems

$$a(q)(w_j, \phi) = -(A'_{q_j}(q)\nabla u, \nabla\phi) \quad \forall \phi \in V, \tag{2.9}$$

with $u = S(q)$. The functions $w_j \in V$ depend continuously on $q \in Q$.

(ii) *The Hessian of $j(\cdot)$ can be expressed by*

$$H = G^*\, G + M, \tag{2.10}$$

where the matrix $M \in \mathbb{R}^{n_p \times n_p}$ is given by

$$M = \sum_{i=1}^{n_m} c_i''(q)(c_i(q) - \hat{C}_i). \tag{2.11}$$

The Hessian of $c_i(q)$ is given by

$$\frac{\partial^2}{\partial q_j\, \partial q_k} c_i(q) = C_i(v_{jk}), \tag{2.12}$$

where the $v_{jk} \in V$ are the solutions of the problems

$$\begin{aligned} a(q)(v_{jk}, \phi) = &-(A'_{q_j}(q)\nabla w_k, \nabla\phi) - (A'_{q_k}(q)\nabla w_j, \nabla\phi) \\ &- (A''_{q_j q_k}(q)\nabla u, \nabla\phi) \quad \forall \phi \in V, \end{aligned} \tag{2.13}$$

with w_j as defined in (2.9)*. The functions $v_{jk} \in V$ depend continuously on $q \in Q$.*

Proof. The derivation of the derivatives of $c(\cdot)$ uses the chain rule,

$$\frac{\partial c_i}{\partial q_j}(q) = \frac{\partial}{\partial q_j} C_i(S(q)) = S'_{q_j}(q)(\xi_i) =: w_j(\xi_i),$$

for $i = 1, \dots, n_m$, $j = 1, \dots, n_p$, where the functions w_j are determined by the relations (2.9). This is seen by considering the limit of difference quotients

$$\begin{aligned} 0 &= \lim_{t \to 0} \frac{1}{t} (a(q + tq_j)(S(q + tq_j), \phi) - (f, \phi) - a(q)(S(q), \phi) + (f, \phi)) \\ &= a(q)(S'_{q_j}(q), \phi) + a'_{q_j}(q)(S(q), \phi) = a(q)(w_j, \phi) + a'_{q_j}(q)(u, \phi). \end{aligned}$$

Analogously, we obtain

$$\frac{\partial^2 c_i}{\partial q_j \partial q_k}(q) = \frac{\partial^2}{\partial q_j \partial q_k} C_i(S(q)) = S''_{q_j q_k}(q)(\xi_i) =: v_{jk}(\xi_i),$$

where the functions v_{jk} are determined by the relations (2.13). To see this, we consider the limit of the difference quotient

$$\begin{aligned} 0 = \lim_{t \to 0} \frac{1}{t} &(a(q + tq_k)(S'_{q_j}(q + tq_k), \phi) + a'_{q_j}(q + tq_k)(S(q + tq_k), \phi) \\ &- a(q)(S'_{q_j}(q), \phi) - a'_{q_j}(q)(S(q), \phi)) \\ = a(q)(S''_{q_j q_k}(q), \phi) &+ a'_{q_k}(q)(S'_{q_j}(q), \phi) + a'_{q_j}(q)(S'_{q_k}(q), \phi) + a''_{q_k q_j}(q)(S(q), \phi). \end{aligned}$$

In Lemma 2.2 below, we will show that all the functions u, w_j, v_{jk} are in $V \cap H^2(\Omega)$. This is due to the fact that they are determined by second-order elliptic boundary value problems on a convex domain, with smooth coefficients and right-hand sides in $L^2(\Omega)$ which depend continuously on the parameter $q \in Q$. This implies that also their solutions depend continuously on $q \in Q$ with respect to the norm of the solution space $H_0^1(\Omega) \cap H^2(\Omega) \subset V$. Hence, the solution operator $S : Q \to V$ is twice continuously differentiable. This completes the proof. ∎

In practice the Hessian H of $j(\cdot)$ is computed using the representation

$$M_{jk} = -(A'_{q_j}(q)\nabla w_k, \nabla z) - (A'_{q_k}(q)\nabla w_j, \nabla z) - (A''_{q_j q_k}(q)\nabla u, \nabla z), \tag{2.14}$$

with the function z determined by the dual equation

$$(A(q)\nabla\phi, \nabla z) = \langle C(u) - \bar{C}, C(\phi)\rangle. \tag{2.15}$$

For later purposes, we provide some a priori bounds for the solutions of the boundary value problems introduced in Theorem 2.1, which follow by standard results of elliptic regularity theory.

LEMMA 2.2. *For the solutions of the elliptic boundary value problems* (2.9) *and* (2.13) *there hold the global* L^2 *a priori estimates*

$$\|u\|_{2,2} + \|w_j\|_{2,2} + \|v_{jk}\|_{2,2} \le c, \tag{2.16}$$

where c *is a generic constant depending only on the data of the problem. Further, for each subdomain* $\Omega_d \subset \Omega$ *with distance* $d > 0$ *to the corner points, there hold the* L^∞ *a priori estimates*

$$\|u\|_{C^{2+\alpha}(\overline{\Omega}_d)} + \|w_j\|_{C^{2+\alpha}(\overline{\Omega}_d)} + \|v_{jk}\|_{C^{2+\alpha}(\overline{\Omega}_d)} \le c_d \tag{2.17}$$

with a generic constant $c_d \approx d^{-1}$.

Proof. The variational equations defining u as well as w_j and v_{jk} can be rewritten in such a form that they represent second-order elliptic boundary value problems with

smooth coefficients and right-hand sides which are bounded functionals on $L^2(\Omega)$ and $C^\alpha_{\mathrm{loc}}(\Omega)$, respectively, as follows:

$$a(q)(u,\phi) = (f,\phi) \quad \forall \phi \in V, \tag{2.18}$$

$$a(q)(w_j,\phi) = (\nabla \cdot A'_{q_j}(q)\nabla u,\phi) \quad \forall \phi \in V, \tag{2.19}$$

$$\begin{aligned} a(q)(v_{jk},\phi) = {} & (\nabla \cdot A'_{q_j}(q)\nabla w_k,\phi) + (\nabla \cdot A'_{q_k}(q)\nabla w_j,\phi) \\ & + (\nabla \cdot A''_{q_j q_k}(q)\nabla u,\phi) \quad \forall \phi \in V. \end{aligned} \tag{2.20}$$

By assumption the coefficient functions $A'_{q_j}(q)$ and $A''_{q_j q_k}(q)$ are smooth. In view of the convexity of the polygonal domain Ω, the H^2-regularity estimates then follow by results from Grisvard [12]. A reference for the corresponding $C^{2+\alpha}$-estimates is Gilbarg and Trudinger [11]. In the first step, from (2.18), we get

$$\begin{aligned} \|u\|_{2,2} &\le c\|f\|_2 \le c, \\ \|u\|_{C^{2+\alpha}(\overline{\Omega}_d)} &\le c_d\big\{\|f\|_{C^\alpha(\overline{\Omega}_{d/2})} + \|u\|_{2,2}\big\} \le c_d. \end{aligned}$$

Then, using this in (2.19), we conclude that

$$\begin{aligned} \|w_j\|_{2,2} &\le c\|u\|_{2,2} \le c, \\ \|w_j\|_{C^{2+\alpha}(\overline{\Omega}_d)} &\le c_d\big\{\|u\|_{C^{2+\alpha}(\overline{\Omega}_{d/2})} + \|u\|_{2,2}\big\} \le c_d. \end{aligned}$$

Finally, this is used in (2.20) and allows us to conclude that

$$\begin{aligned} \|v_{jk}\|_{2,2} &\le c\max\big\{\|w_j\|_{2,2}, \|w_k\|_{2,2}\big\} + c\|u\|_{2,2} \le c, \\ \|v_{jk}\|_{C^{2+\alpha}(\overline{\Omega}_d)} &\le c_d\big\{\|w_j\|_{C^{2+\alpha}(\overline{\Omega}_{d/2})} + \|u\|_{C^{2+\alpha}(\overline{\Omega}_{d/2})} + \|w_j\|_{2,2}\big\} \le c_d. \end{aligned}$$

This completes the proof. □

Similar to the continuous case, we introduce a discrete solution operator S_h: $Q_0 \to V_h$ by the equation

$$a(q_h)(S_h(q_h),\phi_h) = (f,\phi_h) \quad \forall \phi_h \in V_h,\ q_h \in Q_0. \tag{2.21}$$

As before, we turn the discrete problem (1.7), (1.8) into an unconstrained minimization problem,

$$\text{Minimize} \quad j_h(q_h) := \frac{1}{2}\|c_h(q_h) - \hat{C}\|^2, \quad q_h \in Q_0, \tag{2.22}$$

where the discrete reduced observation operator c_h is defined by

$$c_h(q_h) = C(S_h(q)). \tag{2.23}$$

Denoting the corresponding Jacobian by $G_h = c'_h(q_h)$, the necessary optimality condition $j'_h(q_h) = 0$ reads

$$G_h^*(c_h(q_h) - \hat{C}) = 0. \tag{2.24}$$

The derivatives of the discrete observation operator c_h can be computed in a way analogous to that in Theorem 2.1.

Problem (2.22) is solved iteratively starting with an initial guess q_h^0 and using the recursive setting $q_h^{k+1} = q_h^k + \delta q_h$. The update δq_h is obtained as the solution of the system of linear equations

$$H_k\,\delta q_h = G_h^*(\hat{C} - c_h(q_h^k)), \tag{2.25}$$

where $G_h = c_h'(q_h^k)$, and H_k is an appropriate symmetric approximation of the Hessian $\nabla^2 j_h(q_h^k)$. The most widely used choice of the matrix $H_k = G_h^* G_h$ leads us to the Gauß–Newton algorithm; see, e.g., Nocedal and Wright [21].

For one step of the Gauß–Newton algorithm the state equation and n_p tangent problems (2.9) have to be solved, which involves the same linear operator but with different right-hand sides. Due to the small dimension n_p of the parameter space Q, the solution of (2.25) is uncritical. For discussing other Newton-type methods and trust-region techniques for globalization of the convergence in this context, see [26].

3. A paradigm for a priori error analysis. In this section we present a general approach to the error analysis of a class of optimization problems such as are considered in this paper. The main result is stated in the following theorem. It is a variant of well-known perturbation theorems for differentiable mappings, which is particularly tailored to the present situation. However, it seems easier to include the elementary proof than to search for the precise reference.

THEOREM 3.1. *Let $F, F_h : \mathbb{R}^n \to \mathbb{R}^n$, for a discretization parameter $h \in \mathbb{R}_+$, be continuously differentiable operators, and $x \in \mathbb{R}^n$ be a solution of $F(x) = 0$. Let the following conditions be fulfilled:*

(i) *The derivative $F'(x)$ is positive definite; i.e., there is a constant $\gamma > 0$ such that*

$$p^* F'(x) p \geq \gamma \|p\|^2, \quad p \in \mathbb{R}^n. \tag{3.1}$$

(ii) *There is a neighborhood U of x and a positive number $L(h) \in \mathbb{R}_+$ such that*

$$\|F_h'(\xi) - F_h'(\eta)\| \leq L(h)\|\xi - \eta\| \quad \forall \xi, \eta \in U. \tag{3.2}$$

(iii) *With the h-dependent constant $L(h)$, there holds*

$$\lim_{h \to 0} L(h)\|F(x) - F_h(x)\| = 0. \tag{3.3}$$

(iv) *There holds*

$$\lim_{h \to 0} \|F'(x) - F_h'(x)\| = 0. \tag{3.4}$$

Then, for h small enough, there exists $x_h \in U$ such that $F_h(x_h) = 0$, and $F_h'(x_h)$ is positive definite uniformly in h. Further, there holds the a priori error estimate

$$\|x - x_h\| \leq \frac{2}{\gamma}\|F(x) - F_h(x)\|. \tag{3.5}$$

Proof. Due to condition (iv), we can choose a positive number $h_1 \in \mathbb{R}_+$ such that for $h \leq h_1$ there holds

$$\|F'(x) - F_h'(x)\| \leq \frac{1}{4}\gamma. \tag{3.6}$$

Moreover, for $\rho = \rho(h) = \frac{\gamma}{kL(h)}$, with some $k \geq 4$ sufficiently large, there holds

$$B_\rho(x) = \{\xi \in \mathbb{R}^n,\ \|x - \xi\| \leq \rho\} \subset U. \tag{3.7}$$

For this choice, we obtain that, for $h \leq h_1$, $F_h'(\cdot)$ is positive definite on $B_\rho(x)$:

$$\begin{aligned} p^* F_h'(\xi) p &= p^* F'(x) p + p^*(F_h'(x) - F'(x))p + p^*(F_h'(\xi) - F_h'(x))p \\ &\geq \gamma\|p\|^2 - \|F_h'(x) - F'(x)\|\,\|p\|^2 - \|F_h'(\xi) - F_h'(x)\|\,\|p\|^2 \\ &\geq \left(\gamma - \frac{1}{4}\gamma - L(h)\rho\right)\|p\|^2 \geq \frac{1}{2}\gamma\|p\|^2. \end{aligned}$$

In a similar way, we conclude that, for $h \le h_1$, $F_h'(\cdot)$ is also bounded on $B_\rho(x)$:

$$\|F_h'(\xi)\| \le \beta := \|F'(x)\| + \frac{1}{2}\gamma.$$

Next, we prove that there exists a unique $x_h \in B_\rho(x)$ with $F_h(x_h) = 0$. To this end, we define an operator $D_s : \mathbb{R}^n \to \mathbb{R}^n$, for $s \in \mathbb{R}_+$, by

$$D_s(\xi) = \xi - sF_h(\xi).$$

For a certain choice of s, we show that D_s is a contraction on $B_\rho(x)$, and we use the Banach fixed point theorem. For $\xi \in B_\rho(x)$, $h \le h_1$, and an arbitrary $p \in \mathbb{R}^n$, there holds

$$\begin{aligned}\|D_s'(\xi)p\|^2 &= \|p - sF_h'(\xi)p\|^2 = \|p\|^2 - 2sp^*F_h'(\xi)p + s^2\|F_h'(\xi)p\|^2 \\ &\le (1 - s\gamma + s^2\beta^2)\|p\|^2.\end{aligned}$$

For the choice $s = \gamma(2\beta^2)^{-1}$, we obtain

$$\|D_s'(\xi)p\|^2 \le \Big(1 - \frac{\gamma^2}{4\beta^2}\Big)\|p\|^2,$$

and consequently,

$$\|D_s'(\xi)\| \le \Big(1 - \frac{\gamma^2}{4\beta^2}\Big)^{1/2} < 1.$$

Moreover, for arbitrary $\xi \in B_\rho(x)$, there holds

$$\begin{aligned}\|x - D_s(\xi)\| &= \|D_s(x) - D_s(\xi) + sF_h(x)\| \\ &\le \|D_s(x) - D_s(\xi)\| + s\|F_h(x) - F(x)\| \\ &\le \|D_s'(\eta)\|\,\|x - \xi\| + s\|F_h(x) - F(x)\|\end{aligned}$$

for a certain $\eta \in B_\rho$. Hence, the above estimate implies

$$\begin{aligned}\|x - D_s(\xi)\| &\le \Big(1 - \frac{\gamma^2}{4\beta^2}\Big)^{1/2}\rho + s\,\|F_h(x) - F(x)\| \\ &= \rho\left\{\Big(1 - \frac{\gamma^2}{4\beta^2}\Big)^{1/2} + s\frac{k}{\gamma}L(h)\|F_h(x) - F(x)\|\right\}.\end{aligned}$$

Due to condition (iii), there is a number $h_2 \in \mathbb{R}_+$ such that, for $h \le h_2$, there holds

$$L(h)\|F_h(x) - F(x)\| \le \frac{\gamma}{ks}\left\{1 - \Big(1 - \frac{\gamma^2}{4\beta^2}\Big)^{1/2}\right\}.$$

Hence, for $h < h_0 := \min\{h_1, h_2\}$,

$$\|x - D_s(\xi)\| \le \rho,$$

and consequently $D_s(\xi) \in B_\rho(x)$. For $h \le h_0$, by the Banach fixed point theorem, we obtain the existence of $x_h \in B_\rho(x)$ with $F_h(x_h) = 0$. By construction of $B_\rho(x)$, the derivative $F_h'(x_h)$ is positive definite with the h-independent constant $\frac{1}{2}\gamma$. This implies that, for a certain $\xi \in B_\rho(x)$,

$$(x - x_h)^*(F_h(x) - F_h(x_h)) = (x - x_h)^*F_h'(\xi)(x - x_h) \ge \frac{\gamma}{2}\|x - x_h\|^2.$$

Hence, using $F(x) = F_h(x_h) = 0$,

$$\|x - x_h\|^2 \le \frac{2}{\gamma}(x - x_h)^*(F_h(x) - F_h(x_h)) = \frac{2}{\gamma}(x - x_h)^*(F_h(x) - F(x))$$
$$\le \frac{2}{\gamma}\|F_h(x) - F(x)\|\,\|x - x_h\|.$$

This completes the proof. ∎

4. A priori error estimation. In this section we apply the paradigm presented in section 3 to the problem under consideration. We prove the following theorem.

THEOREM 4.1. *Let $q \in Q$ be a stable solution of (2.5). Then, for h small enough, there exists a stable solution $q_h \in Q$ of (2.22), and there holds the following a priori error estimate:*

$$\|q - q_h\| = \mathcal{O}(h^2|\log(h)|^2). \tag{4.1}$$

On the basis of the estimate (4.1), we can also derive optimal-order estimates for the error $u - u_h$ in the corresponding states. However, since this would be a simple exercise using the arguments developed below, and since the optimal states are of only minor practical interest in parameter estimation problems, we do not state these estimates.

The proof of Theorem 4.1 is given by checking the conditions from Theorem 3.1 for the operators

$$F(\xi) := \nabla j(\xi), \quad F_h(\xi) := \nabla j_h(\xi).$$

The constant in (4.1) turns out to depend in a reciprocal way on the distance

$$\delta := \min_{i=1,\dots,n_m} \operatorname{dist}(\xi_i, \Sigma)$$

of the set of measurement points ξ_i to the set Σ of corner points of $\partial\Omega$. Therefore, we will use generic constants c and c_δ, where c depends only on the domain Ω, the force f, and the characteristics of the mesh family $\{\mathcal{T}_h\}_h$, while c_δ may additionally depend on the distance δ like $c_\delta \approx \delta^{-1}$. Further, by $L^p(\Omega)$ and $W^{m,p}(\Omega)$, for $m \in \mathbb{N}$ and $1 \le p \le \infty$, we denote the standard Lebesgue and Sobolev spaces, respectively, and by $\|\cdot\|_p$ and $\|\cdot\|_{m,p}$ the corresponding norms. The restriction of such a norm to a subset $\Omega' \subset \Omega$ is indicated by $\|\cdot\|_{m,p;\Omega'}$.

By $i_h : C(\overline{\Omega}) \to V_h$ we denote the usual (linear) operator of nodal interpolation for which the following cellwise estimate is well known (see Brenner and Scott [5]):

$$h_K^{-2}\|v - i_h v\|_{p;K} + h_K^{-1}\|\nabla(v - i_h v)\|_{p;K} + \|\nabla^2 i_h v\|_{p;K} \le c\|\nabla^2 v\|_{p;K}, \tag{4.2}$$

for $1 \le p \le \infty$, with constants c independent of h.

An important ingredient of the proof of Theorem 4.1 is the following L^∞-stability theorem. For $d > 0$, we define the subset $\Omega_d \subset \Omega$ by

$$\Omega_d := \{x \in \Omega,\ \operatorname{dist}(x, \Sigma) > d\}.$$

THEOREM 4.2 (stability theorem). *Let $q \in Q_0$, $\psi \in H_0^1(\Omega) \cap C(\bar{\Omega})$, and a matrix $B = B(x) \in W^{1,\infty}(\Omega)^{2\times 2}$ be given. Moreover, let $v_h \in V_h$ be a solution of*

$$a(q)(v_h, \phi_h) = (B\nabla\psi, \nabla\phi_h) \quad \forall \phi_h \in V_h. \tag{4.3}$$

Then, there hold the L^2-stability estimate

$$\|v_h\|_2 + h\|\nabla v_h\|_2 \le c\,|||B|||_{1,\infty}\big\{\|\psi\|_2 + h\|\nabla\psi\|_2\big\} \tag{4.4}$$

and the local L^∞-stability estimate

$$\|v_h\|_{\infty;\Omega_d} \le c_d\,|||B|||_{1,\infty}\big\{|\log(h)|\,\|\psi\|_{\infty;\Omega_{d/2}} + \|\psi\|_2 + h\|\nabla\psi\|_2\big\}, \tag{4.5}$$

with a constant $c_d \approx d^{-1}$.

The L^2 estimate (4.4) is a standard result from finite element analysis, while the L^∞ estimate (4.5) can be concluded by estimates of discrete Green functions such as these developed in Frehse and Rannacher [10] and Rannacher and Scott [24] (see also Brenner and Scott [5, Chapter 7]). The proof for this is given in section 5 below. A similar L^∞-stability result has been proven in Rannacher [22] in the time-dependent parabolic case. For the solution q of problem (2.5), we introduce $\bar u_h \in V_h$ determined by

$$a(q)(\bar u_h, \phi_h) = (f, \phi_h) \quad \forall \phi_h \in V_h. \tag{4.6}$$

Further, we define $w_{j,h} \in V_h$ and $v_{jk,h} \in V_h$, for $j,k = 1, 2, \dots, n_p$, as the solutions of the problems

$$a(q)(w_{j,h}, \phi_h) = -(A'_{q_j}(q)\nabla\bar u_h, \nabla\phi_h) \quad \forall \phi_h \in V_h \tag{4.7}$$

and

$$\begin{aligned} a(q)(v_{jk,h}, \phi_h) = &-(A'_{q_j}(q)\nabla w_{k,h}, \nabla\phi_h) - (A'_{q_k}(q)\nabla w_{j,h}, \nabla\phi_h) \\ &- (A''_{q_j q_k}(q)\nabla\bar u_h, \nabla\phi_h) \quad \forall \phi_h \in V_h, \end{aligned} \tag{4.8}$$

respectively. The next lemma provides necessary estimates for the errors $u - \bar u_h$, $w_j - w_{j,h}$, and $v_{jk} - v_{jk,h}$. We recall the notation $\delta := \min_{i=1,\dots,n_m} \operatorname{dist}(\xi_i, \Sigma)$.

LEMMA 4.3. *Under the above assumptions the following estimates hold:*

$$\|C(u - \bar u_h)\| \le c_\delta h^2 |\log(h)|, \tag{4.9}$$

$$\|C(w_j - w_{j,h})\| \le c_\delta h^2 |\log(h)|^2, \quad j = 1, 2, \dots, n_p, \tag{4.10}$$

$$\|C(v_{jk} - v_{jk,h})\| \le c_\delta h^2 |\log(h)|^3, \quad j,k = 1, 2, \dots, n_p. \tag{4.11}$$

Proof. The proof uses the a priori bounds (2.16) and (2.17) provided in Lemma 2.2 for u, and the auxiliary functions w_j, v_{jk}, $j,k = 1, \dots, n_p$, corresponding to arbitrary $q \in Q_0$.

(i) By definition, $\bar u_h$ is the Ritz projection of u corresponding to the energy form $a(q)(\cdot,\cdot)$, i.e.,

$$a(q)(\bar u_h, \phi_h) = a(q)(u, \phi_h) \quad \forall \phi_h \in V_h.$$

By the standard L^2-error estimate for finite elements, there holds

$$\|u - \bar u_h\|_2 + h\|\nabla(u - \bar u_h)\|_2 \le ch^2. \tag{4.12}$$

Further, applying the L^∞-stability estimate (4.5) of Theorem 4.2 for the equation

$$a(q)(i_h u - \bar u_h, \phi_h) = a(q)(i_h u - u, \phi_h) \quad \forall \phi_h \in V_h,$$

with the nodal interpolant $i_h u \in V_h$ of u, yields the estimate

$$\|i_h u - \bar{u}_h\|_{\infty;\Omega_\delta} \le c_\delta \big\{ |\log(h)|\, \|i_h u - u\|_{\infty;\Omega_{\delta/2}} + \|i_h u - u\|_2 + h\|\nabla(i_h u - u)\|_2 \big\}.$$

From this, using the approximation properties (4.2) of i_h, we conclude the error estimate

$$\|u - \bar{u}_h\|_{\infty;\Omega_\delta} \le \|u - i_h u\|_{\infty;\Omega_\delta} + \|i_h u - \bar{u}_h\|_{\infty;\Omega_\delta} \le c_\delta h^2 |\log(h)|. \tag{4.13}$$

Here, the constant c_δ depends on the global H^2 norm and the local $W^{2,\infty}$ norm of the solution, which are both known to be bounded in view of the a priori bounds (2.16) and (2.17). Since $\xi_i \in \overline{\Omega}_\delta$, we obtain the estimate (4.9).

(ii) For proving (4.10), we introduce an additional discrete variable $\bar{w}_{j,h}$ determined by the equation

$$a(q)(\bar{w}_{j,h}, \phi_h) = -(A'_{q_j}(q)\nabla u, \nabla\phi_h) \quad \forall \phi_h \in V_h.$$

The error $e = w_j - w_{j,h}$ is split like $e = e_1 + e_2$, with $e_1 = w_j - \bar{w}_{j,h}$ and $e_2 = \bar{w}_{j,h} - w_{j,h}$. For the Ritz-projection error e_1, as before, there holds the L^2-error estimate

$$\|e_1\|_2 + h\|\nabla e_1\|_2 \le ch^2 \|w_j\|_{2,2} \le ch^2$$

and the pointwise error estimate

$$\|e_1\|_{\infty;\Omega_\delta} \le c_\delta h^2 \big\{ |\log(h)|\, \|\nabla^2 w_j\|_{\infty;\Omega_{\delta/2}} + \|w_j\|_{2,2} \big\} \le c_\delta h^2 |\log(h)|.$$

For $e_2 \in V_h$, we have

$$a(q)(e_2, \phi_h) = -(A'_{q_j}(q)\nabla(u - \bar{u}_h), \nabla\phi_h) \quad \forall \phi_h \in V_h.$$

Hence, the L^2-stability estimate (4.4) of Theorem 4.2 and the estimate (4.12) imply

$$\|e_2\|_2 + h\|\nabla e_2\|_2 \le c\big\{ \|u - \bar{u}_h\|_2 + h\|\nabla(u - \bar{u}_h)\|_2 \big\} \le ch^2.$$

This shows that, for $j = 1, \dots, n_p$,

$$\|w_j - w_{j,h}\|_2 + h\|\nabla(w_j - w_{j,h})\|_2 \le ch^2. \tag{4.14}$$

Further, the L^∞-stability estimate (4.5) of Theorem 4.2 yields

$$\|e_2\|_{\infty;\Omega_\delta} \le c_\delta \big\{ |\log(h)|\, \|u - \bar{u}_h\|_{\infty;\Omega_{\delta/2}} + \|u - \bar{u}_h\|_2 + h\|\nabla(u - \bar{u}_h)\|_2 \big\},$$

which, by (4.12) and (4.13), implies $\|e_2\|_{\infty;\Omega_\delta} \le c_\delta |\log(h)|^2 h^2$. We obtain

$$\|w_j - w_{j,h}\|_{\infty;\Omega_\delta} \le c_\delta |\log(h)|^2 h^2, \quad j = 1, \dots, n_p, \tag{4.15}$$

which implies the desired estimate (4.10).

(iii) The proof of (4.11) uses the same line of argument as before. Using the additional discrete variable $\bar{v}_{jk,h}$ determined by the equation

$$\begin{aligned} a(q)(\bar{v}_{jk,h}, \phi_h) = &-(A'_{q_j}(q)\nabla w_k, \nabla\phi_h) - (A'_{q_k}(q)\nabla w_j, \nabla\phi_h) \\ &- (A''_{q_j q_k}(q)\nabla u, \nabla\phi_h) \quad \forall \phi_h \in V_h, \end{aligned}$$

the error $e = v_{jk} - v_{jk,h}$ is split like $e = e_1 + e_2$, with $e_1 = v_{jk} - \bar{v}_{jk,h}$ and $e_2 = \bar{v}_{jk,h} - v_{jk,h}$. For the Ritz-projection error e_1, as before, we conclude the pointwise error estimate

$$\|e_1\|_{\infty;\Omega_\delta} \le c_\delta h^2 \big\{ |\log(h)|\, \|\nabla^2 v_{jk}\|_{\infty;\Omega_{\delta/2}} + \|v_{jk}\|_{2,2} \big\} \le c_\delta h^2 |\log(h)|.$$

For $e_2 \in V_h$, we have

$$\begin{aligned} a(q)(e_2,\phi_h) = &-(A'_{q_j}(q)\nabla(w_k - w_{k,h}), \nabla\phi_h) - (A'_{q_k}(q)\nabla(w_j - w_{j,h}), \nabla\phi_h) \\ &-(A''_{q_j q_k}(q)\nabla(u - \bar{u}_h), \nabla\phi_h) \quad \forall \phi_h \in V_h, \end{aligned}$$

and therefore, again by the L^∞-stability estimate (4.5) of Theorem 4.2,

$$\begin{aligned} \|e_2\|_{\infty;\Omega_\delta} \le c_\delta |\log(h)| \Big\{ &\max_{j=1,\dots,n_p} \|w_j - w_{j,h}\|_{\infty;\Omega_{\delta/2}} + \|u - \bar{u}_h\|_{\infty;\Omega_{\delta/2}} \Big\} \\ &+ c \max_{j=1,\dots,n_p} \big\{ \|w_j - w_{j,h}\|_2 + h\|\nabla(w_j - w_{j,h})\|_2 \big\} \\ &+ c\big\{ \|u - \bar{u}_h\|_2 + h\|\nabla(u - \bar{u}_h)\|_2 \big\}. \end{aligned}$$

Then, by the foregoing error estimates, we obtain $\|e_2\|_{\infty;\Omega_\delta} \le c_\delta h^2 |\log(h)|^3$, and consequently,

$$\|v_{jk} - v_{jk,h}\|_{\infty;\Omega_\delta} \le c_\delta |\log(h)|^3 h^2, \quad j,k = 1,\dots,n_p. \tag{4.16}$$

This eventually yields the desired estimated (4.11). □

A direct application of Lemma 4.3 leads to the following result.

LEMMA 4.4. *Under the above assumptions, there holds*

$$\Big| \frac{\partial}{\partial q_j}(j - j_h)(q) \Big| \le c_\delta h^2 |\log(h)|^2, \quad j = 1,2,\dots,n_p, \tag{4.17}$$

$$\Big| \frac{\partial^2}{\partial q_j \partial q_k}(j - j_h)(q) \Big| \le c_\delta h^2 |\log(h)|^3, \quad j,k = 1,2,\dots,n_p. \tag{4.18}$$

Proof. We have the representation

$$\begin{aligned} \frac{\partial}{\partial q_j}(j - j_h)(q) &= \langle C(u) - \hat{C}, C(w_j)\rangle - \langle C(\bar{u}_h) - \hat{C}, C(w_{j,h})\rangle \\ &= \langle C(u) - \hat{C}, C(w_j - w_{j,h})\rangle + \langle C(u - \bar{u}_h), C(w_{j,h})\rangle, \end{aligned}$$

from which we obtain

$$\Big| \frac{\partial}{\partial q_j}(j - j_h)(q) \Big| \le \|C(u) - \hat{C}\|\, \|C(w_j - w_{j,h})\| + \|C(u - \bar{u}_h)\|\, \|C(w_{j,h})\|.$$

By the a priori bounds (2.16) and (2.17) and the Sobolev embedding theorem, we see that

$$\|C(u)\| + \|C(w_j)\| + \|C(v_{jk})\| \le c. \tag{4.19}$$

Combining this with the error estimate (4.10) implies $\|C(w_{j,h})\| \le c$. Then, we can conclude the first estimate (4.17) from the error estimates of Lemma 4.3. To prove (4.18), we write

$$\begin{aligned}\frac{\partial^2}{\partial q_j q_k}(j-j_h)(q) &= \langle C(w_j), C(w_k)\rangle + \langle C(u)-\hat{C}, C(v_{jk})\rangle \\ &\quad - \langle C(w_{j,h}), C(w_{k,h})\rangle - \langle C(\bar{u}_h)-\hat{C}, C(v_{jk,h})\rangle \\ &= \langle C(w_j - w_{j,h}), C(w_k)\rangle + \langle C(w_{j,h}), C(w_k - w_{k,h})\rangle \\ &\quad + \langle C(u-\bar{u}_h), C(v_{jk})\rangle + \langle C(\bar{u}_h)-\hat{C}, C(v_{jk}-v_{jk,h})\rangle.\end{aligned}$$

Using as before the bounds (4.19) and the error estimates of Lemma 4.3 completes the proof. ∎

For the application of Theorem 3.1 it remains to check the Lipschitz condition (3.2). For two arbitrary parameter sets $\xi, \eta \in Q_0$, we set $u_\xi = S_h(\xi)$ and $u_\eta = S_h(\eta)$. Correspondingly, we define $w_{j,\xi}, w_{j,\eta} \in V_h$ and $v_{jk,\xi}, v_{jk,\eta} \in V_h$ similarly to $w_{j,h}$ and $v_{j,h}$ for $q = \xi$ and $q = \eta$, respectively.

LEMMA 4.5. *For $\xi, \eta \in Q_0$, there hold*

$$\|C(u_\xi - u_\eta)\| \le c_\delta |\log(h)|\, \|\xi - \eta\|, \tag{4.20}$$

$$\|C(w_{j,\xi} - w_{j,\eta})\| \le c_\delta |\log(h)|^2\, \|\xi - \eta\|, \tag{4.21}$$

$$\|C(v_{jk,\xi} - v_{jk,\eta})\| \le c_\delta |\log(h)|^3\, \|\xi - \eta\|. \tag{4.22}$$

Proof. Due to the definition of u_ξ and u_η, we have

$$(A(\xi)\nabla(u_\xi - u_\eta), \nabla\phi_h) = -((A(\xi) - A(\eta))\nabla u_\eta, \nabla\phi_h) \quad \forall \phi_h \in V_h.$$

Using Theorem 4.2, with $d = \delta$, we obtain

$$\|C(u_\xi - u_\eta)\| \le c\, |||(A(\xi) - A(\eta))|||_{1,\infty} \left\{ |\log(h)|\, \|u_\eta\|_{\infty;\Omega_{\delta/2}} + \|u_\eta\|_2 + h\|\nabla u_\eta\|_2 \right\}.$$

Since u_η is the Ritz projection of an H^2 function, all its norms occurring on the right-hand side are bounded independent of h and $\eta \in Q_0$ by standard estimates from finite element analysis. This implies (4.20) since

$$|||A(\xi) - A(\eta)|||_{1,\infty} \le c\|\xi - \eta\|.$$

The estimates (4.21) and (4.22) are obtained in a similar way. ∎

LEMMA 4.6. *For $\xi, \eta \in Q_0$, there holds*

$$\left| \frac{\partial^2}{\partial q_j q_k} j_h(\xi) - \frac{\partial^2}{\partial q_j q_k} j_h(\eta) \right| \le L(h)\|\xi - \eta\|, \tag{4.23}$$

where $L(h) = c_\delta |\log(h)|^3$.

Proof. We have

$$\begin{aligned}\frac{\partial^2}{\partial q_j q_k} j_h(\xi) - \frac{\partial^2}{\partial q_j q_k} j_h(\eta) &= \langle C(w_{j,\xi}), C(w_{k,\xi})\rangle + \langle C(u_\xi) - \hat{C}, C(v_{jk,\xi})\rangle \\ &\quad - \langle C(w_{j,\eta}), C(w_{k,\eta})\rangle - \langle C(u_\eta) - \hat{C}, C(v_{jk,\eta})\rangle \\ &= \langle C(w_{j,\xi} - w_{j,\eta}), C(w_{k,\xi})\rangle + \langle C(w_{j,\eta}), C(w_{k,\xi} - w_{k,\eta})\rangle \\ &\quad + \langle C(u_\xi - u_\eta), C(v_{jk,\xi})\rangle + \langle C(u_\eta) - \hat{C}, C(v_{jk,\xi} - v_{jk,\eta})\rangle,\end{aligned}$$

and, consequently,

$$\begin{aligned}\Big|\frac{\partial^2}{\partial q_j q_k} j_h(\xi) - \frac{\partial^2}{\partial q_j q_k} j_h(\eta)\Big| &\le \|C(w_{j,\xi} - w_{j,\eta})\|\,\|C(w_{k,\xi})\| \\ &\quad + \|C(w_{j,\eta})\|\,\|C(w_{k,\xi} - w_{k,\eta})\| + \|C(u_\xi - u_\eta)\|\,\|C(v_{jk,\xi})\| \\ &\quad + \|C(u_\eta) - \hat{C}\|\,\|C(v_{jk,\xi} - v_{jk,\eta})\|.\end{aligned}$$

Now, the assertion follows by the estimates of Lemma 4.5 if we can bound the terms $C(u_\eta)$, $C(w_{j,\eta})$, and $C(v_{jk,\xi})$. This is achieved by using the bounds for $C(u)$, $C(w_j)$, and $C(v_{jk})$ in (4.19) together with the error estimates of Lemma 4.3. ∎

To complete the proof of Theorem 4.1 we check the conditions of Theorem 3.1. Condition (3.1) is fulfilled due to the stability of the solution q of the problem (2.5). Condition (3.2) is shown in Lemma 4.6. Condition (3.3) is obtained by Lemma 4.4 and Lemma 4.6 using $\lim_{h\to 0} h^2|\log(h)|^5 = 0$. Finally, condition (3.4) holds due to Lemma 4.4. Hence, the estimate (3.5) of Theorem 3.1 completes the proof.

5. Proof of Theorem 4.2. (i) We begin with the L^2-stability estimate. Taking $\phi_h := v_h$ in (4.3), we obtain

$$\|\nabla v_h\|_2 \le c|||B|||_{1,\infty}\|\nabla\psi\|_2. \tag{5.1}$$

To estimate $\|v_h\|_2$, we use the solution $z \in V \cap H^2(\Omega)$ of the auxiliary equation

$$a(q)(\phi, z) = (\phi, v_h)\|v_h\|_2^{-1} \quad \forall \phi \in V.$$

Taking $\phi := v_h$ as test function and integrating by parts, we have

$$\begin{aligned}\|v_h\|_2 &= a(q)(v_h, z) = a(q)(v_h, z - i_h z) + a(q)(v_h, i_h z) \\ &= a(q)(v_h, z - i_h z) + (B\nabla\psi, \nabla i_h z) \\ &= a(q)(v_h, z - i_h z) + (B\nabla\psi, \nabla(i_h z - z)) - (\psi, \nabla\cdot B^T\nabla z).\end{aligned}$$

Then, using the approximation properties (4.2) of the interpolant $i_h z \in V_h$, we conclude that

$$\begin{aligned}\|v_h\|_2 &\le c\|\nabla v_h\|_2\|\nabla(z - i_h z)\|_2 + |||B|||_{1,\infty}\|\nabla\psi\|_2\|\nabla(z - i_h z)\|_2 \\ &\quad + |||B|||_{1,\infty}\|\psi\|_2\|z\|_{2,2} \\ &\le c\big\{h\|\nabla v_h\| + |||B|||_{1,\infty}h\|\nabla\psi\|_2 + |||B|||_{1,\infty}\|\psi\|_2\big\}\,\|z\|_{2,2}.\end{aligned}$$

Hence, observing (5.1) and the bound $\|z\|_{2,2} \le c$, we obtain

$$\|v_h\|_2 \le c|||B|||_{1,\infty}\big\{\|\psi\|_2 + h\|\nabla\psi\|_2\big\}. \tag{5.2}$$

(ii) Next, we prove the L^∞-stability estimate. Let $a \in \Omega_\delta$ be an arbitrary point lying in a cell K. For any fixed h, there exists a cellwise polynomial function δ_h with $\mathrm{supp}(\delta_h) \subset K$ such that

$$(\phi_h, \delta_h) = \phi_h(a) \quad \forall \phi_h \in V_h.$$

The function δ_h plays the role of an approximate Dirac function. Correspondingly, we introduce a regularized Green function $g \in V \cap H^2(\Omega)$ by

$$(A(q)\nabla\phi, \nabla g) = (\delta_h, \phi) \quad \forall \phi \in V,$$

and the corresponding Ritz projection $g_h \in V_h$ by

$$(A(q)\nabla\phi_h, \nabla g_h) = (\delta_h, \phi_h) \quad \forall \phi_h \in V_h.$$

For functions which are only cellwise defined, we will use the "broken" norm $\|v\|'_p := \sum_{K\in\mathcal{T}_h} \|v\|_{p;T}$. The following three lemmas provide the key estimates for the proof of the theorem.

LEMMA 5.1. *The following global L^2 estimates hold:*

$$\|g\|_2 + |\log(h)|^{-1/2}\|\nabla g\|_2 + h\|\nabla^2 g\|_2 \le c, \tag{5.3}$$

$$h^{-1}\|g - g_h\|_2 + \|\nabla(g - g_h)\|_2 + h\|\nabla^2 g_h\|'_2 \le c. \tag{5.4}$$

Proof. The assertion follows by standard L^2 a priori and error estimates for g and $g - g_h$, respectively. We skip the details and refer to [10]. Note that $\|\nabla^2 g_h\|'_2$ vanishes for linear finite elements. In the case of bilinear elements, we estimate using the interpolant $i_h g$ as follows:

$$\|\nabla^2 g_h\|'_2 \le \|\nabla^2(g_h - i_h g)\|'_2 + \|\nabla^2(g - i_h g)\|'_2 + \|\nabla^2 g\|_2.$$

For the first term, we obtain, using an inverse inequality,

$$\begin{aligned}\|\nabla^2(g_h - i_h g)\|'_2 &\le ch^{-1}\|\nabla(g_h - i_h g)\|_2 \\ &\le ch^{-1}\big\{\|\nabla(g - i_h g)\|_2 + \|\nabla(g - g_h)\|_2\big\}\end{aligned}$$

and obtain by the interpolation estimate (4.2), with $p = 2$, and by the other estimates derived before,

$$\|\nabla^2 g_h\|'_2 \le c\big\{h^{-1}\|\nabla(g - g_h)\|_2 + \|\nabla^2 g\|_2\big\} \le ch^{-1}.$$

This completes the proof. ∎

LEMMA 5.2. *For sufficiently small $h \ll \delta$, the following local L^2 estimate holds:*

$$\|\nabla(g - g_h)\|_{2;\Omega\setminus\Omega_{\delta/2}} + h\|\nabla^2 g\|_{2;\Omega\setminus\Omega_{\delta/2}} \le c_\delta h, \tag{5.5}$$

with a constant $c_\delta \approx \delta^{-1}$ but independent of h.

Proof. The assertion follows by standard local elliptic a priori estimates and by arguments from the local L^2 error analysis for finite elements, as provided in Nitsche and Schatz [20]:

$$\begin{aligned}\|\nabla^2 g\|_{2;\Omega\setminus\Omega_{\delta/2}} &\le c\|\Delta g\|_{2;\Omega\setminus\Omega_{3\delta/4}} + c_\delta\|g\|_2, \\ \|\nabla(g - g_h)\|_{2;\Omega\setminus\Omega_{\delta/2}} &\le c\|\nabla(g - i_h g)\|_{2;\Omega\setminus\Omega_{3\delta/4}} + c_\delta\|g - g_h\|_2,\end{aligned}$$

with constants $c_\delta \approx \delta^{-1}$. Now, the assertion follows by the interpolation estimate (4.2) and the other estimates already proven. ∎

LEMMA 5.3. *The following L^1 a priori and error estimates hold:*

$$\|\nabla g\|_1 + \|\nabla^2 g\|'_1 \le c|\log(h)|, \tag{5.6}$$

$$\|\nabla(g - g_h)\|_1 + h\|\nabla^2(g - g_h)\|'_1 \le ch|\log(h)|, \tag{5.7}$$

with a constant c independent of h and δ.

Proof. The proof can be found in [10]. ∎

For the point $a \in \Omega_d$, there holds

$$v_h(a) = (v_h, \delta_h) = (A(q)\nabla v_h, \nabla g_h) = (B\nabla\psi, \nabla g_h).$$

We employ a standard localization argument. Let $\omega \in C_0^\infty(\Omega)$ be a smooth function with the properties

$$0 \le \omega \le 1, \quad \omega_{|\Omega_{\delta/2}} \equiv 1, \quad \omega_{|\Omega\setminus\Omega_{\delta/4}} \equiv 0.$$

With this notation, we have

$$(B\nabla\psi, \nabla g_h) = (B\nabla(\omega\psi), \nabla g_h) + (B\nabla((1-\omega)\psi), \nabla g_h) =: \Sigma_1 + \Sigma_2.$$

First, we estimate the term Σ_1. By integration by parts and observing that $\psi_{|\partial\Omega} = 0$, we obtain

$$(B\nabla(\omega\psi), \nabla g_h) = \sum_{K\in\mathcal{T}_h} \big\{(\omega\psi, -\nabla\cdot(B\nabla g_h))_T + (\omega\psi, n\cdot B\nabla g_h)_{\partial K\setminus\partial\Omega}\big\},$$

where n is the outward unit normal vector to ∂K. Let $[\nabla g_h]$ denote the jump of the gradient across the interior faces $\Gamma \subset \partial K$. Using this notation, we obtain

$$\Sigma_1 \le \|\psi\|_{\infty;\Omega_{\delta/2}} \sum_{K\in\mathcal{T}_h} \left\{\|\nabla\cdot(B\nabla g_h)\|_{1;K} + \frac{1}{2}\|n\cdot[B\nabla g_h]\|_{1;\partial K\setminus\partial\Omega}\right\}.$$

First, the estimates of Lemma 5.3 yield

$$\sum_{K\in\mathcal{T}_h} \|\nabla\cdot(B\nabla g_h)\|_{1;K} \le c\,|||B|||_{1,\infty}\big\{\|\nabla g_h\|_1 + \|\nabla^2 g_h\|_1'\big\} \le c\,|||B|||_{1,\infty}\,|\log(h)|.$$

Next, observing that $g \in H^2(\Omega)$ and therefore $[B\nabla g] = 0$, we obtain by a trace theorem

$$\begin{aligned}\sum_{K\in\mathcal{T}_h} \|n\cdot[B\nabla g_h]\|_{1;K\setminus\partial\Omega} &= \sum_{K\in\mathcal{T}_h} \|n\cdot[B\nabla(g_h - g)]\|_{1;\partial K\setminus\partial\Omega}\\ &\le c\,|||B|||_{1,\infty} \sum_{K\in\mathcal{T}_h} \big\{h^{-1}\|\nabla(g-g_h)\|_{1;K} + \|\nabla^2(g-g_h)\|_{1;K}'\big\}.\end{aligned}$$

Hence, collecting the foregoing estimates,

$$\Sigma_1 \le c\,|||B|||_{1,\infty}\,\|\psi\|_{\infty;\Omega_{\delta/2}}\big\{|\log(h)| + h^{-1}\|\nabla(g-g_h)\|_1 + \|\nabla^2(g-g_h)\|_1'\big\}.$$

Again using the estimates of Lemma 5.3, we obtain

$$\Sigma_1 \le c\,|||B|||_{1,\infty}\,\|\psi\|_{\infty;\Omega_{\delta/2}}|\log(h)|. \tag{5.8}$$

For the term Σ_2, we estimate as follows:

$$\begin{aligned}\Sigma_2 &= (B\nabla((1-\omega)\psi), \nabla(g_h - g)) + (B\nabla((1-\omega)\psi), \nabla g)\\ &\le |||B|||_{1,\infty}\big\{\|\nabla\psi\|_2\,\|\nabla(g_h-g)\|_{2;\Omega\setminus\Omega_{\delta/2}} + c_\delta\|\psi\|_2\,\|\nabla(g_h-g)\|_2\big\}\\ &\quad + c\,|||B|||_{1,\infty}\|\psi\|_2\,\big\{\|\nabla^2 g\|_{2;\Omega\setminus\Omega_{\delta/2}} + \|\nabla g\|_{2;\Omega\setminus\Omega_{\delta/2}}\big\}.\end{aligned}$$

Then, by the L^2 estimates of Lemmas 5.1 and 5.2, it follows that

$$\Sigma_2 \le c_\delta |||B|||_{1,\infty} \big\{ \|\psi\|_2 + h\|\nabla\psi\|_2 \big\}. \tag{5.9}$$

This completes the proof of the theorem.

6. Numerical results. In this section, we discuss a sample problem confirming the a priori error estimate of Theorem 4.1. The state equation is given by

$$\begin{aligned} -\nabla\cdot(A(q)\nabla u) &= 2 \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{6.1}$$

where Ω is the unit square. The matrix $A(q)$ is a function of the parameter $q = (q_1, q_2) \in Q = \mathbb{R}^2$, given by

$$A(q) = \begin{pmatrix} q_1^2 & q_1 q_2 \\ q_1 q_2 & \exp(q_2) \end{pmatrix}.$$

In this case the admissible set of parameters is $Q_0 = \{(q_1, q_2) \in Q : q_1 \neq 0, e^{q_2} > q_2^2\}$. The parameters are estimated from the measurements of the state variable at nine different points $\xi_i \in \Omega$; see Figure 6.1.

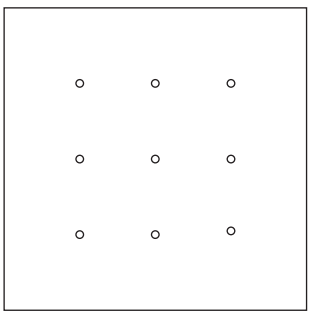


FIG. 6.1. *The computational domain with measurement points marked by circles.*

The vector of measurements $\hat{C}$ is given by

$$\hat{C}_i = C_i(S(\hat{q}))(1 + \varepsilon_i), \quad i = 1, \dots, 9,$$

where the reference parameter is $\hat{q} = (5, 6)$ and $\varepsilon = (\varepsilon_i)$ describes the data perturbation. We consider two cases:

(a) $\varepsilon \approx 0$, (b) $\varepsilon \approx (0.12, -0.26, 0.29, -0.37, -0.49, 0.13, -0.04, -0.45, 0.20)$.

Since the values of $C_i(S(\hat{q}))$ are not available analytically, they are computed approximately by solving state equation (6.1) on a very fine mesh with about 10^6 degrees of freedom. For case (a) the solution $q^{(a)}$ of the parameter identification problem matches the reference parameter $\hat{q}$, and hence the cost functional $J(u)$ almost vanishes in $q^{(a)}$. Case (b) is more realistic because of the "measurement errors" modeled by a randomly chosen ε. Moreover, in this case in contrast to case (a), the solution $q^{(b)}$ of the corresponding parameter identification problem and the reference parameter $\hat{q}$ differ.

The parameter identification problem is discretized using bilinear finite elements on uniformly refined meshes. The results are listed in Tables 6.1 and 6.2. For both cases the theoretically predicted orders of convergence are achieved.

TABLE 6.1
Case (a)*: the error and the order of convergence with respect to the components of q without data perturbation; $N \sim h^{-2}$ number of unknowns.*

N	$q_1^{(a)} - q_{1,h}^{(a)}$	$q_2^{(a)} - q_{2,h}^{(a)}$
81	5.955e-1	9.902e-4
289	1.407e-1	1.731e-4
1089	3.436e-2	4.343e-5
4225	8.509e-3	1.080e-5
16641	2.098e-3	2.668e-6
66049	4.993e-4	6.352e-7
Order	2.05	2.04

TABLE 6.2
Case (b)*: the error and the order of convergence with respect to the components of q with data perturbation, $N \sim h^{-2}$ number of unknowns.*

N	$q_1^{(b)} - q_{1,h}^{(b)}$	$q_2^{(b)} - q_{2,h}^{(b)}$
81	2.059e-0	1.874e-2
289	5.172e-1	2.999e-3
1089	1.467e-1	8.341e-4
4225	3.771e-2	2.111e-4
16641	9.832e-3	5.640e-5
66049	2.350e-3	1.348e-5
Order	2.01	1.98

7. Conclusions and extensions. In this paper we have derived an a priori error estimate for the finite element discretization of an elliptic discrete parameter identification problem with pointwise measurements. The crucial point in our argument is the stability estimate of Theorem 4.2. The result of Theorem 4.1 can be extended to situations in which such a stability estimate is available. We list some possible directions of generalization.

1. *More general meshes.* For simplicity, we have assumed a quasi-uniform mesh family $\{\mathcal{T}_h\}_h$. The analysis can be extended to locally refined meshes, provided that the ratio of $h_{\min}$ and $h = h_{\max}$ is polynomial, $h_{\min} \approx h^p$, with some $p \geq 1$. For such meshes the stability result of Theorem 4.2 holds true with $|\log(h_{\min})| \approx p|\log(h)|$. This will be shown in the forthcoming paper [23]. Related results for L^∞-error estimates can be extracted (with some work) from Schatz and Wahlbin [25].

2. *More general domains.* Our argument uses that the solution operator $S(\cdot)$ maps Q into $H_0^1(\Omega) \cap H^2(\Omega)$, which is guaranteed on smoothly bounded or convex domains. In the case of a domain with reentrant corners or edges this regularity property is lost. This lack of regularity of the solution can be compensated by an appropriate refinement of the mesh near the critical corner points or edges. The stability estimate of Theorem 4.2 also holds in this situation, in two as well as in three dimensions. This will be shown in a forthcoming paper.

3. *Higher-order approximation.* The result of Theorem 4.1 can be also extended to the case of higher-order finite elements, similar to the analysis of Nitsche [19]. In this case the logarithmic factor $|\log(h)|$ can be dropped in the stability Theorem 4.2.

4. *More general equations.* Theorem 4.1 can also be extended to more general elliptic equations or systems of the form

$$-\nabla \cdot (A(q)\nabla u + b_1(q)u) + b_2(q)u = f,$$

with parameter-dependent coefficients $A(q)$, $b_1(q)$, $b_2(q)$. Corresponding L^∞-error estimates for very general (nonlinear) elliptic systems have been derived in Dobrowolski and Rannacher [7].

REFERENCES

[1] N. Arada, E. Casas, and F. Tröltzsch, *Error estimates for the numerical approximation of a semilinear elliptic control problem*, J. Comput. Optim. Appl., 23 (2002), pp. 201–229.

[2] H. T. Banks and K. Kunisch, *Estimation Techniques for Distributed Parameter Systems*, Birkhäuser Boston, Cambridge, MA, 1989.

[3] R. Becker, M. Braack, and B. Vexler, *Numerical parameter estimation for chemical models in multidimensional reactive flows*, Combust. Theory Modelling, 8 (2004), pp. 661–682.

[4] R. Becker and B. Vexler, *A posteriori error estimation for finite element discretization of parameter identification problems*, Numer. Math., 96 (2004), pp. 435–459.

[5] S. Brenner and R. L. Scott, *The Mathematical Theory of Finite Element Methods*, Springer, Berlin, Heidelberg, New York, 1994.

[6] K. Deckelnick and M. Hinze, *Error estimates in space and time for tracking-type control of the instationary Stokes system*, in Proceedings of the 8th Conference on Control of Distributed Parameter Systems, Graz, Austria, 2001, Internat. Ser. Numer. Math., 143, Birkhäuser, Basel, Switzerland, 2002, pp. 87–104.

[7] M. Dobrowolski and R. Rannacher, *Finite element methods for nonlinear elliptic problems of second order*, Math. Nachr., 94 (1980), pp. 155–172.

[8] R. S. Falk, *Approximation of a class of optimal control problems with order of convergence estimates*, J. Math. Anal. Appl., 44 (1973), pp. 28–47.

[9] R. S. Falk, *Error estimates of the numerical identification of a variable coefficient*, Math. Comp., 40 (1983), pp. 537–546.

[10] J. Frehse and R. Rannacher, *Eine L^1-Fehlerabschätzung für diskrete Grundlösungen in der Methode der finiten Elemente*, in Tagungsband Finite Elemente, Bonner Math. Schrifter 89, University of Bonn, Bonn, Germany, 1976, pp. 92–114.

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

[12] P. Grisvard, *Singularities in Boundary Value Problems*, Masson, Paris, and Springer, Berlin, 1992.

[13] M. D. Gunzburger and L. S. Hou, *Finite-dimensional approximation of a class of constrained nonlinear optimal control problems*, SIAM J. Control Optim., 34 (1996), pp. 1001–1043.

[14] C. Johnson, *Numerical Solution of Partial Differential Equations by Finite Element Method*, Cambridge University Press, Cambridge, UK, 1987.

[15] T. Kärkkäinen, *Error estimates for distributed parameter identification in linear elliptic equations*, J. Math. Systems Estim. Control, 6 (1996), pp. 117–120.

[16] C. Kravaris and J. H. Seinfeld, *Identification of parameters in distributed parameter systems by regularization*, SIAM J. Control Optim., 23 (1985), pp. 217–241.

[17] V. G. Litvinov, *Optimization in Elliptic Problems with Applications to Mechanics of Deformable Bodies and Fluid Mechanics*, Oper. Theory Adv. Appl. 119, Birkhäuser, Basel, Switzerland, 2000.

[18] P. Neittaanmäki and X.-C. Tai, *Error estimates for numerical identification of distributed parameters*, J. Comput. Math., 10 (1992), pp. 66–78.

[19] J. Nitsche, *Über L^∞-Abschätzungen von Projektoren auf Finite Elemente*, in Tagungsband Finite Elemente, Bonner Math. Schrifter 89, University of Bonn, Bonn, Germany, 1976, pp. 13–30.

[20] J. Nitsche and A. Schatz, *Interior estimates for Ritz–Galerkin methods*, Math. Comput., 28 (1976), pp. 937–958.

[21] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Ser. Oper. Res., Springer, New York, 1999.

[22] R. Rannacher, *L^∞-stability estimates and asymptotic error expansion for parabolic finite element equations*, in Tagungsband Extrapolation and Defect Correction, Bonner Math. Schrifter 228, University of Bonn, Bonn, Germany, 1991, pp. 74–94.

[23] R. Rannacher, *On L^∞-Stability in the Finite Element Method on Meshes without Uniform Size and Shape Property*, Preprint SFB 359, Institute of Applied Mathematics, University of Heidelberg, Heidelberg, Germany, 2005.
[24] R. Rannacher and L. R. Scott, *Some optimal error estimates for piecewise linear finite element approximations*, Math. Comp., 38 (1982), pp. 437–445.
[25] A. H. Schatz and L. B. Wahlbin, *Maximum norm estimates in the finite element method on plane polygonal domains. Part* I, Math. Comp., 32 (1978), pp. 73–109.
[26] B. Vexler, *Adaptive Finite Element Methods for Parameter Identification Problems*, Ph.D. thesis, Institute of Applied Mathematics, University of Heidelberg, Heidelberg, Germany, 2004.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 5, pp. 1864–1892

DIRAC STRUCTURES AND BOUNDARY CONTROL SYSTEMS ASSOCIATED WITH SKEW-SYMMETRIC DIFFERENTIAL OPERATORS*

Y. LE GORREC†, H. ZWART‡, AND B. MASCHKE†

Abstract. Associated with a skew-symmetric linear operator on the spatial domain $[a, b]$ we define a Dirac structure which includes the port variables on the boundary of this spatial domain. This Dirac structure is a subspace of a Hilbert space. Naturally, associated with this Dirac structure is an infinite-dimensional system. We parameterize the boundary port variables for which the C_0-semigroup associated with this system is contractive or unitary. Furthermore, this parameterization is used to split the boundary port variables into inputs and outputs. Similarly, we define a linear port controlled Hamiltonian system associated with the previously defined Dirac structure and a symmetric positive operator defining the energy of the system. We illustrate this theory on the example of the Timoshenko beam.

Key words. port Hamiltonian systems, strongly continuous semigroup, boundary control systems, Dirac structures

AMS subject classifications. 37K05, 35B37, 93C05, 35B30, 35G15, 47D03, 47D60

DOI. 10.1137/040611677

1. Introduction.

Port Hamiltonian systems have been introduced in the finite-dimensional case as an analytical frame for the modeling and control of open physical systems [12, 14, 18, 28]. The key concepts are the definition of pairs of power conjugated variables and the geometric structure defined on them. This geometric structure is called the Dirac structure [2, 5]. These Dirac structures also define the internal geometric structure of the physical system as the structure of their interaction with the environment [12, 29]. It reflects the (discrete) topology and the geometry of the physical system under consideration such as the port connection graph, constraints, or interdomain coupling [4, 15, 18]. Furthermore, it is the geometric structure which allows us to define *implicit* Hamiltonian systems and Hamiltonian systems with *port variables* [4, 27, 28]. Port Hamiltonian systems have been used for the design of stabilizing control laws; see, e.g., [13, 21, 22].

Recently, an extension of port Hamiltonian systems to infinite-dimensional systems has been proposed for distributed parameter systems with energy flow at their boundary; see [16, 30]. The state space is a vector space of differential forms defined on the spatial domain and the port variables are defined on the boundary of the spatial domain. The port Hamiltonian system is defined with respect to a so-called *Stokes–Dirac structure*, which in turn is uniquely defined by the exterior derivatives and the order of the differential forms. The Stokes–Dirac structure represents the canonical

*Received by the editors July 15, 2004; accepted for publication (in revised form) April 25, 2005; published electronically December 6, 2005.

http://www.siam.org/journals/sicon/44-5/61167.html

†LAGEP, UCB Lyon 1 - CNRS UMR 5007, CPE Lyon - Bâtiment 308 G, Université Claude Bernard Lyon-1, 43, bd du 11 Novembre 1918, F-69622 Villeurbanne cedex, France (legorrec@lagep.univ-lyon1.fr, maschke@lagep.univ-lyon1.fr). The contribution of these authors has been done in the context of the European sponsored project GeoPlex with reference code IST-2001-34166. Further information is available at http://www.geoplex.cc.

‡Department of Applied Mathematics, University of Twente, P.O. Box 217, 7500 AE, Enschede, The Netherlands (h.j.zwart@math.utwente.nl).

interdomain coupling in physical systems [19]. Finally, the Stokes–Dirac structures have been extended in order to encompass fluid dynamics and beam models [30].

Associated with linear skew-symmetric differential operators, we define Dirac structures and port Hamiltonian systems. Our definition extends the definition of Stokes–Dirac structures in which the operator needed to have differential degree one. We use an alternative definition of a Dirac structure on Hilbert spaces as proposed in [23] and [7]. In [7] Dirac structures on Hilbert spaces have also been used for the study of their composition (interconnection) and the definition of scattering representations. In this paper, we are restricting ourselves to one-dimensional spatial domains.

A major motivation of this work is to provide a theoretic formulation of open Hamiltonian systems, i.e., systems which are subject to some energy flow at their boundary. This formulation is *acausal*, i.e., a priori there is no distinction between inputs and outputs. The acausal formulation is obtained by first introducing boundary port variables. Second, these boundary port variables together with the (formal) skew-symmetric operator lead to the Dirac structure associated with the system.

The second motivation of this paper is to study the existence of solutions for our class of systems. This immediately implies some causality conditions among the port variables. Namely, for a (more or less) free choice of inputs there should exist a solution and the outputs should follow from it. In order to show existence of solutions, we relate our system to the class of boundary control systems. We remark that there are other general system classes which we could have chosen, e.g., system nodes [26]. We have chosen the class of boundary control systems, since this fits most naturally to our class of PDEs with their control at the boundary. As a result, we derive a parameterizing of the port variables such that the semigroup associated with the boundary control system is a contraction semigroup.

This paper is organized as follows. In section 2 we recall the definition of Dirac structures on Hilbert spaces. In section 3 we define Dirac structures associated with skew-symmetric linear differential operators and its conjugated port variables on the boundary of the spatial domain. In section 4 we associate with our Dirac structure a family of boundary control systems. The input of this boundary control system is chosen to lie in a subspace of the boundary port variables. The semigroup associated with this system is a contraction semigroup. By choosing the output to lie in the complementary of the "input subspace" we get a power balance system. The above construction gives the parameterization of all systems for which the associated semigroup is contractive and/or unitary. In section 5 we define a port Hamiltonian system associated with a skew-symmetric differential operator and with a Hamiltonian function. This Hamiltonian is a function defined by a symmetric and coercive linear operator and represents the energy in the system.

2. Dirac structures defined on Hilbert spaces. In this section, we recall the definition of Dirac structures defined on Hilbert spaces proposed by Parsian and Shafei Deh Abad in [23] and by Golo and coauthors in [7, 8]. We shall follow the definitions and notation of [7, 8] for the purpose of analyzing and treating the composition of Dirac structures in the frame of port-based modeling and control. This notation is borrowed partially from the bond graph language, which has been a major source of inspiration for the model definition of port Hamiltonian systems [7, 19, 18].

Let us first define the space of *bond variables* which is constituted of pairs of conjugated variables endowed with a pairing. For models of physical systems this corresponds to an associated instantaneous power; see [1, 11]. Let the *space of flow variables,* denoted by $\mathcal{F}$, and the *space of effort variables,* denoted by $\mathcal{E}$, be real Hilbert

spaces endowed with the inner products $\langle .\,,\,. \rangle_{\mathcal{F}}$ and $\langle .\,,\,. \rangle_{\mathcal{E}}$, respectively. Assume moreover that $\mathcal{F}$ and $\mathcal{E}$ are isometrically isomorphic, that is, there exists an isometry: $r_{\mathcal{F},\mathcal{E}} : F \longrightarrow \mathcal{E}$. Denote furthermore its inverse by $r_{\mathcal{E},\mathcal{F}}$. Define now the *space of bond variables* as the Hilbert space $\mathcal{B} = \mathcal{F} \times \mathcal{E}$ endowed with the natural inner product

$$\langle b^1, b^2 \rangle = \langle f^1, f^2 \rangle_{\mathcal{F}} + \langle e^1, e^2 \rangle_{\mathcal{E}}, \quad b^1 = (f^1, e^1), b^2 = (f^2, e^2) \in \mathcal{B}.$$

In order to define a Dirac structure, let us endow the bond space $\mathcal{B}$ with a *canonical symmetrical pairing*, i.e., a bilinear form defined as follows:

$$\langle b^1, b^2 \rangle_+ = \langle f^1, r_{\mathcal{E},\mathcal{F}} e^2 \rangle_{\mathcal{F}} + \langle e^1, r_{\mathcal{F},\mathcal{E}} f^2 \rangle_{\mathcal{E}}, \quad b^1 = (f^1, e^1), b^2 = (f^2, e^2) \in \mathcal{B}. \tag{2.1}$$

We define a Dirac structure on the bond space $\mathcal{B}$ using this canonical pairing. Denote by $\mathcal{D}^\perp$ the orthogonal subspace to $\mathcal{D}$ with respect to the symmetrical pairing (2.1):

$$\mathcal{D}^\perp = \{ b \in \mathcal{B} | \langle b, b' \rangle_+ = 0 \ \text{ for all } \ b' \in \mathcal{D} \}. \tag{2.2}$$

Definition 2.1. *A Dirac structure $\mathcal{D}$ on the bond space $\mathcal{B} = \mathcal{F} \times \mathcal{E}$ is a subspace of $\mathcal{B}$ which is maximally isotropic with respect to the canonical symmetrical pairing* (2.1)*, i.e.,*

$$\mathcal{D}^\perp = \mathcal{D}. \tag{2.3}$$

One may find different examples of such Dirac structures as well as some properties concerning their representations and their composition in [7, Chapter 5]. We shall now give a canonical example of a Dirac structure in the context of the port-based modeling of physical systems. Therefore, we consider the example of a lossless vibrating string. First, we recall the port-based model structure [19, 30] which gives rise to the definition of a Stokes–Dirac structure on Hilbert spaces of functions with a one-dimensional domain [7]. Second, we recall the formulation of the evolution equation as a port Hamiltonian system.

Example 2.2. Consider an elastic string defined on the one-dimensional spatial domain $Z = [a, b] \subset \mathbb{R}$ and subject to boundary conditions which allow some energy flow. Let us denote by $u(t, z)$ the displacement of the string at time t and position z. Let us first recall the port Hamiltonian formulation of its dynamics. This differs from the classical formulation based on the displacement $u(t, z)$ by the choice of the state variables [17, 30]. In this frame, the state variables are called *energy variables* and are chosen in such a way that the total energy of the string does not depend on their derivatives. The elastic potential energy is a function of the *strain*, and the energy variable is defined by

$$\epsilon(t, z) = \frac{\partial u}{\partial z}(t, z). \tag{2.4}$$

The associated coenergy variable is the *stress* given by

$$\sigma(t, z) = T(z)\, \epsilon(t, z) \tag{2.5}$$

with T denoting the elasticity modulus. Hence the potential energy is the quadratic function of the strain:

$$U(\epsilon(t, \cdot)) = \frac{1}{2} \int_a^b T(z)\, \epsilon(t, z)^2\, dz. \tag{2.6}$$

The kinetic energy K is a function of the kinetic *momentum*, $p(t,z)$, and it is defined by the quadratic function

$$K(p(t,\cdot)) = \frac{1}{2}\int_a^b \frac{p^2(t,z)}{\mu(z)}\,dz. \tag{2.7}$$

The associated coenergy variable is the *velocity* given by

$$v(t,z) = \frac{1}{\mu(z)}p(t,z), \tag{2.8}$$

where μ denotes the mass density.

The dynamical model of the vibrating string is obtained by coupling the elastic energy domain and the kinetic domain through the following relations. Consider the time variation of the energy variables, called flow variables,

$$\frac{\partial}{\partial t}\begin{pmatrix} p \\ \epsilon \end{pmatrix} = \begin{pmatrix} f_K \\ f_U \end{pmatrix}. \tag{2.9}$$

The canonical interdomain coupling between the elastic-potential energy and the kinetic energy relates the flow variables with the coenergy variables. This interdomain coupling is given by the differential operator [19]

$$\begin{pmatrix} f_K \\ f_U \end{pmatrix} = \begin{pmatrix} 0 & \frac{\partial}{\partial z} \\ \frac{\partial}{\partial z} & 0 \end{pmatrix}\begin{pmatrix} v \\ \sigma \end{pmatrix}. \tag{2.10}$$

Finally, the interaction of the vibrating string through its boundary is expressed by the definition of the boundary port variables, i.e., the velocity and stress at the boundaries of the string

$$\begin{pmatrix} w_K \\ w_U \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}\begin{pmatrix} v|_{a,b} \\ \sigma|_{a,b} \end{pmatrix}. \tag{2.11}$$

The canonical interdomain coupling equation (2.10) and the boundary coupling equation (2.11) actually define a Dirac structure [7, Chapter 5] called the *Stokes–Dirac structure*. Let us explain this in more detail. Consider the Hilbert spaces of the flow variables $\mathcal{F} = L^2([a,b],\mathbb{R}) \times L^2([a,b],\mathbb{R}) \times \mathbb{R}^2 \ni (f_K, f_U, w_K)$ and of the effort variables $\mathcal{E} = L^2([a,b],\mathbb{R}) \times L^2([a,b],\mathbb{R}) \times \mathbb{R}^2 \ni (v,\sigma,w_U)$. Furthermore, endow the bond space $\mathcal{B} = \mathcal{F}\times\mathcal{E}$ with the following pairing:

$$\begin{aligned}
&\left\langle \left(f_K^1, f_U^1, w_K^1, v^1, \sigma^1, w_U^1\right),\ \left(f_K^2, f_U^2, w_K^2, v^2, \sigma^2, w_U^2\right)\right\rangle_+ \\
&\quad = \int_a^b f_K^1\, v^2\, dz + \int_a^b f_K^2\, v^1\, dz \\
&\qquad + \int_a^b f_U^1\, \sigma^2\, dz + \int_a^b f_U^2\, \sigma^1\, dz + {w_K^1}^T\, \Lambda w_U^2 + {w_K^2}^T\, \Lambda w_U^1,
\end{aligned}$$

where $\Lambda = \left(\begin{smallmatrix} I & 0 \\ 0 & -I \end{smallmatrix}\right)$.

This pairing on the bond space corresponds to the general definition given in (2.1) where both the flow and the effort vector space is a product space given by $\mathcal{F} = \mathcal{F}_{(a,b)} \times \mathcal{F}_\partial$ and $\mathcal{E} = \mathcal{E}_{(a,b)} \times \mathcal{E}_\partial$, respectively. The subspace of flow variables defined on the domain $[a,b]$ is $\mathcal{F}_{(a,b)} = L^2([a,b],\mathbb{R}) \times L^2([a,b],\mathbb{R}) \ni (f_K, f_U)$ and

the conjugated subspace of variables is $\mathcal{E}_{(a,b)} = L^2([a,b],\mathbb{R}) \times L^2([a,b],\mathbb{R}) \ni (v,\sigma)$. These Hilbert spaces are equal and hence the isometry $r_{\mathcal{F}_{(a,b)},\mathcal{E}_{(a,b)}}$ is the identity. On the contrary, for the pairing on the boundary port variables, the matrix Λ actually corresponds to the definition of an isometry $r_{\mathcal{F}_\partial,\mathcal{E}_\partial}$ between the boundary port spaces $\mathcal{F}_\partial = \mathbb{R}^2 \ni w_K$ and $\mathcal{E}_\partial = \mathbb{R}^2 \ni w_U$ endowed with the canonical Euclidean metric.

It has been shown in [7, 8] that (2.10) and (2.11) define a Dirac structure, namely, the Stokes–Dirac structure on $\mathcal{B}$ associated with the differential operator given in (2.10). We shall denote this Dirac structure by $\mathcal{D}_1$.

The system of two conservation laws (2.10), with the closure equations (2.5), (2.8), and (2.9), may be rewritten as the following Hamiltonian system [20]:

$$\frac{\partial}{\partial t}\begin{pmatrix} p \\ \epsilon \end{pmatrix} = \begin{pmatrix} 0 & \frac{\partial}{\partial z} \\ \frac{\partial}{\partial z} & 0 \end{pmatrix}\begin{pmatrix} \delta_p \mathcal{H} \\ \delta_\epsilon \mathcal{H} \end{pmatrix}, \tag{2.12}$$

where $\mathcal{H} = U + K$ denotes the Hamiltonian function corresponding to the total energy of the system and $\delta_p \mathcal{H}(x) = v$, $\delta_\epsilon \mathcal{H}(x) = \sigma$ denote the *variational derivatives* [20] of $\mathcal{H}$ with respect to the momentum p and the strain ϵ, respectively. This system is indeed a Hamiltonian system [20] if the differential operator in (2.12) is skew-symmetric, i.e., if the boundary variables are such that there is *no energy flow at the boundary* of the system:

$$w_K^{1\,T}\,\Lambda w_U^2 + w_K^{2\,T}\,\Lambda w_U^1 = 0. \tag{2.13}$$

In order to account for some energy flow at the boundary, the evolution equation (2.12) may be completed using the *port boundary variables* defined in (2.11), i.e., the velocity and the strain at the boundary

$$\begin{pmatrix} w_K \\ w_U \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}\begin{pmatrix} \delta_p \mathcal{H}\,|_{a,b} \\ \delta_\epsilon \mathcal{H}\,|_{a,b} \end{pmatrix}. \tag{2.14}$$

The system composed of (2.12) and (2.14) defines a port Hamiltonian system with respect to the Stokes–Dirac structure. This port Hamiltonian system is generated by the Hamiltonian $\mathcal{H}$ [7, 8, 30] and it may be written in the following implicit way:

$$\left(\frac{\partial p}{\partial t}, \frac{\partial \epsilon}{\partial t}, w_K, \delta_p \mathcal{H}, \delta_\epsilon \mathcal{H}, w_U\right) \in \mathcal{D}_1. \tag{2.15}$$

Let us briefly compare the port Hamiltonian formulation with the formulation as a PDE. The evolution equations (2.12), with the closure equations (2.9), (2.5), may also be written in the form of the wave equations (in terms of the displacement of the string):

$$\mu \frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial z}\left(T \frac{\partial u}{\partial z}\right). \tag{2.16}$$

The relation between the boundary conditions of this PDE and the port variables is given by

$$\begin{pmatrix} w_K \\ w_U \end{pmatrix} = \begin{pmatrix} v|_{a,b} \\ \sigma|_{a,b} \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial t} u|_{a,b} \\ T(z)\,\frac{\partial u}{\partial z}|_{a,b} \end{pmatrix}. \tag{2.17}$$

This shows clearly that the PDE (2.16) does not reflect the physical structure of the system in the sense that it is not written as a system of conservation laws and

that the total energy appears clearly. Relation (2.17) shows the difference between port variables and boundary control systems in terms of physical elementary interface variables. One may not express the static equilibrium $\sigma_{|a,b}$(stress) in terms of $\frac{\partial u}{\partial z}$ without knowing $T(z)$.

Finally, we compare very briefly the port Hamiltonian formulation with a classical symplectic Hamiltonian formulation (see also [17]). Using the displacement $u(z,t)$ and the velocity $v(z,t) = \frac{\partial u}{\partial t}$ as state variables, one obtains the following infinite-dimensional Hamiltonian system (with energy flows being zero at the boundary):

$$\frac{\partial x}{\partial t} = \begin{pmatrix} 0 & \frac{1}{\mu} \\ -\frac{1}{\mu} & 0 \end{pmatrix} \begin{pmatrix} \frac{\delta H}{\delta u} \\ \frac{\delta H}{\delta v} \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{\mu} \\ -\frac{1}{\mu} & 0 \end{pmatrix} \begin{pmatrix} -\frac{\partial}{\partial z}\left(T\ \frac{\partial u}{\partial z}\right) \\ \mu v \end{pmatrix}, \tag{2.18}$$

where the Hamiltonian function is

$$H(u,v) = \int_a^b \left(\frac{1}{2}\ T\ \left(\frac{\partial u}{\partial z}\right)^2 + \frac{1}{2}\ \mu v^2 \right) dz. \tag{2.19}$$

Contrary to the port Hamiltonian formulation, this formulation does not make the physical structure of conservation laws appear. One may furthermore note that the Hamiltonian system is defined with respect to a *symplectic* Poisson bracket. This bracket is not canonical (it depends on the mass distribution of the string) and cannot be extended in a canonical way to a Dirac structure including boundary variables.

This example has shown that the Stokes–Dirac structure $\mathcal{D}_1$, associated with the canonical interdomain coupling, is derived from a skew-symmetric differential operator of order one. In section 3, we consider a generalization of this differential operator by considering skew-symmetric operators of any order and we derive Dirac structures on Hilbert spaces from them. In Example 2.2, we have also seen how the dynamics can be defined by using the canonical Dirac structure and the Hamiltonian; namely, the dynamics lives on the Dirac structure $\mathcal{D}_1$ and the total energy is defined by a Hamiltonian function. In section 4, we consider energy functions which are equal to the norm of the Hilbert space. Hence there the coenergy variables and the state variables are identical. We show how to parameterize the contractive semigroups associated with the Dirac structures defined in section 3. In section 5, finally, we distinguish between the state and the coenergy variables by introducing more general Hamiltonian functions and define port Hamiltonian systems associated with skew-symmetric differential operators of any order.

3. Dirac structure associated with a skew-symmetric operator. In this section, we extend the definition of Stokes–Dirac structures to skew-symmetric differential operators of any order. Therefore, we first recall how one may extend the Stokes theorem to such operators and how the Stokes theorem induces a symmetric pairing on the boundary variables. Second, we define boundary port variables as a linear combination of the boundary variables associated with the differential operator. Using these boundary port variables, we define a bond space and a Dirac structure associated with the differential operator.

Consider the differential operator $\mathcal{J}$ of order N

$$\mathcal{J}e = \sum_{i=0}^{N} P(i) \frac{d^i e}{dz^i}(z), \quad z \in [a,b], \tag{3.1}$$

where $e \in C^\infty((a,b);\mathbb{R}^n)$ and $P(i)$, $i = 0,\dots,N$, is an $n \times n$ real matrix. The formal adjoint $\mathcal{J}^*$ of $\mathcal{J}$ is given by

$$\mathcal{J}^* e = \sum_{i=0}^{N} P(i)^T (-1)^i \frac{d^i e}{dz^i}(z), \quad z \in [a,b].$$

Now assume that $\mathcal{J}$ is skew-symmetric, i.e., $\mathcal{J} = -\mathcal{J}^*$. From the above expression of $\mathcal{J}^*$ we see that this is equivalent to

$$P(i) = P(i)^T (-1)^{i+1}. \tag{3.2}$$

Using this property, we show that the bilinear symmetric pairing of e and $\mathcal{J}e$ depends only on the boundary values. Thus if the boundary values are zero, then $\langle e_1, \mathcal{J}e_2\rangle + \langle e_2, \mathcal{J}e_1\rangle = 0$, which corresponds to the fact that $\mathcal{J}$ is (formally) skew-symmetric.

THEOREM 3.1. *Let $\mathcal{J}$ be a skew-symmetric operator defined by* (3.1)*, and let $H^N((a,b);\mathbb{R}^n)$ denote the Sobolev space of N times differentiable functions on the interval (a,b). Then for any two functions $e_i \in H^N((a,b);\mathbb{R}^n)$, $i \in \{1,2\}$, we have that*

$$\begin{aligned} &\int_a^b e_1^T(z)(\mathcal{J}e_2)(z) + e_2^T(z)(\mathcal{J}e_1)(z)dz \\ &\quad = \left[\left(e_1^T(z), \dots, \frac{d^{N-1}e_1^T}{dz^{N-1}}(z) \right) Q \begin{pmatrix} e_2(z) \\ \vdots \\ \frac{d^{N-1}e_2}{dz^{N-1}}(z) \end{pmatrix} \right]_a^b, \end{aligned} \tag{3.3}$$

where

$$Q = (Q_{ij}), \qquad i,j = 1,\dots,N,$$

with

$$Q_{ij} = \begin{cases} 0, & i+j > N+1, \\ P(k)(-1)^{i-1}, & i+j-1 = k. \end{cases} \tag{3.4}$$

Furthermore, Q is a symmetric matrix.

Proof. The result can easily be derived from iterative integration by parts; see [10] for details. □

The above theorem shows that any skew-symmetric differential operator $\mathcal{J}$ gives rise to a symmetric bilinear product on the space of boundary conditions $e(a),\dots,\frac{d^{N-1}e}{dz^{N-1}}(a)$, $e(b),\dots,\frac{d^{N-1}e}{dz^{N-1}}(b)$. The coefficients of this symmetric product, captured in the matrix Q, are uniquely defined by the coefficients of the skew-symmetric differential operator $\mathcal{J}$. In what follows, we shall define port boundary variables and a bond space in such a way that the Stokes theorem applied to the differential operator may be expressed using the canonical symmetric pairing defined in (2.1). Therefore, let us focus, in a first step, on the properties of Q and define the matrix R_{ext} which is used for defining the port variables. First of all, note that Q has the following form:

$$Q = \begin{pmatrix} P(1) & P(2) & P(3) & \cdots & P(N-1) & P(N) \\ -P(2) & -P(3) & -P(4) & \cdots & -P(N) & 0 \\ P(3) & P(4) & \iddots & \iddots & 0 & 0 \\ -P(4) & \iddots & \iddots & \iddots & & \vdots \\ \vdots & \iddots & & & & \vdots \\ (-1)^{N-1}P(N) & 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix}. \tag{3.5}$$

From the form of Q, the proof of the following lemma is immediate.

LEMMA 3.2. *The matrix Q introduced in Theorem* 3.1 *is symmetric and*

$$\ker Q = \{0\}$$

if and only if $\ker P(N) = \{0\}$.

From now on we assume that Q is nonsingular.

DEFINITION 3.3. *The matrix Q_{ext} in $\mathbb{R}^{2nN\times 2nN}$ associated with the differential operator $\mathcal{J}$ is defined by*

$$Q_{\text{ext}} = \begin{pmatrix} Q & 0 \\ 0 & -Q \end{pmatrix}. \tag{3.6}$$

Looking at (3.1), one can easily see that it is necessary to proceed to an appropriate change of variables to make this relation equivalent to the desired canonical symmetrical pairing defined in (2.1). This change of variables is done using the matrix R_{ext} detailed in Lemma 3.4.

LEMMA 3.4. *The matrix R_{ext} defined as*

$$R_{\text{ext}} = \frac{1}{\sqrt{2}} \begin{pmatrix} Q & -Q \\ I & I \end{pmatrix} \tag{3.7}$$

is invertible and satisfies

$$\begin{pmatrix} Q & 0 \\ 0 & -Q \end{pmatrix} = R_{\text{ext}}^T \Sigma R_{\text{ext}}, \tag{3.8}$$

where

$$\Sigma = \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix}. \tag{3.9}$$

All possible matrices R which satisfy (3.8) *are given by the formula*

$$R = U R_{\text{ext}}$$

with U satisfying $U^T \Sigma U = \Sigma$.

Proof. We have that

$$\frac{1}{\sqrt{2}} \begin{pmatrix} Q & I \\ -Q & I \end{pmatrix} \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix} \begin{pmatrix} Q & -Q \\ I & I \end{pmatrix} \frac{1}{\sqrt{2}} = \begin{pmatrix} Q & 0 \\ 0 & -Q \end{pmatrix}.$$

Thus using the fact that Q is symmetric $R_{\text{ext}} := \frac{1}{\sqrt{2}} \left(\begin{smallmatrix} Q & -Q \\ I & I \end{smallmatrix}\right)$ satisfies (3.8). Since Q is invertible, the invertibility of R_{ext} follows from (3.8).

Let W be another solution of (3.8). Hence

$$W^T \Sigma W = \begin{pmatrix} Q & 0 \\ 0 & -Q \end{pmatrix} = R_{\text{ext}}^T \Sigma R_{\text{ext}}.$$

This can be written in the equivalent form

$$R_{\text{ext}}^{-T} W^T \Sigma W R_{\text{ext}}^{-1} = \Sigma.$$

Calling $W R_{\text{ext}}^{-1} = U$, we have that $U^T \Sigma U = \Sigma$ and $W = U R_{\text{ext}}$, which proves the assertion. ∎

The crucial step in defining the Dirac structure associated with the operator $\mathcal{J}$ is to define the boundary port variables. These are the following linear combinations of the boundary conditions.

DEFINITION 3.5. *The boundary port variables associated with the differential operator $\mathcal{J}$ are the vectors $e_\partial, f_\partial \in \mathbb{R}^{nN}$ defined by*

$$\begin{pmatrix} f_\partial \\ e_\partial \end{pmatrix} = R_{\text{ext}} \begin{pmatrix} e(b) \\ \vdots \\ \frac{d^{N-1}e}{dz^{N-1}}(b) \\ e(a) \\ \vdots \\ \frac{d^{N-1}e}{dz^{N-1}}(a) \end{pmatrix}, \tag{3.10}$$

where R_{ext} is defined by (3.7).

Consider the effort and flow spaces $\mathcal{E} = \mathcal{F} = L^2((a,b);\mathbb{R}^n) \times \mathbb{R}^{nN}$ with their natural inner product. We define the bond space $\mathcal{B}$ as $\mathcal{F} \times \mathcal{E}$ with the canonical symmetrical pairing

$$\begin{aligned} &\left\langle \left(f^1, f^1_\partial, e^1, e^1_\partial\right), \left(f^2, f^2_\partial, e^2, e^2_\partial\right) \right\rangle_+ \\ &\quad = \langle e^1, f^2 \rangle_{L^2} + \langle e^2, f^1 \rangle_{L^2} - \langle e^1_\partial, f^2_\partial \rangle - \langle e^2_\partial, f^1_\partial \rangle, \end{aligned} \tag{3.11}$$

where

$$(f^i, f^i_\partial, e^i, e^i_\partial) \in \mathcal{B}, \qquad i = \{1, 2\}.$$

Let us emphasize that this pairing on the bond space corresponds to the general definition given in (2.1), where the pairing on the bond space is defined modulo an isometry $r_{\mathcal{F},\mathcal{E}}$. The space of flow variables is the product space $\mathcal{F} = L^2((a,b);\mathbb{R}^n) \times \mathbb{R}^N$. Thus every flow element is a pair with the top element a function, and the bottom element is a part of the (boundary) port variable. The same description holds for the space of effort variables. The spaces $\mathcal{F}$ and $\mathcal{E}$ are equal and the natural isometry would be the identity. However, we choose

$$r_{\mathcal{F},\mathcal{E}} = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}.$$

It is easy to see that this is an isometry, which is equal to its own inverse. Furthermore, with this choice (2.1) equals (3.11).

On the bond space $\mathcal{B}$ with the symmetrical pairing (3.11) we define the Dirac structure, $\mathcal{D}_\mathcal{J}$, associated with the linear skew symmetric operator $\mathcal{J}$.

This Dirac structure is nothing else but the expression of the Stokes theorem (recalled in Theorem 3.1) with respect to the port variables defined in Definition 3.5.

THEOREM 3.6. *Let $H^N((a,b);\mathbb{R}^n)$ denote the Sobolev space of N times differentiable functions on the interval (a,b). The subspace $\mathcal{D}_\mathcal{J}$ of $\mathcal{B}$ defined as*

$$\mathcal{D}_\mathcal{J} = \left\{ \begin{pmatrix} f \\ f_\partial \\ e \\ e_\partial \end{pmatrix} \;\middle|\; e \in H^N((a,b);\mathbb{R}^n), \mathcal{J}e = f, \begin{pmatrix} f_\partial \\ e_\partial \end{pmatrix} = R_{\text{ext}} \begin{pmatrix} e(b) \\ \vdots \\ \frac{d^{N-1}e}{dz^{N-1}}(b) \\ e(a) \\ \vdots \\ \frac{d^{N-1}e}{dz^{N-1}}(a) \end{pmatrix} \right\} \tag{3.12}$$

is a Dirac structure.

Proof. The Dirac structure is defined by the fact that $\mathcal{D}_{\mathcal{J}} = \mathcal{D}_{\mathcal{J}}^{\perp}$.

Step 1. Recall that $\mathcal{D}_{\mathcal{J}} \subset \mathcal{D}_{\mathcal{J}}^{\perp}$ is equivalent to the canonical product $\langle b, b\rangle_+$ being zero for all $b \in \mathcal{D}_{\mathcal{J}}$. From (3.11) we have that

$$\begin{aligned}
&\langle (f, f_\partial, e, e_\partial,), (f, f_\partial, e, e_\partial)\rangle_+ \\
&\quad = \langle e, \mathcal{J}e\rangle_{L^2} + \langle e, \mathcal{J}e\rangle_{L^2} - e_\partial^T f_\partial - e_\partial^T f_\partial \\
&\quad = \left[\left(e^T(z), \dots, \frac{d^{N-1}e^T}{dz^{N-1}}(z) \right) Q \begin{pmatrix} e(z) \\ \vdots \\ \frac{d^{N-1}e}{dz^{N-1}}(z) \end{pmatrix} \right]_a^b - 2e_\partial^T f_\partial \\
&\quad = \left(e^T(b), \dots, \frac{d^{N-1}e^T}{dz^{N-1}}(a) \right) \begin{pmatrix} Q & 0 \\ 0 & -Q \end{pmatrix} \begin{pmatrix} e(b) \\ \vdots \\ \frac{d^{N-1}e}{dz^{N-1}}(a) \end{pmatrix} - 2e_\partial^T f_\partial \\
&\quad = \left(f_\partial^T, e_\partial^T \right) \Sigma \begin{pmatrix} f_\partial \\ e_\partial \end{pmatrix} - 2e_\partial^T f_\partial = 0,
\end{aligned}$$

where we have used Theorem 3.1 and (3.8).

Step 2. Let $(\phi, \phi_\partial, \varepsilon, \varepsilon_\partial) \in \mathcal{D}_{\mathcal{J}}^{\perp}$. Choose $e \in H^N((a,b); \mathbb{R}^n)$ with compact support strictly included in (a,b). Thus $\frac{d^k e}{dz^k}$, $k \in \{0, \dots, N-1\}$, are zero in a and b. Then it is easy to see that $(\mathcal{J}e, 0, e, 0) \in \mathcal{D}_{\mathcal{J}}$. Using (3.11) we have

$$0 = \langle e, \phi\rangle + \langle \varepsilon, f\rangle = \langle e, \phi\rangle + \langle \varepsilon, \mathcal{J}e\rangle$$

for all such e. This implies that $\varepsilon \in H^N((a,b); \mathbb{R}^n)$ and $\mathcal{J}\varepsilon = \phi$.

Step 3. Let $(\phi, \phi_\partial, \varepsilon, \varepsilon_\partial) \in \mathcal{D}_{\mathcal{J}}^{\perp}$ and let $(f, f_\partial, e, e_\partial) \in \mathcal{D}_{\mathcal{J}}$. From Step 2 and (3.11) we obtain

$$\begin{aligned}
0 &= \langle e, \mathcal{J}\varepsilon\rangle + \langle \varepsilon, \mathcal{J}e\rangle - e_\partial^T \phi_\partial - \varepsilon_\partial^T f_\partial \\
&= \left[\left(e^T(z), \dots, \frac{d^{N-1}e^T(z)}{dz^{N-1}}(z) \right) Q \begin{pmatrix} \varepsilon(z) \\ \vdots \\ \frac{d^{N-1}\varepsilon(z)}{dz^{N-1}}(z) \end{pmatrix} \right]_a^b - e_\partial^T \phi_\partial - \varepsilon_\partial^T f_\partial \\
&= \left(f_\partial^T, e_\partial^T \right) \Sigma R_{\text{ext}} \begin{pmatrix} \varepsilon(b) \\ \vdots \\ \frac{d^{N-1}\varepsilon}{dz^{N-1}}(a) \end{pmatrix} - e_\partial^T \phi_\partial - \varepsilon_\partial^T f_\partial \\
&= \left(e_\partial^T, f_\partial^T \right) \left[R_{\text{ext}} \begin{pmatrix} \varepsilon(b) \\ \vdots \\ \frac{d^{N-1}\varepsilon}{dz^{N-1}}(a) \end{pmatrix} - \begin{pmatrix} \phi_\partial \\ \varepsilon_\partial \end{pmatrix} \right].
\end{aligned}$$

By a proper choice of e, we can let the vectors e_∂ and f_∂ have arbitrary values. Thus the above equality has to hold for all $e_\partial \in \mathbb{R}^{nN}$ and $f_\partial \in \mathbb{R}^{nN}$. Consequently, we have that

$$R_{\text{ext}} \begin{pmatrix} \varepsilon(b) \\ \vdots \\ \frac{d^{N-1}\varepsilon}{dz^{N-1}}(a) \end{pmatrix} = \begin{pmatrix} \phi_\partial \\ \varepsilon_\partial \end{pmatrix}.$$

In conclusion, we have that $\mathcal{D}_{\mathcal{J}} = \mathcal{D}_{\mathcal{J}}^{\perp}$, and so $\mathcal{D}_{\mathcal{J}}$ is a Dirac structure. ∎

4. Contraction semigroups, boundary control systems, and their parameterization. In the previous section we have associated with the skew-symmetric operator $\mathcal{J}$ a Dirac structure $\mathcal{D}_{\mathcal{J}}$. In this section, we shall define dynamic systems with inputs, states, and outputs with respect to this Dirac structure. These systems will be boundary control systems in the sense of the semigroup theory [3], which implies that the controls and observations act on the boundary of the spatial domain. With respect to the Dirac structure $\mathcal{D}_{\mathcal{J}}$ it is possible to define many systems. However, we consider only those systems for which the energy does not grow when the input is zero. This implies that the associated semigroup is contractive. We parameterize all these systems by nN-dimensional linear subspaces of the port variables. As a consequence of this parameterization, we identify those systems for which the associated semigroup is unitary.

We begin by showing that $\mathcal{J}$ is the infinitesimal generator of a contraction semigroup for appropriate choices of the boundary conditions.

4.1. Contraction semigroups associated with $\mathcal{D}_{\mathcal{J}}$. We begin by studying the differential operator $\mathcal{J}$ for different boundary conditions. As stated above, we want to characterize those boundary conditions for which the associated differential operator is the infinitesimal generator of a strongly continuous semigroup. Furthermore, this semigroup must be contractive, i.e., $\|T(t)e\| \leq \|e\|$ for all $t \geq 0$ and $e \in L^2((a,b);\mathbb{R}^n)$. We obtain this characterization of the boundary conditions by using Theorem 3.1.6 of [9]. For the history of this result, we refer to the bibliographical comments at the end of [9].

Before stating this result, we recall the following parameterization. Let $\Sigma = \left(\begin{smallmatrix} 0 & I \\ I & 0 \end{smallmatrix}\right)$. A full rank matrix W of size $nN \times 2nN$ satisfies

$$W\Sigma W^T \geq 0 \tag{4.1}$$

if and only if

$$W = S\begin{pmatrix} I+V & I-V \end{pmatrix} \tag{4.2}$$

with S an invertible matrix, and V satisfying $VV^T \leq I$; see the appendix for a proof.

In the following theorem it is shown that if the port variables are restricted to the kernel of W, then this defines the domain of a contraction semigroup associated with the operator $\mathcal{J}$.

THEOREM 4.1. *Let W be a full rank matrix of size $k \times 2nN$. Define the operator J_W and its domain, $D(J_W)$, as*

$$J_W e = \mathcal{J} e \tag{4.3}$$

and

$$\begin{aligned} D(J_W) = \Bigg\{ & e \in L^2((a,b),\mathbb{R}^n) \mid \textit{the port variable associated with } e, \\ & \begin{pmatrix} f_\partial \\ e_\partial \end{pmatrix}, \textit{ is in } \ker W \textit{ and there exists} \\ & \textit{an } f \in L^2((a,b);\mathbb{R}^n) \textit{ such that } (f, f_\partial, e, e_\partial) \in \mathcal{D}_{\mathcal{J}} \Bigg\}. \end{aligned} \tag{4.4}$$

Then J_W generates a contraction semigroup $(T(t))_{t\geq 0}$ on $L^2((a,b);\mathbb{R}^n)$ if and only if $k = nN$ and (4.1) holds.

Furthermore, J_W is the infinitesimal generator of a unitary semigroup on $L^2((a,b);\mathbb{R}^n)$ if and only if $k = nN$ and $W\Sigma W^T = 0$.

Proof. It is well known that operator A is the infinitesimal generator of a contraction semigroup if and only if it is maximally dissipative, i.e., it is dissipative:

$$\mathrm{Re}\langle Az, z\rangle \leq 0$$

for all $z \in D(A)$, and it is not a proper restriction of any other dissipative operator [24].

In [9] a characterization of maximal dissipative differential operators in terms of their boundary conditions is given. However, their formulation is not precisely the one we are using. Hence in the next step we relate their notation with ours. Once that is done, the proof of the theorem is straightforward.

Step 1. Define on $D(\mathcal{J}) = H^N((a,b),\mathbb{C}^n)$ the operator

$$\mathcal{A}_0^* = i\mathcal{J}.$$

Then it is easy to see that $\mathcal{A}_0 = i\mathcal{J}$ with $D(\mathcal{A}_0) = \{e \in H^N((a,b),\mathbb{C}^n) \mid \frac{d^{N-1}e}{dz^{N-1}}(b) = \cdots e(b) = 0$ and $\frac{d^{N-1}e}{dz^{N-1}}(a) = \cdots e(a) = 0\}$.

Using Theorem 3.1 and Definition 3.6 we have for all $e_1, e_2 \in D(\mathcal{J})$

$$\begin{aligned}
\langle \mathcal{A}_0^* e_1, e_2\rangle - \langle e_1, \mathcal{A}_0^* e_2\rangle &= i\langle \mathcal{J}e_1, e_2\rangle + i\langle e_1, \mathcal{J}e_2\rangle \\
&= i\left\langle \begin{pmatrix} e_1(b) \\ \vdots \\ \frac{d^{N-1}e_1}{dz^{N-1}}(a) \end{pmatrix}, Q_{\mathrm{ext}} \begin{pmatrix} e_2(b) \\ \vdots \\ \frac{d^{N-1}e_2}{dz^{N-1}}(a) \end{pmatrix} \right\rangle_{\mathbb{C}^{2nN}} \\
&= i\left\langle \begin{pmatrix} f_{1,\partial} \\ e_{1,\partial} \end{pmatrix}, \Sigma \begin{pmatrix} f_{2,\partial} \\ e_{2,\partial} \end{pmatrix} \right\rangle_{\mathbb{C}^{2nN}} \\
&= i\left[\langle f_{1,\partial}, e_{2,\partial}\rangle_{\mathbb{C}^{nN}} + \langle e_{1,\partial}, f_{2,\partial}\rangle_{\mathbb{C}^{nN}}\right],
\end{aligned}$$

where we have used (3.10). Define operators Γ_1 and Γ_2 from $H^N((a,b),\mathbb{C}^n)$ to $\mathbb{C}^{nN}$ as

$$\Gamma_1 e = if_\partial, \quad \Gamma_2 e = e_\partial. \tag{4.5}$$

It is clear that these mappings are onto. Furthermore, we have that

$$\langle \mathcal{A}_0^* e_1, e_2\rangle - \langle e_1, \mathcal{A}_0^* e_2\rangle = \langle \Gamma_1 e_1, \Gamma_2 e_2\rangle_{\mathbb{C}^{nN}} - \langle \Gamma_2 e_1, \Gamma_1 e_2\rangle_{\mathbb{C}^{nN}}. \tag{4.6}$$

Step 2. Using (4.6), Theorem 3.1.6 of [9] characterizes all maximally accumulative extensions of $\mathcal{A}_0$. An operator A is defined to be accumulative if $\mathrm{Im}\langle Az, z\rangle \leq 0$ for all $z \in D(A)$. It is maximally accumulative, when it has no nontrivial accumulative extension. It is easy to see that A is a maximal accumulative extension of $\mathcal{A}_0$ if and only if $-iA$ is a maximal dissipative extension of $\mathcal{J}$ with the domain $D(\mathcal{A}_0)$.

Step 3. Theorem 3.1.6 of [9] states that any maximally accumulative extension of $\mathcal{A}_0$ is given by

$$A_K = \mathcal{A}_0^*$$

with

$$D(A_K) = \{e \in D(\mathcal{A}_0^*) \mid (K-I)\Gamma_1 e - i(K+I)\Gamma_2 e = 0\},$$

where $K : \mathbb{C}^{nN} \mapsto \mathbb{C}^{nN}$ with $\|K\| \leq 1$. Using the result obtained in the previous step and (4.5), we see that any maximal dissipative extension of $\mathcal{J}, D(\mathcal{A}_0)$, is given by

$$-iA_K = -i\mathcal{A}_0^* = \mathcal{J}$$

with the domain

$$\begin{aligned} D(A_K) &= \{e \in D(\mathcal{A}_0^*) \mid (K-I)\Gamma_1 e - i(K+I)\Gamma_2 e = 0\} \\ &= \{e \in H^N((a,b),\mathbb{C}^n) \mid (K-I)f_\partial - (K+I)e_\partial = 0\}, \end{aligned}$$

where $K : \mathbb{C}^{nN} \mapsto \mathbb{C}^{nN}$ with $\|K\| \leq 1$. Using Lemma A.1 we see that $-iA_K = J_W$ with $W = S(I-K, I+K)$, where S is an arbitrary, invertible matrix. Since we are interested only in real conditions, we have to take K real valued. Hence we have obtained a complete characterization of all boundary conditions for which the differential operator is the infinitesimal generator of a contraction semigroup.

Step 4. The proof of the unitary case is done very similarly. ∎

One may wonder why we have parameterized the boundary port variables using the W instead of the V, since if two W's have the same V, then the associated semigroups are the same. In the following subsection, the boundary variables are decomposed into inputs and outputs. For this splitting the W is important. More specifically, different W's lead to different systems, although the semigroup may be the same.

4.2. Boundary control system and port conjugated output. In the previous subsection we have derived the family of contraction semigroups from the Dirac structure $\mathcal{D}_\mathcal{J}$ associated with a skew-symmetric differential operator $\mathcal{J}$. More precisely, we have parameterized these semigroups by a family of subspaces of the port boundary variables defined as the kernel of a class of matrices W (matrices of size $nN \times 2nN$ satisfying (4.1)). In the following theorem, we use this W to define boundary inputs/controls. Since the rank of W is nN and since we have $2nN$ boundary variables, we see that we use half of the set of boundary variables to define inputs. We show that the other half may be regarded as outputs. Note that the terms input and output are used here to make the relation with infinite-dimensional systems theory. It does not necessarily mean that the input is completely free, i.e., it can be chosen arbitrarily in $L^2_{\text{loc}}((0,\infty);\mathbb{R}^{nN})$, nor does it imply that for every initial condition in $L^2((a,b);\mathbb{R}^n)$ the output is well defined. The system class which is considered is the class of boundary control systems. For more information on this class, we refer the reader to section 3.3 of [3].

Using the splitting of the boundary ports into inputs and outputs, we consider in Theorems 4.4 and 4.6 some special choices which lead to classical power balance equations in the so-called impedance and scattering variables form.

THEOREM 4.2. *For the differential operator $\mathcal{J}$ and the associated Dirac structure $\mathcal{D}_\mathcal{J}$ (see Theorem* 3.6*), we consider the dynamical system*

$$(\dot{x}(t), f_\partial(t), x(t), e_\partial(t)) \in \mathcal{D}_\mathcal{J}, \quad t \geq 0, \tag{4.7}$$

where $(f_\partial(t), e_\partial(t))$ are the boundary port variables associated with $x(t)$; see Definition 3.5*.*

Let W be a full rank matrix of size $nN \times 2nN$ satisfying (4.1)*, and define $\mathcal{B} : H^N((a,b),\mathbb{R}^n) \to \mathbb{R}^{nN}$ as*

$$\mathcal{B}x(t) := W \begin{pmatrix} f_\partial(t) \\ e_\partial(t) \end{pmatrix}. \tag{4.8}$$

Then system (4.7) *with the input defined as*

$$u(t) = \mathcal{B}x(t) \tag{4.9}$$

is a boundary control system.

Furthermore, let $\tilde{W}$ *be a full rank matrix of size* $nN \times 2nN$ *with* $\binom{W}{\tilde{W}}$ *invertible. If we define the linear mapping* $\mathcal{C} : H^N((a,b),\mathbb{R}^n) \to \mathbb{R}^{nN}$ *as*

$$\mathcal{C}x(t) := \tilde{W}\begin{pmatrix} f_\partial(t) \\ e_\partial(t) \end{pmatrix} \tag{4.10}$$

and the output as

$$y(t) = \mathcal{C}x(t), \tag{4.11}$$

then for $u \in C^2((0,\infty);\mathbb{R}^{nN})$, $x(0) \in H^N((a,b),\mathbb{R}^n)$, *and* $\mathcal{B}x(0) = u(0)$, *the following balance equation is satisfied:*

$$\frac{1}{2}\frac{d}{dt}\|x(t)\|^2 = \frac{1}{2}\begin{pmatrix} u^T(t) y^T(t) \end{pmatrix} P_{W,\tilde{W}} \begin{pmatrix} u(t) \\ y(t) \end{pmatrix}, \tag{4.12}$$

where

$$P^{-1}_{W,\tilde{W}} = \begin{pmatrix} W\Sigma W^T & W\Sigma \tilde{W}^T \\ \tilde{W}\Sigma W^T & \tilde{W}\Sigma \tilde{W}^T \end{pmatrix}. \tag{4.13}$$

Furthermore, we have that the matrix $\left(\begin{smallmatrix} W\Sigma W^T & W\Sigma \tilde{W}^T \\ \tilde{W}\Sigma W^T & \tilde{W}\Sigma \tilde{W}^T \end{smallmatrix}\right)$ *is invertible if and only if* $\binom{W}{\tilde{W}}$ *is invertible.*

REMARK 4.3. *The system defined by* (4.7)–(4.9) *may be equivalently written in the more usual form of a boundary control system:*

$$\begin{aligned} \dot{x}(t) &= \mathcal{J}x(t), \\ \mathcal{B}x(t) &= u(t). \end{aligned} \tag{4.14}$$

Proof of Theorem 4.2. In Steps 1 and 2 we show that we have a boundary control system. In Steps 3 and 4, we prove (4.12) and (4.13), respectively. For a boundary control system we have to show that for zero inputs, the system is a C_0-semigroup, and furthermore that there exists a bounded operator B mapping onto the domain of $\mathcal{B}$ and such that $\mathcal{B}Bu = u$ for all $u \in \mathbb{R}^{nN}$.

Step 1. As mentioned above, we have to show that J_W defined as

$$J_W x = \mathcal{J}x$$

on

$$D(J_W) = D(\mathcal{J}) \cap \ker \mathcal{B}$$

is an infinitesimal generator. This follows directly from Theorem 4.1.

Step 2. We have to find a bounded linear operator B such that $Bu \in D(\mathcal{B}) = H^N((a,b);\mathbb{R}^{nN})$ and $\mathcal{B}Bu = u$ for all $u \in \mathbb{R}^{nN}$.

Let $\{u^1,\dots,u^{nN}\}$ be the standard basis of the input space $\mathbb{R}^{nN}$, i.e., $u^i = (\delta_{ij})^T_{j=1,\dots,nN}$. Since R_{ext} is invertible, and since W has rank nN there exists for every u^i a $v^i \in \mathbb{R}^{2nN}$ such that

$$WR_{\text{ext}}v^i = u^i. \tag{4.15}$$

Let v_k^i denote the kth block of v^i, $k = 1, \ldots, 2N$. Using functions $f_{r,j}$ and $f_{l,j}$ introduced in Lemma A.3, we define the ith column of B as

$$B_i = \sum_{k=1}^{N} v_k^i f_{r,k-1}(z) + \sum_{k=1}^{N} v_{k+N}^i f_{l,k-1}(z).$$

It is straightforward that B is a bounded operator mapping onto the domain of $\mathcal{J}$. Furthermore, by Definition 3.5 we have that

$$\mathcal{B}B_i = WR_{\text{ext}} \begin{pmatrix} B_i(b) \\ \vdots \\ \frac{d^{N-1}B_i(b)}{dz^{N-1}} \\ B_i(a) \\ \vdots \\ \frac{d^{N-1}B_i(a)}{dz^{N-1}} \end{pmatrix}.$$

Now by definition

$$\frac{d^p B_i}{dz^p}(z) = \sum_{k=1}^{N} v_k^i f_{r,k-1}^{(p)}(z) + \sum_{k=1}^{N} v_{k+N}^i f_{l,k-1}^{(p)}(z).$$

From (A.3) and (A.4) of Lemma A.3, we have that

$$\frac{d^p B_i}{dz^p}(b) = v_{p+1}^i \quad \text{and} \quad \frac{d^p B_i}{dz^p}(a) = v_{p+N+1}^i,$$

and so B satisfies

$$\mathcal{B}Bu = WR_{\text{ext}} \begin{pmatrix} v_1^i \\ \vdots \\ v_{2N}^i \end{pmatrix} = u^i.$$

Step 3. By the definition of B and $D(J_W)$, we see that the conditions stated in the theorem are the same as $x(0) - Bu(0) \in D(J_W)$. Hence by Theorem 3.3.3 of [3] we have that there exists a classical solution of (4.7)–(4.9). Hence, in particular, $x(t) \in H^N((a,b), \mathbb{R}^n)$ holds pointwise in t, $x(t)$ is differentiable as a function of t, and $\dot{x}(t) = \mathcal{J}x(t)$. Using this, we obtain

$$\begin{aligned} \frac{d}{dt}\|x(t)\|^2 &= \frac{d}{dt}\langle x(t), x(t)\rangle \\ &= \langle \dot{x}(t), x(t)\rangle + \langle x(t), \dot{x}(t)\rangle \\ &= \langle \mathcal{J}x(t), x(t)\rangle + \langle x(t), \mathcal{J}x(t)\rangle \\ &= \begin{pmatrix} f_\partial^T(t) & e_\partial^T(t) \end{pmatrix} \Sigma \begin{pmatrix} f_\partial(t) \\ e_\partial(t) \end{pmatrix}. \end{aligned} \tag{4.16}$$

On the other hand, we have that

$$\begin{pmatrix} u \\ y \end{pmatrix} = \begin{pmatrix} W \\ \tilde{W} \end{pmatrix} \begin{pmatrix} f_\partial \\ e_\partial \end{pmatrix}. \tag{4.17}$$

Combining this with (4.16) gives that

$$\begin{aligned}\frac{d}{dt}\|x(t)\|^2 &= \begin{pmatrix} u^T(t) & y^T(t)\end{pmatrix}\begin{pmatrix} W \\ \tilde{W}\end{pmatrix}^{-T}\Sigma\begin{pmatrix} W \\ \tilde{W}\end{pmatrix}^{-1}\begin{pmatrix} u(t) \\ y(t)\end{pmatrix} \\ &= \begin{pmatrix} u^T(t) & y^T(t)\end{pmatrix} P_{W,\tilde{W}}\begin{pmatrix} u(t) \\ y(t)\end{pmatrix}. \qquad (4.18)\end{aligned}$$

Hence we have proved (4.12).

Step 4. By the definition of $P_{W,\tilde{W}}$, we see that

$$P_{W,\tilde{W}}^{-1} = \begin{pmatrix} W \\ \tilde{W}\end{pmatrix}\Sigma\begin{pmatrix} W \\ \tilde{W}\end{pmatrix}^T = \begin{pmatrix} W\Sigma W^T & W\Sigma\tilde{W}^T \\ \tilde{W}\Sigma W^T & \tilde{W}\Sigma\tilde{W}^T\end{pmatrix},$$

which shows (4.13). From this equality the last assertion of the theorem follows directly. □

Now we consider two particular cases which are canonical. For the first choice of inputs and outputs, the system becomes a lossless system. For the second choice of inputs and outputs, the balance (4.12) becomes canonical for scattering variables. We begin by characterizing the case when the boundary control system becomes a lossless system.

THEOREM 4.4. *Let W and $\tilde{W}$ be $nN \times 2nN$ matrices with W having full rank and satisfying* (4.1). *Associate with these matrices the following system:*

$$\dot{x}(t) = \mathcal{J}x(t), \qquad (4.19)$$

$$u(t) = W\begin{pmatrix} f_\partial(t) \\ e_\partial(t)\end{pmatrix}, \qquad (4.20)$$

$$y(t) = \tilde{W}\begin{pmatrix} f_\partial(t) \\ e_\partial(t)\end{pmatrix}, \qquad (4.21)$$

where $(f_\partial(t), e_\partial(t))$ are the boundary port variables associated with $x(t)$; see Definition 3.5.

The above system is a boundary control system. Furthermore, it satisfies for all $u \in C^2((0,\infty);\mathbb{R}^{nN})$, $x(0) \in H^N((a,b);\mathbb{R}^n)$ with $u(0) = W\left(\begin{smallmatrix} f_\partial(0) \\ e_\partial(0)\end{smallmatrix}\right)$ the balance equation

$$\frac{1}{2}\frac{d}{dt}\|x(t)\|^2 = u(t)^T y(t) \qquad (4.22)$$

if and only if the following conditions are satisfied:

$$W = S\begin{pmatrix} I+V & I-V\end{pmatrix} \quad \text{with } S \text{ invertible and } V \text{ unitary}, \qquad (4.23)$$

$$\tilde{W} = \tilde{S}\begin{pmatrix} I+\tilde{V} & I-\tilde{V}\end{pmatrix} \quad \text{with } \tilde{S} \text{ invertible and } \tilde{V} \text{ unitary}, \qquad (4.24)$$

$$I = 2\tilde{S}(I-\tilde{V}V^T)S^T. \qquad (4.25)$$

Furthermore, under condition (4.23) *the associated semigroup is unitary.*

Proof. Looking at Theorem 4.2 we see that we only have to check that $P_{W,\tilde{W}}$ equals $\left(\begin{smallmatrix} 0 & I \\ I & 0\end{smallmatrix}\right)$. By (4.13) this is equivalent to $W\Sigma W^T = \tilde{W}\Sigma\tilde{W}^T = 0$ and $\tilde{W}\Sigma W^T = I$. By Lemma A.1 we see that the first conditions are equivalent to (4.23) and (4.24), respectively. Direct calculation gives that $\tilde{W}\Sigma W^T = I$ is the same as (4.25). □

Taking in the above theorem, $V = I$, $\tilde{V} = -I$, and $S = \tilde{S} = \frac{1}{2}I$, we obtain the following special case.

COROLLARY 4.5. *Under the general conditions as stated in Theorem* 4.4 *consider the system defined by*

$$\dot{x}(t) = \mathcal{J}x(t), \tag{4.26}$$

$$u(t) = f_\partial(t), \tag{4.27}$$

$$y(t) = -e_\partial(t), \tag{4.28}$$

where $(f_\partial(t), e_\partial(t))$ *are the boundary port variables associated with* $x(t)$*; see Definition* 3.5.

The above system is a boundary control system with the associated semigroup unitary. Furthermore, it satisfies for all $u \in C^2((0,\infty);\mathbb{R}^{nN})$, $x(0) \in H^N((a,b);\mathbb{R}^n)$, *and* $u(0) = f_\partial(0)$ *the following balance equation:*

$$\frac{1}{2}\frac{d}{dt}\|x(t)\|^2 = u(t)^T y(t). \tag{4.29}$$

In the following theorem we characterize the scattering case.

THEOREM 4.6. *Let* W *and* $\tilde{W}$ *be* $nN \times 2nN$ *matrices with* W *having full rank and satisfying* (4.1). *Associate with these matrices the following system:*

$$\dot{x}(t) = \mathcal{J}x(t), \tag{4.30}$$

$$u(t) = W\begin{pmatrix} f_\partial(t) \\ e_\partial(t) \end{pmatrix}, \tag{4.31}$$

$$y(t) = \tilde{W}\begin{pmatrix} f_\partial(t) \\ e_\partial(t) \end{pmatrix}, \tag{4.32}$$

where $(f_\partial(t), e_\partial(t))$ *are the boundary port variable associated with* $x(t)$*; see Definition* 3.5.

The above system is a boundary control system. Furthermore, it satisfies for all $u \in C^2((0,\infty);\mathbb{R}^{nN})$, $x(0) \in H^N((a,b);\mathbb{R}^n)$ *with* $u(0) = W\left(\begin{smallmatrix} f_\partial(0) \\ e_\partial(0) \end{smallmatrix}\right)$ *the balance equation*

$$\frac{1}{2}\frac{d}{dt}\|x(t)\|^2 = \|u(t)\|^2 - \|y(t)\|^2 \tag{4.33}$$

if and only if the following conditions are satisfied:

$$W = S\begin{pmatrix} I+V & I-V \end{pmatrix} \quad \text{with } 4S(I+VV^T)S^T = I, \tag{4.34}$$

$$\tilde{W} = \tilde{S}\begin{pmatrix} -I-V^T & I-V^T \end{pmatrix} \quad \text{with } 4\tilde{S}(I-V^TV)\tilde{S}^T = I. \tag{4.35}$$

Proof. Looking at Theorem 4.2 we see that we only have to check that $P_{W,\tilde{W}}$ equals $\left(\begin{smallmatrix} 2I & 0 \\ 0 & -2I \end{smallmatrix}\right)$. By (4.13) this is equivalent to $W\Sigma W^T = \frac{1}{2}I$, $\tilde{W}\Sigma\tilde{W}^T = -\frac{1}{2}I$, and $W\Sigma\tilde{W}^T = 0$.

It is easy to show that if both (4.34) and (4.35) hold, then (4.33) holds. So it remains to show the converse. Using the standard representation of W (see (4.2)), we get that $W\Sigma W^T = \frac{1}{2}I$ is equivalently written as (4.34).

Since $\tilde{W}\Sigma\tilde{W}^T = -\frac{1}{2}I$, we see that $\tilde{W}$ is of full rank. Furthermore, from the relation $W\Sigma\tilde{W}^T = 0$, we obtain that the range of $\Sigma\tilde{W}^T$ is contained in the kernel of

W. By Lemma A.2 we have that the kernel of W equals the range of $\left(\begin{smallmatrix} I-V \\ -I-V \end{smallmatrix}\right)$. Since both the range of this matrix and that of $\Sigma\tilde{W}^T$ is of dimension nN, we find that

$$\Sigma\tilde{W}^T = \begin{pmatrix} I-V \\ -I-V \end{pmatrix} \tilde{S}^T$$

for some invertible $\tilde{S}$. Hence we have shown that the representation of (4.35) holds. The last part of this equation follows directly from the fact that $\tilde{W}\Sigma\tilde{W}^T = -\frac{1}{2}I$. □

Choosing in the above theorem $V = 0$ and $S = \tilde{S} = \frac{1}{2}I$ gives the following corollary.

COROLLARY 4.7. *Consider the system defined as*

$$\dot{x}(t) = \mathcal{J}x(t), \tag{4.36}$$

$$u(t) = \frac{1}{2}\left(f_\partial(t) + e_\partial(t)\right), \tag{4.37}$$

$$y(t) = \frac{1}{2}\left(f_\partial(t) - e_\partial(t)\right), \tag{4.38}$$

where $(f_\partial(t), e_\partial(t))$ are the boundary port variable associated with $x(t)$; see Definition 3.5.

The above system is a boundary control system with the associated semigroup a contraction. Furthermore, for $u \in C^2((0,\infty);\mathbb{R}^{nN})$, $x(0) \in H^N((a,b);\mathbb{R}^n)$, and $u(0) = \frac{1}{2}\left(f_\partial(0) + e_\partial(0)\right)$ we have that

$$\frac{1}{2}\frac{d}{dt}\|x(t)\|^2 = \|u(t)\|^2 - \|y(t)\|^2. \tag{4.39}$$

In the previous theorems we have seen that for the same Dirac structure the properties of the PDE, obtained by a choice of the inputs and outputs, can be completely different. Hence for the same underlying Dirac structure, many different system theoretic properties are possible. It is even possible that the PDE has no solution for the trivial input signal. Let us illustrate this situation in more detail in the following simple example.

Example 4.8. Consider the PDE on $[a,b]$

$$\frac{\partial x}{\partial t}(t,z) = \frac{\partial x}{\partial z}(t,z). \tag{4.40}$$

Following section 3, we see that $N = n = 1$, and $P(1) = 1$. The boundary port variables (see Definition 3.5) are

$$\begin{pmatrix} f_\partial(t) \\ e_\partial(t) \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} x(t,b) - x(t,a) \\ x(t,b) + x(t,a) \end{pmatrix}.$$

A short calculation gives that the PDE as discussed in Theorem 4.2 is (4.40) with the boundary input

$$u(t) = s\sqrt{2}\left[x(t,b) - vx(t,a)\right] \tag{4.41}$$

and output

$$y(t) = \frac{1}{\sqrt{2}}\left(\tilde{w}_2 + \tilde{w}_1\right)x(t,b) + \frac{1}{\sqrt{2}}\left(\tilde{w}_2 - \tilde{w}_1\right)x(t,a), \tag{4.42}$$

where s is a nonzero scalar, v is an element of $[-1,1]$, and $\tilde{w}_1, \tilde{w}_2$ are such that $\tilde{w}_2(1+v) - \tilde{w}_1(1-v) \neq 0$. As shown in Theorem 4.2 for any choice of $v \in [-1,1]$, we have that the PDE (4.40) with input (4.41) and output (4.42) has a unique classical solution provided the initial condition and the input are sufficiently smooth. Although the underlying Dirac structure stays the same, the system theoretic properties may be different for different choices of v. For instance, if $v = 1$, then the associated semigroup is unitary, whereas for $v = 0$, the associated semigroup is zero for $t \geq (b-a)$; see also [31].

Now one may wonder which (linear combination) of the boundary port variables may serve as an input, by which we only mean that it may be chosen in some sufficiently large (linear) space. Note that the choice, $u(t) = f_\partial(t) - e_\partial(t)$, gives that the input is located at $z = a$. Since (4.40) represents the left shift, it may be clear that the value of x at a cannot be an input. Even more, for $u \equiv 0$, the PDE does not have a solution.

5. Port Hamiltonian system. In this section, we define port Hamiltonian systems associated with (constant) skew-symmetric matrix operators. These systems are defined in terms of network-based modeling [1, 18, 28] which is based on the definition of two objects: the interconnection structure defined by a Dirac structure and the Hamiltonian function representing the total energy of the system. First, using the definition of the Dirac structure associated with a skew-symmetric operator given in section 3, we define a port Hamiltonian system with boundary port variables. Second, using the results of section 4, we formulate these port Hamiltonian systems as boundary control systems. In subsection 5.2 we treat extensively the example of the Timoshenko beam.

5.1. Linear port Hamiltonian systems with boundary port variables. We now extend the definition of linear port Hamiltonian systems as defined for finite-dimensional state spaces [28] to infinite-dimensional state spaces. The interconnection structure is defined by a Dirac structure associated with the skew-symmetric differential operator according to Theorem 3.6. The Hamiltonian function, generating this port Hamiltonian system, is defined by a coercive operator relating the state variable to the effort variable.

In the introductory example of the section 2, the skew-symmetric operator was the 2×2 matrix differential operator of differential order 1 corresponding to the canonical interdomain coupling, and the Dirac structure was the Stokes–Dirac structure. The symmetric operator was defined by the elasticity modulus and the mass distribution defining the elastodynamic energy of the string.

DEFINITION 5.1. *Consider the domain $Z = (a,b) \subset \mathbb{R}$. Let the space of flow variables $\mathcal{F}_Z$ be equal to $L^2((a,b);\mathbb{R}^n)$ and let the space of effort variables $\mathcal{E}_Z$ be equal to $\mathcal{F}_Z$. Consider an $n \times n$ matrix skew-symmetric differential operator of differential order N denoted as $\mathcal{J}$ defined by* (3.1) *and* (3.2). *Define the bond space $\mathcal{B} = \mathcal{F}_Z \times \mathbb{R}^{nN} \times \mathcal{E}_Z \times \mathbb{R}^{nN}$ and the Dirac structure $\mathcal{D}_\mathcal{J}$ associated with the skew-symmetric differential operator $\mathcal{J}$ as defined in Theorem* 3.6. *Let $\mathcal{L}$ be a coercive-symmetric operator on $\mathcal{E}_Z$. The port Hamiltonian system with the boundary port variables associated with $\mathcal{J}$ and generated by $\mathcal{L}$ is defined by*

$$(\dot{x}(t), f_\partial(t), \mathcal{L}x(t), e_\partial(t)) \in \mathcal{D}_\mathcal{J}, \quad t \geq 0, \tag{5.1}$$

where $\left(\begin{smallmatrix} f_\partial \\ e_\partial \end{smallmatrix}\right)$ is the boundary port associated with $e := \mathcal{L}x$ according to Definition 3.5.

REMARK 5.2. *It may be noted that the system in Definition* 5.1 *corresponds to*

the abstract system $\dot{x}(t) = A\,x(t)$ *defined by the differential operator*

$$A = \mathcal{J}\mathcal{L} \tag{5.2}$$

which need not be skew-symmetric nor have constant coefficients.

It is also worth making explicit the Hamiltonian function representing the energy of the system

$$H(x) = \frac{1}{2}\langle x,\, \mathcal{L}x\rangle, \tag{5.3}$$

where $\langle\cdot,\cdot\rangle$ denotes the natural inner product on the space $\mathcal{E}_Z$. The port Hamiltonian system of section 4 may hence be seen as a particular case with $\mathcal{L} = I$.

Noting that $\frac{dH(x(t))}{dt} = \langle\dot{x}(t),\, \mathcal{L}x(t)\rangle$, by the definition of Dirac structure, one obtains the following energy balance equation:

$$\frac{dH(x(t))}{dt} = \frac{1}{2}\left(f_\partial^T(t), e_\partial^T(t)\right)\Sigma\begin{pmatrix} f_\partial(t) \\ e_\partial(t)\end{pmatrix}.$$

This expresses that the variation of the energy of the boundary port Hamiltonian system is equal to the flow of energy at the boundary of the system's domain.

This also motivates us to take the state space equal to those x for which the Hamiltonian is finite. Since $\mathcal{L}$ is coercive on $\mathcal{E}_Z = L^2((a,b);\mathbb{R}^n)$, we see that the state space $\mathcal{X}$ is $L^2((a,b);\mathbb{R}^n)$ with the new inner product

$$\langle x_1, x_2\rangle_{\mathcal{X}} = \langle x_1, \mathcal{L}x_2\rangle_{L^2((a,b);\mathbb{R}^n)}. \tag{5.4}$$

In the previous definition we have defined linear port Hamiltonian systems with boundary port variables using the definition of Dirac structure for which the port variables are not split into input and output variables. However, we have seen in section 4 that using a specific subspace of the port variables, one may define input and output variables as belonging to complementary subspaces of the boundary port variables. Moreover, by choosing in an appropriate way these subspaces, one may define a boundary control system with its associated semigroup being a contraction. In the following, we reformulate the boundary port Hamiltonian system of Definition 5.1 as a boundary control system. We use the parameterization of the input and output variables and the contractive semigroups associated with the Dirac structure $\mathcal{D}_\mathcal{J}$ given in section 4. The state variables have become the image of the effort variables through the coercive operator $\mathcal{L}^{-1}$.

THEOREM 5.3. *The port Hamiltonian system of Definition* 5.1 *may be formulated as a boundary control system on the state space* $\mathcal{X}$*:*

$$(\dot{x}(t), f_\partial(t), \mathcal{L}x(t), e_\partial(t)) \in \mathcal{D}_\mathcal{J}, \quad t \geq 0, \tag{5.5}$$

with the input variables defined by choosing some full rank matrix W *of size* $nN \times 2nN$ *satisfying* (4.1) *and the map*

$$\mathcal{B}x(t) = W\begin{pmatrix} f_\partial(t) \\ e_\partial(t)\end{pmatrix} = u(t) \tag{5.6}$$

on the domain

$$D(\mathcal{B}) = D(\mathcal{J}). \tag{5.7}$$

Furthermore, define the port conjugated output

$$y(t) = \tilde{W} \begin{pmatrix} f_\partial(t) \\ e_\partial(t) \end{pmatrix}$$

with $\tilde{W}$ *a full rank matrix of size* $nN \times 2nN$ *with* $\begin{pmatrix} W \\ \tilde{W} \end{pmatrix}$ *invertible. Then for* $u \in C^2((0,\infty);\mathbb{R}^{nN})$, $x(0) \in H^N((a,b),\mathbb{R}^n)$, *and* $u(0) = \mathcal{B}x(0)$, *the following balance equation is satisfied:*

$$\frac{d}{dt}H(x(t)) = \frac{1}{2}(u^T(t) \quad y^T(t))P_{W,\tilde{W}} \begin{pmatrix} u(t) \\ y(t) \end{pmatrix}, \tag{5.8}$$

where $P_{W,\tilde{W}}$ *is defined in* (4.13).

The proof is a straightforward extension of the proof of Theorem 4.1 using the following lemma.

LEMMA 5.4. *Assume that* W *satisfies* (4.1). *The differential operator* $A_W = \mathcal{J}\mathcal{L}$ *with the domain* $D(A_W) = \{x \in \mathcal{X} \mid \mathcal{L}x \in D(J_W)\}$ (*see* (4.4)) *generates a contraction semigroup on* $\mathcal{X}$.

Proof. We first show that A_W is dissipative. For $x \in D(A_W)$, we have that

$$\langle x, A_W x\rangle_{\mathcal{X}} = \langle x, \mathcal{J}\mathcal{L}x\rangle_{\mathcal{X}} = \langle x, \mathcal{L}\mathcal{J}\mathcal{L}x\rangle_{L_2} = \langle e, \mathcal{J}e\rangle_{L_2},$$

where $e = \mathcal{L}x$. Since $e \in D(J_W)$ and since J_W is a restriction of $\mathcal{J}$, we find that

$$\langle x, A_W x\rangle_{\mathcal{X}} = \langle e, J_W e\rangle_{L_2},$$

which is nonpositive, since J_W generates a contraction semigroup on $L^2((a,b);\mathbb{R}^n)$.

It is not hard to show that $A_W^* = J_W^*\mathcal{L}$ with $D(A_W^*) = \{x \in \mathcal{X} \mid \mathcal{L}x \in D(J_W^*)\}$. Using an argument similar to that above, we find that on $D(A_W^*)$

$$\langle x, A_W^* x\rangle_{\mathcal{X}} \leq 0.$$

Hence we conclude that A_W generates a contraction semigroup on $\mathcal{X}$. ∎

5.2. Example: The Timoshenko's beam model. Timoshenko's beam model describes the infinitesimal planar deformations of a flexible beam reduced to its neutral fiber with some particular geometrical assumptions. We briefly recall the Hamiltonian formulation as proposed by Golo, Talasila, and van der Schaft [6]. Note that this corresponds to taking the Legendre transform of the usual Lagrangian formulation. Consider the spatial domain $Z = [a,b]$. Denote the angular displacement by q_θ, the transversal displacement of the beam by q_y, and the conjugated momenta by p_θ and p_y. The elastic potential energy density is given by $\mathcal{U}(q) = \frac{1}{2}\int_Z F^T q\, dz$, where the strain wrench (torque and force) is $F = K\,q$. Let $K = \operatorname{diag}(c_\theta, c_y)$ denote the positive definite compliance matrix which depends on the elasticity properties of the material and its geometry. The kinetic energy is given by $\mathcal{K}(p) = \frac{1}{2}\int_Z v^T p\, dz$, where the coenergy variable is the velocity $v = M^{-1}\,p$. M denotes the positive definite inertia matrix which is given as $M = \operatorname{diag}(\iota, \mu)$ with ι the momentum of inertia of the beam per unit length and μ the mass per unit length. It is immediate that $F = \delta_q\mathcal{U}(q)$ and $v = \delta_p\mathcal{K}(q)$, where δ denotes the variational derivative [20].

Choose the state vector x as

$$x = \begin{pmatrix} q_\theta \\ q_y \\ p_\theta \\ p_y \end{pmatrix} = \begin{pmatrix} q \\ p \end{pmatrix}.$$

The Timoshenko beam model may be expressed as the following Hamiltonian evolution equations [6, 7]:

$$\frac{\partial x}{\partial t} = \mathcal{J} \begin{pmatrix} \frac{\partial \mathcal{H}}{\partial q} \\ \frac{\partial \mathcal{H}}{\partial p} \end{pmatrix}, \tag{5.9}$$

where $\mathcal{H}(q,p) = \mathcal{U}(q) + \mathcal{K}(p)$ is the total elastodynamic energy of the beam, and the skew-symmetric differential operator $\mathcal{J}$ is

$$\mathcal{J} = \begin{pmatrix} 0_2 & \begin{pmatrix} \frac{\partial}{\partial z} & 0 \\ -1 & \frac{\partial}{\partial z} \end{pmatrix} \\ \begin{pmatrix} \frac{\partial}{\partial z} & 1 \\ 0 & \frac{\partial}{\partial z} \end{pmatrix} & 0_2 \end{pmatrix}. \tag{5.10}$$

We now derive the port Hamiltonian formulation of this system. The time variation of the energy variables is defined as flow variables:

$$\frac{\partial}{\partial t} \begin{pmatrix} q \\ p \end{pmatrix} := \begin{pmatrix} f_q \\ f_p \end{pmatrix}.$$

The variational derivative of the total energy $\delta_x \mathcal{H}$ defines the effort variables:

$$\begin{pmatrix} e_q \\ e_p \end{pmatrix} := \mathcal{L} \begin{pmatrix} q \\ p \end{pmatrix} = \begin{pmatrix} K & 0 \\ 0 & M^{-1} \end{pmatrix} \begin{pmatrix} q \\ p \end{pmatrix}. \tag{5.11}$$

Note that

$$\mathcal{L} \begin{pmatrix} q \\ p \end{pmatrix} = \begin{pmatrix} \frac{\partial \mathcal{H}}{\partial q} \\ \frac{\partial \mathcal{H}}{\partial p} \end{pmatrix}.$$

More precisely,

$$e_q = \begin{pmatrix} c_\theta & 0 \\ 0 & c_y \end{pmatrix} \begin{pmatrix} q_\theta \\ q_y \end{pmatrix} = \begin{pmatrix} T \\ F_y \end{pmatrix} \tag{5.12}$$

is the vector composed of the torque and the force, and

$$e_p = \begin{pmatrix} \iota^{-1} & 0 \\ 0 & \mu^{-1} \end{pmatrix} \begin{pmatrix} p_\theta \\ p_y \end{pmatrix} = \begin{pmatrix} \omega \\ v_y \end{pmatrix} \tag{5.13}$$

is the vector composed of the angular and longitudinal velocities.

Hence, according to the evolution equation (5.9), the flow variables are related to the coenergy variables by the skew-symmetric differential operator $\mathcal{J}$ defined in (3.1)

$$\begin{pmatrix} f_q \\ f_p \end{pmatrix} = \mathcal{J} \begin{pmatrix} e_q \\ e_p \end{pmatrix}.$$

This differential operator may be written as

$$\mathcal{J} = P(0) + P(1)\frac{\partial}{\partial z},$$

where

$$P(0) = \begin{pmatrix} 0_2 & \begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix} \\ \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} & 0_2 \end{pmatrix}, \quad P(1) = \begin{pmatrix} 0_2 & \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} & 0_2 \end{pmatrix}.$$

The symmetric matrix Q corresponding to the bilinear term on the boundary variables in Theorem 3.1 and given in (3.5) reduces to $Q = P(1)$. The matrix R_{ext} defining the boundary port variables equals (see (3.7))

$$R_{\text{ext}} = \frac{\sqrt{2}}{2} \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{5.14}$$

According to Definition 3.5 the port variables are

$$\begin{pmatrix} f_\partial \\ e_\partial \end{pmatrix} = R_{\text{ext}} \begin{pmatrix} \mathcal{L} & 0 \\ 0 & \mathcal{L} \end{pmatrix} \begin{pmatrix} q(b) \\ p(b) \\ q(a) \\ p(a) \end{pmatrix}.$$

Considering relations (5.11), (5.12), and (5.13),

$$\begin{pmatrix} f_\partial \\ e_\partial \end{pmatrix} = \frac{\sqrt{2}}{2} \begin{pmatrix} \omega(b) - \omega(a) \\ v_y(b) - v_y(a) \\ T(b) - T(a) \\ F_y(b) - F_y(a) \\ T(b) + T(a) \\ F_y(b) + F_y(a) \\ \omega(b) + \omega(a) \\ v_y(b) + v_y(a) \end{pmatrix}.$$

The associated Dirac structure is given by

$$\mathcal{D}_{\mathcal{J}} = \left\{ \begin{pmatrix} f_q \\ f_p \\ f_\partial \\ e_q \\ e_p \\ e_\partial \end{pmatrix} \;\middle|\; \begin{pmatrix} e_q \\ e_p \end{pmatrix} \in H^1((a,b);\mathbb{R}^4), \mathcal{J} \begin{pmatrix} e_q \\ e_p \end{pmatrix} = \begin{pmatrix} f_q \\ f_p \end{pmatrix}, \quad \begin{pmatrix} f_\partial \\ e_\partial \end{pmatrix} = R_{\text{ext}} \begin{pmatrix} T(b) \\ F_y(b) \\ \omega(b) \\ v_y(b) \\ T(a) \\ F_y(a) \\ \omega(a) \\ v_y(a) \end{pmatrix} \right\}.$$

We now illustrate the derivation of boundary control systems from the port Hamiltonian system using two different choices of the matrix W defining them according to Theorem 5.3. The first choice corresponding to the boundary control system is

associated with a unitary semigroup and in the other choice corresponds to a system in the scattering representation.

For the unitary case let us choose the matrix W given in (4.2) with the invertible matrix S and matrix V satisfying $VV^T = I$ chosen as follows:

$$S = \frac{1}{2\sqrt{2}} \begin{pmatrix} -I_2 & I_2 \\ I_2 & I_2 \end{pmatrix} \text{ and } V = \begin{pmatrix} 0 & I_2 \\ -I_2 & 0 \end{pmatrix}.$$

This choice corresponds to define the inputs

$$\begin{aligned} u &= S \begin{pmatrix} I_4 + V & I_4 - V \end{pmatrix} \begin{pmatrix} f_\partial \\ e_\partial \end{pmatrix} \\ &= \frac{1}{\sqrt{2}} \begin{pmatrix} -I_2 & 0 & 0 & I_2 \\ 0 & I_2 & I_2 & 0 \end{pmatrix} \begin{pmatrix} f_\partial \\ e_\partial \end{pmatrix} = \begin{pmatrix} \omega(a) \\ v_y(a) \\ T(b) \\ F_y(b) \end{pmatrix}. \end{aligned}$$

The unitary semigroup associated with the boundary control $u = 0$ corresponds to the following boundary conditions:

$$\omega(a,t) = v_y(a,t) = M(b,t) = F_y(b,t) = 0,$$

which are the so-called *clamped-free* boundary conditions. According to Theorem 4.4 the output conjugated to this input is

$$y = \tilde{S}(I_4 + \tilde{V} \quad I_4 - \tilde{V}) \begin{pmatrix} f_\partial \\ e_\partial \end{pmatrix}$$

with $\tilde{V}$ unitary, $\tilde{S}$ invertible, and

$$2\tilde{S}\big(I_4 - \tilde{V}V^T\big)S = I_4.$$

For example, choosing $\tilde{V} = -V = \left(\begin{smallmatrix} 0 & -I_2 \\ I_2 & 0 \end{smallmatrix}\right)$ and $\tilde{S} = S = \frac{1}{2\sqrt{2}} \left(\begin{smallmatrix} -I_2 & I_2 \\ I_2 & I_2 \end{smallmatrix}\right)$ we obtain

$$y = \begin{pmatrix} -T(a) \\ -F_y(a) \\ \omega(b) \\ v_y(b) \end{pmatrix}.$$

For the contractive case, let us choose the matrix W given in (4.2) with the invertible matrix S and matrix V, satisfying $VV^T \leq I$, chosen as follows:

$$S = \frac{\sqrt{2}}{4} \begin{pmatrix} I_2 & I_2 \\ -I_2 & I_2 \end{pmatrix} \text{ and } V = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

According to Theorem 5.3 the inputs are

$$u = \frac{\sqrt{2}}{4} \begin{pmatrix} I_2 & I_2 & I_2 & I_2 \\ -I_2 & I_2 & -I_2 & I_2 \end{pmatrix} \begin{pmatrix} f_\partial \\ e_\partial \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \omega(b) + T(b) \\ v_y(b) + F_y(b) \\ \omega(a) - T(a) \\ v_y(a) - F_y(a) \end{pmatrix}.$$

For $\tilde{S}$ (see Theorem 4.6) we choose S and so the outputs are

$$y = \frac{\sqrt{2}}{4}\begin{pmatrix} I_2 & I_2 & -I_2 & -I_2 \\ -I_2 & I_2 & I_2 & -I_2 \end{pmatrix}\begin{pmatrix} f_\partial \\ e_\partial \end{pmatrix} = -\frac{1}{2}\begin{pmatrix} \omega(a) + T(a) \\ v_y(a) + F_y(a) \\ \omega(b) - T(b) \\ v_y(b) - F_y(b) \end{pmatrix}.$$

In this case, the boundary inputs and outputs correspond to the *scattering variables* and $\frac{1}{2}\|x(t)\|^2_{\mathcal{X}} = \|u(t)\|^2 - \|y(t)\|^2$.

6. Conclusion and further work. The work presented in this paper relates the structure of a class of linear infinite-dimensional dynamical models induced by the physical modeling (existence of energy function, power continuous interconnection structure) with system theoretical properties (passivity, etc.). More precisely, we have defined a class of infinite-dimensional linear systems associated with skew-symmetric differential operators and we have related them to boundary control systems. Knowing the underlying physical structure and the system theoretical notions will be very useful in the further analysis and design for our class of infinite-dimensional systems, for instance, in the construction of stabilizing feedbacks.

Therefore, we have, in the first instance, defined a Dirac structure on a Hilbert space associated with skew-symmetric differential operators with constant coefficients. Using the Stokes theorem, we have defined port boundary variables as the image of the boundary values under a linear map, which is derived from the differential operator. Then we have shown that the differential operator together with the boundary port variables defines a Dirac structure on a vector space (the space of bond variables) endowed with a canonical symmetric pairing. This defines the geometrical structure associated with the initial PDE.

In the second instance, we have shown that one may derive from the Dirac structure infinitesimal generators of contraction semigroups. These infinitesimal generators are obtained by restricting the domain of the skew-symmetric operator to parameterized subspaces. More precisely, we have shown that we have obtained a parameterization of all the contraction semigroups which are associated with the skew-symmetric operator.

In the third instance, we have derived a formulation of our class of infinite-dimensional systems as boundary control systems associated with the class of contraction semigroups obtained from the Dirac structure. We have defined outputs conjugated to the inputs of the boundary control systems in such a way that the system satisfies a power balance equation in a way similar to dissipative systems [25].

In the fourth instance, these results are used to define infinite-dimensional port Hamiltonian systems. These systems are defined with respect to the Dirac structure associated with a skew-symmetric differential operator and a coercive operator defining the Hamiltonian functional, i.e., the total energy of the system. Again from such a port Hamiltonian system one may derive a class of boundary control system associated with contraction semigroups. This is illustrated by the example of Timoshenko's beam.

A natural question is the relation of our class of systems, especially the Hamiltonian systems with the systems nodes (see [26]), and with the class of well-posed linear systems. This has been partially done in [31] for the system nodes and in [32] for well-posed systems (using the idea of feedback). Another issue is the generalization of this work to PDEs on an n-dimensional spatial domain.

Finally, this works also opens the way for the generalization to infinite-dimensional systems of the synthesis of stabilizing controllers using the immersion and Hamiltonian reduction proposed in [13, 22].

Appendix. Technical lemmas.

LEMMA A.1. *Let W be an $nN \times 2nN$ matrix and let $\Sigma = \left(\begin{smallmatrix} 0 & I \\ I & 0 \end{smallmatrix}\right)$. Then W has rank nN and $W\Sigma W^T \geq 0$ if and only if there exist a matrix $V \in \mathbb{R}^{nN \times nN}$ and an invertible matrix $S \in \mathbb{R}^{nN \times nN}$ such that*

$$W = S(I + V \quad I - V) \tag{A.1}$$

with $VV^T \leq I$.

Furthermore, $W\Sigma W^T = 0$ if and only if V is unitary.

Proof. If W is of the form (A.1), then we find

$$W\Sigma W^T = S\begin{pmatrix} I + V & I - V \end{pmatrix} \Sigma \begin{pmatrix} I + V^T \\ I - V^T \end{pmatrix} S^T = S[2I - 2VV^T]S^T,$$

which is nonnegative, since $VV^T \leq I$.

Now we prove that if W is of full rank and is such that $W\Sigma W^T \geq 0$, then (A.1) holds. Writing W as $W = (W_1 \quad W_2)$, we have that $W\Sigma W^T \geq 0$ is equivalent to $W_1 W_2^T + W_2 W_1^T \geq 0$. Hence

$$(W_1 + W_2)(W_1 + W_2)^T \geq (W_1 - W_2)(W_1 - W_2)^T \geq 0. \tag{A.2}$$

If $x \in \ker((W_1 + W_2)^T)$, then the above inequality implies that $x \in \ker((W_1 - W_2)^T)$. Thus $x \in \ker(W_1^T) \cap \ker(W_2^T)$. Since W has full rank, this implies that $x = 0$. Hence $W_1 + W_2$ is invertible.

Using (A.2) once again, we see that

$$(W_1 + W_2)^{-1}(W_1 - W_2)(W_1 - W_2)^T(W_1 + W_2)^{-T} \leq I$$

and thus $V := (W_1 + W_2)^{-1}(W_1 - W_2)$ satisfies $VV^T \leq I$. Summarizing, we have

$$\begin{aligned}(W_1 \quad W_2) &= \frac{1}{2}(W_1 + W_2 + W_1 - W_2 \quad W_1 + W_2 - W_1 + W_2) \\ &= \frac{1}{2}(W_1 + W_2)(I + V \quad I - V).\end{aligned}$$

Defining $S := \frac{1}{2}(W_1 + W_2)$, we have shown the representation (A.1).

If instead of inequality we have equality for W, then it is easy to show that we have equality in the equation for V as well. Thus V is unitary. ∎

LEMMA A.2. *Suppose that the $nN \times 2nN$ matrix W can be written in the format of (A.1), i.e., $W = S(I+V \quad I-V)$ with S and V square matrices, and S is invertible. Then the kernel of W equals the range of $\left(\begin{smallmatrix} I-V \\ -I-V \end{smallmatrix}\right)$.*

If V is unitary, then the kernel of W equals the range of ΣW^T.

Proof. Let $\left(\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right)$ be in the range of $\left(\begin{smallmatrix} I-V \\ -I-V \end{smallmatrix}\right)$. By equality (A.1), we have that

$$\begin{aligned}W\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} &= S\begin{pmatrix} I + V & I - V \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= S\begin{pmatrix} I + V & I - V \end{pmatrix}\begin{pmatrix} I - V \\ -I - V \end{pmatrix} l = 0.\end{aligned}$$

Hence we see that the range of $\left(\begin{smallmatrix} I-V \\ -I-V \end{smallmatrix}\right)$ lies in the kernel of W. It is easy to show that W has rank nN, and so the kernel of W has dimension nN. Thus, if we can show that the $2nN \times nN$ matrix $\left(\begin{smallmatrix} I-V \\ -I-V \end{smallmatrix}\right)$ has full rank, then we have proved the first assertion. If this matrix would not have full rank, then there should be a nontrivial element in its kernel. It is easy to see that the kernel consists of zero only, and so we have proved the first part of the lemma.

Suppose now that V is unitary, then

$$\begin{pmatrix} I-V \\ -I-V \end{pmatrix} = \begin{pmatrix} -I+V^T \\ -I-V^T \end{pmatrix} V = -\Sigma W^T S^{-T} V.$$

Since the range of ΣW^T equals the range of $-\Sigma W^T S^{-T} V$, we have proved the second assertion. ∎

LEMMA A.3. *Given the interval $[a,b]$ and a positive number $N \in \mathbb{N}$. There exist polynomials $f_{lj}(z), f_{rj}(z), j = 0, \ldots, N-1$, such that*

$$\begin{aligned} \frac{d^k f_{l,j}}{dz^k}(a) &= \delta_{kj}, \quad k = 0, \ldots, N-1, \\ \frac{d^k f_{l,j}}{dz^k}(b) &= 0, \quad k = 0, \ldots, N-1, \end{aligned} \tag{A.3}$$

and

$$\begin{aligned} \frac{d^k f_{r,j}}{dz^k}(a) &= 0, \quad k = 0, \ldots, N-1, \\ \frac{d^k f_{r,j}}{dz^k}(b) &= \delta_{kj}, \quad k = 0, \ldots, N-1. \end{aligned} \tag{A.4}$$

Proof. Since the construction of $f_{r,j}$ is very similar to that of $f_{l,j}$, we show only how $f_{l,j}$ is constructed. These functions are constructed using backward induction. It is easily seen that

$$f_{l,N-1}(z) := \frac{1}{(N-1)!}(z-a)^{N-1}(z-b)^N \frac{1}{(a-b)^N}$$

satisfies condition (A.3). Suppose next that we have constructed the functions $f_{l,j}(z)$ for $j = j_0+1, \ldots, N-1$. We next construct $f_{l,j_0}(z)$. Define $\tilde{f}_{l,j_0}(z)$ as

$$\tilde{f}_{l,j_0}(z) = \frac{1}{j_0!}(z-a)^{j_0}(z-b)^N \frac{1}{(a-b)^N}.$$

It is easy to see that

$$\frac{d^k \tilde{f}_{l,j_0}}{dz^k}(a) = \delta_{kj_0}, \quad k = 0, \ldots, j_0,$$

and

$$\frac{d^k \tilde{f}_{l,j_0}}{dz^k}(b) = 0, \quad k = 0, \ldots, N-1.$$

If we define the function $f_{l,j_0}(z)$ as

$$f_{l,j_0}(z) = \tilde{f}_{l,j_0}(z) - \sum_{i=j_0+1}^{N-1} \frac{d^i f_{l,j_0}}{dz^i}(a) f_{l,i}(z),$$

then it is straightforward to see that it satisfies (A.3). ∎

REFERENCES

[1] P.C. Breedveld, *Physical Systems Theory in Terms of Bond Graphs*, Ph.D. thesis, Technische Hogeschool Twente, Enschede, The Netherlands, 1984.

[2] T.J. Courant, *Dirac manifolds*, Trans. Amer. Math. Soc., 319 (1990), pp. 631–661.

[3] R.F. Curtain and H.J. Zwart, *An Introduction to Infinite-Dimensional Linear System Theory*, 1st ed., Springer-Verlag, Berlin, 1995.

[4] M. Dalsmo and A.J. van der Schaft, *On representations and integrability of mathematical structures in energy-conserving physical systems*, SIAM J. Control Optim., 37 (1999), pp. 54–91.

[5] I. Dorfman, *Dirac Structures and Integrability of Nonlinear Evolution Equations*, Wiley, New York, 1993.

[6] G. Golo, V. Talasila, and A.J. van der Schaft, *A Hamiltonian formulation of the Timoshenko beam*, in Proceedings of the Mechatronics, Drebbel Institute for Mechatronics, University of Twente, Enschede, The Netherlands, 2002, pp. 544–553. Full paper on accompanying CD-ROM.

[7] G. Golo, *Interconnection Structures in Port-Based Modelling: Tools for Analysis and Simulation*, Ph.D. thesis, University of Twente, Enschede, The Netherlands, 2002.

[8] G. Golo, O.V. Iftime, H. Zwart, and A.J. van der Schaft, *Tools for Analysis of Dirac Structures on Hilbert Spaces*, Memorandum Faculteit TW 1729, Universiteit Twente, Enschede, 2004, http://www.math.utwente.nl/ssb/annrep04/publs04.htm.

[9] V.I. Gorbachuk and M.L. Gorbachuk, *Boundary Value Problems for Operator Differential Equations*, Math. Appl. (Soviet Ser.) 48, Kluwer Academic Publishers, Norwell, MA, 1991. (Expanded and revised translation of Granichnye zadachi dia differentsial'no-operatornykh uravneii.)

[10] Y. Le Gorrec, H. Zwart, and B.M. Maschke, *Dirac Structures and Boundary Control Systems Associated with Skew-Symmetric Differential Operators,* Memorandum Faculteit TW 1730, Universiteit Twente, Enschede, 2004, http://www.math.utwente.nl/ssb/annrep04/publs04.htm.

[11] D.C. Karnopp, D.L. Margolis, and R.C. Rosenberg, *System Dynamics: A Unified Approach*, Wiley, New York, 1990.

[12] B.M. Maschke, *Interconnection and structure in physical systems' dynamics*, in Proceedings of the 5th NOLCOS, NOLCOS'98, Enschede, 1998, pp. 291–296.

[13] B.M. Maschke, R. Ortega, and A.J. van der Schaft, *Energy-based Lyapunov functions for forced Hamiltonian systems with dissipation*, IEEE Trans. Automat. Control, 45 (2000), pp. 1498–1502.

[14] B.M. Maschke and A.J. van der Schaft, *Port controlled Hamiltonian systems: Modeling origins and system theoretic properties*, in Proceedings of the 3rd NOLCOS, NOLCOS'92, Bordeaux, 1992, pp. 282–288.

[15] B.M. Maschke and A.J. van der Schaft, *Interconnected mechanical systems, Part* 2: *The dynamics of spatial mechanical networks*, in Modelling and Control of Mechanical Systems, Imperial College Press, London, 1997, pp. 17–30.

[16] B.M. Maschke and A.J. van der Schaft, *Port controlled Hamiltonian representation of distributed parameter systems*, in Workshop on Modeling and Control of Lagrangian and Hamiltonian Systems, Princeton, NJ, 2000.

[17] B.M. Maschke and A.J. van der Schaft, *Compositional modelling of distributed-parameter systems*, in Advanced Topics in Control Systems Theory, Lecture Notes in Control and Inform. Sci. 311, F. Lamnabhi-Lagarrigue, A. Loria, and E. Panteley, eds., Springer-Verlag, Berlin, 2005, pp. 115–154.

[18] B.M. Maschke, A.J. van der Schaft, and P.C. Breedveld, *An intrinsic Hamiltonian formulation of network dynamics: Non-standard poisson structures and gyrators*, J. Franklin Inst., 329 (1992), pp. 923–966.

[19] B.M. Maschke and A.J. van der Schaft, *Canonical interdomain coupling in distributed parameter systems: An extension of the symplectic gyrator*, in Proceedings of the International Mechanical Engineering Congress and Exposition, New York, 2001.

[20] P.J. Olver, *Applications of Lie groups to differential equations*, 2nd ed., Grad. Texts in Math. 107, Springer-Verlag, New York, 1993.

[21] R. Ortega, A.J. van der Schaft, I. Mareels, and B. Maschke, *Putting energy back in control*, IEEE Control Syst. Mag., 21 (2001), pp. 18–32.

[22] R. Ortega, A.J. van der Schaft, B. Maschke, and G. Escobar, *Interconnection and damping assignment: Passivity-based control of port-controlled Hamiltonian systems*, Automatica, 38 (2002), pp. 585–596.

[23] A. Parsian and A. Shafei Deh Abad, *Dirac structures on Hilbert spaces*, Int. J. Math. Math. Sci., 22 (1999), pp. 97–108.
[24] R.S. Phillips, *Dissipative operators and hyperbolic systems of partial differential equations*, Trans. Amer. Math. Soc., 90 (1959), pp. 193–254.
[25] H.K. Pillai and J. Willems, *Lossless and dissipative distributed systems*, SIAM J. Control Optim., 40 (2002), pp. 1406–1430.
[26] O.J. Staffans, *Well-Posed Linear Systems*, Cambridge University Press, London, 2005.
[27] A.J. van der Schaft, *L_2-Gain and Passivity Techniques in Nonlinear Control*, 2nd revised and enlarged ed., Comm. Control Engrg. Ser., Springer-Verlag, London, 1999; 1st ed., Lecture Notes in Control and Inform. Sci. 218, Springer-Verlag, Berlin, 1996.
[28] A.J. van der Schaft and B.M. Maschke, *The Hamiltonian formulation of energy conserving physical systems with external ports*, Archiv. Elektronik Übertragungstechnik, 49 (1995), pp. 362–371.
[29] A.J. van der Schaft and B.M. Maschke, *Interconnected mechanical systems, Part* 1: *Geometry of interconnection and implicit Hamiltonian systems*, in Modelling and Control of Mechanical Systems, Imperial College Press, London, 1997, pp. 1–16.
[30] A.J. van der Schaft and B.M. Maschke, *Hamiltonian formulation of distributed parameter systems with boundary energy flow*, J. Geom. Phys., 42 (2002), pp. 166–174.
[31] J.A. Villegas, Y. Le Gorrec, H. Zwart, and A.J. van der Schaft, *Boundary control systems and the system nodes*, in 16th IFAC World Congress, Praha, 2005.
[32] H. Zwart, Y. Le Gorrec, B.M. Maschke, and J.A. Villegas, *Well-posedness of a class of boundary control systems*, in 44th IEEE CDC-ECC, Seville, Spain, submitted.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 5, pp. 1893–1922

MEAN-VARIANCE HEDGING WHEN THERE ARE JUMPS*

ANDREW E. B. LIM†

Abstract. In this paper, we consider the problem of mean-variance hedging in an *incomplete market* where the underlying assets are jump diffusion processes which are driven by Brownian motion and doubly stochastic Poisson processes. This problem is formulated as a stochastic control problem, and closed form expressions for the optimal hedging policy are obtained using methods from stochastic control and the theory of backward stochastic differential equations. The results we have obtained show how backward stochastic differential equations can be used to obtain solutions to optimal investment and hedging problems when discontinuities in the underlying price processes are modeled by the arrivals of Poisson processes with stochastic intensities. Applications to the problem of hedging default risk are also discussed.

Key words. jump diffusion, stochastic intensity, doubly stochastic Poisson process, mean-variance hedging, incomplete markets, backward stochastic differential equations, default risk

AMS subject classifications. 91B28, 91B30, 60J75, 30A36

DOI. 10.1137/040610933

1. Introduction.

Much of the literature on asset price modeling is motivated by the observation that simple models, like Black–Scholes, fail to account for important features of price processes that are observed in data. For example, the log-returns process of real-world asset prices are not normally distributed but exhibit higher peaks and heavier tails, implying a greater probability of extreme price movements than predicted by Black–Scholes. In addition, the price processes of real-world assets are typically not continuous, but may jump (in a nonpredictable way) in response to news or other surprise events. For a number of years, researchers have focused on developing a richer class of asset price models that include jumps as well as stochastic parameters; see, for example, [3, 12, 20]. While the adoption of these models in asset pricing (where simulation can be used) is fairly widespread, their use in dynamic optimization problems like hedging and optimal investment, when the market is incomplete, has been quite limited. This paper is concerned with the problem of *dynamic mean-variance hedging* in an *incomplete market* when there are *random parameters* and *discontinuities* in the price processes. We assume that uncertainty is modeled by Brownian motion and a doubly stochastic Poisson process with intensity that is predictable with respect to the Brownian filtration. We derive expressions for the optimal hedging strategy using methods from stochastic control and the theory of *backward stochastic differential equations* (BSDEs).

While the theory of BSDEs has played an important role in the analysis and solution of mean-variance hedging problems with random parameters (see [25, 27]), it is typically assumed that price processes are continuous and driven by Brownian motion. (We note, however, that generalizations to the continuous semimartingale setting have recently appeared; see Bobrovnytska and Schweizer [6].) One contribution of this paper is to show how BSDEs can be used when there are jumps. In this regard,

*Received by the editors July 2, 2004; accepted for publication (in revised form) April 24, 2005; published electronically December 6, 2005. This work was supported in part by National Science Foundation CAREER Award DMI-0348746.

http://www.siam.org/journals/sicon/44-5/61093.html

†Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA 94720 (lim@ieor.berkeley.edu).

we determine conditions under which the relevant BSDEs have unique solutions, and derive an expression for the optimal hedging strategy in terms of these. In particular, we show how the hedging strategy should respond to news that indicates a higher or lower probability of a sudden price change (i.e., an increase or decrease in the intensity of the jump process).

An alternative approach to mean-variance hedging uses the projection theorem and convex duality and typically assumes that price processes are *continuous* semimartingales; see, for example, [9, 14, 23, 30, 32]. Exceptions include [22], which considers the problem of local risk minimization for a model with jumps under the assumption that the stochastic intensity is independent of the processes driving the stock price processes, and the recent paper by Arai [1], which generalizes the methods in [30] to the discontinuous case. Some key differences between [1] and this paper include the generality of the price processes (discontinuous semimartingales v's processes driven by Brownian motion and a doubly stochastic Poisson process) and the methods that are used to solve the problem (duality v's stochastic control). A key issue in both [1] and this paper concerns the so-called *variance optimal (signed) martingale measure*. In the continuous semimartingale case, the variance optimal signed martingale measure is actually a probability measure, but this is not necessarily the case when there are jumps. (We show this in our example.) For this reason, additional assumptions are needed when dealing with discontinuous problems. In the context of this paper, some of these assumptions are required to prove solvability of one of the BSDEs. On the other hand, the additional structure in our model allows us to dispense with some of the assumptions imposed in [1]. Furthermore, under additional assumptions on the liability, we also prove solvability of the hedging problem, even when the variance optimal martingale measure is *not* a probability measure.

The bulk of the literature on optimal portfolio choice and dynamic hedging has focused, primarily, on market models with *continuous* price processes, and relatively little has been done using models with price discontinuities (some recent exceptions include [2, 15, 18, 26, 28, 29]). This paper may be regarded as a contribution to this literature. In the papers [2, 18], the problem of utility maximization when there are discontinuous price processes is solved using convex duality. Unlike the model in [29] as well as in the present paper, however, the market models in [2, 18] are *complete*. Similar methods are used in [28] to solve a continuous time mean-variance problem with a bankruptcy prohibition when there are price discontinuities, but once again, market completeness is assumed. Finally, the paper [15] discusses the issues of model calibration and optimal portfolio computation in a discontinuous price setting while the recent paper [26] solves a portfolio choice problem with regime switching and price discontinuities.

Finally, the results in this paper may also be regarded as a contribution to the literature on hedging default risk in an *incomplete* market. In particular, doubly stochastic Poisson processes have recently been used to model events default [4, 11, 21] and for this reason, the problem of optimal investment or hedging with default sensitive assets and/or liabilities may be formulated as an optimal investment/hedging problem with asset prices modeled as jump-diffusions. (For further discussion on this issue, the reader can consult [24].) The problem of hedging in a complete market with default risk is studied in Blanchet-Scalliet and Jeanblanc [5]. It should be noted, however, that the market model in [5] is different from ours in a number of ways, and for this reason our result cannot really be regarded as a *faithful* generalization of theirs. For example, we assume that parameters (and in particular

the default intensity) are predictable with respect to Brownian motion, whereas the results in [5] allow for a more general class of parameters. Also, we are assuming that assets remain tradable after a jump occurs, whereas the results in [5] apply to the case when the underlying asset (a zero-coupon bond) ceases to be tradable the instant a jump (i.e., default) occurs.

The outline of this paper is as follows. In section 2, we present the model for the financial market, and formulate the hedging problem as a stochastic control problem. In section 3, the optimal hedging portfolio is derived. In particular, the results in this section depend on the solvability of a certain BSDE that is driven by Brownian motion and the Poisson process. In section 4, solvability of this backwards equation is discussed in greater detail. In particular, we prove solvability under the assumption that a certain local martingale is a positive martingale (the *"martingale condition"*), and derive necessary and sufficient conditions for this to hold. This condition is difficult to check, however, due to its complicated dependence on the problem parameters, which motivates our analysis in section 5 where simple conditions under which the "martingale condition" can be checked are derived. In particular, we show that the "martingale condition" holds when the market is complete, and it is easy to check when the parameters are deterministic. In addition, we also derive conditions on the *liability* under which the hedging problem will still have a solution, *even if the "martingale condition" is not satisfied*. In particular, we show in section 5.4 that solvability can be guaranteed, irrespective of the *"martingale condition," whenever the liability is measurable with respect to the Brownian motion*, which is the case for the continuous time version of Markowitz's mean-variance portfolio selection problem. In section 6, we compare the assumptions made in Arai [1] with those in this paper. In section 7, we present an example where an explicit expression of the optimal hedging strategy can be calculated. We conclude in section 8.

This paper is a substantially expanded version of the conference paper [24]. In particular, the detailed proof of optimality, existence of solutions of the associated BSDEs, necessary and sufficient conditions for the "martingale condition" to be satisfied, comparisons with the paper [1], and the example are not found in the earlier version.

2. Formulation. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space. We assume throughout that all stochastic processes are defined on a finite time horizon $[0, T]$. Suppose that $W(t) \triangleq (W_1(t), \ldots, W_d(t))'$ is a d-dimensional standard Brownian motion on this space defined on $[0, T]$ and $\mathbb{F} \triangleq \{\mathcal{F}_t\}_{t\geq 0}$ is the filtration generated by $W(t)$ augmented by the null sets of $\mathbb{P}$. Let $N(t) \triangleq (N_1(t), \ldots, N_n(t))'$, where $N_i(t)$ is a *doubly stochastic Poisson process* (or a *Cox process*) with an $\mathbb{F}$-predictable nonnegative intensity $\lambda_i(t)$. In relation to $N(t)$, we denote by $\mathbb{D} \triangleq \{\mathcal{D}_t\}_{t\geq 0}$ the filtration generated by $N(t)$ augmented by the $\mathbb{P}$-null sets. We assume throughout that conditional on $\mathcal{F}_T$, $N_i(\cdot)$ is a nonhomogeneous Poisson process with intensity $\lambda_i(t)$, and (conditional on $\mathcal{F}_T$) $N_i(\cdot)$ and $N_j(\cdot)$ are independent when $i \neq j$. It should be noted that the construction of such processes $N_i(t)$ is fairly standard; see, for example, [4]. Finally, let $\mathbb{G}$ denote the filtration $\{\mathcal{G}_t\}_{t\geq 0}$, where $\mathcal{G}_t \triangleq \mathcal{F}_t \vee \mathcal{D}_t$, the smallest filtration containing $\mathbb{F}$ and $\mathbb{D}$. Here, $\mathcal{G}_t$ may be regarded as the information available to the investor at time t. The filtrations $\mathbb{F}$ and $\mathbb{G}$ satisfy the following property (see [7] for a detailed study).

PROPOSITION 2.1 (martingale invariance property). *Every $\mathbb{F}$-martingale under $\mathbb{P}$ is also a $\mathbb{G}$-martingale under $\mathbb{P}$.*

Proof. By the construction of doubly stochastic Poisson processes with $\mathbb{F}$-predictable intensity, $\mathcal{G}_t$ and $\mathcal{F}_T$ are independent, given $\mathcal{F}_t$. The result now follows from the observation that this property is equivalent to $\mathbb{E}[X|\mathcal{F}_t] = \mathbb{E}[X|\mathcal{G}_t]$ for every $\mathcal{F}_T$-measurable random variable X. □

The *martingale invariance property* is studied in detail in [7] and is a common *assumption* in the literature on default risk modeling [4, 13] as well as hedging and portfolio choice with jumps [5]. It holds in the setting of this paper due to the structure of our stochastic model.

We introduce the following notation.

- $\mathcal{P}^2(\mathbb{G}, \mathbb{R}^m)$: the set of $\mathbb{G}$-predictable, $\mathbb{R}^m$-valued processes on $[0, T]$ under $\mathbb{P}$ with norm
$$\|f\|_2 := \left(\mathbb{E}^{\mathbb{P}} \int_0^T |f(t)|^2 \, dt\right)^{\frac{1}{2}} < \infty.$$
- $\mathcal{L}^\infty(\mathbb{G}, \mathbb{R}^m)$: the set of $\mathbb{G}$-adapted $\mathbb{P}$-essentially bounded $\mathbb{R}^m$-valued processes on $[0, T]$.

We refer to processes belonging to $\mathcal{P}^2(\mathbb{G}, \mathbb{R}^m)$ as being *square integrable*, while those that belong to $\mathcal{L}^\infty(\mathbb{G}, \mathbb{R}^m)$ are *uniformly bounded*.

Suppose that there are $m + 1$ tradable assets with prices $B(t)$, $P_1(t), \ldots, P_m(t)$, where $B(t)$ is the price of the money market account with interest rate $r(t)$, and $P_i(t)$ is the price of the ith risky asset. We assume throughout that $B(t)$ and $P_i(t)$ are solutions of the following stochastic differential equations:

$$\begin{cases} dB(t) = r(t)B(t)\, dt, \quad B(0) = 1, \\ dP_i(t) = P_i(t)\mu_i(t)\, dt + P_i(t)\sigma_i(t)dW(t) + P_i(t)\theta_i(t)dN(t), \quad P_i(0) = P_i^0. \end{cases} \tag{2.1}$$

The process $\theta_{ij}(t)$ determines the relative change in the price $P_i(t)$ given an arrival of the jth doubly stochastic Poisson process $N_j(t)$. On the other hand, since the sum of doubly stochastic Poisson processes is itself a doubly stochastic Poisson process, an equivalent interpretation of (2.1) replaces $N(t) = [N_1(t), \ldots, N_n(t)]'$ with a single doubly stochastic Poisson process $\bar{N}(t)$ with intensity $\bar{\lambda}(t) \triangleq \lambda_1(t) + \cdots + \lambda_n(t)$. Conditional on an arrival of $\bar{N}(t)$, the relative change in the price of asset j is $\theta_{ij}(t)$ with probability $\lambda_j(t)/\bar{\lambda}(t)$. In this regard, the components of $\theta_i(t)$ represent possible "jump sizes" with the distribution of the jumps determined by the intensities $\lambda_1(t), \ldots, \lambda_n(t)$. The $\mathbb{F}$-predictability of the $\theta_i(t)$ and $\lambda_i(t)$ implies that the possible jump sizes as well as their distributions depend on "available information," as captured by $\mathbb{F}$.

We assume that the investor in this financial market faces some liability, which we model by a random variable ξ. (For example, ξ may be a contingent claim written on a default event, which itself affects the price of the underlying asset.) Broadly speaking, the investor would like to reduce the uncertainty by investing in the financial market to minimize his/her risk. We shall assume throughout that the following assumptions are satisfied.

Assumption (A).

- $r(t)$, $\mu_i(t)$, $\sigma_{ik}(t)$, $\theta_{ij}(t)$, and $\lambda_j(t)$ are *uniformly bounded* and $\mathbb{F}$-predictable on $[0, T]$ for $i = 1, \ldots, m$, $j = 1, \ldots, n$, and $k = 1, \ldots, d$. That is, there is a constant K such that $|\mu_i(t)| \leq K$ for all $t \in [0, T]$, $\mathbb{P}$-a.s. (and likewise for the other parameters).
- There exists a constant $\delta > 0$ such that $\lambda_i(t) \geq \delta$ for all $t \in [0, T]$, $\mathbb{P}$-a.s.

- $\xi \in L^\infty(\mathcal{G}_T)$, where

$$L^\infty(\mathcal{G}_T) = \{Y : \Omega \to \mathbb{R} \,|\, Y \text{ is } \mathcal{G}_T\text{-measurable and } |Y| < K \ \mathbb{P}\text{-a.s. for some constant } K < \infty\}.$$

  Throughout this paper, random variables satisfying this property are said to be *uniformly bounded*.
- There exists a constant $\delta > 0$ such that

$$\Sigma(t) \triangleq \sigma(t)\sigma(t)' + \theta(t)D(t)\theta(t)' \geq \delta I \quad \text{for all } t \in [0, T], \tag{2.2}$$

  where $D(t) \triangleq \operatorname{diag}(\lambda_1(t), \ldots, \lambda_n(t))$.

The uniform bound on $\lambda_i(t)$ implies that $\mathbb{E}(\int_0^t \lambda_i(s)ds) < \infty$ for all $t \in [0,T]$ from which it follows that the *compensated Poisson process* $M_i(t) \triangleq N_i(t) - \int_0^t \lambda_i(s)\,ds$ is a $\mathbb{G}$-martingale (see Lemma 6.6.3 in [4]). We define the vector process $M(t) \triangleq [M_1(t), \ldots, M_n(t)]'$.

We emphasize again the parameters in our market model (2.1), and in particular the arrival rate intensities $\lambda_i(t)$ of the Poisson processes, are $\mathbb{F}$-predictable processes. Such an assumption is common in the literature on default risk modeling (and particularly in pricing applications) and the reader may consult [4, 11, 21] for more details. Finally, since the market (2.1) is incomplete, *perfect replication* is generally not possible. For this reason, as in [6, 9, 14, 25, 28, 32], we adopt the mean-square error as a measure of *closeness* between the terminal wealth and the liability; see (2.6) below.

Observing that the price of the risky assets can also be written in the form

$$dP_i(t) = P_i(t)[\mu_i(t) + \theta_i(t)\lambda(t)]\,dt + P_i(t)\sigma_i(t)\,dW(t) + P_i(t)\theta_i(t)\,dM(t), \tag{2.3}$$

where $\lambda(t) \triangleq [\lambda_1(t), \ldots, \lambda_n(t)]'$, and denoting by $\pi(t) \triangleq [\pi_1(t), \ldots, \pi_m(t)]'$ the vector of *dollar amounts* invested in the *risky* assets at time t, it is easy to show that the wealth process associated with self-financing investment in (2.3) is

$$\begin{cases} dx(t) = [r(t)x(t) + \pi(t)'b(t)]\,dt + \pi(t)'\sigma(t)\,dW(t) + \pi(t)'\theta(t)\,dM(t), \\ x(0) = x_0, \end{cases} \tag{2.4}$$

where

$$\begin{aligned} b(t) &\triangleq [b_1(t), \ldots, b_m(t)]', \quad b_i(t) \triangleq \mu_i(t) + \theta_i(t)\lambda(t) - r(t), \\ \sigma(t) &\triangleq [\sigma_1(t)', \ldots, \sigma_m(t)']', \\ \theta(t) &\triangleq [\theta_1(t)', \ldots, \theta_m(t)']'. \end{aligned}$$

Note that $\pi_0(t)$, the amount invested in the bond $B(t)$, does not need to be specified since it is determined by the amounts $\pi_1(t), \ldots, \pi_m(t)$ invested in the risky asset and the wealth $x(t)$ at time t through the equation $\pi_0(t) = x(t) - \sum_{i=1}^m \pi_i(t)$. The class of admissible policies is

$$\mathcal{U} = \left\{ \pi : [0, T] \times \Omega \to \mathbb{R}^m \,\middle|\, \pi(t) \text{ is } \mathbb{G}\text{-predictable and } \mathbb{E}\int_0^t |\pi(t)|^2\,dt < \infty \right\}. \tag{2.5}$$

Consider an agent who faces a time T liability ξ. Throughout this paper, we assume that the value of ξ is contingent on the history of the Poisson processes $N(t)$ as well as the Brownian motion $W(t)$. By virtue of this dependence, the investor

faces uncertainty in the value of the liability ξ. One method of reducing this risk is to invest in assets (or *hedging instruments*) that depend, as much as possible, on the same sources of uncertainty $N(t)$ and $W(t)$ that affect the liability. In doing this, a natural objective is to find a *hedging/investment portfolio* $\pi(t)$ such that the terminal value of this investment $x(T)$ is as "close as possible" to the value of ξ. This motivates our model of asset prices (2.1) which are driven by $N(t)$ and $W(t)$, and the following stochastic control problem:

$$\begin{cases} \min\limits_{\pi(\cdot)\in\mathcal{U}} \mathbb{E}[\xi - x(T)]^2 \\ \text{subject to} \\ \quad dx(t) = [r(t)x(t) + \pi(t)'b(t)]\,dt + \pi(t)'\sigma(t)\,dW(t) + \pi(t)'\theta(t)\,dM(t), \\ \quad x(0) = x_0, \\ \quad \pi(\cdot) \in \mathcal{U}. \end{cases} \tag{2.6}$$

In a *complete market* (see section 5.1), an investor *with the appropriate initial wealth* x_0 can eliminate all the risk by replicating ξ; that is, there is a unique value of x_0 and an associated trading strategy $\pi(\cdot)$ such that an investor, starting with x_0 and investing according to $\pi(\cdot)$, will have a terminal wealth satisfying $x(T) = \xi$, $\mathbb{P}$-a.s.; see, for example, [5], which deals with this issue in the context of hedging default risk in a complete market. In the case of an incomplete market, however, perfect replication is usually not possible, no matter what the value of the investor's initial wealth. On the other hand, superreplication (i.e., finding a portfolio such that $x(T) \geq \xi$, $\mathbb{P}$-a.s.) may be possible, but is typically infeasible since the initial wealth required to superreplicate a claim is often too large to be of practical use. As a compromise, an investor in an incomplete market (or, for that matter, in a complete market but with insufficient initial capital to replicate the claim) may seek to solve (2.6).

3. Optimal hedging portfolio. Our solution of the optimal hedging problem (2.6) will involve, in an essential way, the following *backward stochastic differential equations*[1] (BSDEs):

$$\begin{cases} dp(t) = -p(t)\left[2r(t) - \left(b(t) + \dfrac{\sigma(t)\Lambda(t)}{p(t)}\right)'\Sigma(t)^{-1}\left(b(t) + \dfrac{\sigma(t)\Lambda(t)}{p(t)}\right)\right]dt \\ \qquad\qquad + \Lambda(t)'\,dW(t), \\ p(T) = 1, \end{cases} \tag{3.1}$$

$$\begin{cases} dh(t) = \left\{r(t)h(t) + \left(b(t) + \dfrac{\sigma(t)\Lambda(t)}{p(t)}\right)'\Sigma(t)^{-1}\Big(\sigma(t)\eta(t) + \theta(t)D(t)\kappa(t)\Big)\right. \\ \qquad\qquad \left. - \dfrac{\eta(t)'\Lambda(t)}{p(t)}\right\}dt + \eta(t)'\,dW(t) + \kappa(t)'\,dM(t), \\ h(T) = \xi. \end{cases} \tag{3.2}$$

Throughout this paper, a *solution of* (3.1) denotes a pair of processes $(p(t), \Lambda(t))$ such that $p(t)$ is $\mathbb{G}$-*adapted*, *strictly positive*, and *uniformly bounded*, and $\Lambda(t) =$

[1] Although in common use, the term *backward stochastic differential equation* is somewhat misleading in that these equations do not involve time reversal in any way. Furthermore, parameters of these equations as well as the solutions are constrained to be adapted to the "forward" filtration.

$(\Lambda_1(t), \ldots, \Lambda_d(t))'$ is $\mathbb{G}$-*predictable* and *square integrable* under $\mathbb{P}$; that is,

$$(p(t), \Lambda(t)) \in \mathcal{L}^\infty(\mathbb{G}, \mathbb{R}) \times \mathcal{P}^2(\mathbb{G}, \mathbb{R}^d).$$

In this paper, we define a *solution of* (3.2) as a triple $(h(t), \eta(t), \kappa(t))$ such that $h(t)$ is $\mathbb{G}$-*adapted* and *uniformly bounded* and $\eta(t) = (\eta_1(t), \ldots, \eta_d(t))'$ and $\kappa(t) = (\kappa_1(t), \ldots, \kappa_n(t))'$ are $\mathbb{G}$-*predictable* and *square integrable* under $\mathbb{P}$; that is,

$$(h(t), \eta(t), \kappa(t)) \in \mathcal{L}^\infty(\mathbb{G}, \mathbb{R}) \times \mathcal{P}^2(\mathbb{G}, \mathbb{R}^d) \times \mathcal{P}^2(\mathbb{G}, \mathbb{R}^n). \tag{3.3}$$

Note that standard existence and uniqueness results for linear BSDEs driven by Brownian motion and jump processes (such as [31]) do not apply in (3.2) since the coefficient of the component $\eta(t)$ in the drift may be unbounded due to dependence on the square integrable term $\Lambda(t)$. In the case of (3.1), however, there are no terms involving the increment $dM(t)$ since the parameters are assumed to be $\mathbb{F}$-predictable. For this reason, the results obtained in Lim [25] can be applied to establish existence of this equation. This can be summarized as follows.

PROPOSITION 3.1. *Suppose that Assumption* (A) *holds. Then there exists a unique solution* $(p(t), \Lambda(t))$ *of* (3.1) *Moreover, there are finite constants* $0 < \delta_1 < \delta_2 < \infty$ *such that* $\delta_1 \le p(t) \le \delta_2$ *for all* $t \in [0, T]$, $\mathbb{P}$*-a.s. Finally, the stochastic differential equation*

$$\begin{cases} d\rho(t) = -\rho(t)\gamma(t)'\, dW(t), \\ \rho(0) = 1, \end{cases} \tag{3.4}$$

where

$$\gamma(t) \triangleq \sigma(t)'\Sigma(t)^{-1}\left(b(t) + \frac{\sigma(t)\Lambda(t)}{p(t)}\right) - \frac{\Lambda(t)}{p(t)},$$

has a unique solution $\rho(t) = e^{-\frac{1}{2}\int_0^t |\gamma(s)|^2 ds - \int_0^t \gamma(s)' dW(s)}$ *and* $\rho(t)$ *is a strictly positive square integrable martingale.*

Proof. Existence and uniqueness of a solution $(p(t), \Lambda(t)) \in \mathcal{L}^\infty(\mathbb{G}, \mathbb{R}) \times \mathcal{P}^2(\mathbb{G}, \mathbb{R}^d)$ of (3.1) follows from Theorem 5.1 of [25]. The existence of positive constants δ_1 and δ_2 such that $\delta_1 \le p(t) \le \delta_2$ is shown in the proof of this same theorem. That $\rho(t)$ is a strictly positive square integrable martingale follows from Theorem 4.1 in [25]. □

The (martingale) density process $\rho(t)$ in Proposition 3.1 is related to the Radon–Nikodým derivative that defines the $\mathbb{P}$-equivalent probability measure known as the *variance optimal martingale measure* (VMM), which is a fundamental object associated with the mean-variance hedging problem; see, for example, [9, 14, 23, 32] for more on the VMM, and see [25] for the connection between the nonlinear BSDE (3.1) and the VMM in the case of Brownian information. In this regard, the density process associated with the hedging problem (2.6) (introduced below in (4.2)) may be regarded as a generalization of (3.4) in the case when there are jumps. Further discussion on this point follows Theorem 4.3.

The remainder of this section will be devoted to proving optimality of the portfolio

$$\pi(t) = \Sigma(t)^{-1}\left[\sigma(t)\eta(t) + \theta(t)D(t)\kappa(t) + \left(b(t) + \frac{\sigma(t)\Lambda(t)}{p(t)}\right)(h(t^-) - x(t^-))\right] \tag{3.5}$$

under the assumption that (3.2) *has a solution.* (Solvability of (3.2) will be addressed in section 4.) In order to prove optimality, a number of issues need to be resolved.

First, we need to show that the stochastic differential equation (2.6) for the wealth process $x(t)$ has a solution under (3.5). This is not immediately obvious since the coefficients of $x(t)$ in (2.6) under (3.5) are generally unbounded due to dependence on the square integrable process $\Lambda(t)$. As a consequence, standard existence and uniqueness results from the theory of linear stochastic differential equations do not immediately apply since boundedness of coefficients is usually required for these results to hold. (See, for example, [19].)

A second important issue concerns the admissibility (and in particular square integrability) of (3.5) (see the definition (2.5)), which is an important part of the proof of optimality in Theorem 3.5. Once again, however, square integrability of (3.5) is not immediately apparent since the product of the square integrable process $\Lambda(t)$ and the wealth process $x(t)$ is not necessarily square integrable.

The following results resolve the technical issues discussed above. Proposition 3.2 shows that the wealth process (2.6) under (3.5) has a solution $x(t)$. Proposition 3.3 is a technical result concerning the integrability of solutions of linear BSDEs which is used in the proof of Proposition 3.4 where square integrability (and hence admissibility) of (3.5) is established. Optimality of (3.5) is proven in Theorem 3.5. (A similar optimality proof is given in Hu and Zhou [16], though for a problem that involves neither jumps nor a random terminal condition.) We mention again that the results below are based on the assumption that (3.2) has a solution. Solvability of (3.2) is discussed in a later section.

PROPOSITION 3.2. *Suppose that* (3.2) *has a solution* $(h(t), \eta(t), \kappa(t))$ *satisfying the conditions* (3.3). *Then the stochastic differential equation* (2.6) *under the portfolio* $\pi(t)$ *given by* (3.5) *has a solution* $\bar{x}(t)$.

Proof. A solution of (2.6) under (3.5) can be constructed as follows. Define

$$\begin{cases} dY(t) = -r(t)Y(t)\,dt - \{A(t) + \gamma(t)Y(t)\}'\,dW(t) \\ \qquad\qquad -\{B(t) + \psi(t)Y(t)\}'\,dM(t), \\ Y(0) = p(0)[h(0) - x(0)], \end{cases} \tag{3.6}$$

where $\gamma(t)$ and $\psi(t)$ are defined by

$$\gamma(t) \triangleq \sigma(t)'\Sigma(t)^{-1}\left(b(t) + \frac{\sigma(t)\Lambda(t)}{p(t)}\right) - \frac{\Lambda(t)}{p(t)}, \tag{3.7}$$

$$\psi(t) \triangleq \theta(t)'\Sigma(t)^{-1}\left(b(t) + \frac{\sigma(t)\Lambda(t)}{p(t)}\right), \tag{3.8}$$

and

$$A(t) \triangleq p(t)\Big[\sigma(t)\Sigma(t)^{-1}\big(\sigma(t)\eta(t) + \theta(t)D(t)\kappa(t)\big) - \eta(t)\Big],$$
$$B(t) = p(t)\Big[\theta(t)'\Sigma(t)^{-1}\big(\sigma(t)\eta(t) + \theta(t)D(t)\kappa(t)\big) - \kappa(t)\Big].$$

Observe that $\gamma(t)$ and $\psi(t)$ are *square integrable*. We denote the components of $\gamma(t)$ and $\psi(t)$ by $\gamma_i(t)$ and $\psi_i(t)$; that is, $\gamma(t) \triangleq [\gamma_1(t), \ldots, \gamma_n(t)]'$ and $\psi(t) \triangleq [\psi_1(t), \ldots, \psi_n(t)]'$. Denoting $\Delta N_j(t) \triangleq N_j(t) - N_j(t^-)$, it can be shown (using Ito's formula) that $Y(t) = \Phi(t)\{Y(0) + Z(t)\}$, where

$$\Phi(t) = e^{\int_0^t [-r(s) - \frac{1}{2}|\gamma(s)|^2 + \psi(s)'\lambda(s)]\,ds - \int_0^t \gamma(s)'\,dW(s)} \times \prod_{i=1}^{n} \prod_{0<s_i\le t} (1 - \psi_i(s_i)\Delta N_i(s_i)) \tag{3.9}$$

and

$$Z(t) \triangleq -\int_0^t \Phi(s)^{-1}\left[\gamma(s)'B(t) + \sum_{i=1}^n \frac{\psi_i(s)}{1-\psi_i(s)}\lambda_i(s)A_i(s)\right]ds$$
$$-\int_0^t \Phi(s)^{-1}B(s)'\,dW(s) - \sum_{i=1}^n \int_0^t \Phi(s)^{-1}\frac{A_i(s)}{1-\psi_i(s)}\,dM_i(s). \tag{3.10}$$

Note that (3.9) and (3.10) are well-defined processes. Finally, it can be shown using Ito's formula that $\bar{x}(t) \triangleq h(t) - Y(t)/p(t)$ is a solution of (2.6) when the portfolio is (3.5), which implies in turn that the wealth process under (3.5) is well defined. □

The following technical result is required in the proof of Proposition 3.4.

PROPOSITION 3.3. *Suppose that $r(t)$, $\alpha(t)$, $\beta(t)$, and $\lambda_1(t), \ldots, \lambda_n(t)$ are uniformly bounded $\mathbb{G}$-predictable processes on $[0, T]$, τ is a $\mathbb{G}$-stopping time, and $Y \in \mathcal{G}_\tau$ satisfies $\mathbb{E}|Y|^2 < \infty$. Then the BSDE*

$$\begin{cases} dy(t) = \Big[r(t)y(t) + \alpha(t)'q(t) + \beta(t)'z(t)\Big]\,dt + q(t)'\,dW(t) + z(t)'\,dM(t), \\ y(\tau) = Y \end{cases} \tag{3.11}$$

has a unique solution

$$(y(t),\, z(t),\, q(t)) \in \mathcal{L}^2(\mathbb{G},\, \mathbb{R}) \times \mathcal{P}^2(\mathbb{G},\, \mathbb{R}^d) \times \mathcal{P}^2(\mathbb{G},\, \mathbb{R}^n).$$

Moreover, there is a constant $c < \infty$ that depends only on $r(t)$, $\alpha(t)$, $\beta(t)$, and $\lambda_1(t), \ldots, \lambda_n(t)$ (but not the stopping time τ) such that

$$\mathbb{E}\int_0^\tau \left[|q(t)|^2 + \sum_{i=1}^n \lambda_i(t)|z_i(t)|^2\right]ds \le 2\mathbb{E}|Y|^2 e^{2c[1+c(n+1)]T}. \tag{3.12}$$

Proof. Existence and uniqueness for (3.11) can be shown as in Theorem 1 of [31] and the bound (3.12) can be derived along the lines of Lemma 1 in [31]. Because of constraints on the length of this paper, details have not been provided but can be obtained from the author upon request. □

The following result establishes admissibility of (3.5).

PROPOSITION 3.4. *Suppose that* (3.2) *has a solution* $(h(t), \eta(t), \kappa(t))$ *such that* (3.3) *is satisfied. Then the portfolio* $\pi(t)$ *given by* (3.5) *is square integrable and hence admissible.*

Proof. Throughout this proof, $\pi(t)$ denotes the portfolio (3.5) and $\bar{x}(t)$ denotes the solution of the wealth process (2.6) under (3.5). (Recall, by Proposition 3.2, that (2.6) has a solution under (3.5).) Since (3.1) and (3.2) have solutions, the process $p(t)[h(t) - \bar{x}(t)]^2$ is well defined and Ito's formula gives

$$p(t)(h(t) - \bar{x}(t))^2 = p(0)(h(0) - \bar{x}(0))^2$$
$$+ \int_0^t p(t)\Big\{\kappa(t)'D(t)\kappa(t) + \eta(t)'\eta(t)$$
$$-[\sigma(t)\eta(t) + \theta(t)D(t)\kappa(t)]'\Sigma(t)^{-1}[\sigma(t)\eta(t) + \theta(t)D(t)\kappa(t)]\Big\}\,dt$$
$$+ \int_0^t \Big[(h(t) - \bar{x}(t))^2\Lambda(t) + 2p(t)(h(t) - \bar{x}(t))(\eta(t) - \sigma(t)'\pi(t))\Big]'\,dW(t)$$

$$+\int_0^t \sum_{i=1}^n p(t)(\kappa(t)-\theta(t)'\pi(t))_i^2\,dM_i(t)$$
$$+\int_0^t 2p(t)(h(t^-)-\bar{x}(t^-))(\kappa(t)-\theta(t)'\pi(t))'\,dM(t).$$

(A similar calculation for the case of general $\pi(t)$ is given in (3.21) below.) Noting that the stochastic integrals are local martingales, there exists an increasing sequence of stopping times $\{\tau_i\}$ such that $\tau_i \uparrow T$ as $i \to \infty$ and

$$\begin{aligned}(3.13)\qquad &\mathbb{E}\{p(T\wedge\tau_i)(h(T\wedge\tau_i)-\bar{x}(T\wedge\tau_i))^2\} = p(0)(h(0)-\bar{x}(0))^2\\ &\quad+\mathbb{E}\int_0^{T\wedge\tau_i} p(t)\Big\{\kappa(t)'D(t)\kappa(t)+\eta(t)'\eta(t)\\ &\quad-[\sigma(t)\eta(t)+\theta(t)D(t)\kappa(t)]'\Sigma(t)^{-1}[\sigma(t)\eta(t)+\theta(t)D(t)\kappa(t)]\Big\}\,dt.\end{aligned}$$

Since there is a constant $\delta>0$ such that $p(t)\geq\delta$ for all $t\in[0,T]$, $\mathbb{P}$-a.s. (Proposition 3.1), it follows that

$$\begin{aligned}(3.14)\qquad &\delta\mathbb{E}[h(T\wedge\tau_i)-\bar{x}(T\wedge\tau_i)]^2\\ &\quad\leq \mathbb{E}[p(T\wedge\tau_i)(h(T\wedge\tau_i)-\bar{x}(T\wedge\tau_i))^2]\\ &\quad\leq p(0)(h(0)-x(0))^2+\mathbb{E}\int_0^T p(t)\Big\{\kappa(t)'D(t)\kappa(t)+\eta(t)'\eta(t)\\ &\quad-[\sigma(t)\eta(t)+\theta(t)D(t)\kappa(t)]'\Sigma(t)^{-1}[\sigma(t)\eta(t)+\theta(t)D(t)\kappa(t)]\Big\}\,dt\end{aligned}$$

(where the second inequality follows from (3.13), the nonnegativity of the integrand, and the fact that $T\wedge\tau_i\leq T$). In other words, $h(T\wedge\tau_i)-\bar{x}(T\wedge\tau_i)\in\mathcal{L}^2(\mathbb{G},\mathbb{R})$. Finally, noting (by assumption) that $h(t)$ is uniformly bounded (since, by assumption, (3.3) is satisfied), it follows that

$$\bar{x}(T\wedge\tau_i)=h(T\wedge\tau_i)-[h(T\wedge\tau_i)-\bar{x}(T\wedge\tau_i)]\in\mathcal{L}^2(\mathbb{G},\mathbb{R}).$$

We have shown that the wealth-portfolio pair $(\bar{x}(t),\pi(t))$ given by (2.6) and (3.5) satisfy the system of equations

$$(3.15)\qquad \begin{cases} dy(t)=\{r(t)y(t)+b(t)'\pi(t)\}\,dt+\pi(t)'\sigma(t)\,dW(t)+\pi(t)'\theta(t)\,dM(t),\\ y(T\wedge\tau_i)=\bar{x}(\tau_i\wedge T),\end{cases}$$

where $\bar{x}(T\wedge\tau_i)$ is a square integrable $\mathcal{G}_{T\wedge\tau_i}$-measurable random variable. Setting

$$q(t)=\sigma(t)'\pi(t),\quad z(t)=\theta(t)'\pi(t)$$

or, equivalently,

$$(3.16)\qquad \pi(t)=\Sigma(t)^{-1}[\sigma(t)q(t)+\theta(t)D(t)z(t)]$$

and substituting into (3.15), it follows that $(y(t),\,q(t),\,z(t))=(\bar{x}(t),\,\sigma(t)'\bar{\pi}(t),\theta(t)'\bar{\pi}(t))$ is the solution of the following BSDE on the random time horizon $[0,\,T\wedge\tau_i]$:

$$(3.17)\qquad \begin{cases} dy(t)=\Big[r(t)y(t)+b(t)'\Sigma(t)^{-1}\sigma(t)q(t)+b(t)'\Sigma(t)^{-1}\theta(t)D(t)z(t)\Big]\,dt\\ \qquad\qquad+q(t)'\,dW(t)+z(t)'\,dM(t),\quad t\in[0,\,T\wedge\tau_i),\\ y(T\wedge\tau_i)=\bar{x}(T\wedge\tau_i).\end{cases}$$

In particular, (3.17) is a linear BSDE with a square integrable terminal condition $y(T\wedge\tau_i)=\bar{x}(T\wedge\tau_i)$ at the stopping time $T\wedge\tau_i$ with (by Assumption (A)) uniformly bounded parameters $r(t)$, $b(t)'\Sigma(t)^{-1}\sigma(t)$, $b(t)'\Sigma(t)^{-1}\theta(t)D(t)$, and $\lambda_1(t),\dots,\lambda_n(t)$. It follows immediately from Proposition 3.3, and particularly the bound (3.12), that there is a constant $c<\infty$ (which depends only on the parameters $r(t)$, $b(t)$, $\sigma(t)$, $\theta(t)$, and $\lambda_i(t)$ but not the stopping time τ_i) such that

$$\begin{aligned}(3.18)\qquad &\mathbb{E}\biggl\{\int_0^{T\wedge\tau_i}|q(s)|^2\,ds+\sum_{i=1}^n\int_0^{T\wedge\tau_i}\lambda_i(s)|z_i(s)|^2\,ds\biggr\}\\&\quad\le 2\mathbb{E}|\bar{x}(T\wedge\tau_i)|^2e^{2c[1+c(1+n)]T}.\end{aligned}$$

Furthermore, since

$$(3.19)\qquad \mathbb{E}|\bar{x}(T\wedge\tau_i)|^2\le 2\mathbb{E}|h(T\wedge\tau_i)|^2+2\mathbb{E}|\bar{x}(T\wedge\tau_i)-h(T\wedge\tau_i)|^2\le K,$$

where $K<\infty$ is a constant independent of i (by virtue of the uniform bound on $h(t)$ and the bound (3.14)), it follows from (3.18) and (3.19) that

$$\mathbb{E}\biggl\{\int_0^{T\wedge\tau_i}|q(t)|^2\,dt+\sum_{i=1}^n\int_0^{T\wedge\tau_i}\lambda_i(t)|z_i(t)|^2\,dt\biggr\}\le 2Ke^{2c[1+c(1+n)]T}<\infty,$$

and the monotone convergence theorem gives

$$\mathbb{E}\biggl\{\int_0^{T}|q(t)|^2\,dt+\sum_{i=1}^n\int_0^{T}\lambda_i(t)|z_i(t)|^2\,dt\biggr\}<\infty.$$

The square integrability of (3.5) follows from the relationship (3.16) between $\pi(t)$ and $(q(t),z(t))$. □

The following result establishes optimality of (3.5).

THEOREM 3.5. *Assume that* (3.2) *has a solution*

$$(h(t),\eta(t),\kappa(t))\in\mathcal{L}^\infty(\mathbb{G},\mathbb{R})\times\mathcal{P}^2(\mathbb{G},\mathbb{R}^d)\times\mathcal{P}^2(\mathbb{G},\mathbb{R}^n).$$

Then (3.5) *is the optimal hedging portfolio for* (2.6). *The optimal cost is*

$$\begin{aligned}J^*=p(0)(h(0)-x(0))^2+\mathbb{E}\int_0^T p(t)\Bigl\{&\eta(t)'\eta(t)+\kappa(t)'D(t)\kappa(t)\\&-[\sigma(t)\eta(t)+\theta(t)D(t)\kappa(t)]'\Sigma(t)^{-1}[\sigma(t)\eta(t)+\theta(t)D(t)\kappa(t)]\Bigr\}\,dt.\end{aligned}\tag{3.20}$$

Proof. Let $\pi(t)$ be an *arbitrary* admissible policy and $x(t)$ the associated wealth process. From Ito's formula,

$$\begin{aligned}&d\{p(t)(h(t)-x(t))^2\}\\&\quad=\biggl\{(h(t)-x(t))^2\\&\qquad\times\Bigl[-2r(t)p(t)+p(t)\Bigl(b(t)+\frac{\sigma(t)\Lambda(t)}{p(t)}\Bigr)'\Sigma(t)^{-1}\Bigl(b(t)+\frac{\sigma(t)\Lambda(t)}{p(t)}\Bigr)\Bigr]\\&\qquad+2r(t)p(t)(h(t)-x(t))^2+2p(t)(h(t)-x(t))\end{aligned}$$

$$
\begin{aligned}
&\times\Big[\Big(b(t)+\frac{\sigma(t)\Lambda(t)}{p(t)}\Big)'\Sigma(t)^{-1}[\sigma(t)\eta(t)+\theta(t)D(t)\kappa(t)]-\frac{\eta(t)'\Lambda(t)}{p(t)}\Big]\\
&+p(t)[\kappa(t)'D(t)\kappa(t)]+p(t)\pi(t)\Sigma(t)\pi(t)\\
&-2p(t)\pi(t)'[\sigma(t)\eta(t)+\theta(t)D(t)\kappa(t)+b(t)(h(t)-x(t))]\\
&+2(h(t)-x(t))(\eta(t)-\sigma(t)'\pi(t))'\Lambda(t)\Big\}\,dt\\
&+\Big[(h(t)-x(t))^2\Lambda(t)+2p(t)(h(t)-x(t))(\eta(t)-\sigma(t)'\pi(t))\Big]'dW(t)\\
&+\sum_{i=1}^{n}p(t)(\kappa(t)-\theta(t)'\pi(t))_i^2\,dM_i(t)\\
&+2p(t)(h(t^-)-x(t^-))(\kappa(t)-\theta(t)'\pi(t))'dM(t).
\end{aligned}
\tag{3.21}
$$

Since the stochastic integrals are local martingales, there is a sequence of stopping times $\{\tau_i\}$ such that $\tau_i\uparrow T$ as $i\uparrow\infty$ and

$$
\begin{aligned}
&\mathbb{E}\,[p(T\wedge\tau_i)(h(T\wedge\tau_i)-x(T\wedge\tau_i))^2]\\
&=p(0)(h(0)-x(0))^2+\mathbb{E}\int_0^{T\wedge\tau_i}p(t)\Big\{\kappa(t)'D(t)\kappa(t)+\eta(t)'\eta(t)\\
&-[\sigma(t)\eta(t)+\theta(t)D(t)\kappa(t)]'\Sigma(t)^{-1}[\sigma(t)\eta(t)+\theta(t)D(t)\kappa(t)]\Big\}\,dt\\
&+\mathbb{E}\int_0^{T\wedge\tau_i}p(t)\Big[\pi(t)-\Sigma(t)^{-1}\Big(\sigma(t)\eta(t)+\theta(t)D(t)\kappa(t)\\
&+\Big(b(t)+\frac{\sigma(t)\Lambda(t)}{p(t)}\Big)(h(t^-)-x(t^-))\Big)\Big]\\
&\times\Sigma(t)\Big[\pi(t)-\Sigma(t)^{-1}\Big(\sigma(t)\eta(t)+\theta(t)D(t)\kappa(t)\\
&+\Big(b(t)+\frac{\sigma(t)\Lambda(t)}{p(t)}\Big)(h(t^-)-x(t^-))\Big)\Big]\,dt,
\end{aligned}
$$

where the integrand in the expression above is obtained, after several (long!) lines of algebra, from the integrand for the finite variation term in (3.21). Finally, noting that $p(t)$ is uniformly bounded (Proposition 3.1), $h(t)$ is uniformly bounded (by assumption), and $\mathbb{E}[\sup_{t\in[0,\,T]}|x(t)|^2]<\infty$, it follows from the dominated convergence theorem (on the left-hand side) and the monotone convergence theorem (on the right) that

$$
\begin{aligned}
&\mathbb{E}p(T)(h(T)-x(T))^2\\
&=p(0)(h(0)-x(0))^2+\mathbb{E}\int_0^{T}p(t)\Big\{\kappa(t)'D(t)\kappa(t)+\eta(t)'\eta(t)\\
&-[\sigma(t)\eta(t)+\theta(t)D(t)\kappa(t)]'\Sigma(t)^{-1}[\sigma(t)\eta(t)+\theta(t)D(t)\kappa(t)]\Big\}\,dt\\
&+\mathbb{E}\int_0^{T}p(t)\Big[\pi(t)-\Sigma(t)^{-1}\Big(\sigma(t)\eta(t)+\theta(t)D(t)\kappa(t)\\
&+\Big(b(t)+\frac{\sigma(t)\Lambda(t)}{p(t)}\Big)(h(t^-)-x(t^-))\Big)\Big]\Sigma(t)\Big[\pi(t)-\Sigma(t)^{-1}\Big(\sigma(t)\eta(t)+\theta(t)D(t)\kappa(t)\\
&+\Big(b(t)+\frac{\sigma(t)\Lambda(t)}{p(t)}\Big)(h(t^-)-x(t^-))\Big)\Big]\,dt.
\end{aligned}
$$

The claim in Theorem 3.5 follows immediately from this equation and the fact that $p(T) = 1$ and $h(T) = \xi$. □

4. Existence of solutions for (3.2): General results. The solution of (2.6), as stated in Theorem 3.5, depends on the solvability of (3.1)–(3.2). While solvability of (3.1) can be established using the results from [25], which can be applied since the parameters are $\mathbb{F}$-predictable and bounded (see Proposition 3.1), solvability of (3.2) is not so clear. In particular, (3.2) may have unbounded parameters (due to dependence on the component $\Lambda(t)$ of the solution of (3.1)), and for this reason standard existence results for BSDEs driven by jump processes (such as [31]) do not apply.

In the following two sections, we address the question of existence of solutions of (3.2). We begin by presenting a general *"martingale condition"* under which solvability of (3.2) can be guaranteed (Theorem 4.3). This condition (which can be stated in terms of a certain local martingale being a strictly positive martingale) is required in order to construct a solution of (3.2), and is analogous to the assumption in [1] that the variance optimal (signed) martingale measure is a $\mathbb{P}$-equivalent *probability* measure. Following this, we show in Theorem 4.4 that strict positivity of the local martingale in the "martingale condition" is not only necessary, but also sufficient for the "martingale condition" to hold.

Recall the processes $\gamma(t)$ and $\psi(t)$ defined in (3.7)–(3.8). Observe that $\gamma(t)$ and $\psi(t)$ are square integrable $\mathbb{G}$-predictable processes under $\mathbb{P}$. We can rewrite (3.2) as

$$\begin{cases} dh(t) = r(t)h(t)\,dt + \eta(t)'[\gamma(t)\,dt + dW(t)] + \kappa(t)'[D(t)\psi(t) + dM(t)], \\ h(T) = \xi. \end{cases} \tag{4.1}$$

We can construct a solution of (4.1) using the Girsanov transformation and the martingale representation theorem for jump-diffusion processes driven by Brownian motion and doubly stochastic Poisson processes (see Propositions 4.1 and 4.2). In this regard, consider the following stochastic differential equation:

$$\begin{cases} dY(t) = -Y(t^-)\Big\{\gamma(t)'\,dW(t) + \psi(t)'\,dM(t)\Big\}, \\ Y(0) = 1. \end{cases} \tag{4.2}$$

It is easy to show that $Y(t) = \rho(t)\zeta(t)$, where

$$\begin{aligned} d\rho(t) &= -\rho(t^-)\gamma(t)'\,dW(t), \quad \rho(0) = 1, \\ d\zeta(t) &= -\zeta(t^-)\psi(t)'\,dM(t), \quad \zeta(0) = 1. \end{aligned}$$

We can write the solution of these equations as

$$\begin{aligned} \rho(t) &= e^{-\frac{1}{2}\int_0^t |\gamma(s)|^2 ds - \int_0^t \gamma(s)' dW(s)}, \\ \zeta(t) &= e^{\int_0^t \psi(s)'\lambda(s) ds} \prod_{i=1}^{n} \prod_{0<s_i\le t} \Big(1 - \psi_i(s_i)\Delta N(s_i)\Big), \end{aligned}$$

where $\ln(\mathbf{1} - \psi(s))$ is an n-dimensional column vector with entries $\ln(1 - \psi_i(s))$. Assuming that $Y(t)$ is a positive $\mathbb{G}$-martingale under $\mathbb{P}$, we can define a probability measure $\mathbb{Q}$ equivalent to $\mathbb{P}$ on $(\Omega, \mathcal{G}_T)$ by

$$\frac{d\mathbb{Q}}{d\mathbb{P}}\Big|_{\mathcal{G}_T} = Y(T), \quad \mathbb{P}\text{-a.s.} \tag{4.3}$$

The following is taken from [4, Proposition 6.6.8] (see also [10, Proposition 6, p. 361]).

PROPOSITION 4.1 (Girsanov). *Assume that $Y(t)$ is a positive $\mathbb{G}$-martingale under $\mathbb{P}$ and that the Radon–Nikodým density of $\mathbb{Q}$ with respect to $\mathbb{P}$ is given by* (4.2)–(4.3). *Then the process $\bar{W}(t) = W(t) + \int_0^t \gamma(s)\,ds$ is a $\mathbb{G}$-Brownian motion under $\mathbb{Q}$, and $\bar{M}(t) = M(t) + \int_0^t D(s)\psi(s)\,ds = N(t) - \int_0^t D(s)[\mathbf{1} - \psi(s)]\,ds$ is a $\mathbb{G}$-martingale under $\mathbb{Q}$. In addition, if $\psi(t)$ is $\mathbb{F}$-predictable, then $N(t)$ is an $\mathbb{F}$-conditional Poisson process with respect to $\mathbb{G}$ under $\mathbb{Q}$ with intensity $D(t)[\mathbf{1} - \psi(t)]$.*

The following result can be obtained by a fairly straightforward extension of the proof of martingale representation theorem for continuous martingales with respect to a Brownian filtration (see, for example, [33]). For more results on martingale representation for processes other than Brownian motion, see [34].

PROPOSITION 4.2 (martingale representation). *Let $\{Z(t)\}_{t\in[0,\,T]}$ be a square integrable $\mathbb{G}$-martingale under $\mathbb{P}$. Then, there are unique square integrable $\mathbb{G}$-predictable processes $f(t)$ and $g_1(t), \ldots, g_n(t)$ such that*

$$Z(t) = Z(0) + \int_0^t f(s)'\,dW(s) + \sum_{i=1}^{n} \int_0^t g_i(s)'\,dM_i(s). \tag{4.4}$$

The following result gives general conditions under which (3.2) can be solved.

THEOREM 4.3. *Suppose that Assumption* (A) *is satisfied. If the solution $Y(t)$ of* (4.2) *is a strictly positive $\mathbb{G}$-martingale under $\mathbb{P}$, then the BSDE* (3.2) *has a unique solution $(h(t), \eta(t), \kappa(t))$ such that $h(t)$ is uniformly bounded and*

$$\mathbb{E}\int_0^T \left\{ |\eta(t)|^2 + \sum_{i=1}^{n} \lambda_i(t)|\kappa_i(t)|^2 \right\} dt < \infty. \tag{4.5}$$

Before presenting the proof of Theorem 4.3, the following remarks are in order. Recall that the set of all $\mathbb{P}$-equivalent probability measures $\mathbb{Q}$ can be represented by (4.3) and a pair of $\mathbb{G}$-predictable processes $(\gamma(t), \psi(t))$ such that $Y(t)$ is a positive martingale. The equivalent *martingale* measures (EMMs) is the set of $\mathbb{P}$-equivalent measures under which discounted price processes $P_i(t)/B(t)$ obtained from (2.1) are martingales. Using this characterization and the model (2.1) for the price processes, it can be shown that any pair $(\gamma(t), \psi(t))$ associated with an EMM can be written in the form

$$\begin{bmatrix} \gamma \\ \psi \end{bmatrix} = \begin{bmatrix} \sigma'\Sigma^{-1}b + (I - \sigma'\Sigma^{-1}\sigma)Z_1 - \sigma'\Sigma^{-1}\theta D^{\frac{1}{2}} Z_2 \\ \theta'\Sigma^{-1}b - \theta'\Sigma^{-1}\sigma Z_1 - D^{\frac{1}{2}}(I - D^{\frac{1}{2}}\theta'\Sigma^{-1}\theta D^{\frac{1}{2}})Z_2 \end{bmatrix} \tag{4.6}$$

for some choice of $\mathbb{G}$-predictable processes $(Z_1(t), Z_2(t))$, where $D(t)^{\frac{1}{2}} \triangleq \operatorname{diag}(\lambda_1(t)^{\frac{1}{2}}, \ldots, \lambda_n(t)^{\frac{1}{2}})$. In particular, the (nonempty) set of EMMs is not a singleton when there are no arbitrage opportunities and the market is incomplete. Comparing (4.6) with (3.7)–(3.8) we see that the SRE chooses the EMM corresponding to

$$Z_1(t) = -\frac{\Lambda(t)}{p(t)}, \quad Z_2(t) = 0.$$

When $\theta \equiv 0$, which corresponds to the case when the price processes (2.1) are driven by Brownian motion and are independent of the jump processes, the EMM induced by $(p(t), \Lambda(t))$ coincides with the so-called *variance optimal martingale measure* associated with the mean-variance hedging when the price processes are driven

by Brownian motion; see [9, 14, 23, 25, 32] as well as the remarks following Proposition 3.1.

The proof of Theorem 4.3 is as follows.

Proof. We prove this result by constructing the solution of (4.1).

By assumption, we have $Y(T) > 0$ a.s. and $\mathbb{E}^{\mathbb{P}}[Y(T)] = 1$ so we can define a probability measure $\mathbb{Q}$ that is equivalent to $\mathbb{P}$ with Radon–Nikodým derivative (4.3). Moreover, it follows from the Girsanov theorem (Proposition 4.1) that $\bar{W}(t) = W(t) + \int_0^t \gamma(s)\,ds$ is a $\mathbb{G}$-Brownian motion under $\mathbb{Q}$ and, from the $\mathbb{F}$-predictability of $\psi_i(t)$, that $N_i(t)$ is a doubly stochastic Poisson process under $\mathbb{Q}$ with $\mathbb{F}$-predictable intensity $\lambda_i(t)(1-\psi_i(t))$.

Define

$$h(t) = B(t)\,\mathbb{E}^{\mathbb{Q}}\left[\frac{\xi}{B(T)} \;\middle|\; \mathcal{G}_t\right].$$

It follows that $h(t)/B(t)$ is a $\mathbb{G}$-martingale with respect to the probability measure $\mathbb{Q}$. Furthermore, since ξ is uniformly bounded, it follows that $h(t)/B(t)$ is uniformly bounded. The uniform boundedness of $h(t)$ now follows from the fact that $B(t)$ is uniformly bounded. From the martingale representation theorem (Proposition 4.2) there are $\mathbb{G}$-predictable $\mathbb{Q}$-square integrable processes $\bar{\eta}(t)$ and $\bar{\kappa}(t)$ such that

$$\frac{h(t)}{B(t)} = \mathbb{E}^{\mathbb{Q}}\Big[\frac{\xi}{B(T)}\Big] + \int_0^t \bar{\eta}(s)'\,d\bar{W}(s) + \int_0^t \bar{\kappa}(s)'\,d\bar{M}(s), \tag{4.7}$$

where $\bar{M}_i(t) \triangleq N_i(t) - \int_0^t \lambda_i(s)(1-\psi_i(s))\,ds$ is a $\mathbb{G}$-martingale with respect to $\mathbb{Q}$. Applying Ito's formula to (4.7), we obtain

$$\begin{cases} dh(t) = r(t)h(t)\,dt + \eta(t)'\,d\bar{W}(t) + \kappa(t)'\,d\bar{M}(t), \\ h(T) = \xi, \end{cases}$$

where $\eta(t) \triangleq B(t)\bar{\eta}(t)$ and $\kappa(t) \triangleq B(t)\bar{\kappa}(t)$. Changing measure from $\mathbb{Q}$ back to $\mathbb{P}$ shows that $(h(t), \eta(t), \kappa(t))$ is the solution of (4.1), as required. Uniqueness can be seen by carrying out the reverse of this procedure and using the uniqueness of the representation (4.7).

Next we show the integrability properties (4.5) are satisfied. (Note that (4.5) involves an expectation under $\mathbb{P}$, whereas $\bar{\eta}(t)$ and $\bar{\kappa}(t)$ are only $\mathbb{Q}$-square integrable.) Since $B(t)$ is uniformly bounded, (4.5) can be shown by establishing the inequality

$$\mathbb{E}^{\mathbb{P}} \int_0^T \left\{ |\bar{\eta}(t)|^2 + \sum_{i=1}^n \lambda_i(t)|\bar{\kappa}_i(t)|^2 \right\} dt < \infty.$$

Let $Z(t) \triangleq h(t)/B(t)$. Since ξ is uniformly bounded under $\mathbb{Q}$ and $\mathbb{P}$ is equivalent to $\mathbb{Q}$, there is a constant $C < \infty$ such that $|Z(t)| < C$ for all $t \in [0, T]$, $\mathbb{P}$-a.s. It follows from (4.7) that

$$\begin{aligned} Z(t) = Z(0) &+ \int_0^t [\bar{\eta}(s)'\gamma(s) + \bar{\kappa}(s)'D(s)\psi(s) - \bar{\kappa}(s)'\lambda(s)]\,ds \\ &+ \int_0^t \bar{\eta}(s)'\,dW(t) + \int_0^t \bar{\kappa}(s)'\,dN(s). \end{aligned}$$

From Ito's formula,

$$Z(t)^2 = Z(0)^2 + \int_0^t \left\{ 2Z(s^-) \left[\bar{\eta}(s)'\gamma(s) + \sum_{i=1}^n \lambda_i(s)\bar{\kappa}_i(s)\psi_i(s) \right] + |\bar{\eta}(s)|^2 + \sum_{i=1}^n \lambda_i(s)\bar{\kappa}_i(s)^2 \right\} ds + \int_0^t 2Z(s^-)\bar{\eta}(s)'\, dW(s) + \sum_{i=1}^n \int_0^t [2\bar{\kappa}_i(s)Z(s^-) + \bar{\kappa}_i(s)^2]\, dM_i(s).$$

The stochastic integrals above are local martingales, and hence there is a sequence of stopping times $\{\tau_i\}$ such that

$$\mathbb{E}[Z(T\wedge\tau_n)^2] = Z(0)^2 + \mathbb{E}\left[\int_0^{T\wedge\tau_n} \left\{ 2Z(s^-) \left[\bar{\eta}(s)'\gamma(s) + \sum_{i=1}^n \lambda_i(s)\bar{\kappa}_i(s)\psi_i(s) \right] + |\bar{\eta}(s)|^2 + \sum_{i=1}^n \lambda_i(s)\bar{\kappa}_i(s)^2 \right\} ds \right].$$

That is,

$$\begin{aligned} (4.8)\qquad & \mathbb{E}\int_0^{T\wedge\tau_n} \left[|\bar{\eta}(s)|^2 + \sum_{i=1}^n \lambda_i(s)\bar{\kappa}_i(s)^2 \right] ds + Z(0)^2 \\ &= \mathbb{E}[Z(T\wedge\tau_n)^2] - \mathbb{E}\int_0^{T\wedge\tau_n} 2Z(s^-) \left[\bar{\eta}(s)'\gamma(s) + \sum_{i=1}^n \lambda_i(s)\bar{\kappa}_i(s)\psi_i(s) \right] ds \\ &\le \mathbb{E}[Z(T\wedge\tau_n)^2] + \mathbb{E}\int_0^{T\wedge\tau_n} 2C|\bar{\eta}(s)|\,|\gamma(s)|\, ds \\ &\quad + \sum_{i=1}^n \int_0^{T\wedge\tau_n} 2C\lambda_i(s)\,|\bar{\kappa}_i(s)|\,|\psi_i(s)|\, ds, \end{aligned}$$

where we have used the fact that $|Z(t)| \le C$ to obtain the inequality in (4.8). Next, using the inequality $2ab \le a^2 + b^2$, it follows that

$$\begin{aligned} (4.9)\qquad & \mathbb{E}\int_0^{T\wedge\tau_n} 2C|\bar{\eta}(s)|\,|\gamma(s)|\, ds \\ &= \mathbb{E}\int_0^{T\wedge\tau_n} 2C\left(\frac{|\bar{\eta}(s)|}{\delta} \right) \delta|\gamma(s)|\, ds \\ &\le \mathbb{E}\int_0^{T\wedge\tau_n} C\left\{ \frac{|\bar{\eta}(s)|^2}{\delta^2} + \delta^2|\gamma(s)|^2 \right\} ds \\ &= \mathbb{E}\int_0^{T\wedge\tau_n} \left\{ \frac{1}{2}|\bar{\eta}(s)|^2 + 2C^2|\gamma(s)|^2 \right\} ds, \end{aligned}$$

where the last equality follows from choosing the constant $\delta = \sqrt{2C}$. A similar calculation again with $\delta = \sqrt{2C}$ gives

$$\mathbb{E}\int_0^{T\wedge\tau_n} 2C\lambda_i(s)\,|\bar{\kappa}_i(s)|\,|\psi_i(s)|\,ds \tag{4.10}$$
$$\leq \mathbb{E}\int_0^{T\wedge\tau_n}\left\{\frac{\lambda_i(s)}{2}|\bar{\kappa}_i(s)|^2+2C^2\lambda_i(s)|\psi_i(s)|^2\right\}ds.$$

Substituting (4.9) and (4.10) into (4.8) it follows that

$$\mathbb{E}\int_0^{T\wedge\tau_n}\left[|\bar{\eta}(s)|^2+\sum_{i=1}^n\lambda_i(s)\bar{\kappa}_i(s)^2\right]ds+Z(0)^2$$
$$\leq \mathbb{E}[Z(T\wedge\tau_n)^2]+2C^2\mathbb{E}\int_0^{T\wedge\tau_n}\left\{|\gamma(s)|^2+\sum_{i=1}^n\lambda_i(s)|\psi_i(s)|^2\right\}ds$$
$$+\frac{1}{2}\mathbb{E}\int_0^{T\wedge\tau_n}\left[|\bar{\eta}(s)|^2+\lambda_i(s)|\bar{\kappa}_i(s)|^2\right]ds.$$

Rearranging and letting $n\to\infty$ it follows from Fatou's lemma that

$$\frac{1}{2}\mathbb{E}\int_0^{T}\left[|\bar{\eta}(s)|^2+\sum_{i=1}^n\lambda_i(s)\bar{\kappa}_i(s)^2\right]ds+Z(0)^2$$
$$\leq \mathbb{E}|\xi|^2+2C^2\mathbb{E}\int_0^{T}\left\{|\gamma(s)|^2+\sum_{i=1}^n\lambda_i(s)|\psi_i(s)|^2\right\}ds<\infty,$$

which implies (4.5). □

By Theorem 4.3, (3.2) has a unique solution if the local martingale $Y(t)$ is a strictly positive martingale. For this to hold, it is clearly *necessary* that $-\infty<\psi_i(t)<1$ for a.e. $t\in[0,T]$, $\mathbb{P}$-a.s. The following result shows that this condition is also sufficient.

THEOREM 4.4. *Suppose that Assumption* (A) *is satisfied. Then the solution* $Y(t)$ *of* (4.2) *is a strictly positive* $\mathbb{G}$*-martingale with respect to* $\mathbb{P}$ *if and only if*

$$\psi_i(t)<1 \text{ for a.e. } t\in[0,T],\ \mathbb{P}\text{-a.s.} \tag{4.11}$$

In particular, there is a unique solution $(h(t),\eta(t),\kappa(t))$ *of* (3.2) *such that* $h(t)$ *is uniformly bounded and* $(\eta(t),\kappa(t))$ *satisfy the integrability conditions* (4.5) *if* (4.11) *is satisfied.*

Proof. For notational convenience we assume that $W(t)$ and $N(t)$ are one-dimensional processes. The extension to the multidimensional case can be done using the same approach (at the cost of more cumbersome notation).

By Ito's formula it can be shown that $Y(t)=\rho(t)\zeta(t)$, where $\rho(t)$ and $\zeta(t)$ denote the solutions of

$$d\rho(t)=-\rho(t^-)\gamma(t)\,dW(t),\quad \rho(0)=1,$$
$$d\zeta(t)=-\zeta(t^-)\psi(t)\,dM(t),\quad \zeta(0)=1.$$

By Proposition 3.1, $\rho(t)=e^{-\frac{1}{2}\int_0^t|\gamma(s)|^2ds-\int_0^t\gamma(s)'dW(s)}$ is a strictly positive $\mathbb{F}$-martingale under $\mathbb{P}$ and hence, by the *martingale invariance property* (Proposition 2.1), is also a strictly positive $\mathbb{G}$-martingale under $\mathbb{P}$. In addition, it is easy to show that

$$\zeta(t)=e^{\int_0^t\psi(s)\lambda(s)ds}\prod_{0<s\leq t}\Big(1-\psi(s)\triangle N(s)\Big).$$

It follows that strict positivity of $Y(t) = \rho(t)\zeta(t)$ implies that (4.11) is satisfied.

Conversely, suppose that (4.11) is satisfied. This implies that $Y(t) = \rho(t)\zeta(t)$ is strictly positive, so we need only show that $Y(t)$ is a martingale. Observe first that $1-\psi(t) > 0$ for a.e. $t \in [0, T]$, $\mathbb{P}$-a.s. Furthermore, we have $0 < \int_0^T (1-\psi(t))\lambda(t)\,dt < \infty$, $\mathbb{P}$-a.s. Indeed, this follows from the square integrability of $\psi(t)$ (which implies in turn that $\int_0^T (1-\psi(t))\,dt < \infty$) and the uniform bound on $\lambda(t)$ (see Assumption (A)). We now show that $\mathbb{E}^{\mathbb{P}}[\zeta(t)|\mathcal{G}_s \vee \mathcal{F}_T] = \zeta(s)$ for $s < t$. First,

$$
\begin{aligned}
(4.12)\quad & \mathbb{E}^{\mathbb{P}}[\zeta(t)|\mathcal{G}_s \vee \mathcal{F}_T] \\
&= e^{\int_0^t \psi(u)\lambda(u)du} \prod_{0<u\le s} (1-\psi(s)\triangle N(s))\mathbb{E}^{\mathbb{P}}\left[\prod_{s<u\le t}(1-\psi(u)\triangle N(u))\Big|\mathcal{G}_s \vee \mathcal{F}_T\right] \\
&= \zeta(s)e^{\int_s^t \psi(u)\lambda(u)du}\mathbb{E}^{\mathbb{P}}\left[\prod_{s<u\le t}(1-\psi(u)\triangle N(u))\Big|\mathcal{F}_T\right],
\end{aligned}
$$

where the second equality follows from the definition of $\zeta(t)$ and the observation that conditional on $\mathcal{F}_T$, $\sigma\{N(u) - N(s),\ s < u \le t\}$ is independent of $\mathcal{G}_s$ by virtue of the independent increment property of Poisson processes. On the other hand, conditioning on $N(t) - N(s)$ we obtain

$$
\begin{aligned}
(4.13)\quad & \mathbb{E}^{\mathbb{P}}\left[\prod_{s<u\le t}(1-\psi(u)\triangle N(u))\,\Big|\,\mathcal{F}_T\right] \\
&= \mathbb{E}^{\mathbb{P}}\left[\mathbb{E}^{\mathbb{P}}\left\{\prod_{s<u\le t}(1-\psi(u)\triangle N(u))\,\Big|\,\sigma\{N(t)-N(s)\}\vee\mathcal{F}_T\right\}\Big|\,\mathcal{F}_T\right].
\end{aligned}
$$

By Lemma 3.1 in [8], the arrival times of $N(u)$ on $(s, t]$, conditional on $\mathcal{F}_T$ and $N(t) - N(s) = k$, are distributed like k independent random variables on $(s, t]$ with density $\lambda(u)/\int_s^t \lambda(v)\,dv$. It follows that

$$
\mathbb{E}^{\mathbb{P}}\left\{\prod_{s<u\le t}(1-\psi(u)\triangle N(u))\,\Big|\,\sigma\{N(t)-N(s)\}\vee\mathcal{F}_T\right\} = \left[\frac{\int_s^t (1-\psi(u))\lambda(u)\,du}{\int_s^t \lambda(u)\,du}\right]^{N(t)-N(s)},
$$

where the right-hand side is finite since $0 < \int_0^T (1-\psi(t))\lambda(t)\,dt < \infty$ $\mathbb{P}$-a.s. Substituting this into (4.13) we obtain

$$
\mathbb{E}^{\mathbb{P}}\left[\prod_{s<u\le t}(1-\psi(u)\triangle N(u))\,\Big|\,\mathcal{F}_T\right] = e^{-\int_s^t \psi(u)\lambda(u)du}
$$

and (4.12) gives $\mathbb{E}^{\mathbb{P}}[\zeta(t)\,|\,\mathcal{G}_s \vee \mathcal{F}_T] = \zeta(s)$. Finally, since

$$
\begin{aligned}
\mathbb{E}^{\mathbb{P}}[Y(t)|\mathcal{G}_s] &= \mathbb{E}^{\mathbb{P}}[\rho(t)\mathbb{E}^{\mathbb{P}}\{\zeta(t)\,|\,\mathcal{G}_s \vee \mathcal{F}_T\}\,|\,\mathcal{G}_s] = \mathbb{E}^{\mathbb{P}}[\rho(t)\zeta(s)|\mathcal{G}_s] \\
&= \mathbb{E}^{\mathbb{P}}[\rho(t)|\mathcal{G}_s]\zeta(s) = \rho(s)\zeta(s) = Y(s),
\end{aligned}
$$

it follows that $Y(t)$ is a martingale, as claimed.

The existence and uniqueness of solutions of (3.2) satisfying the boundedness and integrability conditions follows from Theorem 4.3. □

From the proof of Theorem 4.4, it is easy to show that $\zeta(t)$ is a strictly positive $\mathbb{G}$-martingale under $\mathbb{P}$ if and only if the condition (4.11) is satisfied.

5. Solvability of (3.2): Special cases. The conditions in Theorems 4.3 and 4.4 for solvability of (3.2) are cumbersome because they involve the solution $(p(t), \Lambda(t))$ of (3.1) in the definition of $\psi(t)$; see (3.8). We now consider some simple special cases where the condition in Theorem 4.4 can be expressed explicitly in terms of the parameters of the problem or otherwise easily checked. Whether there are easily verifiable general conditions remains an open question. In the case of *continuous* price processes, the simplifications resulting from the assumption of a complete market (section 5.1) or deterministic parameters (section 5.2) are well known, being situations where the VMM coincides with the so-called *minimal martingale measure*; see [14, 23, 30]. In sections 5.3 and 5.4, we show that (3.2) and the hedging problem (2.6) may still be solvable even when the martingale condition is not satisfied, so long as the liability ξ is suitably restricted.

5.1. Complete market. In this section, we assume conditions which guarantee completeness of the financial market (2.1) and show, under these assumptions, that (3.2) is solvable. More specifically, we shall assume that $m+d = n$ (that is, the number of risky assets m is equal to the number of independent sources of uncertainty $n+d$) and that the matrix

$$\Gamma(t) \triangleq \begin{bmatrix} \sigma(t) & \theta(t)D(t)^{\frac{1}{2}} \end{bmatrix} \tag{5.1}$$

is invertible. These assumptions imply that the linear equation

$$b(t) = \sigma(t)\gamma^*(t) + \theta(t)D(t)\psi^*(t) = \Gamma(t) \begin{bmatrix} \gamma^*(t) \\ D(t)^{\frac{1}{2}}\psi^*(t) \end{bmatrix} \tag{5.2}$$

has a unique solution $(\gamma^*(t), \psi^*(t))$. In addition, we shall assume that the unique solution $Y^*(t)$ of the stochastic differential equation

$$\begin{cases} dY^*(t) = -Y^*(t^-)\{\gamma^*(t)'\, dW(t) + \psi^*(t)'\, dM(t)\}, \\ Y^*(0) = 1, \end{cases} \tag{5.3}$$

where $(\gamma^*(t), \psi^*(t))$ is the solution of (5.2), is a positive martingale. Under these assumptions, one can show that the market is complete. More specifically, since $\rho^*(t)$ is a positive martingale, we can define a $\mathbb{P}$-equivalent probability measure $\mathbb{Q}$ via the Radon–Nikodým derivative

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = Y^*(T) \tag{5.4}$$

such that $W^*(t) \triangleq W(t) + \int_0^t \gamma^*(s)\, ds$ is a $\mathbb{G}$-Brownian motion and $M^*(t) \triangleq M(t) + \int_0^t D(s)\psi^*(s)\, ds$ is a $\mathbb{G}$-martingale under $\mathbb{Q}$ (Proposition 4.1). Moreover, it is easy to show that the discounted price processes $P_i(t)/B(t)$ are $\mathbb{G}$-martingales under $\mathbb{Q}$, and hence $\mathbb{Q}$ is a $\mathbb{P}$-*equivalent martingale measure* (EMM). To see that this $\mathbb{Q}$ is unique, observe that any positive martingale $Y^*(t)$ satisfying $\mathbb{E}^{\mathbb{P}}Y^*(T) = 1$ is also the solution of an equation of the form (5.3) for appropriately chosen $\mathbb{G}$-predictable processes $(\gamma^*(t), \psi^*(t))$. (See [17, Proposition 6.20] and also p. 162 of [4].) In addition, $(\gamma^*(t), \psi^*(t))$ is necessarily a solution of (5.2) in order for the discounted price processes $P_i(t)/B(t)$ to be martingales. The invertibility of $\Gamma(t)$ implies that there is exactly one solution of (5.2) and hence at most one EMM. That is, invertibility of $\Gamma(t)$ together with the property that the solution $Y^*(t)$ of (5.3) is a positive martingale

imply that the market is *complete*. The following result shows that these conditions imply that (3.2) with parameters (3.7)–(3.8) has a unique solution.

PROPOSITION 5.1. *If Assumption* (A) *holds,* $\Gamma(t)$ *is invertible, and the solution* $Y^*(t)$ *of* (5.3) *is a positive martingale, then* (3.2) *has a unique solution.*

Proof. Denoting

$$X(t) = \begin{bmatrix} \gamma^*(t) \\ D(t)^{\frac{1}{2}}\psi^*(t) \end{bmatrix} \tag{5.5}$$

and noting the invertibility of $D(t)^{\frac{1}{2}} = \operatorname{diag}(\lambda_1(t)^{\frac{1}{2}}, \ldots, \lambda_n(t)^{\frac{1}{2}})$ (see Assumption (A)) it follows that the unique solution $(\gamma^*(t), \psi^*(t))$ of (5.2) can be constructed from (5.5) and the solution $X(t)$ of the linear equation

$$b(t) = \Gamma(t)X(t), \quad \text{a.e. } t \in [0, T], \ \mathbb{P}\text{-a.s.} \tag{5.6}$$

Hence, we shall focus on (5.6) and construct the solution of (5.2) once the solution of (5.6) has been found.

The solution $X(t)$ of (5.6) can be written in the form

$$X(t) = \Gamma(t)'K(t) + \big[I - \Gamma(t)(\Gamma(t)\Gamma(t)')^{-1}\Gamma(t)\big]Z(t) \tag{5.7}$$

for appropriate choices of $K(t)$ and $Z(t)$. In particular, $\Gamma(t)'K(t)$ is the projection of $X(t)$ into the space spanned by the columns of $\Gamma(t)'$, while the vector $[I - \Gamma(t)(\Gamma(t)\Gamma(t)')^{-1}\Gamma(t)]Z(t)$ is the projection of $X(t)$ onto its orthogonal complement. Invertibility of $\Gamma(t)$ implies that

$$I - \Gamma(t)(\Gamma(t)\Gamma(t)')^{-1}\Gamma(t) = 0. \tag{5.8}$$

Substituting (5.7) into (5.6) (and noting (5.8)) gives

$$b(t) = \Gamma(t)\Gamma(t)'K(t) = \Sigma(t)K(t),$$

implying in turn that

$$K(t) = \Sigma(t)^{-1}b(t),$$

where $\Sigma(t)$ is defined by (2.2). It follows from (5.7) that

$$X(t) = \Gamma(t)'\Sigma(t)^{-1}b(t) = \begin{bmatrix} \sigma(t)'\Sigma(t)^{-1}b(t) \\ D(t)^{\frac{1}{2}}\theta(t)'\Sigma(t)^{-1}b(t) \end{bmatrix},$$

and hence, by (5.5), we have

$$\gamma^*(t) = \sigma(t)'\Sigma(t)^{-1}b(t), \quad \psi^*(t) = \theta(t)'\Sigma(t)^{-1}b(t). \tag{5.9}$$

On the other hand, substituting (5.1) into (5.8) and using the definition (5.1) of $\Gamma(t)$ implies

$$\begin{bmatrix} \sigma'\Sigma^{-1}\sigma & \sigma'\Sigma^{-1}\theta D^{\frac{1}{2}} \\ D^{\frac{1}{2}}\theta'\Sigma^{-1}\sigma & D^{\frac{1}{2}}\theta'\Sigma^{-1}\theta D^{\frac{1}{2}} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$

and hence

$$\sigma(t)'\Sigma(t)^{-1}\sigma(t) = I, \quad \theta(t)'\Sigma(t)^{-1}\sigma(t) = 0.$$

It follows from (3.7)–(3.8) that

$$\gamma(t) = \sigma(t)'\Sigma(t)^{-1}b(t), \quad \psi(t) = \theta(t)'\Sigma(t)^{-1}b(t). \tag{5.10}$$

Comparing (5.10) with (5.9) it is clear that $(\gamma(t), \psi(t)) = (\gamma^*(t), \psi^*(t))$. In other words, the density process $Y(t)$ defined by (4.2) and (3.7)–(3.8) coincides with the density process $Y^*(t)$ corresponding to the unique EMM. Therefore, $Y(t)$ is a positive martingale (since $Y^*(t)$ is a positive martingale) and hence, by Theorem 4.3, (3.2) has a solution. □

The following result gives a condition for solvability of (3.2) in terms of the parameters of the problem.

PROPOSITION 5.2. *If Assumption* (A) *holds and* $\Gamma(t)$ *is invertible, then* $\gamma(t)$ *and* $\psi(t)$*, given by* (3.7) *and* (3.8)*, respectively, simplify to*

$$\gamma(t) = \sigma(t)'\Sigma(t)^{-1}b(t), \quad \psi(t) = \theta(t)'\Sigma(t)^{-1}b(t). \tag{5.11}$$

Furthermore, if $\psi_i(t) < 1$ *for a.e.* $t \in [0, T]$, $\mathbb{P}$*-a.s.,* $i = 1, 2, \ldots, n$*, then* (3.2) *has a unique solution.*

Proof. We have already shown in the proof of Proposition 5.1 that invertibility of $\Gamma(t)$ implies (5.11); see (5.10). By Corollary 4.4 and Theorem 4.3, the boundedness assumption of $\ln(1 - \psi_i(t))$ implies solvability of (3.2). □

5.2. Deterministic parameters. If the coefficients $r(t)$, $\mu_i(t)$, $\sigma_i(t)$, $\theta_i(t)$, and $\lambda_i(t)$ are all deterministic, then $\Lambda(t) \equiv 0$, and (3.1)–(3.2) become

$$\begin{cases} \dot{p}(t) = -p(t)[2r(t) - b(t)'\Sigma(t)^{-1}b(t)], \\ p(T) = 1, \end{cases} \tag{5.12}$$

$$\begin{cases} dh(t) = \Big\{ r(t)h(t) + b(t)'\Sigma(t)^{-1}[\sigma(t)\eta(t) + \theta(t)D(t)\kappa(t)] \Big\}\, dt \\ \qquad\qquad + \eta(t)'\, dW(t) + \kappa(t)'\, dM(t), \\ h(T) = \xi. \end{cases}$$

In addition, it follows from (3.7)–(3.8) that

$$\gamma(t) = \sigma(t)'\Sigma(t)^{-1}b(t), \quad \psi(t) = \theta(t)'\Sigma(t)^{-1}b(t).$$

The following result is an immediate consequence of Theorems 4.3 and 4.4.

PROPOSITION 5.3. *Suppose that the coefficients* $r(t)$*,* $\mu_i(t)$*,* $\sigma_i(t)$*,* $\theta_i(t)$*, and* $\lambda_i(t)$ *are all deterministic. If* $\psi_i(t) < 1$ *for a.e.* $t \in [0, T]$*,* $\mathbb{P}$*-a.s. for* $i = 1, \ldots, n$*, then* (3.2) *has a unique solution.*

Although the solvability condition in Proposition 5.3 resembles that in Proposition 5.2, Proposition 5.3 applies to complete and incomplete markets (with deterministic parameters), while Proposition 5.2 applies to complete markets (with possibly random parameters).

5.3. Case $Y(t)$ is not a positive martingale. If the process $Y(t)$ defined by (4.2) is not a strictly positive martingale, which occurs for instance if $\psi_i(t) \not< 1$, as required in Theorem 4.4, then the construction in the proof of Theorem 4.3 cannot be used in general to obtain a solution of (3.2). In this section, we show that while (3.2) may not be solvable for arbitrary ξ, it may nevertheless have a solution if ξ is restricted to an appropriate class of random variables.

Suppose that the vector $\psi(t)$ given by (3.8) is partitioned such that $\psi(t) = \big[\psi^1(t)',\ \psi^2(t)'\big]'$, where $\psi^1(t) = [\psi_1(t), \ldots,\ \psi_L(t)]'$ denotes the first L entries of $\psi(t)$, and $\psi^2(t) = [\psi_{L+1}(t), \ldots,\ \psi_n(t)]'$ denotes the remaining $n - L$ entries. Let $N(t) = \big[N^1(t),\ N^2(t)\big]$ and $M(t) = \big[M^1(t),\ M^2(t)\big]$ be partitioned similarly. Throughout this section (as well as in the next), $\mathbb{D}^i = \mathcal{D}_t^i$ denotes the filtration generated by $N^i(t)$ augmented by the $\mathbb{P}$-null sets of $\mathcal{F}$, and $\mathbb{G}^i = \{\mathcal{G}_t^i\}_{t\geq 0}$, where $\mathcal{G}_t^i \triangleq \mathcal{D}_t^i \vee \mathcal{F}_t$ is the smallest σ-algebra containing $\mathcal{D}_t^i$ and $\mathcal{F}_t$ for $i = 1, 2$.

Suppose that $Y(t)$ is *not* a positive $\mathbb{G}$-martingale, but that $Y^1(t)$ defined by

$$\begin{cases} dY^1(t) = -Y^1(t^-)\{\gamma(t)'\,dW(t) + \psi^1(t)'\,dM^1(t)\}, \\ Y^1(0) = 1 \end{cases} \tag{5.13}$$

is a positive $\mathbb{G}$-martingale under $\mathbb{P}$. Such a situation may arise, for instance, if $\psi_i(t) < 1$ and is uniformly bounded away from 1 for $i = 1, \ldots, L$, while $\psi_i \not< 1$ for $i = L+1, \ldots, n$. While (3.2) will not generally be solvable in this situation, there is a solution if ξ is restricted as follows.

PROPOSITION 5.4. *If $Y^1(t)$ is a positive $\mathbb{G}$-martingale under $\mathbb{P}$ and $\xi \in \mathcal{G}_T^1$, then there exists a solution $(h(t), \eta(t), \kappa(t))$ of (3.2) such that $h(t)$ is $\mathbb{G}^1$-adapted (and hence $\mathbb{G}$-adapted), $\eta(t)$ and $\kappa(t)$ are $\mathbb{G}^1$-predictable (and hence $\mathbb{G}$-predictable), and $\kappa(t) = [\kappa^1(t),\ \kappa^2(t)]$, where $\kappa^2(t) \equiv 0$.*

Proof. Since by assumption $Y^1(t)$ is a positive $\mathbb{G}$-martingale under $\mathbb{P}$, we can define a $\mathbb{P}$-equivalent measure $\mathbb{Q}^1$ via

$$\frac{d\mathbb{Q}^1}{d\mathbb{P}} = Y^1(T).$$

By the Girsanov theorem, $\bar{W}(t) \triangleq W(t) + \int_0^t \gamma(s)\,ds$ is a $\mathbb{G}$-Brownian motion under $\mathbb{Q}^1$ and the components $N_i(t)$ of $N(t)$ are doubly stochastic Poisson process with $\mathbb{F}$-predictable intensities $\lambda_i(t)(1 - \psi_i(t))$ when $i = 1, \ldots, L$, and $\lambda_i(t)$ when $i = L+1, \ldots, n$ (see Proposition 4.1); in particular, this implies that $M^1(t) = N^1(t) - \int_0^t D^1(s)(\mathbf{1} - \psi^1(s))\,ds = M^1(t) + \int_0^t D^1(s)\psi^1(s)\,ds$ is a $\mathbb{G}$-martingale under $\mathbb{Q}$.

Let $h(t)$ be the $\mathbb{G}^1$-adapted process defined via

$$\frac{h(t)}{B(t)} = \mathbb{E}^{\mathbb{Q}^1}\left[\frac{\xi}{B(T)}\Big|\mathcal{G}_t^1\right]. \tag{5.14}$$

Since $h(t)/B(t)$ is a $\mathbb{G}^1$-martingale under $\mathbb{Q}^1$ and $\mathbb{G}^1$ is generated by $\{W(t),\ 0 \leq t \leq T\}$ and $\{N^1(t),\ 0 \leq t \leq T\}$, it follows from the martingale representation theorem that there are $\mathbb{G}^1$-predictable processes $\bar{\eta}(t)$ and $\bar{\kappa}^1(t)$ such that

$$\frac{h(t)}{B(t)} = \mathbb{E}^{\mathbb{Q}^1}\left[\frac{\xi}{B(T)}\right] + \int_0^t \bar{\eta}(s)'\,d\bar{W}(s) + \int_0^t \bar{\kappa}^1(s)'\,d\bar{M}^1(s).$$

Substituting $\eta(t) \triangleq B(t)\bar{\eta}(t)$ and $\kappa^1(t) \triangleq B(t)\bar{\kappa}^1(t)$ and changing measure from $\mathbb{Q}^1$ back to $\mathbb{P}$ gives

$$\begin{cases} dh(t) = r(t)h(t)\,dt + \eta(t)'[\gamma(t)\,dt + dW(t)] \\ \quad\ +\kappa^1(t)'[D^1(t)\psi^1(t)\,dt + dM^1(t)], \\ h(T) = \xi, \end{cases} \tag{5.15}$$

where $D^1(t) \triangleq \operatorname{diag}\{\lambda_1(t), \ldots, \lambda_L(t)\}$. Setting $\kappa(t) = [\kappa^1(t)', \kappa^2(t)']'$, where the components $\kappa^2(t) = [\kappa_{L+1}(t), \ldots, \kappa_n(t)]' \equiv 0$, it follows from (5.15) that the triple $(h(t), \eta(t), \kappa(t))$ is also a solution of (3.2). Square integrability of $(\eta(t), \kappa(t))$ can be proven along the same lines as in the proof of Theorem 4.3. □

The following necessary and sufficient condition for $Y^i(t)$ to be a strictly positive martingale can be shown along the lines of Theorem 4.4.

PROPOSITION 5.5. *Suppose that Assumption* (A) *is satisfied. Then* $Y^1(t)$ *is a strictly positive* $\mathbb{G}$*-martingale under* $\mathbb{P}$ *if and only if for* $i = 1, \ldots, L$, $\psi_i^1(t) < 1$ *for a.e.* $t \in [0, T]$, $\mathbb{P}$*-a.s.*

5.4. "No common jumps." We say that the liability ξ and the asset prices $P_1(t), \ldots, P_m(t)$ have *no common jumps* if there is no jump component $N_j(t)$ that affects both the value of the liability ξ *and* the asset prices $P_i(t)$ $(i = 1, \ldots, m)$. The extreme case of this corresponds to the liability ξ being constant, which arises in the problem of *mean-variance portfolio selection*, but is also satisfied if ξ is measurable with respect to the history of the Brownian motion.

More generally, if $\mathcal{G}_T^1$ denotes the σ-algebra generated by $\{W(t),\ 0 \le t \le T\}$ and $\{N^1(t),\ 0 \le t \le T\}$, where without loss of generality we take $N^1(t)$ to be the first L components $(N_1(t), \ldots, N_L(t))$ of $N(t)$, then there are no common jumps if $\xi \in \mathcal{G}_T^1$ and the price processes $P_i(t)$ given by (2.1) are such that the matrix $\theta(t) = [\theta^1(t),\ \theta^2(t)]$ satisfies

$$\theta^1(t) \equiv 0, \tag{5.16}$$

where $\theta^1(t) = [\theta_1^1(t),\ \ldots, \theta_L^1(t)]$ and $\theta^2(t) = [\theta_{L+1}^2(t), \ldots,\ \theta_n^2(t)]$ are the first L columns and the remaining $n - L$ columns of $\theta(t)$, respectively. Proposition 5.6 states that the BSDE (3.2) *always has a solution when there are no common jumps*. (This result is stronger than existence results established in Theorems 4.3 and 4.4, where strict positivity of $Y(t)$ (or equivalently that $\psi_i(t) < 1$) must be satisfied in order to guarantee existence.)

PROPOSITION 5.6. *Suppose that Assumption* (A) *holds. If* $\xi \in \mathcal{G}_T^1$ *and* $\theta^1(t) \equiv 0$, *then* (3.2) *has a solution* $(h(t), \eta(t), \kappa(t)) \in \mathcal{L}^\infty(\mathbb{G}, \mathbb{R}) \times \mathcal{P}^2(\mathbb{G}, \mathbb{R}^d) \times \mathcal{P}^2(\mathbb{G}, \mathbb{R}^n)$.

Proof. We prove this result by constructing a solution of (3.2). Consider first the following BSDE:

$$\begin{cases} dh(t) = r(t)h(t)\,dt + \eta(t)'[\gamma(t)\,dt + dW(t)] + \kappa(t)'\,dM(t), \\ h(T) = \xi. \end{cases} \tag{5.17}$$

Let $\rho(t)$ denote the solution of

$$\begin{cases} d\rho(t) = -\rho(t^-)\gamma(t)'\,dW(t), \\ \rho(0) = 1, \end{cases}$$

where $\gamma(t)$ is given by (3.7) (see also (3.4)). By Proposition 3.1, we know that $\rho(t)$ is a positive $\mathbb{F}$-martingale. It follows from the *martingale invariance property* (see Proposition 2.1) that $\rho(t)$ is a positive $\mathbb{G}$-martingale under $\mathbb{P}$ and hence defines a $\mathbb{P}$-equivalent probability measure $\mathbb{Q}$ via

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \rho(T).$$

By the Girsanov theorem, $\bar{W}(t) = W(t) + \int_0^t \gamma(s)\,ds$ is a $\mathbb{G}$-Brownian motion under $\mathbb{Q}$ and $N_i(t)$ $(i = 1, \dots, n)$ are doubly stochastic Poisson processes under $\mathbb{Q}$ with (unchanged) $\mathbb{F}$-predictable intensities $\lambda_i(t)$; see Proposition 4.1. Therefore, we can use the same approach as in the proof of Theorem 4.3 to construct a solution of (5.17). In particular, by the martingale representation theorem (Proposition 4.2), we can (uniquely) define $(h(t), \bar{\eta}(t), \bar{\kappa}^1(t))$ via

$$\frac{h(t)}{B(t)} = \mathbb{E}^{\mathbb{Q}}\Big[\frac{\xi}{B(T)}\Big|\mathcal{G}_t^1\Big] = \mathbb{E}^{\mathbb{Q}}\Big[\frac{\xi}{B(T)}\Big] + \int_0^t \bar{\eta}(t)'\,d\bar{W}(t) + \int_0^t \bar{\kappa}^1(t)'\,dM^1(t), \tag{5.18}$$

where $\bar{\eta}(t)$ and $\bar{\kappa}^1(t)$ are $\mathbb{G}^1$-predictable (and hence $\mathbb{G}$-predictable) processes (of dimension d and L, respectively) and $M^1(t)$ is the compensated Poisson process associated with $N^1(t)$. Moreover, setting $\bar{\kappa}(t) = [\bar{\kappa}^1(t)',\, \bar{\kappa}^2(t)']' = [\bar{\kappa}^1(t),\, 0]$, where $\bar{\kappa}^2(t)$ is an $(n-L)$-dimensional vector of zeros, it is clear that (5.18) implies

$$\frac{h(t)}{B(t)} = \mathbb{E}^{\mathbb{Q}}\Big[\frac{\xi}{B(T)}\Big] + \int_0^t \bar{\eta}(t)'d\bar{W}(t) + \int_0^t \bar{\kappa}(t)'dM(t).$$

By Ito's formula, it can be shown that

$$(h(t),\, \eta(t),\, \kappa(t)) = (h(t),\, B(t)\bar{\eta}(t),\, B(t)\bar{\kappa}(t))$$

is a solution of (5.17). Furthermore, this solution is unique due to the uniqueness of martingale representation. Finally, since $\kappa^2(t) = B(t)\bar{\kappa}^2(t) \equiv 0$ it follows from (5.16) that

$$\theta(t)'D(t)\kappa(t) = \theta^1(t)D^1(t)\kappa^1(t) + \theta^2(t)D^2(t)\kappa^2(t) \equiv 0, \tag{5.19}$$

and hence, by (3.8), that

$$\kappa(t)'D(t)\psi(t) \equiv 0.$$

Therefore, the solution of (5.17) is also a solution of (3.2), which establishes our result. □

An important special case is the one where the liability ξ is deterministic, which arises in the problem of mean-variance portfolio selection. Propositions 3.5 and 5.6 imply that (3.2) *always has a solution* and the mean-variance hedging problem (2.6) can be solved, with optimal policy (3.5), without having to assume that the solution of (4.2) is a strictly positive martingale.

COROLLARY 5.7. *Suppose that Assumption* (A) *holds and ξ is $\mathcal{F}_T$-measurable. (In particular, ξ may be deterministic.) Then* (3.2) *has a unique solution* $(h(t), \eta(t), \kappa(t))$, *where* $\kappa(t) \equiv 0$.

6. Some remarks about assumptions. In Arai [1], the problem of mean-variance hedging when underlying prices are discontinuous semimartingales is addressed. We now comment briefly on the assumptions made in [1] and how these compare with those in this paper.

Several differences between the problem formulations in [1] and the present paper should be clarified before comparing the assumptions. First, the price processes in [1] are discontinuous semimartingales, while those in this paper (2.1) have additional structure in that they are driven by Brownian motion and a doubly stochastic Poisson process with $\mathbb{F}$-predictable parameters. Second, it is assumed in [1] that all price

processes have been discounted using the money market account, so stochastic interest rates are not addressed explicitly (they are combined with the other prices through the discounting). This contrasts with (2.1) and (2.6), where price and wealth processes are not discounted. Third, the contingent claim in [1] (what we call ξ) is square integrable, whereas we are assuming that it is bounded. (This was used in the proof of Theorem 4.3. It should be possible to relax this assumption, but this is left for future work.) Finally, the methods adopted in [1] are different from those in the present paper. More specifically, [1] extends the analysis in [30], based on convex duality, to the discontinuous setting, whereas we use stochastic control methods together with the theory of BSDEs.

These differences being understood, we now examine the relationship between the assumptions in [1] and those in this paper. Not surprisingly, the additional structure in our model means that some of the assumptions invoked in [1] are not required here.

Assumptions in Arai [1].

(A1) The variance optimal (signed) martingale measure $\mathbb{Q}$ is a $\mathbb{P}$-equivalent measure; equivalently, $\mathbb{Q}$ has a density process $Z(t)$ that is a strictly positive $\mathbb{P}$-martingale.

(A2) $Z(t)$ satisfies the *reverse Hölder inequality*; that is, there is a constant $C > 0$ such that

$$\mathbb{E}\left[\left(\frac{Z(T)}{Z(\tau)}\right)^2 \Big| \mathcal{H}_\tau\right] \le C$$

for every $\mathbb{H}$-stopping time τ, where $\mathbb{H} = \{\mathcal{H}_t\}_{t\ge 0}$ denotes the underlying filtration.

(A3) There exists a constant $C > 0$ such that $Z(t^-) \le CZ(t)$ for all $t \in [0, T]$, $\mathbb{P}$-a.s.

The density process $Y(t)$ in (4.2) plays an analogous role to the density process $Z(t)$ in [1]. A priori, neither $Z(t)$ nor $Y(t)$ is positive, and hence, only define signed measures. This is the reason why (A1) is needed in [1]. The assumption that $Y(t)$ is a positive martingale in Theorem 4.3 is analogous to this.

Because of the structure of our model, neither (A2) nor (A3) is required in this paper. In particular, it can be shown that the solution $(p(t), \Lambda(t))$ of (3.1) satisfies

$$\frac{1}{p(t)} = \mathbb{E}\left[\left(\frac{Y(T)}{Y(t)}\right)^2 \Big| \mathcal{G}_t\right].$$

It follows that the bound $\delta_1 \le p(t) \le \delta_2$ (Proposition 3.1) guarantees that the reverse Hölder inequality is *automatically* satisfied in this paper and does not need to be assumed. Again, this comes from the structure that we impose on our model.

7. Example. We now present an example with constant parameters (but random liability) for which explicit solutions of (3.1)–(3.2) can be obtained. In particular, we show how formulas for the solution $(h(t), \eta(t), \kappa(t))$ of (3.2) can be derived, and the insights that this gives into the structure of the optimal hedging portfolio.

For this example, we assume that the interest rate r for the money market account $B(t)$ is constant, and a single underlying risky asset

$$dP(t) = P(t)\{\mu\, dt + \sigma\, dW(t) + \theta\, dN(t)\} \tag{7.1}$$

with deterministic parameters μ, σ, and θ. We assume that the arrival rate λ is constant.

Observe first that the market is arbitrage free: the measure $\bar{\mathbb{Q}}$ that makes

$$\bar{W}(t) \triangleq -\frac{t}{\sigma}[r-\mu-\theta\lambda]+W(t)$$

a Brownian motion but leaves the rate λ of $N(t)$ unchanged is an equivalent martingale measure. On the other hand, it follows from Proposition 5.3 that (3.2) has a solution if $\psi<1$, where

$$\gamma=\sigma'\Sigma^{-1}b=\frac{\sigma b}{\sigma^2+\theta^2\lambda},\quad \psi=\theta\Sigma^{-1}b=\frac{\theta b}{\sigma^2+\theta^2\lambda}.$$

In particular, it is possible for the market to be arbitrage free but for $\psi<1$ not to be satisfied. We assume for the remainder that $\psi<1$.

Consider a liability of the form $\xi=\mathbf{1}_{\tau\le T}[A-y(\tau)]^+e^{r(T-\tau)}$, where τ denotes the first arrival time of $N(t)$ and $y(t)$ is the solution of the stochastic differential equation

$$dy(t)=y(t)\{r\,dt+q\,d\bar{W}(t)\}=y(t)\{(r+q\gamma)\,dt+q\,dW(t)\},$$

where $\bar{W}(t)=W(t)+\gamma t$ is a Brownian motion under the equivalent probability measure $\mathbb{Q}$ associated with (γ,ψ) via (4.2)–(4.3). In this model, we assume that "default" occurs at a random time τ that corresponds to the arrival time of the first Poisson event, and that the value of the liability is determined by the value of $[A-y(t)]^+$ at the time of this arrival (i.e., the default event). In particular, if $y(t)$ is significantly smaller than A, then the event of default is more costly for the investor.

By Theorem 3.5 the optimal policy is given by (3.5). Since the parameters μ, σ, θ, and λ of the price process (7.1) in this example are deterministic, it follows that $\Lambda\equiv 0$, so computing the optimal policy (3.5) boils down to finding $(h(t),\,\eta(t),\,\kappa(t))$ by solving (3.2). As in the proof of Theorem 4.3, the solution of (3.2) can be represented in the form

$$h(t)=e^{-r(T-t)}\mathbb{E}^{\mathbb{Q}}[\xi\,|\,\mathcal{G}_t]=e^{-r(T-t)}\mathbb{E}^{\mathbb{Q}}\Big[\mathbf{1}_{\tau\le T}(A-y(\tau))^+e^{r(T-\tau)}\,\Big|\,\mathcal{G}_t\Big]. \tag{7.2}$$

To get something more explicit for $h(t)$, it is convenient to introduce the process

$$z(t)=\int_{s=0}^{t}e^{r(t-s)}[A-y(s)]^+\mathbf{1}_{N(s^-)=0}\,dN(s).$$

Note that

$$z(t)=\mathbf{1}_{\tau\le t}[A-y(\tau)]^+e^{r(t-\tau)}=\mathbf{1}_{N(t)\ge 1}[A-y(\tau)]^+e^{r(t-\tau)},$$

which coincides with the value of the liability given that default occurred before time t. Observing that

$$\mathbf{1}_{\tau\le T}[A-y(\tau)]^+e^{r(T-\tau)}=\mathbf{1}_{\tau\le t}(A-y(\tau))^+e^{r(T-\tau)}+\mathbf{1}_{t<\tau\le T}(A-y(\tau))^+e^{r(T-\tau)},$$

we see from (7.2) that

$$h(t)=\mathbf{1}_{\tau\le t}e^{-r(\tau-t)}(A-y(\tau))^+ +\mathbf{1}_{\tau>t}e^{-r(T-t)}\mathbb{E}^{\mathbb{Q}}\Big[\mathbf{1}_{t<\tau\le T}(A-y(\tau))^+e^{r(T-\tau)}\,\Big|\,\mathcal{G}_t\Big].$$

Observing that

$$\begin{aligned}
&e^{-r(T-t)}\mathbb{E}^{\mathbb{Q}}\Big[\mathbf{1}_{t<\tau\le T}(A-y(\tau))^+e^{r(T-\tau)}\,\Big|\,\mathcal{G}_t\Big]\\
&=e^{-r(T-t)}\mathbb{E}^{\mathbb{Q}}\Big[\mathbb{E}^{\mathbb{Q}}\Big\{\mathbf{1}_{t<\tau\le T}(A-y(\tau))^+e^{r(T-\tau)}\,\Big|\,\mathcal{G}_t\vee\mathcal{F}_T\Big\}\,\Big|\mathcal{G}_t\Big]\\
&=e^{-r(T-t)}\mathbb{E}^{\mathbb{Q}}\left[\int_{s=t}^{T}\lambda(1-\psi)e^{-\lambda(1-\psi)(s-t)}(A-y(s))^+e^{r(T-s)}\,ds\,\Big|\,\mathcal{G}_t\right]\\
&=\int_{s=t}^{T}\lambda(1-\psi)e^{-\lambda(1-\psi)(s-t)}P(t,\,s,\,y(t))\,ds,
\end{aligned}$$

where

$$\begin{aligned}
P(t,\,s,\,y)=Ae^{-r(s-t)}\Phi\Big(\frac{\log(A/y)-(r-\frac{1}{2}q^2)(s-t)}{q\sqrt{s-t}}\Big)\\
-y\Phi\Big(\frac{\log(A/y)-(r+\frac{1}{2}q^2)(s-t)}{q\sqrt{s-t}}\Big),
\end{aligned}$$

denotes the price of a European put option on $y(\cdot)$ with strike A maturing at $s>t$, when $y(t)=y$. It follows that

$$h(t)=\mathbf{1}_{\tau\le t}\,e^{r(t-\tau)}(A-y(\tau))^+ +\mathbf{1}_{\tau>t}\int_{s=t}^{T}\lambda(1-\psi)e^{-\lambda(1-\psi)(s-t)}P(t,\,s,\,y(t))\,ds$$

or, equivalently,

$$\begin{aligned}
h(t)&=\mathbf{1}_{N(t)\ge 1}\,e^{r(t-\tau)}(A-y(\tau))^+\\
&\quad+\mathbf{1}_{N(t)=0}\int_{s=t}^{T}\lambda(1-\psi)e^{-\lambda(1-\psi)(s-t)}P(t,\,s,\,y(t))\,ds\\
&=z(t)+\mathbf{1}_{N(t)=0}\int_{s=t}^{T}\lambda(1-\psi)e^{-\lambda(1-\psi)(s-t)}P(t,\,s,\,y(t))\,ds\qquad(7.3)\\
&=f(t,\,y(t),\,z(t),\,N(t)),
\end{aligned}$$

where

$$f(t,\,y,\,z,\,i)=z+\mathbf{1}_{i=0}\int_{s=t}^{T}\lambda(1-\psi)e^{-\lambda(1-\psi)(s-t)}P(t,\,s,\,y)\,ds.$$

In order to compute the optimal portfolio, we need to find the components $(\eta(t),\,\kappa(t))$ of the solution of (3.2) or (4.1), which in this example is

$$\begin{cases}dh(t)=rh(t)\,dt+\eta(t)\,d\bar{W}(t)+\kappa(t)\,d\bar{M}(t),\\ h(T)=\mathbf{1}_{\tau\le T}[A-y(\tau)]^+e^{r(T-\tau)}.\end{cases}\qquad(7.4)$$

Applying Ito's formula to $h(t)=f(t,\,y(t),\,z(t),\,N(t))$ gives

$$\begin{aligned}
dh(t)=\Big\{&f_t(t,\,y(t),\,z(t^-),\,N(t^-))+ry(t)f_y(t,\,y(t),\,z(t^-),\,N(t^-))\\
&+\frac{1}{2}\sigma^2y^2f_{yy}(t,\,y(t),\,z(t^-),\,N(t^-))
\end{aligned}$$

$$
\begin{aligned}
&+\lambda(1-\psi)\Big[\mathbf{1}_{N(t^-)=0}\Big(f(t,\,y(t),\,(A-y(t))^+,\,1)-f(t,\,y(t),\,0,\,0)\Big)\\
&\quad+\mathbf{1}_{N(t^-)\geq 1}\Big(f(t,\,y(t),\,z(t),\,N(t^-)+1)-f(t,\,y(t),\,z(t^-),\,N(t^-))\Big)\Big]\Big\}\,dt\\
&+\sigma y f_y(t,\,y(t),\,z(t^-),\,N(t^-))\,d\bar{W}(t)\\
&+\Big\{\mathbf{1}_{N(t^-)=0}\Big[f(t,\,y(t),\,(A-y(t))^+,\,1)-f(t,\,y(t),\,0,\,0)\Big]\\
&\quad+\mathbf{1}_{N(t^-)\geq 1}\Big[f(t,\,y(t),\,z(t),\,N(t^-)+1)-f(t,\,y(t),\,z(t^-),\,N(t^-))\Big]\Big\}\,d\bar{M}(t).
\end{aligned}
$$

Comparing coefficients with (7.4) we obtain

$$
\begin{aligned}
\eta(t) &= \eta(t,\,y(t),\,z(t^-),\,N(t^-)) = \sigma y f_y(t,\,y(t),\,z(t^-),\,N(t^-))\\
\kappa(t) &= \kappa(t,\,y(t),\,z(t^-),\,N(t^-))\\
&= \mathbf{1}_{N(t^-)=0}\Big[f(t,\,y(t),\,(A-y(t))^+,\,1)-f(t,\,y(t),\,0,\,0)\Big]\\
&\quad+\mathbf{1}_{N(t^-)\geq 1}\Big[f(t,\,y(t),\,z(t),\,N(t^-)+1)-f(t,\,y(t),\,z(t^-),\,N(t^-))\Big].
\end{aligned}
$$

In particular, by (7.3) we have

$$
\eta(t) = \begin{cases} \sigma y(t)\displaystyle\int_{s=t}^{T}\lambda(1-\psi)e^{-\lambda(1-\psi)(s-t)}P_y(t,\,s,\,y(t))\,ds, & N(t^-)=0,\\ 0, & N(t^-)\geq 1,\end{cases}
$$

$$
\kappa(t) = \begin{cases} (A-y(t))^+ - \displaystyle\int_{s=t}^{T}\lambda(1-\psi)e^{-\lambda(1-\psi)(s-t)}P(t,\,s,\,y(t))\,ds, & N(t^-)=0,\\ 0, & N(t^-)\geq 1.\end{cases}
$$

Finally, it follows from Theorem 3.5 that the optimal portfolio is

$$
\pi(t) = \Sigma^{-1}\Big[\sigma\eta(t)+\theta\lambda\kappa(t)+b(h(t^-)-x(t^-))\Big].
$$

The term $\sigma\,\eta(t)$ hedges the uncertainty due to the Brownian motion fluctuations, while $\theta\lambda\kappa(t)$ hedges the loss at the instant of a jump. The last term which involves $f(t^-,\,y(t^-),\,N(t^-))-x(t^-)$ is concerned with minimizing the difference between the value of the portfolio at $x(t^-)$ and the value of the liability $h(t^-)$. Note that $\kappa(t)\equiv 0$ immediately after the first jump occurs. This is natural since the value of the liability ξ does not depend on jumps that may occur after the first. On the other hand, prior to the first jump, the "size" of $\kappa(t)$ depends both on the *time to maturity* $T-t$ as well as $P(t,\,y(t))$, the (stochastic) *cost of default*. In particular, if $y(t)$ is "close to" (or bigger than) A, then the impact of default is small and $(A-y(t))^+=0$ and $P(t,\,y(t))\approx 0$. In such a case, $\theta\lambda\kappa(t)$ will also be small.

8. Conclusion. In recent years, much effort has gone into the task of developing more sophisticated asset price models which are capable of reproducing some of the empirical features of price processes that are observed in data. One product of this endeavor has been the introduction of processes with jumps and stochastic parameters as an alternative to the Black–Scholes model. In this paper, we consider the problem of

mean-variance hedging in an incomplete market where the price processes have jumps (modeled by a doubly stochastic Poisson process) and parameters may be stochastic. This differs from almost all papers in the literature on mean-variance hedging where continuity of the tradable asset prices is usually assumed. We formulate this problem as a stochastic control problem and derive closed form expressions for the optimal hedging portfolio using the theory of backward stochastic differential equations. In particular, we show how the BSDE theory can be extended from the Brownian motion setting to handle problems with jumps.

Another application of this work relates to the problem of hedging default risk when the default is modeled in the reduced form setting. Considered in this way, the results of this paper extend, in some sense, those of Blanchet-Scalliet and Jeanblanc [5], where the problem of hedging default risk in a *complete* market is discussed.

Acknowledgments. The author would like to thank Kiseop Lee, the associate editor, and two anonymous reviewers for their feedback, as well as the corresponding editor Tyrone Duncan for his efficient handling of this paper. All errors are the responsibility of the author.

REFERENCES

[1] T. Arai, *An extension of mean-variance hedging to the discontinuous case,* Finance Stoch., 9 (2005), pp. 129–139.

[2] I. Bardhan and X. Chao, *Martingale analysis for assets with discontinuous returns,* Math. Oper. Res., 20 (1995), pp. 243–256.

[3] D. Bates, *Post-'87 crash fears and S&P 500 futures options,* J. Econometrics, 94 (2000), pp. 181–238.

[4] T.R. Bielecki and M. Rutkowski, *Credit Risk: Modeling, Valuation and Hedging,* Springer-Verlag, Berlin, 2002.

[5] C. Blanchet-Scalliet and M. Jeanblanc, *Hazard rate for credit risk and hedging defaultable contingent claims,* Finance Stoch., 8 (2004), pp. 145–159.

[6] O. Bobrovnytska and M. Schweizer, *Mean-variance hedging and stochastic control: Beyond the Brownian setting,* IEEE Trans. Automat. Control, 49 (2004), pp. 396–408.

[7] P. Bremaud and M. Yor, *Changes of filtrations and of probability measures,* Z. Wahrsch. Verw. Gebiete, 45 (1978), pp. 269–295.

[8] R. Cont and P. Tankov, *Financial Modelling with Jump Processes,* Chapman & Hall/CRC Financ. Math. Ser., Chapman & Hall/CRC, Boca Raton, FL, 2004.

[9] F. Delbaen and W. Schachermayer, *The variance-optimal martingale measure for continuous processes,* Bernoulli, 2 (1996), pp. 81–105.

[10] D. Duffie, *Dynamic Asset Pricing Theory,* 3rd ed., Princeton University Press, Princeton, NJ, 2001.

[11] D. Duffie and K.J. Singleton, *Credit Risk,* Princeton University Press, Princeton, NJ, 2003.

[12] B. Erakar, M.S. Johannes, and N.G. Polson, *The impact of jumps in returns and volatility,* J. Finance, 58 (2003), pp. 1269–1300.

[13] R.J. Elliott, M. Jeanblanc, and M. Yor, *On models of default risk,* Math. Finance, 10 (2000), pp. 179–195.

[14] C. Gourieroux, J.P. Laurent, and H. Pham, *Mean-variance hedging and numéraire,* Math. Finance, 8 (1998), pp. 179–200.

[15] F.B. Hanson and J.J. Westman, *Jump-diffusion models in finance: Stochastic process density with uniform jump amplitude,* in Proceedings of the 15th International Symposium on Mathematical Theory of Networks and Systems, 2002.

[16] Y. Hu and X.Y. Zhou, *Constrained stochastic LQ control with random coefficients, and application to portfolio selection,* SIAM J. Control Optim., 44 (2005), pp. 444–466.

[17] J. Jacod, *Calcul Stochastique et Problèmes de Martingales,* Lecture Notes in Math. 714, Springer-Verlag, New York, 1979.

[18] M. Jeanblanc-Picque and M. Pontier, *Optimal portfolio for a small investor in a market model with discontinuous prices,* Appl. Math. Optim., 22 (1990), pp. 287–310.

[19] I. Karatzas and S.E. Shreve, *Brownian Motion and Stochastic Calculus,* Springer-Verlag, New York, 1988.

[20] S.G. Kou, *A jump-diffusion model for option pricing,* Manag. Sci., 48 (2002), pp. 1086–1101.
[21] D. Lando, *On Cox processes and credit-risky securities,* Rev. Derivatives Res., 2 (1998), pp. 99–120.
[22] K. Lee and S. Song, *Insider's Hedging in Jump Diffusion Models,* working paper, Department of Mathematics, University of Louisville.
[23] J.P. Laurent and H. Pham, *Dynamic programming and mean-variance hedging,* Finance Stoch., 3 (1999), pp. 83–110.
[24] A.E.B. Lim, *Hedging default risk in an incomplete market,* in Mathematics of Finance, G. Yin and Q. Zhang, eds., AMS Contemp. Math. 351, Providence, RI, 2004, pp. 231–246.
[25] A.E.B. Lim, *Quadratic hedging and mean-variance portfolio selection with random parameters in an incomplete market,* Math. Oper. Res., 29 (2002), pp. 132–161.
[26] A.E.B. Lim and T. Watewai, *Optimal Portfolio Choice with Discontinuous Price Processes and Multiple Regimes,* working paper, IEDR Department, University of California, Berkeley.
[27] A.E.B. Lim and X.Y. Zhou, *Mean-variance portfolio selection with random parameters in a complete market,* Math. Oper. Res., 27 (2002), pp. 101–120.
[28] A.E.B. Lim and X.Y. Zhou, *Mean-variance portfolio choice with discontinuous asset prices and nonnegative wealth processes,* in Mathematics of Finance, G. Yin and Q. Zhang, eds., AMS Contemp. Math. 351, Providence, RI, 2004, pp. 247–258.
[29] J. Liu, F. Longstaff, and J. Pan, *Dynamic asset allocation with event risk,* J. Finance, 58 (2003), pp. 231–259.
[30] T. Rheinländer and M. Schweizer, *On L^2-projections on a space of stochastic integrals,* Ann. Probab., 25 (1997), pp. 1810–1831.
[31] S. Rong, *On solutions of backward stochastic differential equations with jumps and applications,* Stoch. Process. Appl., 66 (1997), pp. 209–236.
[32] M. Schweizer, *Approximation pricing and the variance-optimal martingale measure,* Ann. Probab., 64 (1996), pp. 206–236.
[33] J.M. Steele, *Stochastic Calculus and Financial Applications,* Springer-Verlag, New York, 2001.
[34] X.X. Xue, *Martingale representation for a class of processes with independent increments and its applications,* in Applied Stochastic Analysis: Proceedings of a U.S.–French Workshop, Lecture Notes in Control and Inform. Sci. 177, Springer-Verlag, New York, 1992, pp. 279–311.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 6, pp. 1923–1949

CONTROLLER DESIGN VIA NONSMOOTH MULTIDIRECTIONAL SEARCH*

PIERRE APKARIAN† AND DOMINIKUS NOLL‡

Abstract. We propose an algorithm which combines multidirectional search (MDS) with nonsmooth optimization techniques to solve difficult problems in automatic control. Applications include static and fixed-order output feedback controller design, simultaneous stabilization, H_2/H_∞-synthesis, and much else. We show how to combine direct search techniques with nonsmooth descent steps in order to obtain convergence certificates in the presence of nonsmoothness. Our technique is efficient when small and medium size controllers for plants with large state dimension are sought. Our numerical testing includes several benchmark examples. For instance, our algorithm needs 0.41 s to compute a static output feedback stabilizing controller for the Boeing 767 flutter benchmark problem [E. E. J. Davison, *IFAC Technical Committee Reports*, Pergamon Press, Oxford, 1990], a system with 55 states. The first static controller without performance specifications for this system was obtained in [J. Burke, A. Lewis, and M. Overton, *SIAM J. Optim.*, 15 (2003), pp. 751–779].

Key words. *NP*-hard design problems, static output feedback, fixed-order synthesis, simultaneous stabilization, mixed H_2/H_∞-synthesis, pattern search algorithm, moving polytope, nonsmooth analysis, spectral bundle method, ε-subgradients, bilinear matrix inequality (BMI)

AMS subject classifications. 93B36, 93B40, 93B50, 93B51, 90C22, 90C56, 90C34, 90C26, 49J52, 49J35

DOI. 10.1137/S0363012904441684

1. Introduction.

Pattern search or moving polytope methods belong to a large class of derivative-free optimization methods referred to as direct search (DS) techniques. In this paper, we present a nonsmooth modification of Virginia Torczon's multidirectional search (MDS) [66, 67] algorithm and apply it to a broad class of problems in automatic control. We aim at several nonconvex and even *NP*-hard problems, for which LMI techniques or algebraic Riccati equations are impractical. In particular, we propose algorithmic solutions for static and fixed-order output feedback control, simultaneous stabilization problems, and mixed H_2/H_∞-control.

1.1. Direct search methods.

The idea of DS methods can be traced back to the pioneering work of Box [11] and Hook and Jeeves [37], who first coined the term "direct search." The MDS algorithm is due to Torczon [66, 67] and is directly inspired by the work of Spendley, Hext, and Himsworth [63], and the popular method of Nelder and Mead [55]. MDS significantly revived the interest in DS methods, because it came with a sound convergence theory [66]. This is in contrast with the Nelder–Mead algorithm, which may fail to converge even for smooth convex objective functions; see [52]. Later, Torczon generalized her work to the entire class of DS techniques [67].

*Received by the editors March 4, 2004; accepted for publication (in revised form) April 22, 2005; published electronically January 6, 2006.

http://www.siam.org/journals/sicon/44-6/44168.html

†ONERA-CERT, Centre d'études et de recherche de Toulouse, Control System Department, 2 av. Edouard Belin, 31055 Toulouse, France, and Mathématiques pour l'Industrie et la Physique, Université Paul Sabatier, Toulouse, France (apkarian@cert.fr).

‡Université Paul Sabatier, Mathématiques pour l'Industrie et la Physique, 118, route de Narbonne, CNRS UMR 5640, 31062 Toulouse, France (noll@mip.ups-tlse.fr).

DS methods compute local minima of unconstrained optimization programs:

$$\text{minimize} \quad f(x),\ x \in \mathbb{R}^n, \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a $\mathcal{C}^1$ function. DS techniques are *derivative-free* in the sense that they do not require gradient information in order to *compute* descent steps. This is a convenient feature if derivatives or their finite difference approximations are not available and/or too expensive to compute or when automatic differentiation is hindered by the presence of for loops in the function evaluation.

However, contrary to what the name suggests, the term derivative-free does not mean that derivatives do not altogether exist. On the contrary, DS methods are designed for $\mathcal{C}^1$ functions, and their convergence theory is heavily based on differentiability [67]. Problems encountered when search methods are used with genuinely nonsmooth criteria are discussed in [46].

DS techniques can also be used for constrained optimization programs. The ideas to attack those range from quadratic or exact penalty techniques over barrier functions to the augmented Lagrangian method.

1.2. Nonsmoothness. In the present paper, we apply the ideas of MDS to several constrained and unconstrained optimization problems in automatic control, where nonsmooth functions like the maximum eigenvalue function, the spectral abscissa, the distance to instability, and the H_∞-norm arise naturally. Due to the failure of convergence under nonsmoothness, DS methods may not be applied in their original form and additional tools from nonsmooth optimization are required. An algorithm combining both ideas is what will eventually emerge. Using nonsmooth techniques in control design is not altogether a new idea; see, e.g., [62, 61, 44, 53, 40]. What has not been tried before is combining nonsmooth techniques with DS strategies.

The lack of a convergence certificate under nonsmoothness has not prevented practitioners from applying DS methods in such cases. It is often argued that the contingency of a failure due to nonsmoothness is a remote one. The argument on which such reasoning is usually based is that even nonsmooth functions are, as a rule, almost everywhere differentiable, so that nonsmooth points are never encountered in practice. Our present work reveals this as an illusory argument. Nonsmoothness may and will cause failure of DS techniques, as we demonstrate by several striking examples.

In response, we show how MDS can be combined with nonsmooth descent steps in order to avoid the typical failure, where simplices shrink and iterates converge to a nonstationary point, which we also call a *dead point*. It is crucial to be able to distinguish dead points from local minima, and this is done by adding a nonsmooth stopping test to the usual hand tools of MDS. Such a test either indicates success or allows one to escape from a dead point, keeping the search algorithm moving.

However, this is not the end of the story. Calling for a nonsmooth stopping test whenever the simplex shrinks below a certain threshold may keep MDS moving, but it is not strong enough to ensure convergence. In order to get a convergence certificate in the presence of nonsmoothness, we need to supply MDS with *quantified descent steps* similar to those employed by nonsmooth optimization techniques to ensure convergence. We will refer to these two types of nonsmooth substrata to MDS as *crisis intervention* and *crisis prevention*. While crisis intervention is done only occasionally, being therefore less costly, crisis prevention is more complex, as it requires that the nonsmooth technique assists the search during the whole process.

We will indicate in which way crisis intervention and crisis prevention should be organized for application in automatic control, but our approach is in principle open to more general nonsmooth objectives.

We mention that a different approach to integrate nonsmoothness into MDS was recently proposed by Audet and Dennis [7, 1] for general locally Lipschitz functions. Their approach and ours are somewhat complementary. While we are more specific as far as the applications are concerned, our combined method can accommodate composite functions with the spectral abscissa, which are not even locally Lipschitz smooth. Also, our intervention technique is applicable to other derivative-free method, like for instance the wedge algorithm of Marrazzi and Nocedal [51].

The paper is organized as follows. We start with an introductory section 2, where three nonsmooth criteria are discussed. We proceed with the central sections 3 and 4, where we indicate why and in which form nonsmoothness arises in automatic control. In section 5 we briefly recall the mode of operation of MDS, including the possibility of the two types of intervention steps, by which the failure at dead points can be avoided. In section 6 we proceed to the implementation of *crisis intervention* and *crisis prevention* for nonsmooth objectives like the maximum eigenvalue function, the spectral abscissa, and the H_∞-norm. Crisis intervention is discussed in section 6, while the more sophisticated crisis prevention is discussed in section 7. Numerical experiments to validate the proposed tools and techniques are discussed in section 8 for a rich set of control applications.

1.3. Notation. Notation from convex and nonsmooth analysis are covered by [35] and [22]. We let $\mathbb{S}^m$ denote the set of $m \times m$ symmetric matrices, equipped with the scalar product $\langle X, Y\rangle = X \cdot Y = \operatorname{Tr}(XY)$. Let $\mathbb{M}_n$ be the space of real $n \times n$ matrices, $\mathbb{M}_{n,m}$ the space of $n \times m$ matrices, equipped with the corresponding scalar product $\langle X, Y\rangle = \operatorname{Tr}(X^{\mathrm{T}}Y)$, where X^{T} is the transpose of the matrix X, $\operatorname{Tr} X$ its trace. For complex matrices X^{H} stands for its transconjugate. For Hermitian or symmetric matrices, $X \succ Y$ means that $X - Y$ is positive definite, $X \succeq Y$ that $X - Y$ is positive semidefinite. We shall use superscripts for the iteration index, lower scripts to indicate vector components. Our notation from feedback control is standard and follows, e.g., [14].

2. Examples of nonsmooth functions in control. In this section we briefly discuss several nonsmooth functions arising in automatic control applications.

Our first example is the maximum eigenvalue function $\lambda_1 : \mathbb{S}^m \to \mathbb{R}$, defined on the space $\mathbb{S}^m$ of symmetric $m \times m$ matrices. We will use composite functions of the form $f(x) = \lambda_1(\mathcal{B}(x))$, where $\mathcal{B} : \mathbb{R}^n \to \mathbb{S}^n$ is usually a bilinear, quadratic, or class $\mathcal{C}^2$-operator. The interest in $f = \lambda_1 \circ \mathcal{B}$ stems from the fact that the matrix inequality $\mathcal{B}(x) \preceq 0$ is equivalent to the scalar constraint $f(x) \leq 0$. Notice that λ_1 is convex, which gives f a lot of structure. For instance, the Clarke subdifferential of f (cf. [22]) is the set

$$\partial f(x) = \mathcal{B}'(x)^\star[\partial\lambda_1(\mathcal{B}(x))] = \{\mathcal{B}'(x)^\star Z : Z = QYQ^{\mathrm{T}},\, Y \succeq 0, \operatorname{Tr}(Y) = 1\}, \tag{2}$$

where the columns of the matrix Q form an orthonormal basis of the eigenspace of $\lambda_1(\mathcal{B}(x))$. Here and in what follows, $\mathcal{B}'(x)$ denotes the derivative of $\mathcal{B}$ at x, understood as a linear operator $\mathbb{R}^n \to \mathbb{S}^m$, while $\mathcal{B}'(x)^\star$ denotes its adjoint, mapping $\mathbb{S}^m \to \mathbb{R}^n$. A case of special interest is when $\mathcal{B}$ is quadratic:

$$\mathcal{B}(x) = A_0 + \sum_{i=1}^n x_i A_i + \sum_{i,j=1}^n x_i x_j B_{ij}.$$

Then $\mathcal{B}'(x)d = \sum_{i=1}^n d_i A_i + \sum_{i,j=1}^n (x_i d_j + x_j d_i) B_{ij}$, and the adjoint is obtained as

$$(\mathcal{B}'(x)^\star Z)_i \;=\; \left(A_i + \sum_{j=1}^n x_j B_{ij} + x_j B_{ji} \right) \cdot Z.$$

Our second example of a nonsmooth function is the pseudospectral abscissa. Following Trefethen [68], the pseudospectral abscissa of a matrix $A \in \mathbb{M}_m$ is defined as

$$\alpha_\varepsilon(A) \;=\; \max \{ \operatorname{Re} \lambda : \lambda \in \Lambda_\varepsilon(A) \},$$

where Λ_ε is the ε-pseudospectrum of A, that is, the set of all eigenvalues of matrices $A + E$ with euclidean norm $\|E\| \le \varepsilon$. For $\varepsilon = 0$ we recover $\alpha = \alpha_0$, the spectral abscissa, $\Lambda = \Lambda_0$ the spectrum of A. Our second class of nonsmooth functions is now of the form $g(x) = \alpha(\mathcal{A}(x))$ or $g(x) = \alpha_\varepsilon(\mathcal{A}(x))$, where $\mathcal{A}$ is a smooth operator defined for $x \in \mathbb{R}^n$ with values in the matrix space $\mathbb{M}_m$. Use of this function for static feedback synthesis was first proposed by Burke, Lewis, and Overton in [17, 18]. We will discuss this particular application in sections 6 and 8.1. The interest in $g = \alpha \circ \mathcal{A}$ is obviously due to the fact that $\mathcal{A}(x) \in \mathbb{M}_m$ is Hurwitz if and only if $g(x) < 0$. Notice that $g = \alpha \circ \mathcal{A}$ is smooth at x when $\alpha(\mathcal{A}(x)) = \operatorname{Re} \lambda_i(\mathcal{A}(x))$ for a single eigenvalue, where complex conjugate pairs are counted once. On the other hand, g is nonsmooth in general for multiple eigenvalues. What is worse is that neither $g = \alpha \circ \mathcal{A}$ nor $g = \alpha_\varepsilon \circ \mathcal{A}$ is locally Lipschitz function in general [17], which makes the functions somewhat delicate to handle.

Notice that function evaluation for α_ε may be based on the criss-cross method in [19], a generically globally quadratically convergent algorithm, which bears some resemblance with the Hamiltonian algorithm [12] to compute the H_∞-norm. For smooth points x, the criss-cross algorithm computes the gradient, while it still gives a subgradient of $\alpha_\varepsilon \circ \mathcal{A}$ at x if x is a nonsmooth point.

Our third example is the H_∞-norm. Notice that the stability requirement $\alpha_\varepsilon(A) < 0$ is equivalent to the estimate $\|(sI - A)^{-1}\|_\infty < \varepsilon^{-1}$. This means that α_ε could be avoided and replaced by composite functions of the H_∞-norm.

Consider the H_∞-norm of a nonzero transfer matrix function $G(s)$:

$$\|G\|_\infty \;=\; \sup_{\omega \in \mathbb{R}} \overline{\sigma}\left(G(j\omega)\right),$$

where G is stable and $\overline{\sigma}(X)$ is the maximum singular value of X. Suppose $\|G\|_\infty = \overline{\sigma}\left(G(j\omega)\right)$ is attained at some frequency ω, where the case $\omega = \infty$ is allowed. Let $G(j\omega) = U \Sigma V^{\mathrm{H}}$ be a singular value decomposition. Pick u the first column of U, v the first column of V, that is, $u = G(j\omega)v/\|G\|_\infty$. Then the linear functional

$$\begin{aligned} \phi(H) = \operatorname{Re}\left(u^{\mathrm{H}} H(j\omega) v\right) \;&=\; \|G\|_\infty^{-1} \operatorname{Re} \operatorname{Tr} v v^{\mathrm{H}} G(j\omega)^{\mathrm{H}} H(j\omega) \\ &= \|G\|_\infty^{-1} \operatorname{Re} \operatorname{Tr} G(j\omega)^{\mathrm{H}} u u^{\mathrm{H}} H(j\omega) \end{aligned}$$

is continuous on the space $\mathbf{H}_\infty$ of stable transfer functions and is a subgradient of $\|\cdot\|_\infty$ at G [13]. More generally, assume the columns of Q_u form an orthonormal basis of the eigenspace of $G(j\omega)G(j\omega)^{\mathrm{H}}$ associated with the largest eigenvalue $\lambda_1\left(G(j\omega)G(j\omega)^{\mathrm{H}}\right) = \overline{\sigma}(G(j\omega))^2$, and assume the columns of Q_v form an orthonormal

basis of the eigenspace of $G(j\omega)^H G(j\omega)$, associated with the same eigenvalue; then for every $Y_v \succeq 0$, $Y_u \succeq 0$ with $\mathrm{Tr}\,(Y_v) = 1$ and $\mathrm{Tr}\,(Y_u) = 1$,

(3)
$$\phi(H) = \|G\|_\infty^{-1}\,\mathrm{Re}\;\mathrm{Tr}\,Q_v Y_v Q_v^{\mathrm{H}} G(j\omega)^{\mathrm{H}} H(j\omega) = \|G\|_\infty^{-1}\mathrm{Re}\,\mathrm{Tr}\,G(j\omega)^{\mathrm{H}} Q_u Y_u Q_u^{\mathrm{H}} H(j\omega)$$

are subgradients of $\|\cdot\|_\infty$ at G, where Y_v and Y_u are (complex) Hermitian matrices. Finally, assume that $G(s)$ is rational, and that there exist finitely many frequencies $\omega_1, \ldots, \omega_p$ where the supremum $\|G\|_\infty = \overline{\sigma}(G(j\omega_\nu))$ is attained. Then the subgradients of $\|\cdot\|_\infty$ at G are precisely of the form

$$\phi(H) = \|G\|_\infty^{-1}\,\mathrm{Re}\;\sum_{\nu=1}^{p} \mathrm{Tr}\,G(j\omega_\nu)^{\mathrm{H}} Q_\nu Y_\nu Q_\nu^{\mathrm{H}} H(j\omega_\nu),$$

where the columns of Q_ν form an orthonormal basis of the eigenspace of $G(j\omega_\nu)$ $G(j\omega_\nu)^{\mathrm{H}}$ associated with the leading eigenvalue $\|G\|_\infty^2$, and where $Y_\nu \succeq 0$, $\sum_{\nu=1}^{p}$ $\mathrm{Tr}(Y_\nu) = 1$. See [22, Prop. 2.3.12 and Thm. 2.8.2] for this.

Suppose now we have a smooth operator $\mathcal{G}$, mapping $\mathbb{R}^n$ onto the space $\mathbf{H}_\infty$ of stable transfer functions G. Then the composite function $n(x) = \|\mathcal{G}(x)\|_\infty$ is Clarke subdifferentiable at x with

$$\partial n(x) \;=\; \mathcal{G}'(x)^\star[\partial\|\cdot\|_\infty\,(\mathcal{G}(x))],$$

where $\partial\|\cdot\|_\infty$ is the subdifferential of the H_∞-norm above. In section 6 we will compute this adjoint $\mathcal{G}'(x)^\star$ in a more specific situation. Suitable chain rules for this case are covered by [22, sect. 2.3].

3. Nonsmoothness in control. In automatic control, difficulties with computing derivatives arise frequently. This happens, for instance, when design specifications include time-domain constraints (settling-time, overshoot) and function evaluations depend on simulations or experiments. But even genuine nonsmoothness arises when criteria like the maximum eigenvalue function, the spectral abscissa, or the H_∞-norm are optimized. For a large class of problems in robust control theory, these nonsmooth criteria can be avoided since a smooth reformulation is available. The price to pay is a significant increase of the number of variables. There are situations where this becomes the major impediment to currently available optimization codes.

The situation we have in mind occurs for problems where bilinear matrix inequalities (BMIs) arise:

(4)
$$\begin{array}{ll} \text{minimize} & a^{\mathrm{T}}x + b^{\mathrm{T}}y, \quad x \in \mathbb{R}^r, \quad y \in \mathbb{R}^s \\ \text{subject to} & A_0 + \displaystyle\sum_{i=1}^{r} x_i A_i + \sum_{j=1}^{s} y_j B_j + \sum_{\ell=1}^{r}\sum_{k=1}^{s} x_\ell y_k C_{\ell k} \preceq 0, \end{array}$$

with $a \in \mathbb{R}^r$, $b \in \mathbb{R}^s$ and $A_i, B_j, C_{\ell k} \in \mathbb{S}^m$ given. Typically in (4) the decision vector splits into $x \in \mathbb{R}^r$, which gathers all free components or gains in the controller to be designed, while $y \in \mathbb{R}^s$ regroups the Lyapunov variables. All our examples discussed in section 8 may be brought to this form. In order to understand the problem better, let us discuss an application of particular importance.

3.1. Static output-feedback synthesis. It is well known that static output H_2- or H_∞-synthesis are NP-hard problems (cf. [56]), which may be cast as BMI-optimization programs. Given the plant

$$\begin{bmatrix} \dot{x} \\ z \\ y \end{bmatrix} = \begin{bmatrix} A & B_1 & B_2 \\ C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & 0 \end{bmatrix} \begin{bmatrix} x \\ w \\ u \end{bmatrix}$$

with $x \in \mathbb{R}^{n_1}$, $u \in \mathbb{R}^{m_2}$, $w \in \mathbb{R}^{m_1}$, $y \in \mathbb{R}^{p_2}$, $z \in \mathbb{R}^{p_1}$, we ask for a static feedback control law $u = Ky$ such that the closed-loop system is internally stable and, moreover, a suitable operator norm of the performance channel $w \to z$ is minimized. For the H_∞-norm, the existence of such a K with the norm estimate $\|T_{w\to z}(K)\|_\infty < \gamma$ is equivalent to the existence of a Lyapunov matrix $Y \in \mathbb{S}^{n_1}$ satisfying $Y \succ 0$ and

$$\begin{bmatrix} (A+B_2KC_2)^{\mathrm{T}}Y + Y(A+B_2KC_2) & Y(B_1+B_2KD_{21}) & (C_1+D_{12}KC_2)^{\mathrm{T}} \\ * & -\gamma I & (D_{11}+D_{12}KD_{21})^{\mathrm{T}} \\ * & * & -\gamma I \end{bmatrix} \prec 0. \tag{5}$$

If we optimize the gain γ, we obtain a BMI program (4) with unknown variables $\gamma \in \mathbb{R}$, $K \in \mathbb{R}^{m_2\times p_2}$, and $Y \in \mathbb{S}^{n_1}$. We may identify $x \in \mathbb{R}^r$ with the true decision variables γ and K, so $r = 1 + m_2p_2$, while $y \in \mathbb{R}^s$ gathers the Lyapunov variables Y, so $s = n_1(n_1+1)/2$. If the system size n_1 is large, the number of Lyapunov variables is dominant. A somewhat extreme example is the Boeing 767 under flutter condition (AC10), treated in section 8, where $n_1 = 55$, while $m_2 = p_2 = 2$. Here the BMI problem has 1490 variables, while there are only 4 true decision parameters (see [47, 24] for details).

The BMI problem (4) can be handled via smooth techniques by exploiting stationarity conditions [41] or via interior-point methods [36] and [49, 48]. An alternative is to use augmented Lagrangian techniques like Mosheyev and Zibulevsky [54]; see also [45] and [65]. Their approach extends naturally to nonlinear SDPs like (4). Unfortunately, all these approaches lead to large-size optimization problems even for control problems of moderate sizes due to the presence of Lyapunov variables y. One way to partly alleviate the difficulty in the nonlinear case is to use the projection lemma [27], whenever possible, to reduce at least the number of variables in x. The new cast is then a program with LMI constraints in tandem with nonlinear equality constraints:

$$\min\left\{ c^{\mathrm{T}}y : A_0 + \sum_{i=1}^{r} y_iA_i \preceq 0,\ h(y) = 0 \right\} \tag{6}$$

where $h : \mathbb{R}^n \to \mathbb{R}^p$ represents a finite number of nonlinear equality constraints. As suggested by our notation, the projection lemma reduces the x part in (4) to size $r = 1$ (to size $r = 0$ for pure stabilization), but gives only a slight reduction of the number s of Lyapunov variables y. The additional benefit of the projection lemma is that it avoids the redundancies of the controller state-space representations. For static output-feedback stabilization (northwest $(1,1)$ block in inequality (5)), a controller-free version is as follows: A stabilizing static controller K exists if there

exist Lyapunov matrices Y_1, $Y_2 \in \mathbb{S}^{n_1}$ such that

$$\begin{aligned} \mathcal{N}_Q^{\mathrm{T}} \left(A^{\mathrm{T}} Y_1 + Y_1 A\right) \mathcal{N}_Q &\prec 0, \\ \mathcal{N}_P^{\mathrm{T}} \left(A Y_2 + Y_2 A^{\mathrm{T}}\right) \mathcal{N}_P &\prec 0, \\ \begin{bmatrix} Y_1 & I \\ I & Y_2 \end{bmatrix} \succ 0, \qquad & Y_1 Y_2 - I = 0, \end{aligned}$$

where $\mathcal{N}_P$ and $\mathcal{N}_Q$ are bases of the nullspaces of C and B^{T}, respectively. A version including H_∞-norm performance has the same form and may be found, e.g., in [59].

Different techniques have been developed to solve problems (5), (6) or problems with more general matrix inequality and equality constraints. Leibfritz and Mustafa [49, 48] use interior-point techniques in tandem with ideas from sequential quadratic programming to separate Lyapunov and true decision variables in the tangent programs. A successive SDP approach is given in [25] and an augmented Lagrangian approach in [6]. These techniques are supported by local and global convergence theory [59], but have shown some limitations:

- Our experiments have revealed size limitations to about 1500 variables [5]. This allows solving problems with up to $n_1 = 40$ states.
- The transformation of (4) into (6) is not always possible. Only a restricted and well-identified class of problems is amenable to the projection lemma. A prominent case where this is *not* possible is simultaneous stabilization, considered in section 8.

In our testing, we have compared the nonsmooth MDS method to the BMI-based methods in [5, 65] (see the corresponding column in Table 2).

4. Nonsmoothness by avoiding Lyapunov variables. For large systems, the number $s = n_1(n_1+1)/2$ of Lyapunov variables y is a serious obstacle to the BMI-optimization approach (4) or (6). It seems natural to consider alternatives where Lyapunov variables y can be avoided, so that the optimization concentrates on the true decision variables $x = (\gamma, K)$. This is possible if one accepts nonsmooth optimization programs. Here we propose to replace (5) by the following constrained program:

$$\begin{array}{ll} \text{minimize} & \|T_{w\to z}(K,s)\|_\infty \\ \text{subject to} & \alpha_\varepsilon\left(A + B_2 K C_2\right) \le 0, \\ & K \in \mathbb{R}^{m_2 \times p_2}, \end{array} \tag{7}$$

for fixed $\varepsilon \ge 0$, where the performance channel $w \to z$ is specified by the transfer function

$$\begin{gathered} T_{w\to z}(K,s) \;=\; \mathcal{C}(K)\left(sI - \mathcal{A}(K)\right)^{-1}\mathcal{B}(K) + \mathcal{D}(K), \\ \mathcal{A}(K) := A + B_2 K C_2,\; \mathcal{B}(K) := B_1 + B_2 K D_{21},\; \mathcal{C}(K) := C_1 + D_{12} K C_2, \\ \mathcal{D}(K) := D_{11} + D_{12} K D_{21}. \end{gathered} \tag{8}$$

An alternative is the constrained program

$$\begin{array}{ll} \text{minimize} & \|T_{w\to z}(K,s)\|_\infty \\ \text{subject to} & \left\|\left(sI - \mathcal{A}(K)\right)^{-1}\right\|_\infty \le \varepsilon^{-1}, \\ & K \in \mathbb{R}^{m_2 \times p_2}. \end{array} \tag{9}$$

Notice that in both programs, the controller K has to be stabilizing, or what is the same, iterates have to be feasible. This requires a feasible initial point K^0, which we compute by the unconstrained optimization program (with $\varepsilon \ge 0$ fixed):

$$\text{minimize} \quad \alpha_\varepsilon\left(A + B_2 K C_2\right), K \in \mathbb{R}^{m_2 \times p_2}. \tag{10}$$

Using (10) for static feedback control has first been proposed in [17, 19].

Remark. Notice an important difference between programs like (7), (9) and program (10), used to initialize the others. While all programs encountered are nonconvex and often exhibit multiple local minima, it is usually satisfactory to accept a local minimum of the H_∞-norm in (7), (9), because the controller K is always stabilizing. This is different in program (10), where a local minimum K is useless as long as it satisfies $\alpha(A + B_2KC_2) \geq 0$, because it does not provide a stabilizing controller. In such a case, we have to restart the algorithm. Notice, however, that this does not mean that we require the full machinery of a global optimization technique, because we are not interested in *the* global minimum of (10). A value $\alpha < 0$ is all what is wanted. □

Similar nonsmooth formulations can be obtained for various other robust control problems, such as static and fixed-order stabilization, H_2- and H_∞-synthesis problems, simultaneous (multimodel) synthesis problems, control design with fixed structure controllers, robust synthesis and synthesis problems involving scaling and multipliers, and linear parameter-varying syntheses, to cite just a few.

Some of these problems are investigated in section 8. Our experiments seem to indicate that as soon as Lyapunov variables y in (4) dominate, nonsmooth programs like (7), (9) in conjunction with nonsmooth techniques are very attractive. The MDS algorithm and more general DS or pattern search techniques, supplemented by nonsmooth techniques, are serious alternatives to BMI- or LMI-based methods. This is most promising when the number of controller variables $x = (\gamma, K)$ is small. In our experiments, small means not more than 30–35 controller variables x. This situation occurs when simple controllers for large systems are sought. For problems with high-order controllers, a pure nonsmooth approach is inevitable. This is investigated in [3].

Remark. We end this paragraph by pointing the reader to a very important feature of optimization programs (7), (9), (10), which seem to invite techniques like MDS. Namely, in MDS and other search algorithms exact function evaluations can often be avoided. All that is needed is that we be able to compare the value of the objective at the different nodes to the current best value. This is in perfect agreement with function evaluations for α_ε, λ_1 and the H_∞-norm, which are all based on iterative procedures. For instance, the bisection algorithm for the H_∞-norm [12] need not be run to completion, a premature stopping criterion can be exploited to enhance efficiency. This renders our present approach open to larger problem sizes. □

5. The MDS algorithm with nonsmooth steps. In this section we give a description of the MDS algorithm and indicate in which way a nonsmooth step may be added to cope with nonsmoothness. For an in-depth discussion of MDS in the smooth case the interested reader is referred to [66].

The MDS algorithm requires a "seed" or base point v_0 and an initial simplex S in $\mathbb{R}^n$ with vertices $v_0, v_1, \ldots, v_n$. The vertices are then relabeled so that v_0 becomes the best vertex, that is, $f(v_0) \leq f(v_i)$ for $i = 1, \ldots, n$. The initial S is chosen from one of three different shapes; see Figure 1. The scaled simplex is used when prior knowledge on the problem scaling is available, but right-angled and regular simplices are generally preferred in the absence of information. The algorithm updates the current simplex S into a new simplex S^+ by performing three types of operations, which drive the search for a better point: reflection, expansion, and contraction; see Figure 2. First vertices $v_1, \ldots, v_n$ are reflected through the current best vertex v_0 to give $r_1, \ldots, r_n$. If a reflected vertex r_i gives a better function value than v_0, the algorithm tries an expansion step. This is done by increasing the distance between v_0 and r_i for

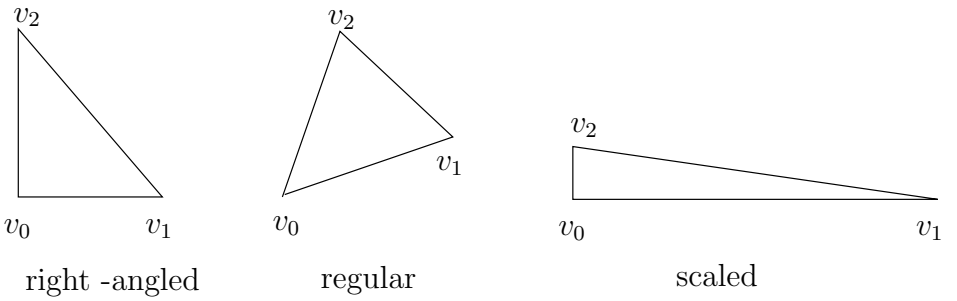


FIG. 1. *Selection of initial simplex.*

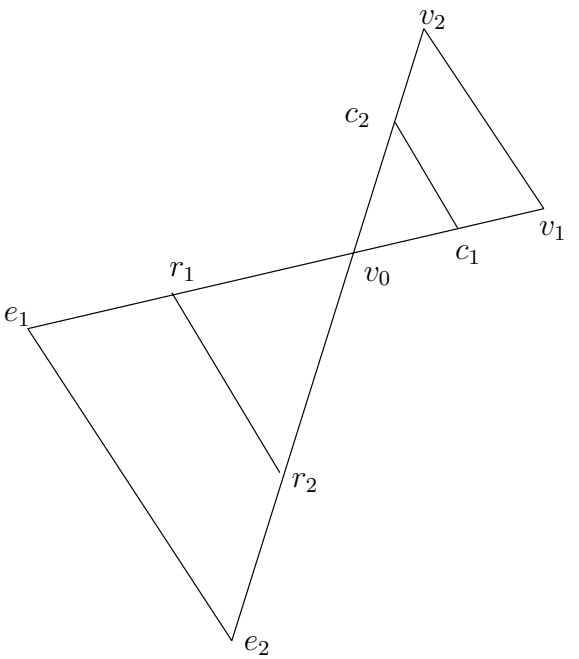


FIG. 2. *Reflection, expansion, and contraction of current simplex.*

$i = 1, \ldots, n$ and yields new expansion vertices e_i for $i = 1, \ldots, n$. The current simplex S is then replaced by either $S^+ = \{v_0, r_1, \ldots, r_n\}$ or $S^+ = \{v_0, e_1, \ldots, e_n\}$, depending on whether the best point was among the reflection or expansion vertices. If neither reflection nor expansion provide a point better than v_0, a contraction step is performed. This is done by decreasing the distances from v_0 to $v_1, \ldots, v_n$. If a point better than v_0 is found among the contraction vertices $c_1, \ldots, c_n$, the simplex S is replaced by $S^+ = \{v_0, c_1, \ldots, c_n\}$. To complete one iteration (or sweep) of the algorithm, v_0^+ is taken to be the best vertex of S^+.

In the presence of nonsmoothness, we endow the MDS algorithm with a fourth element. MDS may take a nonsmooth step w away from the current best node v_0 under consideration. In our applications, w will typically be the result of a nonsmooth descent step away from v_0, computed at the beginning of each sweep of MDS. If the sweep produces a new vertex v_0^+ better than w, MDS ignores w and keeps moving as

1. Select initial simplex $S = \{v_0, \ldots, v_n\}$, where v_0 is the best vertex. Fix an expansion factor $\mu \in (1, \infty)$ and a contraction factor $\theta \in (0, 1)$, and an intervention tolerance $\omega > 0$.
2. Given the current simplex S with best vertex v_0, call for a nonsmooth step w if the size of S is below threshold ω. If $w = v_0$ stop at critical point v_0.
3. Perform a reflection step $r_i = v_0 - (v_i - v_0)$. Compute $f(r_i)$.
4. If improvement $f(r_i) < f(v_0)$
   perform expansion step $e_i = (1 - \mu)v_0 + r_i$. Compute $f(e_i)$.
   If improvement $f(e_i) < f(v_0)$
   put $S^+ = \{v_0, e_1 \ldots, e_n\}$. Goto step 5.
   else
   put $S^+ = \{v_0, r_1 \ldots, r_n\}$. Goto step 5.
   else
   perform contraction step $c_i = (1 + \theta)v_0 - \theta r_i$. Compute $f(c_i)$. Put $S^+ = \{c_0, \ldots, c_n\}$.
5. Compare best vertex in S^+ to $f(w)$. If w is better, replace S^+ by new simplex containing w as a vertex. Otherwise accept S^+. Go back to step 2 to loop on.

FIG. 3. *MDS with nonsmooth steps.*

planned. On the other hand, if w is better than all the nodes tested by MDS during reflection, expansion, and contraction, we include w among the vertices of the new simplex S^+. In that event, we have to decide in which way the old vertices produced by MDS are recycled, or whether new nodes need to be created. This will obviously depend on geometrical properties. One possibility is to abandon the worst among the nodes of S^+ found by MDS and add the new node w as best point. If this produces angles below a certain threshold, one has to (partly) abandon S^+ and add new vertices to avoid bad geometry. In such a case, one can also build a completely new simplex with right-angled or regular geometry, using w as seed point. In our tests, we have observed that it is beneficial in such a situation to switch between the geometries (regular, right angled) in order to give MDS some additional help to move on. But all these considerations are clearly heuristic, depend on the context, and will need further testing.

In order to avoid serious slowdown of MDS, the nonsmooth step w is only solicited when the size of the simplex is below a certain threshold ω. Large S indicate that MDS is making good progress, so a costly nonsmooth step should be avoided. The situation we expect is that most of the time the point w is not better than the new best point v_0^+ of S^+ found by MDS. In that case, w plays a role similar to the Cauchy point in trust region methods. That is, it is hardly ever taken as the new iterate, but gives a convergence certificate. In our case, this will be made precise in Theorem 1. The different ways in which w may be computed will be explained subsequently. We sum up the above discussion in the pseudocode shown in Figure 3.

The following sections will show how the nonsmooth steps $v_0 \to w$ may be computed. From step 2 of the algorithm it is clear that the minimal requirement any w should satisfy is that $0 \notin \partial f(v_0)$ should give $f(w) < f(v_0)$, so when $w = v_0$, the algorithm stops with $0 \in \partial f(v_0)$.

The choice of the intervention tolerance ω should be compared to the usual stopping tests for smooth versions of MDS. Modern implementations use the relative size of the current simplex as a stopping test:

$$\frac{1}{\max(1,\ \|v_0\|)}\ \max_{1\le i\le n}\|v_i - v_0\| < \varepsilon, \tag{11}$$

where v_0 is the current best vertex of $S = \{v_0, \dots, v_n\}$ and $\varepsilon > 0$ is a prescribed tolerance. If a crisis intervention strategy is used, ω should be chosen slightly larger than the size (11). In the case of crisis prevention, an even larger ω is chosen.

The choice of the initial simplex S is a relatively unexplored topic. The convergence proof in [66] requires only that S be nondegenerate, which means that the $n+1$ points $\{v_0, v_1, \dots, v_n\}$ defining the simplex must span $\mathbb{R}^n$. Otherwise, MDS would only search over the subspace spanned by the degenerate simplex.

6. Nonsmooth stopping tests. Our first strategy is crisis intervention and uses a very small threshold ω. This means that the nonsmooth descent step $v_0 \to w$ is called for only when MDS gets stalled. What this essentially amounts to is a nonsmooth optimality test, which will either show that we are at a local minimum (or critical point) or give us a descent step $v_0 \to w$ to escape from the current point v_0, allowing MDS to move on. This strategy is preferable if nonsmooth descent steps are expensive. During the following we compute these steps for the criteria presented in section 2 and for the programs in section 4.

6.1. Maximum eigenvalue function. This case is well known. From the formula (2) of the Clarke subdifferential of $f = \lambda_1 \circ \mathcal{B}$ we see that $0 \in \partial f(x^*)$ if and only if the value t of the following semidefinite program is zero:

$$\min\{t : Q^{\mathrm{T}}[\mathcal{B}'(x^*)d]Q \preceq tI, \|d\| \le 1\}.$$

On the other hand, when the value is negative, the optimal solution (t, d) of this SDP gives the steepest descent direction d for $f = \lambda_1 \circ \mathcal{B}$ at x^*. If x^* is the current best vertex v_0 in MDS, then the nonsmooth stopping test either shows $0 \in \partial f(x^*)$ or produces w with $f(w) < f(x^*)$ of the form $w = x^* + \tau d$, where $\tau > 0$ is found by a suitable line search.

6.2. Spectral abscissa. This is a more difficult case. Consider the minimization program

$$\min_{x\in\mathbb{R}^n} g(x) = \alpha\left(\mathcal{F}(x)\right),$$

where $\mathcal{F} : \mathbb{R}^n \to \mathbb{M}_m$ is smooth. Since α is not even locally Lipschitz in general, we need a more elaborate way to obtain a stopping test.

Suppose MDS gets stalled at x^* and we want to know whether x^* is a local minimum of g or a dead point. We use the following lemma.

LEMMA 1. *Let $F \in \mathbb{M}_m$. Then $\alpha(F) \le t$ if and only if there exists $Y \in \mathbb{S}^m$, $0 \prec Y \prec I$, such that $F^{\mathrm{T}}Y + YF - 2tY \preceq 0$.* □

For a bounded set of matrices F, the condition number of Y is bounded. The inequality $Y \succ 0$ can therefore be replaced by $Y \succeq \theta I$ for a fixed small enough $\theta > 0$, uniformly over all F in that bounded set. Assume now that we have chosen an initial iterate x_0 such that $L = \{x \in \mathbb{R}^n : g(x) \le g(x_0)\}$ is bounded. Since we use a method of descent type, all our iterates x lie in L, so that the condition number of the

Lyapunov matrices Y arising at the corresponding $F = \mathcal{F}(x)$ are uniformly bounded: $\theta I \preceq Y \preceq I$ for some $0 < \theta \ll 1$. This allows us to consider the optimization program

$$\text{(P)} \qquad \begin{array}{ll} \text{minimize} & t \\ \text{subject to} & Y \succeq \theta I,\ Y \preceq I, \\ & \mathcal{F}(x)^{\mathrm{T}} Y + Y\mathcal{F}(x) - 2tY \preceq 0, \end{array}$$

with decision vector $(x, t, Y) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{S}^m$. Let $x^* \in L$. Define $F^* = \mathcal{F}(x^*)$ and $t^* = \alpha(F^*)$. Correspondingly, compute Y^* with $\theta I \preceq Y^* \preceq I$ such that $F^{*T}Y^* + Y^*F^* - 2t^*Y^* \preceq 0$. As a consequence of Lemma 1 we have the following proposition.

PROPOSITION 1. *$x^* \in L$ is a local minimum of $g = \alpha \circ \mathcal{F}$ if and only if (x^*, t^*, Y^*) is a local minimum of program* (P). □

In order to decide whether the latter is the case, we use a general result from [9]. Define $f(x, t, Y) = t$ and

$$\mathcal{G}(x,t,Y) = \begin{bmatrix} Y - I & 0 & 0 \\ 0 & \theta I - Y & 0 \\ 0 & 0 & \mathcal{F}(x)^{\mathrm{T}} Y + Y\mathcal{F}(x) - 2tY \end{bmatrix}. \tag{12}$$

Then (P) is equivalent to the abstract program

$$\min f(x,t,Y) \ \text{ subject to } \ \mathcal{G}(x,t,Y) \in \mathbb{S}^{3m}_{-}.$$

Assume that Robinson's constraint qualification [9] is satisfied for this program. Then if (x^*, t^*, Y^*) is a local minimum, the tangent program

$$\begin{array}{ll} \text{minimize} & f'(x^*,t^*,Y^*)^{\mathrm{T}}(\delta x, \delta t, \delta Y) \\ \text{subject to} & \mathcal{G}'(x^*,t^*,Y^*)(\delta x, \delta t, \delta Y) \in T(\mathbb{S}^{3m}_{-}, \mathcal{G}(x^*,t^*,Y^*)) \end{array} \tag{13}$$

has the unique solution $(\delta x, \delta t, \delta Y) = (0, 0, 0)$. Here $T(\mathbb{S}^{3m}_{-}, G)$ is the Clarke tangent cone, which according to [9] is $T(\mathbb{S}^{3m}_{-}, G) = \{Z \in \mathbb{S}^{3m} : Q^{\mathrm{T}} Z Q \preceq 0\}$ if $\lambda_1(G) = 0$, where the columns of the matrix Q are an orthonormal basis of the eigenspace of G associated with the maximum eigenvalue $\lambda_1(G) = 0$, while $T(\mathbb{S}^{3m}_{-}, G) = \mathbb{S}^{3m}$ if $\lambda_1(G) < 0$, $T(\mathbb{S}^{3m}_{-}, G) = \emptyset$ if $\lambda_1(G) > 0$.

It turns out that optimality of $(0, 0, 0)$ in (13) is a condition which may be checked by solving an SDP. Indeed, observe that

$$f'(x^*,t^*,Y^*)^{\mathrm{T}}(\delta x, \delta t, \delta Y) = \delta t$$

and

$$\mathcal{G}'(x^*,t^*,Y^*)(\delta x, \delta t, \delta Y) \ = \ \begin{bmatrix} \delta Y & 0 & 0 \\ 0 & -\delta Y & 0 \\ 0 & 0 & \delta Z \end{bmatrix},$$

where as before $\mathcal{G}'$ denotes the differential of the operator $\mathcal{G}$, and where we use the shorthand notation

$$\begin{aligned} Z^* &:= \mathcal{F}(x^*)^{\mathrm{T}} Y^* + Y^* \mathcal{F}(x^*) - 2t^* Y^*, \\ \delta Z &:= [\mathcal{F}'(x^*)\delta x]^{\mathrm{T}} Y^* + \mathcal{F}(x^*)^{\mathrm{T}} \delta Y + Y^* [\mathcal{F}'(x^*)\delta x] + \delta Y \mathcal{F}(x^*) - 2t^* \delta Y - 2\delta t Y^*. \end{aligned}$$

Clearly, the tangent cone in question is

$$T(\mathbb{S}^{3m}_{-}, \mathcal{G}(x^*,t^*,Y^*)) \ = \ T(\mathbb{S}^{m}_{-}, Y^* - I) \times T(\mathbb{S}^{m}_{-}, \theta I - Y^*) \times T(\mathbb{S}^{m}_{-}, Z^*),$$

so we have to compute these three tangent cones.

Let Q_1 be an orthonormal basis of the eigenspace of $Y^* - I$ associated with the eigenvalue 0, and let Q_θ be a basis of the eigenspace of $\theta I - Y^*$ associated with the eigenvalue 0. Finally, let P be a basis of the eigenspace of Z^* associated with the eigenvalue 0. Then the tangent program becomes

$$\begin{array}{ll} \text{minimize} & \delta t \\ \text{subject to} & Q_1^{\mathrm{T}} \delta Y Q_1 \preceq 0, \\ & Q_\theta^{\mathrm{T}} \delta Y Q_\theta \succeq 0, \\ & P^{\mathrm{T}} \delta Z P \preceq 0, \\ & \|\delta x\| \le 1,\ |\delta t| \le 1,\ \|\delta Y\| \le 1. \end{array} \tag{14}$$

This is an SDP in the unknown variable $(\delta x, \delta t, \delta Y)$. The decision is now as follows. If our tangent program reveals (x^*, t^*, Y^*) as a critical point, we stop and thereby accept the solution proposed by MDS. Otherwise, δx will show us the way to escape from the current point x^*. In terms of the MDS algorithm, when $x^* = v_0$, the nonsmooth descent step will be $w = x^* + \tau \delta x$ for some $\tau > 0$ found by a line search.

6.3. Stopping test for the H_∞-norm. For constrained programs like those in section 4, the situation is principally the same as in the unconstrained case. When we get stalled at some iterate K^*, we would like to know whether we have a local minimum (a KKT point), or whether we could keep making progress by avoiding a dead point.

In this section, we consider a stopping test for the nonsmooth program (9), which is based on the frequency domain representation of the H_∞-norm.

Suppose we have reached an iterate K^* such that $\|T_{w\to z}(K^*)\|_\infty = \gamma^*$ and $\|(sI - \mathcal{A}(K^*))^{-1}\|_\infty = \varepsilon^{-1}$. We want to decide whether K^* is a critical point of the program

$$\min\{\|T_{w\to z}(K)\|_\infty : \|(sI - \mathcal{A}(K))^{-1}\|_\infty \le \varepsilon^{-1}\}.$$

This may be based on a nonsmooth stationarity test, which checks whether or not $0 \in \partial n(K^*) + \mathbb{R}_+ \partial m(K^*)$, where $n(K) = \|T_{w\to z}(K)\|_\infty$, $m(K) = \max(0, \|\,(sI - \mathcal{A}(K))^{-1}\|_\infty - \varepsilon^{-1})$ (see [22, Thm. 6.1.1, Prop. 3.3.1]). We therefore need to compute the subdifferentials $\partial n(K^*)$ and $\partial m(K^*)$.

Let us start with $\partial n(K^*)$, which is more general. The subdifferential $\partial m(K^*)$ will then follow as a special case. Recall that $T_{w\to z}(K, s)$ is of the form

$$T_{w\to z}(K, s) = \mathcal{C}(K)(sI - \mathcal{A}(K))^{-1}\mathcal{B}(K) + \mathcal{D}(K),$$

where $\mathcal{A}(K)$, $\mathcal{B}(K)$, $\mathcal{C}(K)$, and $\mathcal{D}(K)$ are given in (8). Defining $\mathcal{F}(K, s) = (sI - \mathcal{A}(K))^{-1}$, we obtain the derivative $T'_{w\to z}$ of $T_{w\to z}$ at K^* as

$$\begin{aligned} T'_{w\to z}(K^*)\,\delta K(s) = {} & D_{12}\,\delta K\, C_2 \mathcal{F}(K^*, s)\mathcal{B}(K^*) \\ & + \mathcal{C}(K^*)\mathcal{F}(K^*, s)B_2\,\delta K\, C_2 \mathcal{F}(K^*, s)\mathcal{B}(K^*) \\ & + \mathcal{C}(K^*)\mathcal{F}(K^*, s)B_2\,\delta K\, D_{21} + D_{12}\,\delta K\, D_{21}. \end{aligned} \tag{15}$$

Now let $\phi = \phi_Y$ be a subgradient of $\|\cdot\|_\infty$ at $T_{w\to z}(K^*)$ of the form (3), specified by $Y \succeq 0$, $\mathrm{Tr}(Y) = 1$ and with $\|T_{w\to z}(K^*)\|_\infty$ attained at frequency ω. We wish

to compute $\Phi_Y := T'_{w\to z}(K^*)^\star \phi_Y \in \mathbb{M}_{m_2,p_2}$. The adjoint $T'_{w\to z}(K^*)^\star$ acts on ϕ_Y through

$$\begin{aligned}
&\langle T'_{w\to z}(K^*)^\star \phi_Y, \delta K\rangle \\
&\quad = \langle T'_{w\to z}(K^*)\delta K, \phi_Y\rangle \\
&\quad = \|T_{w\to z}(K^*)\|_\infty^{-1}\mathrm{Re}\,\mathrm{Tr}\,\left(T_{w\to z}(K^*,j\omega)^{\mathrm{H}} QYQ^{\mathrm{H}} T'_{w\to z}(K^*)\delta K\,(j\omega)\right) \\
&\quad = \|T_{w\to z}(K^*)\|_\infty^{-1}\mathrm{Re}\,\mathrm{Tr}\,\big[C_2\mathcal{F}(K^*,j\omega)\mathcal{B}(K^*)T_{w\to z}(K^*,j\omega)^{\mathrm{H}} QYQ^{\mathrm{H}} D_{12} \\
&\qquad + C_2\mathcal{F}(K^*,j\omega)\mathcal{B}(K^*)T_{w\to z}(K^*,j\omega)^{\mathrm{H}} QYQ^{\mathrm{H}}\mathcal{C}(K^*)\mathcal{F}(K^*,j\omega)B_2 \\
&\qquad + D_{21}T_{w\to z}(K^*,j\omega)^{\mathrm{H}} QYQ^{\mathrm{H}}\mathcal{C}(K^*)\mathcal{F}(K^*,j\omega)B_2 \\
&\qquad + \ D_{21}T_{w\to z}(K^*,j\omega)^{\mathrm{H}} QYQ^{\mathrm{H}} D_{12}\,\big]\,\delta K.
\end{aligned}$$

In consequence, the Clarke subgradients of $n = \|\cdot\|_\infty \circ T_{w\to z}$ at K^* are of the form

$$\begin{aligned}
\Phi_Y = \|T_{w\to z}(K^*)\|_\infty^{-1}\mathrm{Re}\,&\big[C_2\mathcal{F}(K^*,j\omega)\mathcal{B}(K^*)T_{w\to z}(K^*,j\omega)^{\mathrm{H}} QYQ^{\mathrm{H}} D_{12} \\
&+ C_2\mathcal{F}(K^*,j\omega)\mathcal{B}(K^*)T_{w\to z}(K^*,j\omega)^{\mathrm{H}} QYQ^{\mathrm{H}}\mathcal{C}(K^*)\mathcal{F}(K^*,j\omega)B_2 \\
&+ D_{21}T_{w\to z}(K^*,j\omega)^{\mathrm{H}} QYQ^{\mathrm{H}}\mathcal{C}(K^*)\mathcal{F}(K^*,j\omega)B_2 \\
&+ D_{21}T_{w\to z}(K^*,j\omega)^{\mathrm{H}} QYQ^{\mathrm{H}} D_{12}\big]^{\mathrm{T}},
\end{aligned}$$

or more simply,

$$\Phi_Y = \|T_{w\to z}(K^*)\|_\infty^{-1}\mathrm{Re}\,\left\{G_{21}(K^*,j\omega)\,T_{w\to z}(K^*,j\omega)^{\mathrm{H}} QYQ^{\mathrm{H}} G_{12}(K^*,j\omega)\right\}^{\mathrm{T}},$$

where

$$\begin{aligned}
G_{21}(K^*,j\omega) &:= C_2\mathcal{F}(K^*,j\omega)\mathcal{B}(K^*) + D_{21}, \\
G_{12}(K^*,j\omega) &:= \mathcal{C}(K^*)\mathcal{F}(K^*,j\omega)B_2 + D_{12}.
\end{aligned}$$

The subdifferential of the function $m(.)$ is obtained through similar calculations. We first note that up to a constant term, the second component of $m(.)$ is $\|\mathcal{F}(K)\|_\infty$, a simplification of $T_{w\to z}(K)$ with $\mathcal{C}(K) = I$, $\mathcal{B}(K) = I$, and $\mathcal{D}(K) = 0$. Assuming this time that the supremum is attained at frequency ω', the Clarke subgradients of $\|\mathcal{F}(K)\|_\infty$ at K^* are of the form

$$\Psi_{\widehat{Y}} := \|\mathcal{F}(K^*)\|_\infty^{-1}\mathrm{Re}\,\left\{C_2\mathcal{F}(K^*,j\omega')\mathcal{F}(K^*,j\omega')^{\mathrm{H}}\widehat{Q}\widehat{Y}\widehat{Q}^{\mathrm{H}}\mathcal{F}(K^*,j\omega')B_2\right\}^{\mathrm{T}},$$

with $\widehat{Y} \succeq 0$, $\mathrm{Tr}\,(\widehat{Y}) = 1$. Since both components of the max function $m(\cdot)$ are active at K^*, the subdifferential of m at K^* is the convex hull of the origin with the subdifferential of $\|\mathcal{F}(K)\|_\infty$ at K^* [22]. Those subgradients are therefore of the form $\Psi_{\widehat{Y}}$, $\widehat{Y} \succeq 0$, and $\mathrm{Tr}\,(\widehat{Y}) \le 1$. These formulae are easily adapted if the first H_∞-norm is attained at frequencies $\omega_1, \ldots, \omega_p$, and the second at $\omega'_1, \ldots, \omega'_q$.

Suppose $\|T_{w\to z}(K^*)\|_\infty$ is attained at a single ω, and $\|\mathcal{F}(K^*)\|_\infty$ at a single ω'. Then the optimality test leads to solving the optimization program

$$\min\{\|\Phi_Y + \Psi_{\widehat{Y}}\|_2 : Y \succeq 0, \mathrm{Tr}\,(Y) = 1, \widehat{Y} \succeq 0\,\}$$

which is a low-dimensional SDP. If the value of this program is 0, then K^* is a critical point.

7. Crisis prevention. The nonsmooth stopping tests developed in the previous section could be adapted to many other programs. We should be aware, however, that the steps $v_0 \to w$ they generate are steepest descent steps, which cannot guarantee convergence under nonsmoothness (see [50] for a discussion). Put differently, even though the stopping test may allow us to move on, we have no guarantee that an accumulation point of the sequence so generated would not be another dead point. In order to exclude this categorically, a more sophisticated strategy, crisis prevention, is required. Here we get a convergence certificate, which is built on the possibility to *quantify* descent.

A well-known tool of convex nonsmooth analysis which allows us to quantify descent is ε-subgradients (see [35, Thm. 1.1.5]). Since our present criteria are nonconvex, those may not be used directly and some modifications are required (see [57, 58]). But the idea is essentially the same.

7.1. Quantitative descent for $f = \lambda_1 \circ \mathcal{B}$. To begin with, let us examine a strategy suited for eigenvalue optimization, used in the simultaneous stabilization problem section 8.3. We consider a nonconvex maximum eigenvalue function of the form

$$f(x) = \lambda_1 \left(\mathcal{B}(x)\right) \tag{16}$$

with a bilinear (or more generally $\mathcal{C}^2$) operator $\mathcal{B}$. We solve the unconstrained optimization problem:

$$\text{minimize} \quad f(x) = \lambda_1 \left(\mathcal{B}(x)\right), \ x \in \mathbb{R}^n.$$

We follow [57, 58], which extends previous work by Cullum, Donath, and Wolfe [23] and Oustry [60], where affine operators were used, to more general functions $f = \lambda_1 \circ \mathcal{B}$. We use an approximation $\delta_\varepsilon f(x)$ of the ε-subdifferential $\partial_\varepsilon f(x)$ of f at the current x, called the *ε-enlarged subdifferential.* We compute the approximate subgradient $g \in \delta_\varepsilon f(x)$, which gives rise to the so-called *steepest ε-enlarged descent direction.* Let us define

$$\delta_\varepsilon f(x) \ = \ \left\{ \mathcal{B}'(x)^\star Z : Z = Q_\varepsilon Y Q_\varepsilon^{\mathrm{T}}, \ Y \succeq 0, \ \mathrm{tr}(Y) = 1, \ Y \in \mathbb{S}^{r(\varepsilon)} \right\},$$

where the first $r(\varepsilon)$ eigenvalues of $\mathcal{B}(x) \in \mathbb{S}^m$ are those which satisfy $\lambda_i > \lambda_1 - \varepsilon$, and where the columns of the $r(\varepsilon) \times m$-matrix Q_ε form an orthonormal basis of the invariant subspace associated with these eigenvalues. Then

$$\partial f(x) \subset \delta_\varepsilon f(x) \subset \partial_\varepsilon f(x),$$

and $\delta_\varepsilon f(x)$ is an inner approximation of $\partial_\varepsilon f(x)$, which has the advantage of being computable. Namely, the direction of steepest ε-enlarged descent d is obtained as

$$d = -\frac{g}{\|g\|}, \quad g = \mathrm{argmin} \left\{ \|g\| : g \in \delta_\varepsilon f(x) \right\}. \tag{17}$$

The solution g of (17) is the projection of the origin onto the compact convex set $\delta_\varepsilon f(x)$. This is in complete analogy with the direction of steepest descent, which is obtained by projecting the origin onto the subdifferential $\partial f(x) = \delta_0 f(x)$. What would be the most useful is the direction of steepest ε-descent, obtained by projecting 0 onto $\partial_\varepsilon f(x)$, but this quantity is difficult to compute (see, however, [35] for some ideas how this may be tried).

1. Given iterate x, stop if $0 \in \partial f(x) = \delta_0 f(x)$, because x is a critical point. Otherwise choose $\varepsilon > 0$.
2. Given $\varepsilon > 0$, compute the direction d of steepest ε-enlarged descent by solving (19). Let (t, d) be the solution.
3. If $d = 0$ (and hence $t = 0$), then $0 \in \delta_\varepsilon f(x)$. Decrease ε and go back to step 2.
4. If $d \neq 0$, then $0 \notin \delta_\varepsilon f(x)$ and we obtain $x^+ = x + \tau d$ with $f(x^+) < f(x)$ using a line search like in [57]. Let $w = x^+$ be the intervention step for MDS and quit.

FIG. 4. *Quantified descent $v_0 \to w$ for $f = \lambda_1 \circ \mathcal{B}$.*

Contrary to $\partial_\varepsilon f(x)$, the support function of the compact convex set $\delta_\varepsilon f(x)$ is known explicitly. We have (cf. [23, 60, 57])

$$\tilde{f}'_\varepsilon(x; d) \;:=\; \max\{g^{\mathrm{T}} d : g \in \delta_\varepsilon f(x)\} \;=\; \lambda_1\left(Q_\varepsilon^{\mathrm{T}}\,[\mathcal{B}'(x)d]\,Q_\varepsilon\right),$$

where $\tilde{f}'_\varepsilon(x; d)$ is the directional derivative considered in [23, 60]. Therefore, the direction of steepest ε-enlarged descent is found by solving the program

$$\min_{\|d\| \le 1} \lambda_1\left(Q_\varepsilon^{\mathrm{T}}\,[\mathcal{B}'(x)d]\,Q_\varepsilon\right), \tag{18}$$

and the solution $\mathbf{d} = -g/\|g\|$ satisfies

$$-\|g\| = -\mathrm{dist}\,(0, \delta_\varepsilon f(x)) = \tilde{f}'_\varepsilon(x; d) < 0.$$

Notice that (18) is equivalent to the SDP

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & Q_\varepsilon^{\mathrm{T}}\,[\mathcal{B}'(x)d]\,Q_\varepsilon \preceq tI, \\ & \|d\| \le 1. \end{array} \tag{19}$$

A descent direction d for $f = \lambda_1 \circ \mathcal{B}$ at x is therefore found as soon as the value of (19) is negative, and the corresponding d gives even a quantifiable descent in the sense of Theorem 1 below. The appealing feature of this method is that the size of the LMI in (18) and (19) is $r(\varepsilon)$, which is usually small. An important consequence is that it can be solved very cheaply if a dual SDP formulation is used. Altogether we have the crisis prevention method shown in Figure 4.

The possible decrease $f(x^+) < f(x)$ is quantified by the following result, whose proof is given in [57] for a spectral bundle algorithm which generates descent steps as above. Since the convergence properties of the nonsmooth MDS method hinge on the properties of the sequence of Cauchy points w, the result carries over to our present situation.

THEOREM 1. *Consider the minimization of $f = \lambda_1 \circ \mathcal{B}$. Suppose x^0 is such that $\{x \in \mathbb{R}^n : f(x) \le f(x^0)\}$ is compact. Let the sequence x^k with starting point x^0 be generated by the MDS method with nonsmooth descent step. Suppose at stage k the parameter ε_k is chosen according to the ε-management of* [57, 58]. *Then there exists a constant $C > 0$ such that the nonsmooth MDS method achieves a decrease of at least $f(x^{k+1}) - f(x^k) \le -C\,\Delta_{\varepsilon_k}\,|\tilde{f}'_{\varepsilon_k}(x^k; d^k)|^2$, where d^k is the direction of steepest ε_k-enlarged descent at x^k and $\Delta_{\varepsilon_k} = \lambda_{r(\varepsilon_k)} - \lambda_{r(\varepsilon_k)+1}$. Moreover, some subsequence of x^k converges to a critical point of f.* ∎

7.2. Quantifiable descent for $g = \alpha \circ \mathcal{F}$. In this section we discuss the difficult case of the spectral abscissa. Due to its highly nonsmooth character, quantified decrease for $g = \alpha \circ \mathcal{F}$ is more difficult to guarantee than for $f = \lambda_1 \circ \mathcal{B}$.

Let us again take recourse to the SDP formulation of α. Suppose $g(x^*) = \alpha(\mathcal{F}(x^*)) = t^*$. We wish to decrease the value of g in a neighborhood U of x^*. Following Lemma 1, for fixed $0 < \theta \ll 1$, there exists $Y^* \in [\theta I, I]$ such that $\lambda_1(\mathcal{B}(x^*, Y^*, t^*)) = 0$, where we define $\mathcal{B}(x, Y, t) := \mathcal{F}(x)^{\mathrm{T}} Y + Y\mathcal{F}(x) - 2tY \preceq 0$. Finding Y^* amounts to solving an SDP. Now let us introduce

$$\widetilde{\mathcal{B}}(x, Y, t) = \begin{bmatrix} Y - I & 0 & 0 \\ 0 & \theta I - Y & 0 \\ 0 & 0 & \mathcal{B}(x, Y, t) \end{bmatrix}.$$

Then decreasing the value $g(x) = t$ below t^* is equivalent to decreasing the value t of the program

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & \widetilde{\mathcal{B}}(x, Y, t) \preceq 0 \end{array}$$

below t^*. We obtain such a decrease $t < t^*$ using Kiwiel's progress function [43], which in our situation may be written as

$$\kappa(x, Y, t; t^*) = \lambda_1 \begin{bmatrix} t - t^* & 0 \\ 0 & \widetilde{\mathcal{B}}(x, Y, t) \end{bmatrix} =: \lambda_1 \left(\widehat{\mathcal{B}}(x, Y, t; t^*) \right).$$

We have the following.

LEMMA 2. *Suppose $g(x^*) = \alpha(\mathcal{F}(x^*)) = t^*$. Then decrease $t = g(x) < g(x^*) = t^*$ is achieved for some x in a neighborhood U of x^* if and only if $\kappa(x, Y, t; t^*) < 0$ for suitable Y.* ∎

What we are interested in is quantified decrease in the same sense as used before, so we use the ε-enlarged subdifferential $\delta_\varepsilon \kappa$ of the maximum eigenvalue function $\kappa = \lambda_1 \circ \widehat{\mathcal{B}}$. The procedure, whose convergence theory is covered by [57], is shown in Figure 5.

Notice that the costly part here is computing Y^*. The second SDP in step 3 is of small size, since the corresponding LMI is in the space of $r(\varepsilon) \times r(\varepsilon)$ matrices. Repeating this step to identify a suitable ε is therefore not expensive. This has the interesting feature that as long as ε-steepest descent steps are taken, the large SDP need not be solved at all. This makes a pure nonsmooth descent method seem attractive. Such an approach is developed in [32] for large SDPs arising as relaxations of integer programs. Similar to that reference, solving the SDP dual of (7.2) is more efficient. Finally, we stress that extending the quantified descent step for the spectral abscissa to a broader class of problems like those in (4) is straightforward and left to the reader.

Remark. As soon as search directions based on the ε-enlarged subdifferential are used, a good choice of ε is required. Based on extensive numerical testing, we have used a very small $\varepsilon = 1e{-}9$ for stopping tests, while good progress in a descent step seems to ask for moderate values $\varepsilon \in [0.01\,; 0.1]$. This is what has been used in section 8.

8. Numerical experiments. In this section, we test the MDS algorithm with nonsmooth descent steps on a wide range of synthesis problems from the literature.

1. Given $g(x^*) = \alpha\left(\mathcal{F}(x^*)\right) = t^*$, quit if the stopping test (14) indicates a critical point. Otherwise:
2. Solve an SDP to compute Y^* such that $\lambda_1\left(\mathcal{B}(x^*, Y^*, t^*)\right) = 0$. Choose $\varepsilon > 0$.
3. Given $\varepsilon > 0$, compute $d = (\delta x, \delta Y, \delta t)$, the direction of steepest ε-enlarged descent of $\kappa(\cdot,\cdot,\cdot; t^*)$ at the point (x^*, Y^*, t^*) by solving the SDP:

$$\begin{aligned} &\text{minimize} && \rho \\ &\text{subject to} && \widehat{Q}_\varepsilon^{\mathrm{T}}\left[\widehat{\mathcal{B}}'(x^*, Y^*, t^*; t^*)\, d\right]\widehat{Q}_\varepsilon \preceq \rho I, \\ &&& \|\delta x\| \le 1,\ \|\delta Y\| \le 1,\ |\delta t| \le 1. \end{aligned}$$

   Here the $r(\varepsilon)$ columns of $\widehat{Q}_\varepsilon$ are an orthonormal basis of the invariant subspace of $\widehat{\mathcal{B}}(x^*, Y^*, t^*; t^*)$ associated with its ε-largest eigenvalues. Let $d = (\delta x, \delta Y, \delta t)$ be the solution.
4. If $d = 0$, then $0 \in \delta_\varepsilon \kappa(x^*, Y^*, t^*; t^*)$. Decrease ε and go back to step 3.
6. Having found $d \neq 0$, decrease the value of κ using a line search as in [57]. The corresponding step $x^+ = x^* + \tau\delta x$ decreases g accordingly. Let $w = x^+$ be the intervention step for MDS, and quit.

FIG. 5. *Quantified descent step $v_0 \to w$ for $g = \alpha \circ \mathcal{F}$.*

Computations were performed on a (low-level) SUN-Blade Sparc with 256 RAM and a 650 MHz sparcv9 processor. LMI-related computations needed for nonsmooth descent steps were performed using either the LMI Control Toolbox [28] or our homemade SDP code [5]. The contraction and expansion parameters were set to $\theta = 0.5$ and $\mu = 2.0$ throughout.

8.1. Static output-feedback stabilization. We start with static output-feedback stabilization without any performance specification. Solving (10) is a pure feasibility problem and somewhat simpler than the problems examined in what follows. It is used to initialize the constrained problem (9).

In our implementation, the MDS code was stopped as soon as a strictly negative spectral abscissa was obtained. Restarts were used as soon as the nonsmooth optimality test indicated a local minimum $\bar{x}$ of $g = \alpha \circ \mathcal{F}$ with positive value $g(\bar{x}) > 0$. We also encountered dead points, where the nonsmooth stopping test indicated a way to move on. What helps in this case is to restart MDS with the new seed proposed by the spectral bundling step, and change the geometry of the simplex. In all tests, the initial seed point was chosen to be the origin of the variable space. The vertices of the initial S are then relabeled so that v_0 is the best vertex. Contrary to what might seem plausible, MDS frequently encounters dead points and fails when run in default mode without nonsmooth steps. We discuss some of these at the end of this section.

As emphasized in the introductory section, the nonsmooth MDS is fairly insensitive to the number of states, since Lyapunov variables are not involved. A striking example is the Boeing 767 flutter problem (AC10), which our algorithm solved in 0.41-s cpu, starting from the initial point $K = 0$. This indicates that this problem is not as difficult as the size would suggest. (In fact, some of our smaller problems turned out more difficult.) The nonsmooth MDS technique appears surprisingly efficient compared to the gradient sampling algorithm proposed in [18], which for this problem required hours of cpu time and several hundreds of restarts. This example is

TABLE 1
Static output-feedback stabilization right-angled simplex.

Problem	(n, m, p)	Iteration	cpu (s)	Reference
Transport airplane	$(9, 1, 5)$	3	0.05	[29]
Horisberger's example	$(9, 1, 4)$	13	0.12	[38]
VTOL helicopter	$(4, 2, 1)$	1	0.01	[42]
Chemical reactor	$(4, 2, 2)$	2	0.02	[39]
Piezoelectric actuator	$(5, 1, 3)$	2	0.17	[21]
AC10	$(55, 2, 2)$	3	0.41	[47]
HF1	$(130, 1, 2)$	Stable	–	[47]

also included in the library [47] and has been solved by the technique of [49, 48].

Example. Let us illustrate a typical difficulty related to nonsmoothness of the spectral abscissa, when MDS stops at an iterate K^* where several eigenvalues of the closed-loop spectrum are active. This happens, e.g., in Horisberger's example with seed point at the origin and with the regular simplex geometry. When nonsmooth descent is switched off, MDS eventually hits such a nonsmooth iterate and starts contracting the simplex. This yields the static (nonstabilizing) gain

$$K = [\,-5.9176e{-}01 \quad 7.1864e{+}00 \quad -3.1396e{+}01 \quad 3.5870e{+}01\,],$$

with closed-loop spectrum

$$\Lambda(A + B_2KC_2) = \begin{cases} -6.6646e{-}01 \pm 6.2303e{+}01j \\ -3.9851e{+}00 \pm 1.8336e{+}01j \\ -7.8086e{+}00 \pm 4.0906e{+}00j \\ 5.4005e{-}01 \pm 8.3040e{-}01j \\ 5.4005e{-}01 \end{cases}.$$

The question is now whether we are at a dead point or at a local minimum. If the technique discussed in section 6 is switched on, a nonsmooth descent step $v_0 \to w$ is performed, which reduces the spectral abscissa from $5.4005e{-}01$ to $5.183e{-}01$. This is followed by a number of reflection/expansion/contraction steps of MDS, yielding the iterate

$$K = [\,-2.0595e-01 \quad 6.4949e+00 \quad -3.1503e+01 \quad 3.6173e{+}01\,]$$

with closed-loop spectrum

$$\Lambda(A + B_2KC_2) = \begin{cases} -6.7523e{-}01 \pm 6.2320e{+}01j \\ -4.1595e{+}00 \pm 1.8393e{+}01j \\ -7.5240e{+}00 \pm 4.9624e{+}00j \\ 4.7250e{-}01 \pm 3.8239e{-}01j \\ 4.7250e{-}01 \end{cases}.$$

The nonsmooth stopping test now clearly identifies this as a local minimum (a critical point), since no descent direction exists. At this stage a restart of MDS is inevitable, because $\alpha(A + B_2KC_2) = 4.7250e{-}01 > 0$.

Our testing has shown that the following simple trick is successful when a restart is due. We keep the current best point but switch geometries, for instance from regular to right-angled or vice versa. In the example, we switched from regular to right-angled

simplices, which generated different search directions. MDS was now successful and reached a stabilizing gain:

$$K = [\,3.1794e{+}01 \quad 6.4949e{+}00 \quad 4.3250e{+}02 \quad 5.1173e{+}01\,],$$

with corresponding spectral abscissa $\alpha(A + B_2KC_2) = -4.1442e{-}01 < 0$. ∎

8.2. Static and fixed-order output-feedback H_∞-synthesis. This section reports experiments with static and fixed-order output-feedback H_∞-synthesis. The out-set is from section 3, the extension to fixed-order problems is standard [6]. We solve program (9), using the corresponding controllers K^0 computed via (10) as initial value.

Results achieved with nonsmooth MDS are based on the infinite barrier

$$B(K) = \begin{cases} \|T_{w\to z}(s,K)\|_\infty & \text{if } \alpha(A + B_2KC_2) \le -\tau, \\ +\infty & \text{otherwise,} \end{cases} \tag{20}$$

where $\tau > 0$ is some small fixed threshold. The infinite barrier function works surprisingly well with the MDS technique, as also witnessed by [10] in different contexts. Function evaluation for the H_∞-norm is based on the efficient bisection algorithm in [12]. See also the MATLAB implementation described in [26]. A catalog of results is displayed in Table 2. The H_∞ performance achieved with the MDS method "H_∞ MDS" as well as with the spectral quadratic SDP method "H_∞ AL" in [5] are described. For completeness, in column "H_∞ full" the performance of the full-order H_∞ controller (computed by LMIs or algebraic Riccati equations) is shown and gives a lower bound for the H_∞-gain.

TABLE 2
Static and fixed-order H_∞-synthesis with MDS algorithm best results with right-angled and regular simplices stopping tolerance $\varepsilon = 1e{-}9$.

Problem	Order	Iteration	cpu (s)	H_∞ MDS	H_∞ AL	H_∞ full
Transport airplane	Static	37	20	2.34	2.22	1.60
VTOL helicopter	Static	10	2.69	0.190	0.157	0.096
Chemical reactor	Static	38	16.96	1.183	1.202	1.141
Piezoelectric actuator	Static	112	8.62	$1.76e{-}4$	$3.05e{-}3$	$9.63e{-}5$
AC10	Static	72	612	14.22	Intractable	0.052
HF1	Static	11	1100	0.447	Intractable	0.449

The choice of the simplex geometry, right-angled or regular, may influence the computed solution. Contrary to what might be guessed, the regular geometry is not always better than the right-angled geometry. We have therefore decided to test both and report the best result. This is reasonably affordable with regard to cpu time, as seen in Table 2 even for high-order systems. The initial seed point was the origin in all examples. As already discussed in [5], the augmented Lagrangian (AL) technique is no longer operational for systems with roughly more than 40 states. Again, we would like to stress the good results obtained with the MDS method for the Boeing 767 problem (AC10). Actually, the projective SDP code of MATLAB ran into difficulties to solve the LMI problem corresponding to the (convex) full-order problem and diagnosed the problem as infeasible after more than 4 hours of execution time in default mode.

The computed static controller obtained by the MDS method for the Boeing 767 flutter problems (AC10) is

$$K_{\text{static}} = \begin{bmatrix} -0.0966 & 0.0000 \\ 3.1681 & 0.0000 \end{bmatrix}.$$

The large-size HF1 problem is taken from the library [47]. It does not require prior stabilization as the plant is open-loop stable. Hence, $K = 0$ may serve as a starting point for the H_∞-optimization in Table 2. Here the static gain $K = [\,1.9943 \quad -3.4943\,]$ is found.

8.3. Simultaneous stabilization problems. Simultaneous stabilization is a longstanding problem in the automatic control literature. It consists in the search of a single controller which stabilizes a finite set of plants. This is of great practical interest in different situations. A system may have several modes of operation, but the controller is required to stabilize all modes. A more challenging situation is when the system may be subject to different failures such as actuator/detector breakdown, which often result in drastic deviations of the plant from its nominal description. The controller is then required to stabilize normal and abnormal operating modes. Unfortunately, the simultaneous stabilization problem has no analytical solution for more than two plants and is classified as NP-hard [8]. Existing techniques usually try to verify sufficient conditions. If successful, this leads to high-order controllers. Our experiments indicate that local optimization techniques and in particular DS methods may be of interest for designing simpler controllers, which is crucial for applications.

For single-input single-output systems $\{G_i(s),\, i = 1, \ldots, q\}$, the simultaneous stabilization problem can be formulated as follows:

- find a controller with transfer function

$$K(s,x) = \frac{N_K(s,x)}{D_K(s,x)} = \frac{x_1 s^m + \cdots + x_m s + x_{m+1}}{s^n + x_{m+2}s^{n-1} + \cdots + x_{m+n}s + x_{m+n+1}}, \tag{21}$$

  where as before $x := [x_1 \cdots x_{m+n+1}]^{\mathrm{T}}$ gathers the decision variables,
- such that the closed-loop characteristic polynomials

$$p_i(s,x) := N_{G_i}(s)N_K(s,x) + D_{G_i}(s)D_K(s,x)$$

  have only stable roots for $i = 1, \ldots, q$.

This may be addressed by the optimization program

$$\underset{x \in \mathbb{R}^{m+n+1}}{\text{minimize}} \; \max_{i=1,\ldots,q} \; \mathrm{Re}\,(\text{roots of } p_i(s,x)) \tag{22}$$

and a simultaneous stabilizing controller is found as soon as the value of this program is < 0. Program (22) resembles the static stabilization formulation discussed in section 8.1 and we follow a similar line of attack.

A challenging variant of this problem is the strong stabilization problem, where the controller itself is required to be stable. This is incorporated into the cast (22) by just adding $D_K(s,x)$ to the family of plant polynomials.

TABLE 3
*Simultaneous stabilization with MDS right-angled simplex * strong stabilization problem.*

Problem	Order	Iteration	cpu (s)	Restart	Reference
F4e aircraft	Static	2	0.71	None	[2]
cao	Static	13	0.28	None	[20]
cao	1	1	0.25	3	[20]
henrion	1	2	0.51	None	[34]
bredemann1*	1	6	2.05	3	[16, p. 68]
bredemann2*	1	3	0.82	3	[15]

In this testing, the nonsmooth MDS was again successful on a list of applications from the literature. Restarts have been used with a different initial seed point when an unsatisfactory local minimum was encountered. Often we obtained simpler controllers than those previously published and derived from constructive sufficient conditions. For example, the method in [16] yields a fifth-order controller for example bredemann1, whereas the MDS technique was able to show that first-order strong simultaneous stabilization is possible. A similar comment applies to example bredemann2.

An alternative cast for simultaneous stabilization is via Hermite–Fujiwara matrices. In this setting, the nonsmooth program (22) reduces to a finite set of quadratic matrix inequality constraints [33]:

$$\mathcal{H}(x) := \sum_{i=1}^{m+n+1} \sum_{j=i}^{m+n+1} x_i x_j H_{ij} \prec 0,$$

where the decision vector x comprises controller parameters in (21). Here MDS is applied to the eigenvalue optimization program

$$\min_x \lambda_1 \left(\mathcal{H}(x)\right). \tag{23}$$

We apply MDS to a problem from [33], which consists in the simultaneous stabilization of four plants. Hence, x is required to be strictly feasible for four quadratic matrix inequalities of the form (23). This problem is of special interest because numerous dead points and unsatisfactory local minima were found if different seed points were used.

TABLE 4

Simultaneous stabilization using Hermite–Fujiwara BMI characterization final spectrum of quadratic SDP results with two starting points and regular simplices.

Seed	$1, 1, -1, -1$	$-1, -1, 1, 1$
Final iterate	$3.5068, 4.2139, 0.0925, 0.0925$	$-4.4420, 0.4275, 0.5059, 0.1618$
Final spectrum of quadratic SDP (23)	$-1.3846e{+}03$	$-2.2926e{+}03$
	$-1.0473e{+}03$	$-3.7468e{+}02$
	$-8.0116e{+}02$	$-3.1919e{+}02$
	$-3.9982e{+}02$	$-2.1174e{+}01$
	$-3.8359e{+}02$	$-6.7302e{+}00$
	$-2.0603e{+}02$	$-6.1145e{+}00$
	$-1.3928e{+}02$	$-2.3900e{+}00$
	$-8.4586e{+}01$	$-2.3453e{+}00$
	$-2.6890e{+}00$	$-5.0609e{-}02$
	$-1.4544e{+}00$	$1.8472e{-}01$
	$-7.6821e{-}01$	$1.8528e{-}01$
	$-6.3137e{-}01$	$1.8528e{-}01$
Controller	$\dfrac{3.5068s + 4.2139}{9.2540e{-}02s + 9.2540e{-}02}$	none

Example. For the purpose of testing, MDS was first run without nonsmooth steps. Table 4 shows two scenarios with default MDS. In column 2 the nonsmooth stopping test from section 7.1 was switched on as soon as MDS got stalled. It reveals that we are at a dead point and not at a local minimum. While nonsmooth steps $v_0 \to w$ allowed MDS to move on, crisis intervention ultimately did not lead to a stabilizing controller

in this case. The procedure gets again stalled and this time achieves convergence to an infeasible local minimum. □

Example (continued). In a second testing, we examined this case more closely. As it is too late to shut the stable door after the horse has gone, we opted to used the ε-descent nonsmooth technique of section 7.1 in order to avoid failure. We call for a nonsmooth step as soon as the MDS simplex shrinks below $\omega = 0.1$ in relative size. Starting with the same initial point, the simultaneous stabilization problem is now satisfactorily solved in a few iterations: four MDS iterations and a single call for the nonsmooth intervention technique of section 7.1. The evolution of the maximum eigenvalue of the quadratic SDP in (23) as a function of the iteration index is the five-element sequence

$$\{6.5072e{+}01,\ 1.5172e{+}01,\ 5.1720e{+}00,\ 2.868e{+}00,\ -1.4778e{+}00\},$$

where the nonsmooth descent step $v_0 \to w$ corresponds to the decrease from $5.1720e{+}00$ to $2.868e{+}00$. Note that since sole stabilization is of interest, the algorithm has been stopped as soon as the maximum eigenvalue was found negative. The associated first-order controller solution is described by the transfer function

$$K(s) = \frac{4.9843s - 4.2577}{4.2783e{-}01s + 6.2861e{-}01}. \qquad \square$$

Example (continued). In our third experiment, we assess the performance of the nonsmooth descent technique alone. We no longer sample the space using MDS. Instead we follow descent steps $v_0 \to w$ proposed by the nonsmooth technique in section 7.1. This option corresponds to a pure spectral bundle method [57]. With the same starting point causing failure of the default MDS, the problem is now solved in seven calls according to the sequence

$$\{65.0718,\ 43.0862,\ 39.8725,\ 20.1852,\ 19.9853,\ 3.5719,\ 2.8877,\ -0.0228\}.$$

The resulting stabilizing controller is

$$K(s) = \frac{0.6029s + 1.115}{0.03361s + 0.1064}.$$

All controllers computed in this application have significantly different pole/zero patterns, but all stabilize the four-plant family. □

8.4. Mixed H_2/H_∞ state-feedback synthesis. Mixed H_2/H_∞-synthesis with state- or output-feedback is one of those archetype problems which cannot be simplified using the projection lemma and resist to linearizing changes of variable like [30]. What remains are special BMI techniques or algorithmic approaches like the one we propose here. The mixed H_2/H_∞ state-feedback synthesis problem is as follows. Given a synthesis state-space representation

$$\begin{cases} \dot{x} = Ax + B_{1,2}w_2 + B_{1,\infty}w_\infty + B_2u, \\ z_2 = C_{1,2}x + D_{12,2}u, \\ z_\infty = C_{1,\infty}x + D_{11,\infty}w_\infty + D_{12,\infty}u, \end{cases}$$

the goal is to compute a state-feedback control law $u = Kx$ such that

- the closed-loop system is asymptotically stable, i.e., $\alpha(A + B_2K) < 0$,
- the H_2-norm of the channel $\|T_{w_2 \to z_2}(K, s)\|_2$ is minimized subject to an H_∞-norm constraint on the channel $\|T_{w_\infty \to z_\infty}(K, s)\|_\infty \le \gamma$.

An example of this type is given in [31], and we reexamine it here using our nonsmooth MDS. We proceed as follows. First a state-feedback gain satisfying both stability and the H_∞ constraint is computed as in section 8.2. In a second phase, the H_2-norm is added and minimized, using an infinite barrier

(24)
$$B(K) = \begin{cases} \|T_{w_2\to z_2}(K,s)\|_2 & \text{if } \alpha(A+B_2K) \le -\tau \text{ and } \|T_{w_\infty\to z_\infty}(K,s)\|_\infty \le \gamma, \\ +\infty & \text{otherwise,} \end{cases}$$

now maintaining the constraints of phase 1.

With data imported from [31] and $\gamma = 2$, MDS computed a gain K in 25 MDS iterations within 11.1 s of cpu time. The solution found is

$$K = [\,1.8236,\ 2.5648e{-}01,\ -2.0453e{-}01\,]$$

with $\|T_{w_2\to z_2}(K,s)\|_2 = 7.502e{-}01$. The H_∞-norm constraint was of course active at this point. Note *en passant* that this result outperforms those achieved via the spectral augmented Lagrangian method in [65], which gave $\|T_{w_2\to z_2}(K,s)\|_2 = 0.8384$, and the successive linearization approach in [31], which found $\|T_{w_2\to z_2}(K,s)\|_2 = 0.8930$. Since this problem has multiple local minima, this fact does not imply that any one of those methods is better than any other, except perhaps for cases where a solution without optimality certificate is presented. The solution in [65] is a local minimum, and we checked optimality of our present K by adapting the nonsmooth frequency domain test for program (9) from section 6. This requires not much extra work, as the H_2-norm is smooth (see [13]). We observed that the H_∞-norm is attained at a single frequency, which seems to be rather the rule than the exception.

9. Conclusion. We have proposed a new algorithmic strategy for difficult and even NP-hard synthesis problems in automatic control, which are inaccessible via convexity methods. Our algorithm combines DS methods like Torczon's MDS with spectral bundle techniques, imported from nonsmooth optimization, in order to cope with typical nonsmooth criteria in control like the spectral abscissa, the maximum eigenvalue function, or the H_∞-norm. Our approach is a serious alternative to nonlinear programming algorithms based on bilinear matrix inequalities, as long as the number of *controller* decision variables is not too large. Since our approach avoids Lyapunov variables, it may be used to design small- or medium-size controllers even for very large systems, as witnessed by the Boeing 767 and Heat Flow (HF1) benchmark examples, systems with 55 and 130 states, respectively. As soon as the number of controller gain parameters gets sizable, the search method is often too slow, and pure nonsmooth approaches like spectral bundling perform better. How those should be organized for control applications is discussed in [3]. In a similar vein, a pure nonsmooth and frequency-domain approach for solving multidisk problems is proposed in [4]. A nonsmooth spectral bundle method for solving state-space BMI programs is developed in [64].

Our approach combines MDS with suitable nonsmooth descent steps. This gives a convergence certificate toward critical points, an important feature lacking in all the heuristic approaches proposed to date. We have observed that MDS is fairly insensitive to noise corrupting the function evaluation. This makes it particularly useful in control applications, where objective functions typically result from iterative procedures to compute the H_∞- or H_2-norm. We have noticed that in the neighborhood of nonsmooth surfaces, gradient directions behave irregularly and are often distorted and unreliable, while progress is still achievable with MDS.

In conclusion, we believe the proposed framework is very versatile and can accommodate a vast array of design problems, expanding on those discussed in this paper. Structured feedback design is near at hand, while robust control is currently under investigation.

Acknowledgment. Fruitful discussions with Adrian Lewis and Michael Overton are gratefully acknowledged.

REFERENCES

[1] M. A. Abramson, C. Audet, and J. E. Dennis Jr., *Generalized pattern searches with derivative information*, Math. Program. Ser. B, 100 (2004), pp. 3–25.
[2] J. Ackermann, *Robust Control. Systems with Uncertain Physical Parameters*, Springer-Verlag, Berlin, Heidelberg, New York, 1993.
[3] P. Apkarian and D. Noll, *Nonsmooth H_∞ Synthesis*, submitted.
[4] P. Apkarian and D. Noll, *Nonsmooth Optimization for Multidisk H_∞ Synthesis*, submitted.
[5] P. Apkarian, D. Noll, J. B. Thevenet, and H. D. Tuan, *A spectral quadratic-SDP method with applications to fixed-order H_2 and H_∞ synthesis*, Eur. J. Control, 10 (2004), pp. 527–538.
[6] P. Apkarian, D. Noll, and H. D. Tuan, *Fixed-order H_∞ control design via an augmented Lagrangian method*, Int. J. Robust Nonlinear Control, 13 (2003), pp. 1137–1148.
[7] C. Audet and J. E. Dennis Jr., *Mesh adaptive direct search algorithms for constrained optimization*, SIAM J. Optim., to appear.
[8] V. Blondel and M. Gevers, *Simultaneous stabilizability question of three linear systems is rationally undecidable*, Math. Control Signals Syst., 6 (1994), pp. 135–145.
[9] J. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer Series in Operations Research, Springer-Verlag, New York, 2000.
[10] A. Booker, J. E. Dennis Jr., P. Frank, D. Serafini, V. Torczon, and M. Trosset, *A Rigorous Framework for Optimization of Expertise Functions by Surrogates*, Technical report, 1998, pp. 1–24.
[11] G. E. P. Box, *Evolutionary operation: A method for increasing industrial productivity*, Appl. Stat., 6 (1957), pp. 81–101.
[12] S. Boyd, V. Balakrishnan and P. Kabamba, *A bisection method for computing the H_∞ norm of a transfer matrix and related problems*, Math. Control Signals Syst., 2 (1989), pp. 207–219.
[13] S. Boyd and C. Barratt, *Linear Controller Design: Limits of Performance*, Prentice-Hall, Englewood Cliffs, NJ, 1991.
[14] S. Boyd, L. ElGhaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in Systems and Control Theory*, SIAM Stud. Appl. Math. 15, SIAM, Philadelphia, 1994.
[15] M. Bredemann, C. T. Abdallah, and P. Dorato, *Polynomial solutions for simultaneous stabilization*, in Proceedings of the 2nd IFAC Symposium on Robust Control Design, Budapest, Hungary, 1997, pp. 193–198.
[16] M. V. Bredemann, *Feedback Controller Design for Simultaneous Stabilization*, Ph.D. thesis, University of New Mexico, 1995.
[17] J. Burke, A. Lewis, and M. Overton, *Two numerical methods for optimizing matrix stability*, Linear Algebra Appl., 351–352 (2002), pp. 147–184.
[18] J. Burke, A. Lewis, and M. Overton, *A robust gradient sampling algorithm for nonsmooth, nonconvex optimization*, SIAM J. Optim., 15 (2003), pp. 751–779.
[19] J. Burke, A. Lewis, and M. Overton, *Robust stability and a criss-cross algorithm for pseudospectra*, IMA J. Numer. Anal., 23 (2003), pp. 1–17.
[20] Y. Y. Cao and Y. X. Sun, *Static output feedback simultaneous stabilization: ILMI approach*, Int. J. Control, 70 (1998), pp. 803–814.
[21] B. M. Chen, *H_∞ Control and Its Applications*, Lecture Notes in Control and Inform. Sci. 235, Springer-Verlag, New York, Heidelberg, Berlin, 1998.
[22] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Canadian Math. Soc. Series, John Wiley & Sons, New York, 1983.
[23] J. Cullum, W. Donath, and P. Wolfe, *The minimization of certain nondifferentiable sums of eigenvalues of symmetric matrices*, Math. Program. Stud., 3 (1975), pp. 35–55.
[24] E. E. J. Davison, *Benchmark Problems for Control System Design*, Technical report, IFAC Technical Committee Reports, Pergamon Press, Oxford, 1990.

[25] B. Fares, D. Noll, and P. Apkarian, *Robust control via sequential semidefinite programming*, SIAM J. Control Optim., 40 (2002), pp. 1791–1820.
[26] P. Gahinet and P. Apkarian, *Numerical computation of the L_∞ norm revisited*, in Proceedings of the IEEE Conference on Decision and Control, 1992.
[27] P. Gahinet and P. Apkarian, *A linear matrix inequality approach to H_∞ control*, Int. J. Robust Nonlinear Control, 4 (1994), pp. 421–448.
[28] P. Gahinet, A. Nemirovski, A. J. Laub, and M. Chilali, *LMI Control Toolbox*, The MathWorks Inc., Natick, MA, 1995.
[29] D. Gangsaas, K. Bruce, J. Blight, and U.-L. Ly, *Application of modern synthesis to aircraft control: Three case studies*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 995–1014.
[30] J. C. Geromel, P. L. D. Peres, and J. Bernussou, *On a convex parameter space method for linear control design of uncertain systems*, SIAM J. Control Optim., 29 (1991), pp. 381–402.
[31] A. Hassibi, J. How, and S. Boyd, *A pathfollowing method for solving BMI problems in control*, in Proceedings of the American Control Conference, 1999, pp. 1385–1389.
[32] C. Helmberg and F. Rendl, *Spectral bundle method for semidefinite programming*, SIAM J. Optim., 10 (2000), pp. 673–696.
[33] D. Henrion and M. Sebek, *LMIs and polynomial methods in control: Illustrative examples*, in IFAC/IEEE Symposium on Advances in Control Education, Australia, 2000.
[34] D. Henrion, S. Tarbouriech, and M. Sebek, *Rank-one LMI approach to simultaneous stabilization of linear systems*, Syst. Control Lett., 11 (1998), pp. 167–172.
[35] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* II: *Advanced Theory and Bundle Methods*, Grundlehren Math. Wiss. 306, Springer-Verlag, New York, 1993.
[36] C. Hol, C. Scherer, E. van der Meché, and O. Bosgra, *A nonlinear SDP approach to fixed-order controller synthesis and comparison with two other methods applied to an active suspension system*, Eur. J. Control, 9 (2003), pp. 13–28.
[37] R. Hook and T. A. Jeeves, *"Direct search" solution of numerical and statistical problems*, J. Assoc. Comput. Mach., 8 (1961), pp. 212–229.
[38] H. P. Horisberger and P. R. Belanger, *Solution of the optimal constant output feedback problem by conjugate gradients*, IEEE Trans. Automat. Control, 19 (1974), pp. 434–435.
[39] Y. S. Hung and A. G. J. MacFarlane, *Multivariable Feedback: A Classical Approach*, Lecture Notes in Control and Inform. Sci. 40, Springer-Verlag, New York, Heidelberg, Berlin, 1982.
[40] J. Imae and T. Furudate, *A design method for fixed-order H_∞ controllers via bilinear matrix inequalities*, in Proceedings of the American Control Conference, San Diego, CA, 1999, pp. 1876–1880.
[41] F. Jarre, *A QQP-Minimization Algorithm Method for Semidefinite and Smooth Nonconvex Programs*, Technical report.
[42] L. H. Keel, S. P. Bhattacharyya, and J. W. Howze, *Robust control with structured perturbations*, IEEE Trans. Automat. Control, 36 (1988), pp. 68–77.
[43] K. C. Kiwiel, *Methods of Descent for Nondifferentiable Optimization*, Lecture Notes in Math. 1133, A. Dold and B. Eckmann, eds., Springer-Verlag, Berlin, Heidelberg, New York, 1985.
[44] K. C. Kiwiel, *A linearization algorithm for optimizing control systems subject to singular value inequalities*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 595–602.
[45] M. Kocvara and M. Stingl, *A code for convex nonlinear and semidefinite programming*, Optim Methods Software, 18 (2003), pp. 317–333.
[46] T. G. Kolda, R. M. Lewis, and V. Torczon, *Optimization by direct search: New perspectives on some classical and modern methods*, SIAM Rev., 45 (2003), pp. 385–482.
[47] F. Leibfritz, *COMPL$_e$IB, Constraint Matrix-Optimization Problem Library: A Collection of Test Examples for Nonlinear Semidefinite Programs, Control System Design and Related Problems*, Technical report, Universität Trier, 2003.
[48] F. Leibfritz and E. M. E. Mostafa, *Trust region methods for solving the optimal output feedback design problem*, Int. J. Control, 76 (2000), pp. 501–519.
[49] F. Leibfritz and E. M. E. Mostafa, *An interior point constrained trust region method for a special class of nonlinear semidefinite programming problems*, SIAM J. Optim., 12 (2002), pp. 1048–1074.
[50] C. Lemarechal, *Nondifferentiable optimization*, in Optimization, G. L. Nemhauser and A. H. G. Rinnooy Kan, eds., Elsevier North-Holland, New York, 1989, pp. 529–572.
[51] M. Marazzi and J. Nocedal, *Wedge trust region methods for derivative free optimization*, Math. Program. Ser. A, 91 (2002), pp. 289–305.
[52] K. I. M. McKinnon, *Convergence of the Nelder–Mead simplex method to a nonstationary*

point, SIAM J. Optim., 9 (1998), pp. 148–158.
[53] K. Miettinen and M. M. Mäkelä, *An interactive method for nonsmooth multiobjective with an application to optimal control*, Optim. Methods Software, 2 (1993), pp. 31–44.
[54] L. Mosheyev and M. Zibulevsky, *Penalty/barrier multiplier algorithm for semidefinite programming*, Optim. Methods Software, 13 (2000), pp. 235–261.
[55] J. A. Nelder and R. Mead, *A simplex method for function minimization*, Comput. J., (1965), pp. 441–461.
[56] A. Nemirovskii, *Several NP-hard problems arising in robust stability analysis*, Math. Control Signals Syst., 6 (1993), pp. 99–105.
[57] D. Noll and P. Apkarian, *Spectral bundle methods for non-convex maximum eigenvalue functions: First-order methods*, Math. Program. Ser. B, to appear.
[58] D. Noll and P. Apkarian, *Spectral bundle methods for non-convex maximum eigenvalue functions: Second-order methods*, Math. Program. Ser. B, to appear.
[59] D. Noll, M. Torki, and P. Apkarian, *Partially augmented Lagrangian method for matrix inequality constraints*, SIAM J. Optim., 15 (2004), pp. 161–184.
[60] F. Oustry, *A second-order bundle method to minimize the maximum eigenvalue function*, Math. Program. Ser. A, 89 (2000), pp. 1–33.
[61] E. Polak and S. Salcudean, *On the design of linear multivariable feedback systems via constrained nondifferentiable optimization in H_∞ spaces*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 268–276.
[62] E. Polak and Y. Wardi, *A nondifferentiable optimization algorithm for the design of control systems subject to singular value inequalities over a frequency range*, Automatica, 18 (1982), pp. 267–283.
[63] W. Spendley, G. R. Hext, and F. R. Himsworth, *Sequential application of simplex designs in optimisation and evolutionary operation*, Technometrics, 4 (1962), pp. 441–461.
[64] J. Thevenet, P. Apkarian, and D. Noll, *A bundle-type method for "large scale" output feedback control design*, SIAM Conference on Optimization, Stockholm, Sweden, 2005.
[65] J. Thevenet, D. Noll, and P. Apkarian, *Nonlinear spectral SDP method for BMI-constrained problems: Applications to control design*, in ICINCO Proceedings, Setúbal, Portugal, 2004, pp. 237–248.
[66] V. Torczon, *On the convergence of the multidirectional search algorithm*, SIAM J. Optim., 1 (1991), pp. 123–145.
[67] V. Torczon, *On the convergence of pattern search algorithms*, SIAM J. Optim., 7 (1997), pp. 1–25.
[68] L. Trefethen, *Pseudospectra of linear operators*, SIAM Rev., 39 (1997), pp. 383–406.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 6, pp. 1950–1972

ON THE CONTROLLABILITY OF A FRACTIONAL ORDER PARABOLIC EQUATION*

SORIN MICU† AND ENRIQUE ZUAZUA‡

Abstract. The null-controllability property of a $1-d$ parabolic equation involving a fractional power of the Laplace operator, $(-\Delta)^\alpha$, is studied. The control is a scalar time-dependent function $g=g(t)$ acting on the system through a given space-profile $f=f(x)$ on the interior of the domain. Thus, the control g determines the intensity of the space control f applied to the system, the latter being given a priori. We show that, if $\alpha \leq 1/2$ and the shape function f is, say, in L^2, no initial datum belonging to any Sobolev space of negative order may be driven to zero in any time. This is in contrast with the existing positive results for the case $\alpha > 1/2$ and, in particular, for the heat equation that corresponds to $\alpha = 1$. This negative result exhibits a new phenomenon that does not arise either for finite-dimensional systems or in the context of the heat equation.

On the contrary, if more regularity of the shape function f is assumed, then we show that there are initial data in any Sobolev space H^m that may be controlled. Once again this is precisely the opposite behavior with respect to the control properties of the heat equation in which, when increasing the regularity of the control profile, the space of controllable data decreases.

These results show that, in order for the control properties of the heat equation to be true, the dynamical system under consideration has to have a sufficiently strong smoothing effect that is critical when $\alpha = 1/2$ for the fractional powers of the Dirichlet Laplacian in $1-d$. The results we present here are, in nature and with respect to techniques of proof, similar to those on the control of the heat equation in unbounded domains in [S. Micu and E. Zuazua, *Trans. Amer. Math. Soc.*, 353 (2000), pp. 1635–1659] and [S. Micu and E. Zuazua, *Portugal. Math.*, 58 (2001), pp. 1–24].

We also discuss the hyperbolic counterpart of this problem considering a fractional order wave equation and some other models.

Key words. null controllability, parabolic equation, fractional power of the Laplace operator

AMS subject classifications. 35B37, 93B05, 93C20

DOI. 10.1137/S036301290444263X

1. Introduction.

It is generally considered that, due to the strong dissipative effect on the high modes, parabolic equations behave like ordinary differential equations (i.e., finite-dimensional dynamical systems) from a control theoretical point of view. This is true for instance for the heat equation concerning the problem of null controllability, i.e., that of driving solutions from a given initial configuration to equilibrium, in several respects: (a) both finite-dimensional systems and the heat equation are controllable in an arbitrarily short time; (b) the controls may be taken to be arbitrarily smooth. In this way, for instance, the heat equation in bounded domains is controllable with L^2-controls for initial data in a Sobolev space of arbitrary negative order, in an arbitrarily short time and with controls supported in an arbitrarily small

*Received by the editors March 26, 2004; accepted for publication (in revised form) April 26, 2005; published electronically January 6, 2006. This work was partially supported by grant BFM 2002-03345 of MCYT (Spain).

http://www.siam.org/journals/sicon/44-6/44263.html

†Facultatea de Matematica, Universitatea din Craiova, Al. I. Cuza, 13, Craiova, 1100, Romania (sd_micu@yahoo.com). The research of this author was partially supported by grant 17 of the Egide-Brancusi Program.

‡Departamento de Matemáticas, Facultad de Ciencias, Universidad Autónoma, Cantoblanco, 28049, Madrid, Spain (enrique.zuazua@uam.es). The research of this author was partially supported by the TMR networks of the EU "Homogenization and Multiple Scales" (HMS2000) and "New Materials, Adaptive Systems and Their Nonlinearities: Modelling, Control and Numerical Simulation" (HPRN-CT-2002-00284).

subdomain. Recently it was proved, however, that this is not true in unbounded domains (see [17] and [18]).

The object of this article is to further investigate to what extent this analogy is systematically true or whether it is related to the intrinsic properties of the heat equation.

To do this we consider the following *null-controllability problem: Given* $T > 0$ *and* $f \in L^2(0,\pi)$*, for any* $u^0 \in L^2(0,\pi)$ *find a control* $g \in L^2(0,T)$ *such that the solution* u *of the problem*

$$\begin{cases} u_t + (-\Delta)^\alpha u = g(t)f(x), & x \in (0,\pi),\ t \in (0,T), \\ u = 0, & x \in \{0,\pi\},\ t \in (0,T), \\ u(0,x) = u^0(x), & x \in (0,\pi), \end{cases} \tag{1.1}$$

satisfies

$$u(T,\,\cdot\,) = 0. \tag{1.2}$$

Here and in what follows $(-\Delta)^\alpha$ denotes the fractional power of order $\alpha > 0$ of the Dirichlet Laplacian that we shall denote by A_α. More precisely,

$$\begin{gathered} A_\alpha : D(A_\alpha) \subset L^2(\Omega) \to L^2(\Omega), \\ D(A_\alpha) = \left\{ u \in L^2(0,\pi) \,:\, u = \sum_{n\geq 1} a_n \sin(nx) \text{ and } \sum_{n\geq 1} |a_n|^2 n^{4\alpha} < \infty \right\}, \\ u(x) = \sum_{n\geq 1} a_n \sin(nx) \;\longrightarrow\; A_\alpha u(x) = \sum_{n\geq 1} a_n n^{2\alpha} \sin(nx). \end{gathered} \tag{1.3}$$

Equation (1.1) is of parabolic type for any $\alpha > 0$. In the absence of control, solutions of (1.1) decay exponentially as $t \to \infty$ in, say, L^2. When $\alpha = 1$ we recover the classical heat equation.

When $0 < \alpha < 1$, (1.1) is a model example of parabolic dynamical system with weaker diffusivity (subdiffusion). Fractional equations of diffusion type are useful models for the description of transport processes in complex systems, slower than the Brownian diffusion. The list of systems displaying such anomalous dynamic behavior is quite extensive: charge carrier transport in amorphous semiconductors, nuclear magnetic resonance diffusometry in percolative and porous media, transport on fractal geometries, diffusion of a scalar tracer in an array of convection rolls, dynamics of a bead in a polymeric network, transport in viscoelastic materials, etc. (see [16] and [12]).

The state of system (1.1) is u and the control, which acts on its right-hand side term as an external source, is given by $g(t)f(x)$, where the shape function $f = f(x)$ is given and the intensity $g = g(t)$ is at our disposal. Such types of controls are sometimes called "lumped" or "bilinear" (see, for instance, [1] and [11]).

The null-controllability problem (1.1)–(1.2) has been considered and solved in [7] for the case $\alpha > 1/2$. The proof in [7] is based on the fact that the null-control problem may be rewritten as a *problem of moments* of the following form: *Find* $g \in L^2(0,T)$ *such that*

$$\int_0^T g(t) e^{\lambda_n t} dt = \beta_n \quad \forall n \geq 1, \tag{1.4}$$

where $\beta_n = -\pi a_n/2f_n$ *depend on the Fourier coefficients* $(a_n)_{n\geq 1}$ *of the initial data to be controlled and those of the control profile* $(f_n)_{n\geq 1}$.

Here λ_n is the sequence of the (real) eigenvalues of the equation under consideration: $\lambda_n = n^{2\alpha}$.

It is by now well known—and this is the second ingredient in the proof of [7]—that if

$$\lambda_n \sim cn^{\gamma} \quad \text{as} \quad n \to \infty \tag{1.5}$$

for some $\gamma > 1$ and a positive constant $c > 0$, then (1.4) has L^2-solutions if the values β_n do not increase too much.

This result may be proved by means of a careful evaluation of the norm of a biorthogonal sequence to the family of exponentials $\{e^{\lambda_n t}\}_{n\geq 1}$ and it is related to the Müntz theorem (see [20]), guaranteeing that the family of exponentials $\{e^{\lambda_n t}\}_{n\geq 1}$ is linearly independent in $L^2(0,T)$ if and only if

$$\sum_{n=1}^{\infty} \frac{1}{|\lambda_n|} < \infty. \tag{1.6}$$

In the context of system (1.3), condition (1.5) and, implicitly, (1.6) are verified if and only if $\alpha > 1/2$.

According to this analysis, it was proved in [7] that, when $\alpha > 1/2$, and when the control profile f satisfies the condition

$$\lim_{n\to\infty} \left(\left| \int_0^{\pi} f(x) \sin(nx) dx \right| e^{\eta\lambda_n} \right) > 0 \quad \forall \eta > 0, \tag{1.7}$$

system (1.1) is null controllable in the sense above for an arbitrarily short time and with smooth time-dependent controls g.

It is important to note that, according to condition (1.7), the shape function f is not "too regular." In particular, its Fourier coefficients may not decay faster than a suitable exponential function. Obviously, one can find control profiles f with such a property in any Sobolev space $H^s(0,\pi)$ and, in particular, in $L^2(0,\pi)$.

The present paper deals with the case $\alpha \leq 1/2$. As we shall see, the behavior of the system from the control theoretical point of view is, surprisingly, the opposite one.

Concerning the growth condition (1.5) on the spectrum, the case $\alpha = 1/2$ is critical and the condition, clearly, does not hold when $0 < \alpha < 1/2$. The same can be said about the summability condition (1.6). In this sense, the situation we are dealing with is similar to that in [17] and [18], where the heat equation in the half-line and half-space was considered. Indeed, in [17] it was proved that when $\lambda_n = n$, the corresponding moment problem (1.4) has a solution only if the β_n grows very fast as n tends to infinity.[1] Since β_n is, essentially, the ratio between the Fourier coefficients of the initial data to be controlled and those of the control profile f, we concluded that no regular nontrivial initial data allow a L^2-solution of the moment problem, when the profile is not too smooth. Accordingly L^2-controls may not exist either. The same can be said about the control problem (1.1) in the whole range $0 < \alpha \leq 1/2$.

[1] Recently, the results of [17] and [18], and more precisely its consequences in the context of unique continuation, were generalized in [5] to the case of parabolic equations with a potential, by means of Carleman inequalities.

This negative result shows that the parabolic nature of the equation and the infinite velocity of propagation do not suffice to guarantee the controllability of the system. On the contrary, we see that, in order for the control properties of the heat equation to be true, very much like in the finite-dimensional theory, the underlying semigroup is required to have a very strong dissipative effect that fails when $\alpha \leq 1/2$.

To be more precise, we shall show that

- if the shape function f satisfies (1.7), no initial data in any negative Sobolev space may be controlled to zero;
- if this function is more regular, for instance, if it satisfies

$$\left| \int_0^\pi f(x) \sin(nx) dx \right| \leq e^{-\eta \lambda_n} \tag{1.8}$$

for some $\eta > T$, then there are initial data in any Sobolev space $H^m(0,\pi)$ that are null controllable in time T with L^2-controls.

As we said above, and contrary to intuition, this behavior is in opposition to the control properties of the heat equation corresponding to $\alpha = 1$.

Let us mention that the property

$$\sum_{n=1}^{\infty} \frac{1}{|\lambda_n|} > \infty \tag{1.9}$$

of the eigenvalues of the differential operator leads to a result of no spectral controllability in the case of the heat equation in multidimensional problems (see [1, Theorem IV.1.3, p. 178]). On the other hand, under hypothesis (1.9), Fattorini [6] shows that for any $T > 0$ there exist a shape function $f \in L^2(0,\pi)$ and an initial datum $u^0 \in L^2(0,\pi)$ which cannot be driven to zero in time T by means of a control of type $f(x)g(t)$. The proofs of these results are based on the fact that an entire function which vanishes at every λ_n is identically zero and are related to the methods we use in our article. However, note that, given a shape function f, we describe the space of the initial data which cannot be controlled to zero in finite time.

It is also interesting to compare the results we obtain in this paper with those that one could expect from the application of the methodology in the articles by Lebeau and Robbiano [13] and Lebeau and Zuazua [14]. In [13] and [14] an iterative method was developed to prove the null controllability of the heat equation when the control acts in an open subset of the domain where the equation holds. The same method can be used to deal with control mechanisms as in (1.1). Their main idea was to split the time interval into a sequence of decreasing consecutive subintervals. In each of these intervals an increasing finite number of Fourier components (determined by a diadic decomposition) is controlled to zero, the control being applied in two steps. In a first step (in half of the subinterval) where a nontrivial control is applied, an estimate based on Carleman inequalities guarantees that the size of the control does not grow faster than an exponential factor, in which the maximal eigenfrequency of the eigenfunctions under consideration enters. It was then shown that the dissipative property of the heat equation in the remaining half of the subinterval was able to compensate this exponential growth. A careful analysis of the method of proof in [13] and [14] shows that it works if $\alpha > 1/2$. The results of the present paper show that the results this method yields are sharp in the sense that completely opposite results hold when $\alpha \leq 1/2$. This fact confirms once more that the control theoretical results of the heat equation do hold because of its very strong dissipative properties.

The paper is organized as follows. In section 2 we present the controllability problem and some equivalent formulations. Some known results are also mentioned. In section 3 our main controllability results for the case $\alpha \leq 1/2$ are stated. They are based on two propositions concerning entire functions that are proven in section 4. In section 5 we give a negative result concerning the dual observability inequality (with respect to the control problem). In section 6 we analyze the controllability properties of a hyperbolic equation involving the same operator $(-\Delta)^\alpha$:

$$u_{tt} + (-\Delta)^\alpha u = g(t)f(x).$$

In this case the situation is even worse since the classical control properties of the $1-d$ wave equation (that correspond to the exponent $\alpha = 1$) fail for all $\alpha < 1$. Some comments and open problems are included in section 7.

2. Problem formulation and existing results. We first observe that the operator A_α in (1.3) is well defined since $(\sqrt{2}\sin(nx)/\sqrt{\pi})_{n\geq 1}$ forms an orthonormal basis in $L^2(0,\pi)$. Moreover, the operator A_α is densely defined and is self-adjoint in $L^2(0,\pi)$.

The eigenvalues of the operator A_α are given by

$$\lambda_n = n^{2\alpha} \quad \forall n \geq 1 \tag{2.1}$$

with eigenfunctions

$$\varphi_n = \sin(nx) \quad \forall n \geq 1. \tag{2.2}$$

With this notation the control problem for system (1.1) can be formulated as follows: Given $T > 0$, $f \in L^2(0,\pi)$ and $u^0 \in L^2(0,\pi)$ find $g \in L^2(0,T)$ such that the solution u of problem

$$\begin{cases} u_t + A_\alpha u = g(t)f(x), & x \in (0,\pi),\ t \in (0,T), \\ u = 0, & x \in \{0,\pi\},\ t \in (0,T), \\ u(0,x) = u^0(x), & x \in (0,\pi), \end{cases} \tag{2.3}$$

satisfies

$$u(T,\,\cdot\,) = 0. \tag{2.4}$$

An initial datum u^0 with such property is said to be *null controllable* in time T. If all initial data in $L^2(0,\pi)$ are null controllable we say that (2.3) is null controllable in $L^2(0,\pi)$.

The goal is to drive the initial datum u^0 to rest by using a control with a given shape $f(x)$ in space at each time. Then the control $g(t)$ determines the intensity of the control profile applied to the system.

Let us first give the following variational characterization of controllable initial data.

LEMMA 2.1. *The initial datum $u^0 \in L^2(0,\pi)$ is null controllable in time T with control $g \in L^2(0,T)$ if and only if the identity*

$$-\int_0^\pi u^0(x)\varphi(0,x)dx = \int_0^T g(t)\left(\int_0^\pi f(x)\varphi(t,x)dx\right)dt \tag{2.5}$$

holds for any $\varphi^T \in L^2(0,\pi)$ *with* $\varphi(t,x)$ *solution of the adjoint equation*

$$\begin{cases} -\varphi_t + A_\alpha \varphi = 0, & x \in (0,\pi),\ t \in (0,T), \\ \varphi = 0, & x \in \{0,\pi\},\ t \in (0,T), \\ \varphi(T,x) = \varphi^T(x), & x \in (0,\pi). \end{cases} \tag{2.6}$$

Proof. The proof follows immediately by multiplying (2.3) by φ, the solution of (2.6), integrating in $(0,T)\times\Omega$, and taking into account that A_α is self-adjoint. ∎

Since $(\sin(nx))_{n\geq 1}$ is complete in $L^2(0,\pi)$, considering $\varphi^T(x) = \sin(nx)$ for each $n \geq 1$ in Lemma 2.1, the following equivalent condition for the null-controllability results.

LEMMA 2.2. *An initial datum* $u^0 \in L^2(0,\pi)$ *of the form*

$$u^0(x) = \sum_{n\geq 1} a_n \sin(nx) \tag{2.7}$$

is null controllable in time T *if and only if there exists* $g \in L^2(0,T)$ *such that, for any* $n \geq 1$,

$$f_n \int_0^T g(t) e^{\lambda_n t} dt = -\frac{\pi}{2} a_n, \tag{2.8}$$

where

$$f_n = \int_0^\pi f(x) \sin(nx) dx. \tag{2.9}$$

Note that (2.8) is a *moment problem.*

Note also that, given an arbitrary initial datum u^0, a necessary condition for this moment problem to have a solution is that

$$f_n = \int_0^\pi f(x) \sin(nx) dx \neq 0 \quad \forall n \geq 1. \tag{2.10}$$

Indeed, if there exists $k \geq 1$ such that $f_k = 0$, the kth equation in (2.8) does not hold except for the case $a_k = 0$. In fact, if $f_k = 0$, it is easy to see that the kth Fourier component of the solution of the controlled problem (1.1) is invariant in time. This makes the controllability property impossible unless $a_k = 0$.

From now on we shall suppose that f verifies (2.10).

Let us now recall the following result from [7].

THEOREM 2.1. *Let* $\alpha > 1/2$ *and suppose that the Fourier coefficients of* f *satisfy* (2.10) *and the following additional condition:*

$$\liminf_{n\to\infty} |f_n| e^{\eta n^{2\alpha}} > 0 \tag{2.11}$$

for any $\eta > 0$.

Then, the initial state $u^0 = \sum_{n\geq 1} a_n \sin(nx)$ *is null controllable in time* $T > 0$ *by means of a control* $g \in L^2(0,T)$ *if, for some* $M, \eta > 0$,

$$|a_n| \leq M e^{n^{2\alpha} T} e^{-(\pi+\eta)n}, \quad n = 1, 2, \ldots. \tag{2.12}$$

Moreover, when this holds, the control g may be chosen to be in $C^m([0,T])$ for all $m \geq 1$.

REMARK 2.1. *The right-hand side term in* (2.12) *tends to infinity as $n \to \infty$. Thus, Theorem* 2.1 *implies, for instance, that any initial data in $L^2(0,\pi)$ are null controllable in any time $T > 0$. This result is in contrast with which we shall prove for the case $\alpha \leq 1/2$ that no initial data in a negative Sobolev space may be driven to zero in finite time with an L^2-control g.*

REMARK 2.2. *Condition* (2.11) *requires the shape function $f = f(x)$ to be not "too regular." Obviously, one can find control profiles f with such a property in any Sobolev space $H^s(0,\pi)$, but a too fast exponential decay rate of the Fourier coefficients of f is incompatible with* (2.11). *In particular, when f is a Gaussian function,* (2.11) *fails for $\alpha = 1$, i.e., for the classical heat equation.*

3. Controllability results in the case $\alpha \leq 1/2$. Let us now address the case $0 < \alpha \leq 1/2$. Throughout this section we will assume that $0 < \alpha \leq 1/2$. However, some of the results we present here are valid for all $\alpha > 0$. This will be indicated explicitly when it is the case.

3.1. The main negative result. The following result is completely different from that obtained in Theorem 2.1.

THEOREM 3.1. *Let $0 < \alpha \leq 1/2$ and suppose that the Fourier coefficients of f satisfy* (2.11). *Then any nontrivial initial state $u^0 = \sum_{n\geq 1} a_n \sin(nx)$ with the property that for any $\mu > 0$ there exists a constant $C_\mu > 0$ such that*

$$|a_n| \leq C_\mu e^{\mu n^{2\alpha}} \quad \forall n \geq 1 \tag{3.1}$$

cannot be driven to zero in time $T > 0$ by means of a control $g \in L^2(0,T)$, whatever $T > 0$ is.

REMARK 3.1. *The right-hand side term in* (3.1) *grows exponentially as $n \to \infty$. Thus, Theorem* 3.1 *implies that there is no initial datum in any Sobolev space of negative order that might be null controllable in any time $T > 0$ with controls g in $L^2(0,T)$.*

Consequently, this result is in opposition to the positive one in Theorem 2.1 *for the case $\alpha > 1/2$.*

In particular, Theorem 3.1 *means that choosing quite irregular control profiles f, as one is required to do when $\alpha > 1/2$ according to* (2.11), *is a very bad choice when $\alpha \leq 1/2$.*

REMARK 3.2. *From* (3.1) *it seems that, as α increases, the class of data for which the null-controllability property fails increases as well. However, a careful analysis of the proof of the theorem and Proposition* 3.2 *shows the contrary. Indeed, for the null-controllability property to fail, not all, but only part, of the Fourier coefficients of the initial datum must satisfy* (3.1). *Indeed, instead of* (3.1) *it is sufficient to have*

$$|a_{n_k}| \leq C_\mu e^{\mu n_k^{2\alpha}} \quad \forall k \geq 1 \tag{3.2}$$

for a suitable subsequence $(n_k)_{k\geq 1}$ (see Lemma 4.2 *and Remark* 4.2) *satisfying*

$$|n_{k+1} - n_k| \geq \frac{1}{2\alpha} k^{\frac{1}{2\alpha}-1} - 2 \quad \forall k \geq 1. \tag{3.3}$$

Note that (3.3) *shows that the distance between two consecutive terms of the sequence $(n_k)_{k\geq 1}$ decreases when α increases. Hence, the same happens to the class of*

data for which the null-controllability property fails. This agrees with the first intuition that suggests that, as the dissipativity of the system increases, its controllability properties improve.

A dramatic change in the controllability properties arises when $\alpha = 1/2$. *For* $\alpha > 1/2$ *the control problem is very well behaved (see Theorem* 2.1). *On the contrary, the controllability properties are very poor when* $\alpha \leq 1/2$. *Note that the same occurs with the spectral property* (1.6). *Something similar happens in* (3.3) *where, when* $\alpha > 1/2$, *the gap condition is fulfilled for all indices* k *without extracting subsequences.*

3.2. Proof of the negative result. According to Lemma 2.2, the property of null controllability of $u^0 = \sum_{n\geq 1} a_n \sin(nx)$ is equivalent to the existence of a function $g \in L^2(0,T)$ such that, for any $n \geq 1$, (2.8) is verified.

Before getting into the proof of Theorem 3.1 let us first give an equivalent condition for the existence of such a control function g.

PROPOSITION 3.1. *The following assertions are equivalent:*

(a) *There exists* $g \in L^2(0,T)$ *such that the following holds:*

$$\int_0^T g(s) e^{n^{2\alpha} s} ds = \alpha_n \quad \forall n \geq 1. \tag{3.4}$$

(b) *There exists an entire function* F *of exponential type* $\leq T/2$, *with*

$$\int_{-\infty}^{\infty} |F(iy)|^2 \, dy < \infty \tag{3.5}$$

and such that

$$F(n^{2\alpha}) = \alpha_n e^{-n^{2\alpha} T/2} \quad \forall n \geq 1. \tag{3.6}$$

Recall that an entire function is said to be of exponential type $\leq B$ if there exists a positive constant $A > 0$ such that (see [21])

$$|F(z)| \leq A e^{B|z|} \quad \forall z \in \mathbb{C}. \tag{3.7}$$

REMARK 3.3. *Several remarks are in order:*

- *Proposition* 3.1 *is a very general result in which the explicit values of the coefficients* α_n *and the eigenvalues* $\lambda_n = n^{2\alpha}$ *do not matter.*
- *The proof of Proposition* 3.1 *uses the Fourier transform and the Paley–Wiener theorem and will be given in the next section.*
- *As the proof of this proposition shows, the function* F *in* (b) *is uniformly bounded along the imaginary axis.*
- *From Proposition* 3.1, *in order to characterize the null-controllable initial data it is necessary and sufficient to characterize the sequences* $\{F(n^{2\alpha})\}_{n\geq 1}$ *that may be obtained by means of entire functions* F *of exponential type* $\leq T/2$ *satisfying* (3.5).

The following proposition provides significant information on the rate of growth of $F(n^{2\alpha})$ for functions F as above.

PROPOSITION 3.2. *Let* $F : \mathbb{C} \to \mathbb{C}$ *be a function satisfying the following properties:*

(i) F *is an entire function of exponential type* $\leq T/2$;
(ii) $\int_{-\infty}^{\infty} |F(iy)|^2 \, dy < \infty$;

(iii) *for any* $\delta > 0$ *there exists* $C_\delta > 0$ *such that*

$$|F(n^{2\alpha})| \leq C_\delta e^{\delta n^{2\alpha}} e^{-n^{2\alpha}T/2} \quad \forall n \geq 1.$$

Then, necessarily, $F \equiv 0$.

The proof of Proposition 3.2 is based on a result of Duffin and Schaeffer (see [4] and also [3, p. 191]) which gives conditions for the boundedness of an analytic function in a sector of the complex plane if its boundedness on a sequence of complex numbers is assumed. In our case, the information we have on the behavior of $F(n^{2\alpha})$ allows us to construct an analytic function in the right half-plane which is bounded on a sequence of complex numbers close to n and to apply the mentioned result. The complete proof of Proposition 3.2 will be given in the next section.

Let us now show how Theorem 3.1 follows from Propositions 3.1 and 3.2.

If u^0 is null controllable, then the existence of a function F as in Proposition 3.1 is ensured with $\lambda_n = n^{2\alpha}$ and $\alpha_n = -a_n/(2f_n)$. Hence, F satisfies conditions (i)–(ii) in Proposition 3.2. Then, condition (3.1) on the Fourier coefficients of u^0 and condition (2.11) on the shape function f imply that

$$|F(n^{2\alpha})| \leq Ce^{(\mu+\eta)n^{2\alpha}} e^{-n^{2\alpha}T/2}. \tag{3.8}$$

Since μ and η are arbitrary, the function F also satisfies property (iii) from Proposition 3.2.

It follows that $F \equiv 0$ and, consequently, under the growth condition (3.1) and with control profiles satisfying (2.11), the only controllable initial datum is the trivial one. □

3.3. Other controllability properties. As we have said before, condition (2.11) indicates that the shape function f is not "too regular." Let us now show that assuming more regularity on f may increase the space of controllable data. This fact is also in opposition to the behavior of the system in the case $\alpha > 1/2$, in which increasing the regularity of the profile f reduces the space of controllable data.

PROPOSITION 3.3. *Let* $\alpha \leq 1/2$ *and suppose that there exists* $\eta > T$ *such that*

$$|f_n| \leq e^{-\eta n^{2\alpha}} \quad \forall n \geq 1. \tag{3.9}$$

Then there are initial data in any Sobolev space $H^m(0,\pi)$ *which are null controllable by means of a control function* $g \in L^2(0,T)$.

REMARK 3.4. *It is important to note that the result in Proposition* 3.3 *holds for* $\alpha > 1/2$ *as well. However, in this case, as mentioned above, one can prove much better results guaranteeing that all initial data in* $L^2(0,\pi)$ *are controllable even if condition* (3.9) *is not satisfied.*

Proof. From Lemma 2.2 and Proposition 3.1, it follows that an initial datum whose Fourier coefficients are given by

$$a_n = -2f_n F(n^{2\alpha}) e^{\frac{T}{2}n^{2\alpha}},$$

where

$$F(z) = \frac{\sin\left(\frac{T}{2}zi\right)}{\frac{T}{2}zi},$$

is null controllable in time T.

The initial datum with these Fourier coefficients belongs to any Sobolev space $H^m(0,\pi)$ with $m \geq 0$. Indeed,

$$\sum_{n\geq 1} |a_n|^2 n^{2m} \leq 4 \sum_{n\geq 1} |f_n|^2 |F(n^{2\alpha})|^2 n^{2m} e^{Tn^{2\alpha}}.$$

We now use in an essential way that F is of exponential type $\leq T/2$. This is obvious in this case in view of the explicit form of F. It follows that

$$\sum_{n\geq 1} |a_n|^2 n^{2m} \leq 4 \sum_{n\geq 1} e^{-2\eta n^{2\alpha}} \frac{e^{Tn^{2\alpha}}}{\left(\frac{Tn^{2\alpha}}{2}\right)^2} e^{Tn^{2\alpha}} n^{2m} < \infty. \qquad \square$$

As we mentioned above, when $\alpha > 1/2$, if the regularity of the shape function increases, the space of controllable initial data diminishes. As we have just proved, this is no longer true if $\alpha \leq 1/2$. In this case, some regular initial data may be controlled only if more regularity is assumed for the shape function f.

REMARK 3.5. *There exists an alternative proof for the above proposition which allows us to construct an explicit null-controllable initial datum under hypothesis* (3.9). *Indeed, let $g \in L^2(0,T)$ such that the solution u_1 of the ordinary differential equation*

$$\begin{cases} u_1' + u_1 = g(t) f_1, & t \in (0,T), \\ u_1(0) = 1 \end{cases} \tag{3.10}$$

satisfies $u_1(T) = 0$.

Now, for each $n \geq 2$, solve the following backward ordinary differential equation:

$$\begin{cases} u_n' + n^{2\alpha} u_n = g(t) f_n, & t \in (0,T), \\ u_n(T) = 0. \end{cases} \tag{3.11}$$

It is easy to see that

$$|u_n(0)| \leq \sqrt{T} |f_n| e^{Tn^{2\alpha}} ||g||_{L^2}.$$

Under hypothesis (3.9) *the initial datum*

$$u^0 = \sin(x) + \sum_{n\geq 2} u_n(0) \sin(nx)$$

belongs to $H^m(0,\pi)$ for any $m \geq 0$ and it is null controllable.

This example can easily be generalized by choosing first the control corresponding to a finite number of Fourier components, and then determining the other Fourier components of the controllable initial datum from the final equilibrium condition in terms of this control.

More precisely, fix a finite $N \geq 1$ and an arbitrary choice of the first N Fourier components of the initial datum to be controlled: $a_1, \ldots, a_N$. Let $g = g(t)$ be such that each of the solutions of

$$\begin{cases} u_n' + n^{2\alpha} u_n = g(t) f_n, & t \in (0,T), \\ u_n(0) = a_n \end{cases} \tag{3.12}$$

satisfies $u_n(T) = 0$ for all $n = 1, \ldots, N$. The existence of this control g is guaranteed. Indeed, system (3.12) *is controllable since the classical Kalman rank condition is satisfied.*

Once this is done for each $n \geq N+1$ we solve the backward problem

$$\begin{cases} u_n' + n^{2\alpha} u_n = g(t) f_n, & t \in (0,T), \\ u_n(T) = 0. \end{cases} \tag{3.13}$$

Under assumption (3.9) *the controlled initial datum*

$$u^0 = \sum_{n=1}^{N} a_n \sin(nx) + \sum_{n \geq N+1} u_n(0) \sin(nx)$$

belongs to $H^m(0,\pi)$ for any $m \geq 0$.

3.4. Partial controllability. In order to better explain the previous result it is convenient to introduce the following definition.

DEFINITION 3.1. *The initial datum $u^0 \in L^2(0,\pi)$ is N-partially controllable in time $T > 0$ if there exists $g = g_N \in L^2(0,T)$ such that the solution u of* (2.3) *verifies*

$$\Pi_N(u(T,\cdot)) = 0, \tag{3.14}$$

where Π_N is the orthogonal projection over the space generated by the first N eigenfunctions $(\sqrt{2}\sin(nx)/\sqrt{\pi})_{1 \leq n \leq N}$.

Arguing as in Lemma 2.2 we can show that the N-partial controllability problem is equivalent to a finite moment problem and more precisely to the existence of $g_N \in L^2(0,T)$ such that

$$f_n \int_0^T g_N(t) e^{\lambda_n t} dt = -\frac{\pi}{2} a_n \quad \text{for any } 1 \leq n \leq N. \tag{3.15}$$

A function g_N with property (3.15) will be called N-partial control. Its existence is easy to prove since, as mentioned above, the Kalman rank condition is satisfied. The lack of controllability properties proved above on the case $\alpha \leq 1/2$ suggests that the controls g_N should diverge as $N \to \infty$. Let us check this fact in a simple but illustrative example.

The system is N-partially controllable if and only if, for all $k \geq 1$, there exists $g_{k,N} \in L^2(0,T)$ such that

$$\int_0^T g_{k,N}(t) e^{\lambda_n t} dt = \delta_{kn} \text{ for any } 1 \leq n \leq N. \tag{3.16}$$

If $(g_{k,N})_{N \geq 1}$ is bounded in $L^2(0,T)$, there exists a subsequence which weakly converges as $N \to \infty$ to $g_k \in L^2(0,T)$ and

$$\int_0^T g_k(t) e^{\lambda_n t} dt = \delta_{kn} \quad \forall n \geq 1. \tag{3.17}$$

But relation (3.17) cannot hold since $(e^{\lambda_n t})_{n \geq 1, n \neq k}$ is complete in $L^2(0,T)$ (the divergence property (1.6) still holds if one exponent is eliminated). Hence, $(g_{k,N})_{N \geq 1}$ may not be bounded in $L^2(0,T)$.

A sequence $(g_k)_{k \geq 1}$ with property (3.17) is called biorthogonal to $(e^{\lambda_n t})_{n \geq 1}$. When $\alpha \leq 1/2$ such a biorthogonal sequence does not exist. From the controllability point of view the fact that, for k fixed, $g_{k,N}$ diverges as $N \to \infty$ means that it is impossible to control to zero one Fourier mode of the initial datum. This is in

agreement with the positive result in Proposition 3.3 indicating that taking a more smooth control profile may increase the space of controllable data. Note that, as the Fourier components of the control profile f decay faster, the impact of the controls on the high frequencies decreases. This does help in building smooth data that are controllable, as the construction of Remark 3.5 shows.

In fact, according to Theorem 3.1, if the Fourier coefficients of the initial datum are not large enough, the sequence of N-partial controls $(g_N)_{N\geq 1}$ diverges and no control exists. □

The previous notion of N-partial controllability can be extended as follows: Given a subset $I \subset \mathbb{N}$ of indices we introduce the subspace H_I of $L^2(0,\pi)$ spanned by the eigenfunctions of the Laplacian with indices in I. More precisely,

$$H_I = \left\{ \varphi \in L^2(0,\pi) \,:\, \varphi(x) = \sum_{j\in I} a_j \sin(jx),\ \sum_{j\in I} |a_j|^2 < \infty \right\}. \tag{3.18}$$

We then introduce the orthogonal projection Π_I from $L^2(0,\pi)$ into H_I.

Definition 3.2. *System* (2.3) *is H_I-partially controllable in time $T > 0$ if for every initial datum $u^0 \in L^2(0,\pi)$ there exists a control $g \in L^2(0,T)$ such that the solution u of* (2.3) *verifies*

$$\Pi_I(u(T,\cdot)) = 0. \tag{3.19}$$

This control property is also equivalent to finding $g_I \in L^2(0,T)$ such that

$$f_j \int_0^T g_I(t) e^{\lambda_j t} dt = -\frac{\pi}{2} a_j \quad \forall j \in I. \tag{3.20}$$

Obviously, this generalizes the N-partial controllability problem that corresponds to the case where $I = \{1, 2, \dots, N\}$.

As mentioned in the introduction, the solvability of (3.19) and/or (3.20) depends on the summability condition

$$\sum_{j\in I} \frac{1}{|\lambda_j|} < \infty. \tag{3.21}$$

In the case under consideration, $\lambda_j = j^{2\alpha}$. Therefore, we see that (3.21) is satisfied under the following conditions:

1. When $\alpha > 1/2$ for $I = \mathbb{N}$. In this case partial controllability turns out to be complete null controllability.
2. When $0 < \alpha \leq 1/2$ for a suitable subsequence I_α of $\mathbb{N}$. It is obvious that one needs to consider a strict subsequence I_α of $\mathbb{N}$. Moreover, as α decreases, the subsequence I_α becomes more and more sparse in $\mathbb{N}$ and, therefore, the property of partial controllability weaker and weaker. This result agrees with a first intuition suggesting that an increase of diffusivity enhances the null-controllability properties of the system.

4. Proofs of some technical results.

4.1. Proof of Proposition 3.1. First of all we observe that

$$\int_0^T g(s)e^{n^{2\alpha}s}ds = \int_{-T/2}^{T/2} g(s+T/2)e^{n^{2\alpha}(s+T/2)}ds$$
$$= e^{n^{2\alpha}T/2}\int_{-T/2}^{T/2} g(s+T/2)e^{n^{2\alpha}s}ds = e^{n^{2\alpha}T/2}\int_{-T/2}^{T/2} h(s)e^{n^{2\alpha}s}ds$$

with $h(s) = g(s+T/2)$.

Hence, statement (a) of Proposition 3.1 is equivalent to the following one:

(4.1)

(a′) $\exists h \in L^2(-T/2, T/2)$ such that $\int_{-T/2}^{T/2} h(s)e^{n^{2\alpha}s}ds = e^{-n^{2\alpha}T/2}\alpha_n \quad \forall n \geq 1.$

We now prove that (a′) and (b) are equivalent.

- (a′) ⇒ (b).
  Let H be the Fourier transform of $h(s)1_{(-T/2,T/2)}$, i.e.,

  $$H(z) = \int_{-T/2}^{T/2} h(s)e^{-izs}ds,$$

  and let $F(z) = H(iz)$. According to the Paley–Wiener theorem (see, for instance, [3] or [21]), we know that $H : \mathbb{C} \to \mathbb{C}$ is an entire function of exponential type $\leq T/2$ and such that $\int_{-\infty}^{\infty} \mid H(x) \mid^2 dx < \infty$. Consequently, F is also an entire function of exponential type $\leq T/2$ such that $\int_{-\infty}^{\infty} \mid F(ix) \mid^2 dx < \infty$.
  Moreover, in view of (4.1),

  $$F(n^{2\alpha}) = H(in^{2\alpha}) = \int_{-T/2}^{T/2} h(s)e^{n^{2\alpha}s}ds = e^{-n^{2\alpha}T/2}\alpha_n.$$

  This shows that (b) holds.
- (b) ⇒ (a′).
  Let F be an entire function of exponential type $\leq T/2$, with $\int_{-\infty}^{\infty} \mid F(ix) \mid^2 dx < \infty$ and such that (3.6) holds.
  We then set $H(z) = F(-iz)$, which is also an entire function of exponential type $\leq T/2$ with $\int_{-\infty}^{\infty} \mid H(x) \mid^2 dx < \infty$.
  From the Paley–Wiener theorem we deduce that there exists $h \in L^2(-\frac{T}{2}, \frac{T}{2})$ such that

  $$H(z) = \int_{-T/2}^{T/2} h(s)e^{-izs}ds.$$

  We have that

  $$\int_{-T/2}^{T/2} h(s)e^{n^{2\alpha}s}ds = H(in^{2\alpha}) = F(n^{2\alpha}) = \alpha_n e^{-n^{2\alpha}T/2}$$

  and (a′) is verified.

This completes the proof of Proposition 3.1. □

4.2. Properties of the eigenvalues. Let us recall that the operator A_α we are dealing with has a sequence of eigenvalues $\lambda_n = n^{2\alpha}$, $n \geq 1$. Recall also that we are dealing with the case $0 < \alpha \leq 1/2$. In this section we deduce some properties of these eigenvalues.

LEMMA 4.1. *The sequence $(\lambda_n)_{n\geq 1}$ has the following properties:*

1. *It is strictly increasing and* $\lim_{n\to\infty} \lambda_n = \infty$.
2. *For any $n \geq 1$,*

$$\lambda_{n+1} - \lambda_n \leq \frac{2\alpha}{n^{1-2\alpha}}. \tag{4.2}$$

Proof. The first part is obvious. For the second one let us note that

$$\lambda_{n+1} - \lambda_n = (n+1)^{2\alpha} - n^{2\alpha} = n^{2\alpha}\left[\left(\frac{1}{n}+1\right)^{2\alpha} - 1\right]$$

and use the fact that for any $x > 0$ there exists ξ in $[0, x]$ such that

$$(x+1)^{2\alpha} = 1 + 2\alpha x + 2\alpha(2\alpha-1)\frac{x^2}{2}(\xi+1)^{2\alpha-2} \leq 1 + 2\alpha x.$$

It follows that

$$\lambda_{n+1} - \lambda_n = n^{2\alpha}\left[\left(\frac{1}{n}+1\right)^{2\alpha} - 1\right] \leq n^{2\alpha}\left[\left(1 + 2\alpha\frac{1}{n}\right) - 1\right] = \frac{2\alpha}{n^{1-2\alpha}}. \qquad \square$$

Concerning the distribution of the sequence $(\lambda_n)_{n\geq 1}$ the following can be said.

LEMMA 4.2. *There exists an increasing sequence $(n_k)_{k\in\mathbb{N}^*}$ in $\mathbb{N}^*$ such that*

1. *there exists $\beta > 0$ such that $0 < \beta < n_{k+1}^{2\alpha} - n_k^{2\alpha}$, for any $k \geq 1$;*
2. *for any $k \geq 1$, $|k - n_k^{2\alpha}| \leq \alpha$.*

Proof. If $\alpha = 1/2$, we may take $n_k = k$ and both properties are verified.

Consider now the case $\alpha < 1/2$. If $k = 1$, we take $n_k = 1$. Suppose that $k \geq 2$. Let $n'_k = \inf\{n \in \mathbb{N}^* : k \leq n^{2\alpha}\}$. We have

$$(n'_k - 1)^{2\alpha} < k \leq (n'_k)^{2\alpha}$$

and $n'_k > 1$.

Define

$$n_k = \begin{cases} n'_k - 1 & \text{if } k - (n'_k-1)^{2\alpha} \leq (n'_k)^{2\alpha} - k, \\ n'_k & \text{if } (n'_k)^{2\alpha} - k < k - (n'_k-1)^{2\alpha}. \end{cases}$$

Taking Lemma 4.1 into account we obtain that

$$\begin{aligned} |k - (n_k)^{2\alpha}| &= \min\{k - (n'_k-1)^{2\alpha},\, (n'_k)^{2\alpha} - k\} \leq \frac{1}{2}\left(k - (n'_k-1)^{2\alpha} + (n'_k)^{2\alpha} - k\right) \\ &= \frac{1}{2}\left((n'_k)^{2\alpha} - (n'_k-1)^{2\alpha}\right) = \frac{1}{2}(\lambda_{n'_k} - \lambda_{n'_k-1}) \leq \frac{\alpha}{(n'_k-1)^{1-2\alpha}} \leq \alpha \end{aligned}$$

and the second property of the statement of the Lemma is verified. On the other hand

$$\begin{aligned} |n_k^{2\alpha} - n_{k-1}^{2\alpha}| &\geq 1 - \left(|n_k^{2\alpha} - k| + |n_{k-1}^{2\alpha} - (k-1)|\right) \\ &\geq 1 - \alpha\left[\frac{1}{(n'_k-1)^{1-2\alpha}} + \frac{1}{(n'_{k-1}-1)^{1-2\alpha}}\right] \geq 1 - 2\alpha > 0 \end{aligned}$$

and the first property is verified as well. □

REMARK 4.1. *Lemma* 4.2 *says that there exists a subsequence* $(\lambda_{n_k})_{k\geq 1}$ *of the sequence of eigenvalues* $(\lambda_n)_{n\geq 1}$ *such that*

- $|\lambda_{n_k} - k| \leq \alpha$ *for all* $k \geq 1$*;*
- $|\lambda_{n_k} - \lambda_{n_{k-1}}| > \beta > 0$ *for all* $k \geq 2$.

This subsequence (λ_{n_k}) *will be used to prove Proposition* 3.2.

REMARK 4.2. *The subsequence* $(n_k)_{k\geq 1}$ *constructed in the proof of Lemma* 4.2 *satisfies*

$$|n_{k+1} - n_k| \geq \frac{1}{2\alpha} k^{\frac{1}{2\alpha}-1} - 2 \quad \forall k \geq 1. \tag{4.3}$$

Indeed,

$$n_{k+1} - n_k \geq n'_{k+1} - n'_k - 1 \geq (k+1)^{\frac{1}{2\alpha}} - k^{\frac{1}{2\alpha}} - 2 \geq \frac{1}{2\alpha} k^{\frac{1}{2\alpha}-1} - 2.$$

4.3. Proof of Proposition 3.2. We introduce the function $G : \mathbb{C} \to \mathbb{C}$:

$$G(z) = e^{Tz/2} F(z). \tag{4.4}$$

In view of properties (i)–(ii) of F it is immediate that

$$G \text{ is an entire function of exponential type } \leq T; \tag{4.5}$$

$$\int_{-\infty}^{\infty} |\, G(iy)\,|^2 \, dy < \infty; \tag{4.6}$$

$$\forall \delta > 0 : |\, G(n^{2\alpha})\,| \leq C_\delta e^{\delta n^{2\alpha}} \quad \forall n \geq 1. \tag{4.7}$$

Moreover, G is bounded on the negative semiaxis, i.e.,

$$\exists L > 0 : |\, G(-x)\,| \leq L \quad \forall x \geq 0. \tag{4.8}$$

Property (4.8) is an immediate consequence of the fact that F is of exponential type $\leq T/2$.

We now introduce

$$G_1(z) = G\left(-ze^{i\pi/4}\right) \tag{4.9}$$

and apply the Phragmén–Lindelöf theorem to G_1 in the sector $|\arg z| < \pi/4$ to deduce that there exists $M_1 > 0$ such that

$$|\, G_1(z)\,| \leq M_1 \quad \forall z \in \mathbb{C} : |\arg z| \leq \pi/4. \tag{4.10}$$

This is possible since

$$G_1 \text{ is analytic on } \mathbb{C}; \tag{4.11}$$

$$G_1 \text{ is bounded when } \arg z = \pm\pi/4; \tag{4.12}$$

$$|\, G_1(z)\,| = O\left(e^{|z|^\beta}\right) \text{ for some } \beta < 2, \text{ as } |\, z\,| \to \infty. \tag{4.13}$$

Note that (4.12) holds because G is bounded along the imaginary axis by (4.6) and on the negative semiaxis by (4.8). On the other hand, (4.13) holds for any $\beta > 1$ since $|\, G_1(z)\,| = O\left(e^{T|z|}\right)$, due to (4.5).

As a consequence of (4.10) we deduce that

$$| G(z) | \leq M_1 \tag{4.14}$$

for all $z \in \mathbb{C}$ with $\arg(z) \in [\pi/2, \pi]$.

In a similar way we may prove the existence of $M_2 > 0$ such that

$$| G(z) | \leq M_2 \tag{4.15}$$

for all $z \in \mathbb{C}$ with $\arg(z) \in [\pi, 3\pi/2]$. Hence, G is bounded in the half complex plane $\operatorname{Re} z \leq 0$.

Let us now consider the function

$$H_\delta(z) = G(z)e^{-\delta z} = e^{Tz/2}e^{-\delta z}F(z) \tag{4.16}$$

defined on the half-plane $\operatorname{Re} z \geq 0$. It is easy to see that H_δ satisfies the following properties:

$$H_\delta \text{ is analytic on the closed half-plane } \operatorname{Re} z \geq 0; \tag{4.17}$$

$$H_\delta \text{ is of exponential type;} \tag{4.18}$$

$$\exists C_\delta > 0 :| H_\delta(n^{2\alpha}) | \leq C_\delta \quad \forall n \geq 1; \tag{4.19}$$

$$H_\delta \text{ is bounded on the imaginary axis.} \tag{4.20}$$

We now introduce the indicator function

$$h_{H_\delta}(\theta) = \limsup_{r\to\infty} \left[\frac{1}{r}\log\left|H_\delta\left(re^{i\theta}\right)\right|\right] \quad \forall\theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]. \tag{4.21}$$

Lemma 4.3. *For any $\delta < T$, there exists a positive constant $A > 0$ such that*

$$h_{H_\delta}(\theta) \leq A\cos\theta \quad \forall\theta \in [-\pi/2, \pi/2]. \tag{4.22}$$

Proof of Lemma 4.3. We have

$$\begin{aligned}\log\left|H_\delta\left(re^{i\theta}\right)\right| &= \log\left|e^{(T-2\delta)re^{i\theta}/2}F\left(re^{i\theta}\right)\right| \\ &= \log\left|e^{(T-2\delta)re^{i\theta}/2}\right| + \log\left|F\left(re^{i\theta}\right)\right| = \frac{(T-2\delta)r\cos\theta}{2} + \log\left|F\left(re^{i\theta}\right)\right|.\end{aligned} \tag{4.23}$$

On the other hand, arguing as in the proof of Proposition 3.1, we deduce from the Paley–Wiener theorem the existence of a function $\psi \in L^2(-T/2, T/2)$ such that

$$F(z) = \int_{-T/2}^{T/2} \psi(s)e^{zs}ds.$$

Therefore

$$\left|F\left(re^{i\theta}\right)\right| \leq \int_{-T/2}^{T/2} | \psi(s) | e^{sr\cos\theta}ds \leq e^{Tr|\cos\theta|/2}\int_{-T/2}^{T/2} | \psi(s) | ds. \tag{4.24}$$

Combining (4.22) and (4.24) we deduce that

$$\log\left|H_\delta\left(re^{i\theta}\right)\right| \leq (T-\delta)r | \cos\theta | + \log \| \psi \|_{L^1(-T/2, T/2)}. \tag{4.25}$$

From (4.25) we easily deduce that (4.21) holds with

$$A = T - \delta. \qquad \square \tag{4.26}$$

Let us now return to the proof of Proposition 3.2. By a result of Duffin and Schaeffer [4] (see also [3, p. 191]) we have the following theorem.

THEOREM 4.1. *Let f be analytic in $\mid \arg(z) \mid \leq \gamma \leq \pi/2$ and suppose that its indicator function h_f satisfies*

$$\mid h_f(\theta) \mid \leq a \mid \cos\theta \mid + b \mid \sin\theta \mid \quad \forall \mid \theta \mid \leq \gamma \tag{4.27}$$

with $a, b > 0$ and $b < \pi$.

If $(\nu_k)_{k\geq 1}$ is an increasing sequence of real numbers such that

$$\nu_{k+1} - \nu_k \geq \beta > 0 \quad \forall k \geq 1, \tag{4.28}$$

$$\mid \nu_k - k \mid \leq L \quad \forall k \geq 1, \tag{4.29}$$

and $f(\nu_k)$ is bounded, then $f(x)$ is bounded for all $x > 0$.

We apply Theorem 4.1 to the function H_δ with $\nu_k = \lambda_{n_k} = n_k^{2\alpha}$, where n_k are given by Lemma 4.2. The sequence $(\nu_k)_{k\geq 1}$ satisfies the hypotheses of Theorem 4.1. Moreover,

$$|H_\delta(\nu_k)| = |G(\nu_k)|e^{-\delta\nu_k} = |G(\lambda_{n_k})|e^{-\delta n_k^{2\alpha}} \leq C_\delta.$$

We deduce from Theorem 4.1 that H_δ is bounded on the positive real axis. Since, by (4.20), H_δ is also bounded on the imaginary axis, we deduce, by the Phragmén–Lindelöf theorem, that H_δ is bounded in the half-plane $\operatorname{Re} z \geq 0$ for all $0 < \delta < S$.

Consequently,

- G is bounded on the half-plane $\operatorname{Re} z \leq 0$;
- $\mid G(z) \mid \leq C(\delta)e^{\delta|z|}$ on the half-plane $\operatorname{Re} z \geq 0$ for all $0 < \delta < S$;
- G is entire;
- $\int_{-\infty}^{\infty} \mid G(iy) \mid^2 dy < \infty$.

According to the Paley–Wiener theorem, these properties are sufficient to guarantee that $G \equiv 0$. $\square$

5. On the lack of observability estimates. A natural approach to the problem of null controllability of heat equations consists in dealing with the dual observability problem for the adjoint system (see, for instance, [8], [22], and [23]).

More precisely, the null controllability of system (2.3) in $L^2(0,\pi)$ with controls in $L^2(0,T)$ is equivalent to the existence of a positive constant $C > 0$ such that

$$\| \varphi(0) \|^2_{L^2(0,\pi)} \leq C \int_0^T \left| \int_0^\pi \varphi(t,x) f(x) dx \right|^2 dt \quad \forall \varphi^T \in L^2(0,\pi), \tag{5.1}$$

where φ is the solution of (2.6).

As we have shown in Theorem 3.1, when $0 < \alpha \leq 1/2$, the null-controllability result is false and therefore (5.1) does not hold. In fact, according to the statement of Theorem 3.1, it turns out that all the possible weaker versions of (5.1) in which the L^2-norm of the left-hand side is replaced by an $H^{-\sigma}$-norm for any $\sigma > 0$ are false as well.

In this section we describe how the lack of observability inequalities of form (5.1) may be proved directly.

In view of the Fourier series expansion of the solution φ of (2.6) we have

$$\varphi(t,x)=\sum_{n\geq 1} a_n e^{-n^{2\alpha}(T-t)}\sin(nx).$$

Thus (5.1) is equivalent to

$$\sum_{n\geq 1}|a_n|^2 e^{-2n^{2\alpha}T}\leq C\int_0^T\left|\sum_{n\geq 1}a_n f_n e^{-n^{2\alpha}t}\right|^2 dt. \tag{5.2}$$

Inequalities of form (5.2) are well known to be true when $\alpha>1/2$ (see, for instance, [7] and [19]). But they fail when $\alpha\leq 1/2$ since the series $\sum_{n\geq 1}1/n^{2\alpha}$ diverges in that case (see [17]). More precisely, the following negative result holds.

PROPOSITION 5.1. *When $0<\alpha\leq 1/2$ there is no sequence $(\rho_n)_{n\geq 1}$ of positive weights, i.e., $\rho_n>0$ for all $n\geq 1$, such that*

$$\sum_{n\geq 1}\rho_n|b_n|^2\leq\int_0^T\left|\sum_{n\geq 1}b_n e^{-n^{2\alpha}t}\right|^2 dt \tag{5.3}$$

for all finite sequence $(b_n)_{n\geq 1}$.

This result excludes inequality (5.2) and any other weaker version of it. Observe that an inequality like (5.2) is equivalent to the null controllability in time T of all initial data in the class

$$H=\left\{u^0=\sum_{n\geq 1}a_n\sin(n\pi):\sum_{n\geq 1}|a_n|^2/\rho_n<\infty\right\},$$

and, according to the result of Theorem 3.1, we know that this is false for all sequences of weights $(\rho_n)_{n\geq 1}$.

Proposition 5.1 is an immediate consequence of the following one.

PROPOSITION 5.2. *Let $(\nu_n\}_{n\geq 1}$ be an increasing sequence of positive real numbers. Assume that there exists a sequence of positive weights $(\rho_n)_{n\geq 1}$ such that*

$$\sum_{n\geq 1}\rho_n\,|a_n|^2\leq\int_0^T\left|\sum_{n\geq 1}a_n e^{-\nu_n t}\right|^2 dt \tag{5.4}$$

for all finite sequence $(a_n)_{n\geq 1}$. Then, necessarily,

$$\sum_{n\geq 1}\frac{1}{\nu_n}<\infty. \tag{5.5}$$

We refer to Proposition 3.5 in [17] for a proof.

The proof of Proposition 5.2 provides in fact a stronger result. Namely, it shows that if the sequence $(\nu_n)_{n\geq 1}$ is such that for some n_0 and $\rho>0$ we have

$$\rho\,|a_{n_0}|^2\leq\int_0^1\left|\sum_{n\geq 0}a_n e^{-\nu_n t}\right|^2 dt \tag{5.6}$$

for all finite sequence $(a_n)_{n\geq 1}$, then, necessarily,

$$\sum_{n\geq 1} \frac{1}{\nu_n} < \infty. \tag{5.7}$$

Inequalities of form (5.6) are related to the so-called spectral controllability problem, which consists in analyzing whether all the eigenfunctions may be driven to zero in finite time. Theorem 3.1 provides a negative answer.[2] Proposition 5.1 provides a second proof of this negative result in which the effect of the divergence of the series $\sum_{n\geq 1} 1/\nu_n$ is clearly seen.

Note that spectral controllability also implies that all finite combinations of eigenfunctions are controllable and also show the controllability property in the infinite-dimensional space generated by the eigenfunctions, with suitable weights as the frequency increases.

The results on partial controllability of section 3.4 can also be understood in terms of observability inequalities. Indeed, the H_I-partial controllability property is equivalent to the observability property (5.1) in the subspace of solutions of the adjoint system (2.6) with initial data φ^T in H_I, i.e., of solutions φ of (2.6) involving only the Fourier coefficients with indices $j \in I$. This turns out to be equivalent to an inequality of the form

$$\sum_{j\in I} |a_j|^2 e^{-2\lambda_j T} \leq C \int_0^T \left| \sum_{j\in I} a_j e^{-\lambda_j t} \right|^2 dt \tag{5.8}$$

for all finite sequence $(a_n)_{n\geq 1}$.

Inequality (5.8) holds provided the subsequence $(\lambda_j)_{j\in I}$ fulfills a gap condition and the summability condition

$$\sum_{j\in I} \frac{1}{|\lambda_j|} < \infty. \tag{5.9}$$

As indicated in section 3.4 these conditions are satisfied provided the sequence I is sparse enough.

6. A hyperbolic problem. In this section we consider a hyperbolic system involving the operator A_α and address the corresponding control problem: Given $T > 0$, $f \in L^2(0,\pi)$, and initial data (u^0, u^1), find $g \in L^2(0,T)$ such that the solution u of the problem

$$\begin{cases} u_{tt} + A_\alpha u = g(t)f(x), & x \in (0,\pi),\ t \in (0,T), \\ u = 0, & x \in \{0,\pi\},\ t \in (0,T), \\ u(0,x) = u^0(x), \quad u_t(0,x) = u^1(x), & x \in (0,\pi), \end{cases} \tag{6.1}$$

satisfies

$$u(T,\cdot) = u_t(T,\cdot) = 0. \tag{6.2}$$

[2] In fact the lack of (5.6) for any index n_0 shows that there is no single eigenfunction that may be driven to zero in final time with $L^2(0,T)$ controls.

Note that system (6.1) is a generalization of the wave equation

$$u_{tt} - u_{xx} = g(t)f(x)$$

that corresponds to the case $\alpha = 1$. In the absence of control (i.e., when $g = 0$) system (6.1) is conservative and generates a group of isometries in the corresponding energy space.

The eigenvalues corresponding to (6.1) are given by $i\nu_n$, $n \in \mathbb{Z}^*$, where

$$\nu_n = \operatorname{sgn}(n)|n|^\alpha \quad \forall n \neq 0.$$

When $\alpha \geq 1$ system (6.1) is controllable provided the control profile is such that all the Fourier components do not vanish. In particular that is the case for the wave equation in time $T = 2\pi$. As we shall see, the situation is even better when $\alpha > 1$, in which case the control property holds for an arbitrarily short time $T > 0$.

More precisely, the following holds.

THEOREM 6.1. *Let $\alpha \geq 1$ and f_k given by* (2.9) *satisfying* (2.10). *Any initial state in the space*

$$H = \left\{ (u^0, u^1) = \sum_{k\in\mathbb{Z}^*} a_k(\frac{1}{k^\alpha\, i}, -1)\sin(kx) : \sum_{k\in\mathbb{Z}^*} \frac{|a_k|^2}{|f_k|^2} < \infty \right\}$$

is controllable in time $T \geq 2\pi$ if $\alpha = 1$ and any time $T > 0$ if $\alpha > 1$, by means of a control $g \in L^2(0,T)$.

Proof. We first claim that the controllability of all initial data from H is equivalent to the inequality

$$C \sum_{n\in\mathbb{Z}^*} |c_n|^2 \leq \int_0^T \left| \sum_{n\in\mathbb{Z}^*} c_n e^{i\nu_n t} \right|^2 \tag{6.3}$$

for every sequence $(c_n)_{n\in\mathbb{Z}^*} \in \ell^2$.

Indeed, as in Lemma 2.2, it is easy to show that the controllability of

$$(u^0, u^1) = \sum_{k\in\mathbb{Z}^*} a_k\Big(\frac{1}{k^\alpha\, i}, -1\Big)\sin(kx)$$

is equivalent to the following moment problem: Find $g \in L^2(0,T)$ such that

$$f_k \int_0^T g(t)e^{i\nu_k t}dt = a_k \quad \forall k \in \mathbb{Z}^*. \tag{6.4}$$

The moment problem (6.4) has a solution for any $(a_n/f_n)_{n\in\mathbb{Z}^*} \in \ell^2$ if and only if[3] the sequence $(e^{i\nu_n t})_{n\in\mathbb{Z}^*}$ is a Riesz–Fischer sequence in $L^2(0,T)$.

On the other hand, from the characterization of the Riesz–Fischer sequences (see [21, Theorem 3, p. 155]), it follows that the sequence $(e^{i\nu_n t})_{n\in\mathbb{Z}^*}$ is a Riesz–Fischer sequence in $L^2(0,T)$ if and only if (6.3) holds. This proves the claim.

We deduce that the moment problem (6.4) has a solution for any $(a_n/f_n)_{n\in\mathbb{Z}^*} \in \ell^2$ if and only if (6.3) holds.

[3]This is an immediate consequence of the definition of *Riesz–Fischer sequence*. Recall that a sequence of vectors $(x_n)_{n\in\mathbb{Z}^*}$ belonging to a Hilbert space H is said to be a Riesz–Fischer sequence if the moment problem $(x, x_n) = c_n$ for all $n \in \mathbb{Z}^*$ has a solution $x \in H$ for any $(c_n)_{n\in\mathbb{Z}^*} \in \ell^2$.

This can also be seen by the so-called HUM method by Lions [15] (see Remark 6.1).

Let us now show that, in the case $\alpha \geq 1$, (6.3) holds under the restrictions on T in the statement of the theorem. Indeed, from [2] (see also [9] and [10]), it follows that (6.3) holds for any $T > 2\pi/\gamma_\infty$ if

$$\liminf_{|n|\to\infty} |\nu_{n+1} - \nu_n| \geq \gamma_\infty > 0. \tag{6.5}$$

Since $|\nu_{n+1} - \nu_n| = (n+1)^\alpha - n^\alpha$, it follows that property (6.5) holds for any $T > 2\pi$ if $\alpha = 1$ (since $\gamma_\infty = 1$) and for any $T > 0$ if $\alpha > 1$ (since $\gamma_\infty = \infty$). Moreover, when $\alpha = 1$, in view of the time-orthogonality of the complex exponentials involved in the Fourier series development of solutions, property (6.5) holds for $T = 2\pi$ as well.

This completes the proof of the theorem. □

REMARK 6.1. *The controllability of* (6.1) *with initial data in* H *and controls in* $L^2(0,T)$ *is equivalent to the existence of a positive constant* $C > 0$ *such that*

$$||(\varphi^0, \varphi^1)||^2_{H'} \leq C \int_0^T \left| \int_0^\pi \varphi(t,x) f(x) dx \right|^2 dt,$$

$$\forall (\varphi^0, \varphi^1) \in H' = \left\{ (\varphi^0, \varphi^1) = \sum_{k\in\mathbb{Z}^*} a_k \left(\frac{1}{k^\alpha\, i}, -1 \right) \sin(kx) : \sum_{k\in\mathbb{Z}^*} |a_k|^2 |f_k|^2 < \infty \right\}, \tag{6.6}$$

where (φ, φ_t) *is the solution of*

$$\begin{cases} \varphi_{tt} + A_\alpha \varphi = 0, & x \in (0,\pi),\ t \in (0,T), \\ \varphi = 0, & x \in \{0,\pi\},\ t \in (0,T), \\ \varphi(0,x) = \varphi^0(x), \quad \varphi_t(0,x) = \varphi^1(x), & x \in (0,\pi). \end{cases} \tag{6.7}$$

Inequality (6.6) *is usually called the* observation inequality.

Using the Fourier expansion of the solutions of (6.7) *it is easy to see that* (6.6) *may be written as* (6.3). *Inequality* (6.3) *may be proved by means of the classical Ingham inequality* (*see* [21]).

Once inequality (6.6) *is known to hold, the control* $g = g(t)$ *can be built by minimizing the quadratic functional*

$$J(\varphi^0, \varphi^1) = \frac{1}{2} \int_0^T \left| \int_0^\pi \varphi(t,x) f(x) dx \right|^2 dt + \langle (u^1, -u^0), (\varphi^0, \varphi^1) \rangle \tag{6.8}$$

in the space H'. *Indeed, under the assumption that* $(u^1, -u^0) \in H$ (*the dual of* H') *the functional* J *is continuous, convex, and coercive in the Hilbert space* H'. *Thus its minimum exists. It is then easy to see that the control* g *we are looking for is* $g(t) = \int_0^\pi \hat{\varphi}(t,x) f(x) dx$, *where* $\hat{\varphi}$ *is the solution of* (6.7) *with the minimizer of* J *as initial datum.*

The proof of Theorem 6.1 is based on inequality (6.3), which holds in the case $\alpha \geq 1$. Nevertheless, if $\alpha < 1$, there exists no uniform gap between two consecutive eigenvalues and (6.3) does not hold. The controllability properties are very different in this case.

In fact, when $0 < \alpha < 1$ system (6.1) is very badly controllable. Even the *spectral control property* fails to hold. We recall that system is said to be spectrally controllable if all initial data consisting in a single eigenfunction of the system may be controlled.

THEOREM 6.2. *If $\alpha < 1$, equation (6.1) is not spectrally controllable in any time $T > 0$.*

Proof. Suppose that (6.1) is spectrally controllable. Hence, every eigenfunction of system (6.1) may be driven to zero by using a control in $L^2(0,T)$.

But an initial datum of the form $(u^0, u^1) = (1/k^\alpha\, i, -1)\sin(kx)$ is controllable if and only if there exists $g \in L^2(0,T)$ such that

$$\int_0^T g(t)e^{i\nu_n t}dt = \delta_{nk}/f_k \quad \forall n \neq 0. \tag{6.9}$$

From the Paley–Wiener theorem we obtain that (6.9) implies the existence of an entire function G of exponential type $T/2$, such that $\int_{-\infty}^{\infty} |G(x)|^2 dx < \infty$ and $G(\nu_n) = 0$ for all $n \neq 0, k$.

Let $n_G(r)$ denote the number of zeros of the function G which belong to the ball of center zero and radius r,

$$n_G(r) = \#\{z \in \mathbb{C} \,:\, G(z) = 0 \text{ and } |z| \leq r\}.$$

We have

$$n_G(r) = 2\#\{n \in \mathbb{N}^* \,:\, n^\alpha \leq r\} = 2\left[r^{\frac{1}{\alpha}}\right].$$

Since $\alpha < 1$ it follows that

$$\lim_{r\to\infty} n_G(r)/r = \infty. \tag{6.10}$$

We need now the following result, which is a consequence of the well-known Jensen formula (see [21, Theorems 2 and 3, pp. 59–61]): *if f is an entire nontrivial function of exponential type, then $n_f(r)/r$ remains bounded as r tends to infinity.*

From (6.10) and the previous theorem it follows that $G \equiv 0$, which contradicts (6.9). □

Our results show that for the hyperbolic equation (6.1) the critical exponent becomes $\alpha = 1$, instead of the exponent $\alpha = 1/2$ we have obtained for the parabolic equations (2.3).

7. Comments.

7.1. More general $1 - d$ problems. In this article we have considered the problem of controllability of a parabolic equation involving the fractional power of the Laplace operator. The control has a fixed shape, given by the function f. The problems of distributed control of the form $v(t,x)1_\omega$, with ω a subinterval of $(0,\pi)$, or of boundary control $v(t)$ may also be considered and will be treated elsewhere by similar techniques. The $1-d$ analysis on the wave equation in section 6 may also be carried out for the Schrödinger and the beam equations.

7.2. Multidimensional problems. In several space dimensions, $N \geq 2$, similar problems can be analyzed. Consider the Dirichlet problem,

$$\begin{cases} u_t + (-\Delta)^\alpha u = g(t)f(x) & \text{in } \Omega \times (0,T), \\ u = 0 & \text{on } \partial\Omega \times (0,T), \\ u(0,x) = u^0(x) & \text{in } \Omega. \end{cases} \tag{7.1}$$

Here Ω is a bounded domain of $\mathbb{R}^N$.

By Weyl's theorem, the spectrum of the Laplacian grows as the frequency increases in the following way: $\lambda_n \sim c(\Omega)n^{2/N}$.

According to this, the spectrum of the α-power of the Laplacian, $(-\Delta)^\alpha$, grows at a rate $n^{2\alpha/N}$ as $n \to \infty$. The critical case is then $\alpha = N/2$. One can then expect to obtain positive results for $\alpha > N/2$ and negative ones, as those presented here, for $\alpha \leq N/2$. An analysis of this multidimensional problem is also to be done.

REFERENCES

[1] S. A. Avdonin and S. A. Ivanov, *Families of Exponentials. The Method of Moments in Controllability Problems for Distributed Parameter Systems*, Cambridge University Press, Cambridge, UK, 1995.

[2] J. M. Ball and M. Slemrod, *Nonharmonic Fourier series and the stabilization of distributed semi-linear control systems*, Commun. Pure Appl. Math., 32 (1979), pp. 555–587.

[3] R. P. Boas, *Entire Functions*, Academic Press, New York, 1954.

[4] R. J. Duffin and A. C. Schaeffer, *Power series with bounded coefficients*, Amer. J. Math., 67 (1945), pp. 141–154.

[5] L. Escauriaza, G. Seregin, and V. Sverak, *On backward uniqueness for parabolic equations*, Arch. Ration. Mech. Anal., 169 (2003), pp. 147–157.

[6] H. O. Fattorini, *Control in finite time of differential equations in Banach space*, Commun. Pure Appl. Math., 19 (1966), pp. 17–34.

[7] H. O. Fattorini and D. L. Russell, *Exact controllability theorems for linear parabolic equations in one space dimension*, Arch. Ration. Mech. Anal., 43 (1971), pp. 272–292.

[8] E. Fernández-Cara and E. Zuazua, *The cost of approximate controllability for heat equations: The linear case*, Adv. Differential Equations, 5 (2000), pp. 465–514.

[9] A. Haraux, *Séries lacunaires et contrôle semi-interne des vibrations d'une plaque rectangulaire*, J. Math. Pure Appl., 68 (1989), pp. 457–465.

[10] A. E. Ingham, *Some trigonometrical inequalities with applications to the theory of series*, Math. Z., 41 (1936), pp. 367–369.

[11] A. Khapalov, *Approximate controllability and its well-posedness for the semilinear reaction-diffusion equation with internal lumped controls*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 83–98.

[12] A. A. Kilbas, H. M. Srivastava, and J. J. Trujillo, *Fractional differential equations: An emergent field in applied and mathematical science*, in Factorisation, Singular Operators and Relative Problems (FSORP'02, Madeira), S. Samko, A. Lebre, and A. F. dos Santos, eds., Kluwer, Boston, 2002, pp. 151–173.

[13] G. Lebeau and L. Robbiano, *Contrôle exact de l'équation de la chaleur*, Commun. Partial Differential Equations, 20 (1995), pp. 335–356.

[14] G. Lebeau and E. Zuazua, *Null controllability of a system of linear thermoelasticity*, Arch. Ration. Mech, Anal., 141 (1998), pp. 297–329.

[15] J. L. Lions, *Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués, Tome* 1: *Contrôlabilité exacte*, Rech. Math. Appl. 8, Masson, Paris, 1988.

[16] R. Metzler and J. Klafter, *The random walk's guide to anomalous diffusion: A fractional dynamics approach*, Phys. Rep., 339 (2000), pp. 1–77.

[17] S. Micu and E. Zuazua, *On the lack of null-controllability of the heat equation on the half line*, Trans. Amer. Math. Soc., 353 (2000), pp. 1635–1659.

[18] S. Micu and E. Zuazua, *On the lack of null-controllability of the heat equation on the half space*, Portugal. Math., 58 (2001), pp. 1–24.

[19] D. L. Russell, *Controllability and stabilizability theory for linear partial differential equations. Recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639–739.

[20] L. Schwartz, *Etude des sommes d'exponentielles*, Hermann, Paris, 1959.

[21] R. M. Young, *An Introduction to Nonharmonic Fourier Series*, Academic Press, New York, 1980.

[22] E. Zuazua, *Some problems and results on the controllability of partial differential equations*, in Second European Congress of Mathematics, Budapest, 1996, Progr. Math. 165, Birkhäuser-Verlag, Basel, 1998, pp. 276–311.

[23] E. Zuazua, *Controllability of partial differential equations and its semi-discrete approximation*, Discrete Contin. Dyn. Syst., 8 (2002), pp. 469–513.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 6, pp. 1973–1991

STATE FEEDBACK H_∞ CONTROL FOR A CLASS OF NONLINEAR STOCHASTIC SYSTEMS*

WEIHAI ZHANG† AND BOR-SEN CHEN‡

Abstract. This paper discusses the H_∞ control problem for a class of nonlinear stochastic systems with both state- and disturbance-dependent noise. By means of Hamilton–Jacobi equations, both infinite and finite horizon nonlinear stochastic H_∞ control designs are developed.

Some results on nonlinear H_∞ control of deterministic systems are generalized to a stochastic setting. We introduce some useful concepts such as "zero-state observability" and "zero-state detectability" which, together with the stochastic LaSalle invariance principle, yield some valuable consequences in infinite horizon nonlinear stochastic H_∞ control.

Key words. H_∞ control, nonlinear stochastic systems, Hamilton–Jacobi equation, zero-state observability, zero-state detectability

AMS subject classifications. 93C10, 93D09, 93E15

DOI. 10.1137/S0363012903423727

1. Introduction.

In practice, when the exogenous disturbance enters the system, an H_∞ control design is often first considered, when a control law is sought to efficiently eliminate the effect of the disturbance; see [12], [13] and the references therein. Theoretically, study of H_∞ control first starts from the deterministic linear systems, and the derivation of the state-space formulation of the standard H_∞ control leads to a breakthrough; details can be found in the prize-winning paper [17]. From the viewpoint of the state space, the linear H_∞ control problem can be converted into the study of a game-theoretic Riccati equation, and the "completion of square methodology," similar to the linear quadratic (LQ) and linear quadratic Gaussian (LQG) theories, can be applied; see [15], [16], [20], [21], [25], and [30].

Soon after the appearance of [17], the nonlinear H_∞ control problem (on deterministic systems) was investigated by many authors; see [5], [14], [18], and [19]. From the time-domain perspective, an H_∞ norm of the transfer function is nothing else but the $\mathcal{L}_2$-induced norm of the input-output operator with initial state zero. This important feature makes it possible to develop nonlinear or stochastic H_∞ theory. Van der Schaft [5] made a contribution to the state feedback H_∞ control for nonlinear deterministic systems with infinite time horizon, where a relatively deep tool, i.e., the strict relation between Hamilton–Jacobi equations (HJEs) and invariant manifolds of Hamiltonian vector fields, was applied. Using this tool, he showed that a local solution to the primal nonlinear H_∞ control exists if its linearized H_∞ control problem is solvable. The authors of [18] and [19] dealt with the output feedback H_∞ control of nonlinear deterministic systems with incomplete state information, and a separation principle was obtained. In [5], [18], and [19], the differential geometric approaches

*Received by the editors February 27, 2003; accepted for publication (in revised form) June 2, 2005; published electronically January 6, 2006.

http://www.siam.org/journals/sicon/44-6/42372.html

†College of Information and Electrical Engineering, Shandong University of Science and Technology, Qingdao 266510, P.R. China, and College of Electronic Information and Control Engineering, Shandong Institute of Light Industry, Jinan 250100, P.R. China (w_hzhang@163.com). The research of this author was supported by NSF of China grant 60474013.

‡Department of Electrical Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan (bschen@ee.nthu.edu.tw). The research of this author was supported by NSC of Taiwan contract NSC 92-2213-E-007-001.

were employed, which cannot be directly extended to stochastic H_∞ control for technical reasons. Chen, Tseng, and Vang [14] present a fuzzy treatment for nonlinear H_2/H_∞ control, and [11] treats the singular H_∞ control of nonlinear systems.

In recent years, stochastic H_∞ control systems, such as Markovian jump systems [22], [23], [24], H_∞ Gaussian control design [3], and Itô differential systems (systems governed by the Itô equation) have received a great deal of attention (see [2], [28], [30], [33], [34], [35], [36], and the references therein). H_∞ control has significant application in other problems, such as filtering theory [4]. Up to now, most of the work on stochastic H_∞ control for the Itô differential systems concentrated on the linear case, while for the general nonlinear systems, few results have been reported.

This paper will follow along the lines of [5] to study stochastic H_∞ control (including infinite and finite time horizons) for a class of affine nonlinear Itô differential systems, mainly using the completion of square methodology. To achieve our goal, some essential difficulties must be overcome. For instance, in order to discuss the relation between external stability and the existence of global solutions of the corresponding HJE, the so-called stochastic dissipative theory is developed, which can be viewed as a generalized version of [10]. Likewise, to present sufficient conditions for a closed-loop system to be internally stable, zero-state observability and detectability are introduced for nonlinear stochastic systems which, combined with the stochastic LaSalle invariance principle [8], yield some valuable consequences.

Section 2 develops the dissipative theory for stochastic systems paralleling that of [10]. Theorem 2.1 is an important result of this section, which extends Theorem 1 of [7], and will be used in section 3. Similar to applications of Theorem 1 of [7] in [6], [7], and [9], we also believe that Theorem 2.1 can be applied to nonlinear stochastic stability and stabilization, which merits further study. Section 3 is concerned with the infinite horizon H_∞ control of nonlinear affine stochastic systems, where the H_∞-norm is expressed by the norm of the nonlinear perturbation operator $\tilde{\mathcal{L}}_{zd}$. Theorem 3.1 and Corollary 3.1 extend Theorem 16 and Corollary 17 of [5], respectively. Lemma 3.2 may be viewed as the "nonlinear stochastic bounded real lemma (SBRL)." Section 4 is devoted to the finite horizon H_∞ control of affine nonlinear time-varying stochastic systems, and necessary and sufficient conditions for nonlinear H_∞ control are derived (Theorems 4.1 and 4.2). It is shown that a finite horizon H_∞ control can be converted into solving a stochastic game problem, while (u_T^*, d_T^*) is in fact the saddle point of this game. The relation between finite and infinite time HJEs has also been clarified, specifically, under some precise conditions; the solution of the finite time HJE converges to that of the infinite time HJE. Section 5 ends this paper with some remarks.

For convenience, we adopt the following notation:

$\mathcal{S}_n$: the set of all real $n \times n$ symmetric matrices;

A': the transpose of the corresponding matrix A;

$A \geq 0 (A > 0)$: the positive semidefinite (positive definite) matrix A;

I: the identity matrix;

$C_0^2(\{t > 0\} \times U)$: the class of functions $V(t, x)$ twice continuously differential with respect to $x \in U$ and once continuously differential with respect to $t > 0$ except possibly at the point $x = 0$;

$C^2(U)$: the class of functions $V(x)$ twice continuously differential with respect to $x \in U$.

2. Dissipative stochastic systems. Consider the following nonlinear stochastic controlled system governed by the Itô differential equation (the time variable t is

suppressed):

$$\begin{cases} dx(t) = (f(x) + g(x)u)dt + (h(x) + l(x)u)dW, \\ \qquad f(0) = 0, h(0) = 0, \\ z = m(x),\ m(0) = 0. \end{cases} \tag{1}$$

In the above, f, g, h, l, and m are uniformly continuous and Lipschitz satisfying a linear growth condition, which guarantees that (1) has a unique strong solution [32]. $x(t) \in \mathcal{R}^n$ is called the system state, and $z(t) \in \mathcal{R}^{n_z}$ is the regulated output. $W(t)$ is the one-dimensional standard Wiener process defined on the complete probability space $(\Omega, \mathcal{F}, \mathcal{P})$, with the natural filter $\mathcal{F}_t$ generated by $W(\cdot)$ up to time t. $u(t) \in \mathcal{R}^{n_u}$ is the control input, which is an adapted process with respect to $\{\mathcal{F}_t\}_{t\geq 0}$, and causes system (1) to have a unique strong solution under the above conditions. The dissipative dynamic system theory was founded in [10] and has become an important tool in studying the stability and stabilization of nonlinear systems; see [6], [7], and [9]. In recent years, the theory of [10] has been extended to stochastic systems in various different ways by many researchers [27], [29], [37]. In the following, we develop a parallel dissipative theory for stochastic systems which is slightly different from the previous references and is self-contained.

An admissible control set $\mathcal{L}^2_{\mathcal{F}}([s,T], \mathcal{R}^{n_u})$ consists of all adapted and measurable processes $u(t)$ with respect to $\mathcal{F}_t$, such that

$$\|u\|^2_{\mathcal{L}^2([s,T])} := E\int_s^T \|u(t)\|^2\, dt < \infty, \qquad s \geq 0.$$

In the terminology of [10], a function $w(\cdot,\cdot) : \mathcal{R}^{n_u} \times \mathcal{R}^{n_z} \mapsto \mathcal{R}$ associated with system (1) is called the supply rate on $[s, \infty)$ if it has the following property: for any $u \in \mathcal{L}^2_{\mathcal{F}}([s,T], \mathcal{R}^{n_u})$, $x(s) \in \mathcal{R}^n$, the controlled output $z(t) = m(x(t))$ of (1) is such that

$$E\int_s^T |w(u(t), z(t))|\, dt < \infty \ \forall\ \ T \geq s \geq 0.$$

Definition 2.1. *System* (1) *with supply rate* w *is said to be dissipative on* $[s, \infty)$, $s \geq 0$, *if there exists a nonnegative continuous function* $V : \mathcal{R}^n \mapsto \mathcal{R}^+$, *called the storage function, such that for all* $t \geq s \geq 0$, $x(s) \in \mathcal{R}^n$,

$$EV(x(t)) - V(x(s)) \leq E\int_s^t w(u(\tau), z(\tau))\, d\tau. \tag{2}$$

As in deterministic systems [10], (2) can be called the dissipative inequality.

Proposition 2.1. *If there exists a Lyapunov function* V *defined on* $\mathcal{R}^n$ *(i.e.,* $V \in \mathcal{C}^2(\mathcal{R}^n)$ *and is positive definite) satisfying*

$$\mathcal{L}_u V(x) \leq w(u, z)\ \ \forall (u, z) \in \mathcal{R}^{n_u} \times \mathcal{R}^{n_z},$$

then system (1) *is dissipative with supply rate* w *on* $[s, \infty)$ *for any* $s \geq 0$, *where* $\mathcal{L}_u$ *is the infinitesimal generator of the equation*

$$dx = (f(x) + g(x)u)dt + (h(x) + l(x)u)dW. \tag{3}$$

Proof. By Itô's formula, for any $t \geq s \geq 0$, $x(s) \in \mathcal{R}^n$,

$$V(x(t)) - V(x(s)) = \int_s^t \mathcal{L}_u V(x)\, dt + \int_s^t \frac{\partial V'(x)}{\partial x}(h(x) + l(x)u)\, dW.$$

Calculating the above expectation, we get

$$EV(x(t)) - V(x(s)) = E\int_s^t \mathcal{L}_u V(x)\,dt \le E\int_s^t w(u(t), z(t))\,dt.$$

This ends the proof. □

DEFINITION 2.2. *An available storage with supply rate w on $[s,\infty)$, $s \ge 0$, is defined by*

$$\begin{aligned} V_{a,s}(x) &= -\inf_{u\in\mathcal{L}^2_{\mathcal{F}}([s,t];\mathcal{R}^{n_u}),t\ge s,x(s)=x\in\mathcal{R}^n} E\int_s^t w(u(s),z(s))\,ds \\ &= \sup_{u\in\mathcal{L}^2_{\mathcal{F}}([s,t];\mathcal{R}^{n_u}),t\ge s,x(s)=x\in\mathcal{R}^n} -E\int_s^t w(u(s),z(s))\,ds. \end{aligned} \tag{4}$$

A stochastic version of Proposition 2.3 of [9] is as follows.

PROPOSITION 2.2. *If system* (1) *with supply rate w is dissipative on $[s,\infty)$, $s \ge 0$, then the available storage $V_{a,s}(x)$ is finite for each $x \in \mathcal{R}^n$. Moreover, for any possible storage function V_s,*

$$0 \le V_{a,s}(x) \le V_s(x) \quad \forall x \in \mathcal{R}^n. \tag{5}$$

$V_{a,s}$ is itself a possible storage function. Conversely, if $V_{a,s}$ is finite for each $x \in \mathcal{R}^n$, then system (1) *is dissipative on $[s,\infty)$.*

Proof. $V_{a,s} \ge 0$ is obvious. Next, by Definition 2.1, if system (1) with supply rate w is dissipative on $[s,\infty)$, then (2) holds for some storage function V_s. So for any $x(s) = x \in \mathcal{R}^n$, $t \ge s \ge 0$,

$$V_s(x) \ge -E\int_s^t w(u(\tau),z(\tau))\,d\tau + EV_s(x(t)) \ge -E\int_s^t w(u(\tau),z(\tau))\,d\tau,$$

which yields

$$\begin{aligned} V_s(x) &\ge \sup_{t\ge s,u\in\mathcal{L}^2_{\mathcal{F}}([s,t],\mathcal{R}^{n_u}),x(s)=x} -E\int_s^t w(u(\tau),z(\tau))\,d\tau \\ &= -\inf_{t\ge s,u\in\mathcal{L}^2_{\mathcal{F}}([s,t],\mathcal{R}^{n_u}),x(s)=x} E\int_s^t w(u(\tau),z(\tau))\,d\tau = V_{a,s}(x). \end{aligned}$$

Therefore, $V_{a,s}$ is finite and (5) holds. The rest can be done along the lines of [10] mainly using the relation

$$V_{a,s}(x) + E\int_s^t w(u(\tau),z(\tau))\,d\tau \ge EV_{a,s}(x(t)). \tag{6}$$ □

The following general theorem with $w(u,z) = z'Qz + 2z'Su + u'Ru$ will be used in section 3, where $Q \in \mathcal{S}_{n_z}$, $S \in \mathcal{R}^{n_z\times n_u}$, and $R \in \mathcal{S}_{n_u}$ are constant matrices. The following assumption is necessary.

Assumption 2.1. The storage function of (4), if it exists, belongs to $C^2(\mathcal{R}^n)$.

THEOREM 2.1. *A necessary and sufficient condition for system* (1) *to be dissipative on $[s,\infty)$ with respect to a supply rate $w(\cdot,\cdot)$ is that there exists $V_s \in C^2(\mathcal{R}^n)$:*

$\mathcal{R}^n \mapsto \mathcal{R}^+, V_s(0) = 0, \tilde{l} : \mathcal{R}^n \mapsto \mathcal{R}^q$, *and* $\tilde{w} : \mathcal{R}^n \mapsto \mathcal{R}^{q\times n_u}$ *for some integer* $q > 0$, *such that*

$$m'Qm - \frac{\partial V_s'}{\partial x} f - \frac{1}{2} h' \frac{\partial^2 V_s}{\partial x^2} h = \tilde{l}'\tilde{l}, \tag{7}$$

$$R - \frac{1}{2} l' \frac{\partial^2 V_s}{\partial x^2} l = \tilde{w}'\tilde{w}, \tag{8}$$

$$2S'm - g' \frac{\partial V_s}{\partial x} - l' \frac{\partial^2 V_s}{\partial x^2} h = 2\tilde{w}'\tilde{l}. \tag{9}$$

Proof. If system (1) is dissipative on $[s, \infty)$ with respect to a supply rate $w(\cdot, \cdot)$, by Proposition 2.2, $V_{a,s}$ is a possible storage function, which satisfies (6). By (6) with any $x(s) = x \in \mathcal{R}^n$ and Assumption 2.1, we have

$$-\frac{EV_{a,s}(x(t)) - V_{a,s}(x)}{t-s} + \frac{E\int_s^t w(u(\tau), z(\tau))\, d\tau}{t-s} \geq 0, \qquad t > s.$$

Let $t \downarrow s$ in the above and note that (applying Itô's formula)

$$EV_{a,s}(x(t)) = V_{a,s}(x) + E\int_s^t \left(\frac{\partial V_{a,s}'}{\partial x}(f + gu) + \frac{1}{2}(h + lu)' \frac{\partial^2 V_{a,s}}{\partial x^2}(h + lu) \right) d\tau,$$

it follows that

$$\begin{aligned} J(x,u) := {} & m'Qm + 2m'Su + u'Ru - \frac{\partial V_{a,s}'}{\partial x}(f + gu) \\ & -\frac{1}{2}(h + lu)' \frac{\partial^2 V_{a,s}}{\partial x^2}(h + lu) \geq 0 \end{aligned} \tag{10}$$

for all x and u. Obviously, by the fact that the right-hand side of (10) is quadratic in u, there exist $\tilde{l} : \mathcal{R}^n \mapsto \mathcal{R}^q$, and $\tilde{w} : \mathcal{R}^n \mapsto \mathcal{R}^{q\times n_u}$ (not necessarily unique), such that

$$J(x,u) = (\tilde{l}(x) + \tilde{w}(x)u)'(\tilde{l}(x) + \tilde{w}(x)u).$$

By comparing the coefficients of the same powers of u, we deduce (7), (8), and (9). The inverse can be very easily shown by noting that for any $x(s) = x \in \mathcal{R}^n$, we have

$$\begin{aligned} E\int_s^t w(u(\tau), z(\tau))\, d\tau = {} & E\int_s^t (\tilde{l}(x) + \tilde{w}(x)u)'(\tilde{l}(x) + \tilde{w}(x)u)\, d\tau \\ & + EV_s(x(t)) - V_s(x) \geq EV_s(x(t)) - V_s(x). \end{aligned}$$

Theorem 2.1 is complete. □

Theorem 2.1 will have important applications in stochastic stabilization, which will be further studied in our future work. In this paper, we mainly study the dissipative systems with $w(u,z) = \gamma^2 u'u - z'z$ called finite gain systems. When $w(u,z) = u'z$, it is called a passive system [9], which is very useful in the study of nonlinear stochastic stability. If (2) is replaced by

$$EV(x(t)) - V(x(s)) = E\int_s^t w(u(s), z(s))\, ds \quad \forall (u,z) \in \mathcal{R}^{n_u} \times \mathcal{R}^{n_z}, \tag{11}$$

system (1) is said to be lossless.

Remark 2.1. For stochastic system $dx = m(x,u)dt+\sigma(x)dW$, a more general definition for stochastic dissipativeness can be found in [29, Defs. 4.1 and 4.2]. However, Definition 2.1 above is sufficient for our purposes. In particular, when $w(u,z) = u'z$, by using the well-known Dynkin's formula, it can be seen that Definition 2.1 extends Definition 4.1 of [27] for stochastic passive systems.

Remark 2.2. If in (2) and (4), $s = 0$ and t is any bounded stopping time, then [37] gave Definitions 2.1 and 2.2 for the following general nonlinear stochastic system:

$$dx = f(x,u)dt + g(x,u)dW, \qquad x(0) = x \in \mathcal{R}^n.$$

Here, we take the terminal time t to be any fixed scalar only for technical reasons.

3. Infinite horizon H_∞ control. To treat the infinite horizon H_∞ control problem, we need the following definitions of stochastic stability and observability.

3.1. Stochastic stability and observability.

DEFINITION 3.1 (see [1]). *Consider the stochastic unforced system*

$$dx = f(x)\,dt + h(x)\,dW, \quad x(0) = x_0 \in \mathcal{R}^n, \quad f(0) = h(0) = 0. \tag{12}$$

(a) $x \equiv 0$ *of* (12) *is said to be stable in probability if for any* $\epsilon > 0$,

$$\lim_{x_0 \to 0} P\left(\sup_{t\geq 0} \|x(t)\| > \epsilon\right) = 0. \tag{13}$$

(b) $x \equiv 0$ *of* (12) *is said to be locally asymptotically stable in probability if* (13) *holds and*

$$\lim_{x_0 \to 0} P\left(\lim_{t\to\infty} x(t) = 0\right) = 1.$$

(c) $x \equiv 0$ *of* (12) *is said to be globally asymptotically stable in probability if* (13) *holds and*

$$P\left(\lim_{t\to\infty} x(t) = 0\right) = 1.$$

(d) $x \equiv 0$ *of* (12) *is said to be asymptotically mean square stable if* $\lim_{t\to\infty} E\|x(t)\|^2 = 0$.

(e) $x \equiv 0$ *of* (12) *is said to be exponentially mean square stable if there exist* ϱ *and* $\rho > 0$, *such that* $E\|x(t)\|^2 \leq \rho\|x_0\|^2 \exp(-\varrho t)$.

It is very well known that for the linear time-invariant stochastic systems (LTISS), (d) *is equivalent to* (e) [1]. *The following definition can be thought of an extension of the observability in nonlinear deterministic systems.*

DEFINITION 3.2. *We say that the following system* (*or* $[f(x), l(x)|h(x)]$)

$$dx = f(x)\,dt + l(x)\,dW, \qquad z = h(x), \tag{14}$$

is locally zero-state detectable if there is a neighborhood N *of* 0 *such that for all* $x(0) = x_0 \in N$,

$$z(t) = h(x(t)) = 0 \text{ a.s. } \forall t \geq 0 \Rightarrow P\left(\lim_{t\to\infty} x(t) = 0, x(0) = x_0\right) = 1.$$

If $N = \mathcal{R}^n$, (14) *is called zero-state detectable. Equation* (14) *is locally (resp., globally) zero-state observable if there is a neighborhood* N *of* 0 *such that for all* $x_0 \in N$ *(resp.,* $\mathcal{R}^n$*),* $z(t) \equiv 0$ *implies* $x_0 \equiv 0$.

Obviously, for the LTISS

$$dx = Fx\,dt + Lx\,dW, \qquad z = Hx, \tag{15}$$

the local zero-state observability (detectability) is equivalent to the global zero-state observability (detectability). In particular, when (15) is observable (detectable), we also call $[F, L|H]$ observable (detectable).

The following lemma, used in the next subsection, can be called the stochastic version of LaSalle's invariance principle [8].

LEMMA 3.1. *Assume there exists a Lyapunov function V such that*

$$\mathcal{L}_{u\equiv 0}V(x) \le 0$$

for any $x \in \mathcal{R}^n$; then the solution $x(t)$ of (12) tends in probability to the largest invariant set whose support is contained in the locus $\Upsilon := \{x : \mathcal{L}_{u\equiv 0}V(x) = 0\}$ for any $t \ge 0$.

3.2. Main results. Consider the nonlinear stochastic system

$$\begin{cases} dx = (f(x) + g(x)u + k(x)d)dt + (h(x) + l(x)d)dW,\ f(0) = 0,\ h(0) = 0, \\ z = \begin{bmatrix} m(x) \\ u \end{bmatrix}, m(0) = 0, \end{cases} \tag{16}$$

where $d(t)$ stands for the exogenous disturbance, which is an adapted process with respect to $\mathcal{F}_t$. Under very mild conditions, (16) has a unique strong solution $x(t)$ or, for clarity, $x(t, u, d, x(t_0), t_0)$ [32] on any finite interval $[t_0, T]$ with initial state $x(t_0) \in \mathcal{R}^n$. Let $\mathcal{L}^2_{\mathcal{F}}(\mathcal{R}_+, \mathcal{R}^{n_y})$ denote a space composed of all nonanticipative stochastic processes $y(t)_{t\ge 0}$ with respect to $\mathcal{F}_t$, such that

$$\|y\|_{\mathcal{L}^2(\mathcal{R}_+)} := \left(E\int_0^\infty \|y(t)\|^2\,dt\right)^{1/2} < \infty.$$

Obviously, $\mathcal{L}^2_{\mathcal{F}}(\mathcal{R}_+, \mathcal{R}^{n_y})$ is a Hilbert space equipped with the inner product

$$\langle y_1, y_2\rangle = E\int_0^\infty y_1'(t)y_2(t)\,dt\ \forall y_1, y_2 \in \mathcal{L}^2_{\mathcal{F}}(\mathcal{R}_+, \mathcal{R}^{n_y}).$$

DEFINITION 3.3 (infinite horizon nonlinear state feedback H_∞ control). *Given $\gamma > 0$, we want to find an admissible control u^*_∞, such that for any $d \ne 0 \in \mathcal{L}^2_{\mathcal{F}}(\mathcal{R}^+, \mathcal{R}^{n_d})$, when $x(0) = 0$, the following inequality holds:*

$$\|z\|_{\mathcal{L}^2(\mathcal{R}_+)} \le \gamma\|d\|_{\mathcal{L}^2(\mathcal{R}_+)}. \tag{17}$$

Equation (17) is equivalent to $\|\tilde{\mathcal{L}}_{zd}\|_\infty \le \gamma$, where the perturbation operator $\tilde{\mathcal{L}}_{zd}$ is defined by $\tilde{\mathcal{L}}_{zd} : \mathcal{L}^2_{\mathcal{F}}(\mathcal{R}_+, \mathcal{R}^{n_d}) \mapsto \mathcal{L}^2_{\mathcal{F}}(\mathcal{R}_+, \mathcal{R}^{n_z})$ as

$$\tilde{\mathcal{L}}_{zd}(d) = z(x(t, u^*_\infty, d, 0, 0)), \quad t \ge 0, \quad d \in \mathcal{L}^2_{\mathcal{F}}(\mathcal{R}_+, \mathcal{R}^{n_d}),$$

$$\begin{aligned} \|\tilde{\mathcal{L}}_{zd}\|_\infty &= \sup_{d\in\mathcal{L}^2_{\mathcal{F}}(\mathcal{R}_+,\mathcal{R}^{n_d}), d\ne 0, x(0)=0} \frac{\|z\|_{\mathcal{L}^2(\mathcal{R}_+)}}{\|d\|_{\mathcal{L}^2(\mathcal{R}_+)}} \\ &= \sup_{d\in\mathcal{L}^2_{\mathcal{F}}(\mathcal{R}_+,\mathcal{R}^{n_d}), d\ne 0} \frac{\{E\int_0^\infty (\|m(x(t, u^*_\infty, 0, 0))\|^2 + \|u^*_\infty\|^2)\,dt\}^{1/2}}{\{E\int_0^\infty \|d\|^2\,dt\}^{1/2}}. \end{aligned} \tag{18}$$

Compared with the definition of linear (uncertain) time-invariant stochastic H_∞ [2], we find that the internal mean square stability is required therein. Naturally, in the

nonlinear case, we expect the closed-loop system to be of some internal stability, which will be guaranteed by (17) *together with zero-state observability or zero-state detectability. As pointed out by* [5], *it is easier to consider first an infinite horizon nonlinear state feedback* H_∞ *control as in Definition* 3.3. *More specifically, if we let* $u \equiv 0$, $z = m(x)$, $\hat{\mathcal{L}}_{zd} := m(x(t,0,d,0,0))$, *and take* d *as a control variable, then when* $\|\hat{\mathcal{L}}_{zd}\|_\infty \le \gamma$ *for some* $\gamma > 0$, *the nonlinear system*

$$\begin{cases} dx = (f(x) + k(x)d)dt + (h(x) + l(x)d)dW,\ f(0) = 0,\ h(0) = 0, \\ z = m(x),\ m(0) = 0 \end{cases} \tag{19}$$

is said to be externally stable or $\mathcal{L}^2$ *input-output stable. We refer the reader to* [2] *for the definition of external stability of linear stochastic systems.*

THEOREM 3.1. *Suppose there exists a nonnegative solution* $V \in C^2(\mathcal{R}^n)$ *to the HJE*

$$\begin{cases} \mathcal{H}^1_\infty(V(x)) := \frac{\partial V'}{\partial x} f + \frac{1}{2}(\frac{\partial V'}{\partial x} k + h' \frac{\partial^2 V}{\partial x^2} l)(\gamma^2 I - l' \frac{\partial^2 V}{\partial x^2} l)^{-1}(k' \frac{\partial V}{\partial x} + l' \frac{\partial^2 V}{\partial x^2} h) \\ \quad - \frac{1}{2} \frac{\partial V'}{\partial x} g g' \frac{\partial V}{\partial x} + \frac{1}{2} m' m + \frac{1}{2} h' \frac{\partial^2 V}{\partial x^2} h = 0, \\ \gamma^2 I - l' \frac{\partial^2 V}{\partial x^2} l > 0, V(0) = 0; \end{cases} \tag{20}$$

then

$$u^*_\infty = -g' \frac{\partial V}{\partial x} \tag{21}$$

is an H_∞ *control for system* (16).

Proof. By Itô's formula,

$$\begin{aligned} dV(x) = &\left[\frac{\partial V'}{\partial x}(f + gu + kd) + \frac{1}{2}(h + ld)' \frac{\partial^2 V}{\partial x^2}(h + ld) \right] dt \\ &+ \frac{\partial V'}{\partial x}(h + ld)\, dW(t). \end{aligned} \tag{22}$$

By completing the square together with (20), we have for any $T > 0$,

$$\begin{aligned} EV(x(T)) - V(0) = EV(x(T)) &= E \int_0^T \left[\frac{\partial V'}{\partial x}(f + gu + kd) \right. \\ &\qquad \left. + \frac{1}{2}(h + ld)' \frac{\partial^2 V}{\partial x^2}(h + ld) \right] dt \\ &= \frac{1}{2} E \int_0^T \left(\left\| u + g' \frac{\partial V}{\partial x} \right\|^2 + 2\mathcal{H}^1_\infty(V(x)) \right. \\ &\qquad - \| d - \left(\gamma^2 I - l' \frac{\partial^2 V}{\partial x^2} l \right)^{-1} \left(k' \frac{\partial V}{\partial x} + l' \frac{\partial^2 V}{\partial x^2} h \right) \|^2_{\gamma,l,V} \\ &\qquad \left. - \|z\|^2 + \gamma^2 \|d\|^2 \right) dt \\ &= \frac{1}{2} E \int_0^T \left(\left\| u + g' \frac{\partial V}{\partial x} \right\|^2 - \|z\|^2 + \gamma^2 \|d\|^2 \right. \\ &\qquad \left. - \| d - \left(\gamma^2 I - l' \frac{\partial^2 V}{\partial x^2} l \right)^{-1} \left(k' \frac{\partial V}{\partial x} + l' \frac{\partial^2 V}{\partial x^2} h \right) \|^2_{\gamma,l,V} \right) dt, \end{aligned} \tag{23}$$

where $\|Z(x)\|^2_{\gamma,l,V} := Z'(x)(\gamma^2 I - l'\frac{\partial^2 V}{\partial x^2} l)Z(x)$. Obviously, when $u = u^*_\infty$, (23) leads to

$$E\int_0^T \|z\|^2\,dt = -E\int_0^T \left\| d - \left(\gamma^2 I - l'\frac{\partial^2 V}{\partial x^2} l\right)^{-1}\left(k'\frac{\partial V}{\partial x} + l'\frac{\partial^2 V}{\partial x^2} h\right)\right\|^2_{\gamma,l,V} dt$$
$$-2EV(x(T)) + 2V(0) + \gamma^2 E\int_0^T \|d\|^2\,dt$$
$$\le \gamma^2 E\int_0^T \|d\|^2\,dt. \tag{24}$$

Let $T \to \infty$ in (24); then $\|\tilde{\mathcal{L}}_{zd}\|_\infty \le \gamma$ follows because of $V \ge 0$ and $V(0) = 0$. This ends the proof of Theorem 3.1. □

Remark 3.1. From the proof of Theorem 3.1, it can be seen that we have in fact obtained the identity

$$\mathcal{L}_{u,d}V(x) = \frac{\partial V'}{\partial x}(f + gu + kd) + \frac{1}{2}(h + ld)'\frac{\partial^2 V}{\partial x^2}(h + ld)$$
$$= \frac{1}{2}(\|u - u^*_\infty\|^2 - \|d - d^*_\infty\|^2_{\gamma,l,V} + 2\mathcal{H}^1_\infty(V(x)) - \|z\|^2 + \gamma^2\|d\|^2), \tag{25}$$

which will be used throughout this paper, where $\mathcal{L}_{u,d}$ is the infinitesimal generator of

$$dx = (f(x) + g(x)u + k(x)d)dt + (h(x) + l(x)d)dW$$

and

$$d^*_\infty = \left(\gamma^2 I - l'\frac{\partial^2 V}{\partial x^2} l\right)^{-1}\left(k'\frac{\partial V}{\partial x} + l'\frac{\partial^2 V}{\partial x^2} h\right).$$

We can also see that Theorem 3.1 still holds if HJE (20) is replaced by the Hamilton–Jacobi inequality

$$\mathcal{H}^1_\infty(V(x)) \le 0,\ \gamma^2 I - l'\frac{\partial^2 V}{\partial x^2} l > 0, \qquad V(0) = 0.$$

Remark 3.2. From inequality (24), it immediately follows that for any $d \in \mathcal{L}^2_{\mathcal{F}}(\mathcal{R}^+, \mathcal{R}^{n_d})$, we have $z \in \mathcal{L}^2_{\mathcal{F}}(\mathcal{R}^+, \mathcal{R}^{n_z})$, $u^*_\infty \in \mathcal{L}^2_{\mathcal{F}}(\mathcal{R}^+, \mathcal{R}^{n_u})$. However, we cannot assert $d^*_\infty \in \mathcal{L}^2_{\mathcal{F}}(\mathcal{R}^+, \mathcal{R}^{n_d})$.

The following result generalizes Corollary 17 of [5] to the stochastic case.

COROLLARY 3.1. *Under the condition of Theorem* 3.1, *if* $[f, h|m]$ *is zero-state observable, then any solution to HJE* (20) *satisfying* $V(x) > 0$ *for* $x \ne 0$, *and the closed-loop system* (*with* $d \equiv 0$)

$$dx = (f(x) + g(x)u^*_\infty)dt + h(x)dW \tag{26}$$

is locally asymptotically stable in probability. If V *is also proper (i.e., for each* $a > 0$, $V^{-1}([0, a])$ *is compact), then it is globally asymptotically stable in probability. Moreover,* $\lim_{t\to\infty} EV(x(t)) = 0$.

Proof. By (25), we have

$$\mathcal{L}_{u=u^*_\infty,d=0}V(x) = -\frac{1}{2}(\|d^*_\infty\|^2_{\gamma,l,V} + \|m(x)\|^2 + \|u^*_\infty\|^2)$$
$$\le -\frac{1}{2}\begin{bmatrix} m'(x) & u^{*\prime}_\infty \end{bmatrix}\begin{bmatrix} m(x) \\ u^*_\infty \end{bmatrix}. \tag{27}$$

If $V(x)$ is not strictly positive definite in Lyapunov's sense, then there exists $x_0 \neq 0$, such that $V(x_0) = 0$. Integrating from zero to T, and then taking expectation on both sides of (27), it follows that

$$0 \leq EV(x(T)) = -\frac{1}{2}E\int_0^T (\|m\|^2 + \|u_\infty^*\|^2)\, dt \leq 0. \tag{28}$$

Equation (28) concludes that $z(t)|_{u=u_\infty^*} \equiv 0, t \in [0,T]$, for any $T > 0$. From the zero-state observability of $[f, h|m]$, it is easy to prove the zero-state observability of $[f + gu_\infty^*, h|[m'\ u^{*\prime}{}_\infty]']$. According to the definition of zero-state observability, we must have $x(t) \equiv 0$ from $z(t)|_{u=u_\infty^*, d=0} \equiv 0$ a.s., which contradicts $x_0 \neq 0$. $V > 0$ is proved.

In addition, by the above analysis, we have

$$\Upsilon = \{x : \mathcal{L}_{u=u_\infty^*, d=0} V(x) = 0\} \subset \{x : m(x) = 0\} = \{0\}.$$

Hence, the asymptotic stability is proved by use of Lemma 3.1.

Finally, to show $\lim_{t\to\infty} EV(x(t)) = 0$, we apply Itô's formula to system (26); then for any $t > s > 0$,

$$\begin{aligned} V(x(t)) &= V(x(s)) + \int_s^t \mathcal{L}_{u=u_\infty^*, d=0} V(x(\tau))\, d\tau + \int_s^t h' \frac{\partial V}{\partial x}\, dW(\tau) \\ &= V(x(s)) - \frac{1}{2}\int_s^t (\|d_\infty^*\|_{\gamma,l,V}^2 + \|m(x)\|^2 + \|u_\infty^*\|^2)\, dt + \int_s^t h' \frac{\partial V}{\partial x}\, dW(\tau). \end{aligned}$$

So

$$\begin{aligned} E[V(x(t))|\mathcal{F}_s] &= E\left[V(x(s))|\mathcal{F}_s\right] - \frac{1}{2}E[\int_s^t (\|d_\infty^*\|_{\gamma,l,V}^2 + \|m(x)\|^2 + \|u_\infty^*\|^2)\, dt|\mathcal{F}_s\right] \\ &\quad + E\left[\int_s^t h' \frac{\partial V}{\partial x}\, dW(\tau)|\mathcal{F}_s\right] \\ &= V(x(s)) - \frac{1}{2}E\left[\int_s^t (\|d_\infty^*\|_{\gamma,l,V}^2 + \|m(x)\|^2 + \|u_\infty^*\|^2)\, dt|\mathcal{F}_s\right] \\ &\leq V(x(s)), \end{aligned}$$

which shows that $\{V(x(t)), \mathcal{F}_t\}$ is a nonnegative supermartingale. By Doob's convergence theorem and asymptotic stability, $V(x(\infty)) = \lim_{t\to\infty} V(x(t)) = 0$ a.s. Moreover, $\lim_{t\to\infty} EV(x(t)) = EV(x(\infty)) = 0$. The proof of this corollary is complete. □

Remark 3.3. By analogous discussions as in Corollary 3.1, if we replace zero-state observability with zero-state detectability, then Corollary 3.1 still holds except for $V > 0$.

We attempt to give an inverse result of Theorem 3.1; however, there remain some technical problems that cannot be overcome at present. Despite all that, we believe the following lemma, which can be called "nonlinear SBRL," will contribute to the inverse result of Theorem 3.1.

LEMMA 3.2. *For system* (19), $\gamma > 0$, *if there exists a nonnegative solution* $V \in C^2(\mathcal{R}^n) : \mathcal{R}^n \mapsto \mathcal{R}^+$ *to the HJE*

$$\begin{cases} \frac{\partial V'}{\partial x} f + \frac{1}{2}(\frac{\partial V'}{\partial x} k + h' \frac{\partial^2 V}{\partial x^2} l)(\gamma^2 I - l' \frac{\partial^2 V}{\partial x^2} l)^{-1} \\ \qquad \cdot (k' \frac{\partial V}{\partial x} + l' \frac{\partial^2 V}{\partial x^2} h) + \frac{1}{2} m'm + \frac{1}{2} h' \frac{\partial^2 V}{\partial x^2} h = 0, \\ \gamma^2 I - l' \frac{\partial^2 V}{\partial x^2} l > 0 \ \forall x \in \mathcal{R}^n, V(0) = 0, \end{cases} \tag{29}$$

then $\|\hat{\mathcal{L}}_{zd}\|_\infty \leq \gamma$ *for* $x(0) = 0$. *Conversely,* (i) *if there exists a positive definite function* $q : x \in \mathcal{R}^n \mapsto \mathcal{R}^+$, $q(0) = 0$, *such that for all* $x(0) = x \in \mathcal{R}^n, d \in \mathcal{L}^2_{\mathcal{F}}(\mathcal{R}_+, \mathcal{R}^{n_d})$,

$$\|z\|^2_{\mathcal{L}^2(\mathcal{R}_+)} \leq \gamma^2 \|d\|^2_{\mathcal{L}^2(\mathcal{R}_+)} + q(x), \qquad d \neq 0. \tag{30}$$

(ii) *The storage function* $V_{a,0} \in C^2(\mathcal{R}^n)$ *exists with* $\gamma^2 I - l' \frac{\partial^2 V_{a,0}}{\partial x^2} l > 0$ *for all* $x \in \mathcal{R}^n$, *where*

$$V_{a,0}(x) = - \inf_{d \in \mathcal{L}^2_{\mathcal{F}}([0,T];\mathcal{R}^{n_d}), T \geq 0, x(0) = x \in \mathcal{R}^n} E \int_0^T w(d, z)\, dt$$

with $w(d, z) = \frac{1}{2}\gamma^2 \|d\|^2 - \frac{1}{2}\|z\|^2$. *Then* $V_{a,0}$ *solves HJE* (29). *Moreover, for any solution* V *of* (29),

$$V \geq V_{a,0} \geq 0, V_{a,0}(0) = 0. \tag{31}$$

Proof. The first part is an immediate corollary of Theorem 3.1 ($g \equiv 0, u \equiv 0$). As to the inverse result, we first note that (30) concludes, for any $T \geq 0$, that

$$\|z\|^2_{\mathcal{L}^2([0,T])} \leq \gamma^2 \|d\|^2_{\mathcal{L}^2([0,T])} + q(x) \forall d \in \mathcal{L}^2_{\mathcal{F}}([0, T], \mathcal{R}^{n_d}).$$

Actually, for any $d \in \mathcal{L}^2_{\mathcal{F}}([0, T], \mathcal{R}^{n_d})$, if we let

$$\hat{d}(t) = \begin{cases} d(t), \ t \in [0, T], \\ 0, \ t \in (T, \infty), \end{cases}$$

then $\hat{d} \in \mathcal{L}^2_{\mathcal{F}}(\mathcal{R}_+, \mathcal{R}^{n_d})$. By (30),

$$\begin{aligned} \|z\|^2_{\mathcal{L}^2([0,T])} &\leq \|z\|^2_{\mathcal{L}^2(\mathcal{R}_+)} = \|z\|^2_{\mathcal{L}^2([0,T])} + \|z\|^2_{\mathcal{L}^2((T,\infty))} \\ &\leq \gamma^2 \|\hat{d}\|^2_{\mathcal{L}^2(\mathcal{R}_+)} + q(x) \\ &= \gamma^2 \|d\|^2_{\mathcal{L}^2([0,T])} + q(x). \end{aligned} \tag{32}$$

So

$$0 \leq V_{a,0}(x) \leq \frac{1}{2} q(x), \qquad V_{a,0}(0) = 0.$$

Take $R = \frac{1}{2}\gamma^2 I, S = 0, Q = -\frac{1}{2} I, V_s = V_{a,0}$, and

$$\tilde{w} = \frac{\sqrt{2}}{2} \left(\gamma^2 I - l' \frac{\partial^2 V_{a,0}}{\partial x^2} l \right)^{1/2}.$$

HJE (29) is derived from Theorem 2.1. It is easy to show that any solution V of (29) is a possible storage function with supply rate w. Therefore, (31) is followed from Proposition 2.2. ∎

For the LTISS

$$\begin{cases} dx = (Ax + Bu + Kd)dt + (Cx + Dd)dW, \\ z = \begin{bmatrix} Mx \\ u \end{bmatrix}. \end{cases} \tag{33}$$

Take $V(x) = \frac{1}{2}x'Px$; then Theorem 3.1 and Corollary 3.1 lead to the following corollary.

COROLLARY 3.2. *Suppose there exists a solution $P \geq 0$ to the generalized algebraic Riccati equation (GARE)*

$$\begin{cases} PA + A'P + C'PC + (PK + C'PD)(\gamma^2 I - D'PD)^{-1} \\ \qquad \cdot (K'P + D'PC) - PBB'P + M'M = 0, \\ \gamma^2 I - D'PD > 0 \end{cases} \tag{34}$$

for some $\gamma > 0$, then $u_\infty^*(x) = \tilde{k}x = -B'Px$ is an H_∞ control, which makes the closed-loop system satisfy $\|\tilde{\mathcal{L}}_{zd}\|_\infty \leq \gamma$. Additionally, if $[A, C|M]$ is observable, then (i) *$P > 0$;* (ii) *system*

$$dx = (A - BB'P)x\,dt + CxdW$$

is asymptotically mean square stable.

Proof. The first part is an immediate corollary of Theorem 3.1. As to the second part, (i) and (ii) are concluded from Corollary 3.1. □

Example 3.1. Consider the following nonlinear stochastic system with state-dependent noise:

$$dx = \left(\begin{bmatrix} x_1^3 - 2x_1 - 4x_2 \\ x_2^3 - 2x_2 \end{bmatrix} + \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} u + \begin{bmatrix} 1 \\ 1 \end{bmatrix} d \right) dt + \begin{bmatrix} x_2^2 \\ x_1 x_2 \end{bmatrix} dW$$
$$z = \begin{bmatrix} 2(x_1 + x_2) \\ u(t) \end{bmatrix}.$$

Assume the disturbance attenuation $\gamma = 1$ for the H_∞ control designed for the above nonlinear stochastic system. Then, by Theorem 3.1 and Remark 3.1, we need to solve the following Hamilton–Jacobi inequality for H_∞ control:

$$\begin{aligned} \mathcal{H}_\infty^1(V(x)) = {} & \frac{\partial V'}{\partial x} \begin{bmatrix} x_1^3 - 2x_1 - 4x_2 \\ x_2^3 - 2x_2 \end{bmatrix} + \frac{1}{2}\frac{\partial V'}{\partial x} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \end{bmatrix} \frac{\partial V}{\partial x} \\ & - \frac{1}{2}\frac{\partial V'}{\partial x} \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} \begin{bmatrix} 2x_1 & 2x_2 \end{bmatrix} \frac{\partial V}{\partial x} + 2(x_1 + x_2)^2 \\ & + \frac{1}{2} \begin{bmatrix} x_2^2 & x_1 x_2 \end{bmatrix} \frac{\partial^2 V}{\partial x^2} \begin{bmatrix} x_2^2 \\ x_1 x_2 \end{bmatrix} \leq 0. \end{aligned}$$

Let us choose the solution as $V(x) = x_1^2 p_1 + x_2^2 p_2$, for $p_1 > 0$, $p_2 > 0$. Then the

Hamilton–Jacobi inequality is given by

$$\begin{aligned}
\mathcal{H}^1_\infty(V(x)) &= \begin{bmatrix} \frac{\partial V}{\partial x_1} & \frac{\partial V}{\partial x_2} \end{bmatrix} \begin{bmatrix} x_1^3 - 2x_1 - 4x_2 \\ x_2^3 - 2x_2 \end{bmatrix} \\
&\quad + \frac{1}{2} \begin{bmatrix} \frac{\partial V}{\partial x_1} & \frac{\partial V}{\partial x_2} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \frac{\partial V}{\partial x_1} \\ \frac{\partial V}{\partial x_2} \end{bmatrix} \\
&\quad - \frac{1}{2} \begin{bmatrix} \frac{\partial V}{\partial x_1} & \frac{\partial V}{\partial x_2} \end{bmatrix} \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} \begin{bmatrix} 2x_1 & 2x_2 \end{bmatrix} \begin{bmatrix} \frac{\partial V}{\partial x_1} \\ \frac{\partial V}{\partial x_2} \end{bmatrix} + 2(x_1 + x_2)^2 \\
&\quad + \frac{1}{2} \begin{bmatrix} x_2^2 & x_1 x_2 \end{bmatrix} \begin{bmatrix} \frac{\partial^2 V}{\partial x_1^2} & \frac{\partial^2 V}{\partial x_1 x_2} \\ \frac{\partial^2 V}{\partial x_2 x_1} & \frac{\partial^2 V}{\partial x_2^2} \end{bmatrix} \begin{bmatrix} x_2^2 \\ x_1 x_2 \end{bmatrix}, \\
\mathcal{H}^1_\infty(V(x)) &= \begin{bmatrix} 2x_1 p_1 & 2x_2 p_2 \end{bmatrix} \begin{bmatrix} x_1^3 - 2x_1 - 4x_2 \\ x_2^3 - 2x_2 \end{bmatrix} \\
&\quad + \frac{1}{2} \begin{bmatrix} 2x_1 p_1 & 2x_2 p_2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 2x_1 p_1 \\ 2x_2 p_2 \end{bmatrix} \\
&\quad - \frac{1}{2} \begin{bmatrix} 2x_1 p_1 & 2x_2 p_2 \end{bmatrix} \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} \begin{bmatrix} 2x_1 & 2x_2 \end{bmatrix} \begin{bmatrix} 2x_1 p_1 \\ 2x_2 p_2 \end{bmatrix} + 2(x_1 + x_2)^2 \\
&\quad + \frac{1}{2} \begin{bmatrix} x_2^2 & x_1 x_2 \end{bmatrix} \begin{bmatrix} 2p_1 & 0 \\ 0 & 2p_2 \end{bmatrix} \begin{bmatrix} x_2^2 \\ x_1 x_2 \end{bmatrix} \\
&= 2x_1^4 p_1 - 4x_1^2 p_1 - 8x_1 x_2 p_1 + 2x_2^4 p_2 - 4x_2^2 p_2 \\
&\quad + 2[x_1^2 p_1^2 + 2x_1 x_2 p_1 p_2 + x_2^2 p_2^2] \\
&\quad - 8[x_1^4 p_1^2 + 2x_1^2 x_2^2 p_1 p_2 + x_2^4 p_2^2] + 2[x_1^2 + x_2^2 + 2x_1 x_2] + [x_2^4 p_1 + x_1^2 x_2^2 p_2].
\end{aligned}$$

If we let $p_1 = 1$, $p_2 = 1$, i.e. $V(x) = x_1^2 + x_2^2$, then

$$\begin{aligned}
\mathcal{H}^1_\infty(V(x)) &= 2x_1^4 - 4x_1^2 - 8x_1 x_2 + 2x_2^4 - 4x_2^2 + 2[x_1^2 + 2x_1 x_2 + x_2^2] \\
&\quad - 8[x_1^4 + 2x_1^2 x_2^2 + x_2^4] \\
&\quad + 2[x_1^2 + x_2^2 + 2x_1 x_2] + [x_2^4 + x_1^2 x_2^2] \\
&= 2x_1^4 - 4x_1^2 - 8x_1 x_2 + 2x_2^4 - 4x_2^2 + 4[x_1^2 + 2x_1 x_2 + x_2^2] \\
&\quad - 8[x_1^4 + 2x_1^2 x_2^2 + x_2^4] + [x_2^4 + x_1^2 x_2^2] \\
&= -6x_1^4 - 5x_2^4 - 15x_1^2 x_2^2 \le 0;
\end{aligned}$$

i.e., if we choose $u^*_\infty(x) = -g' \frac{\partial V}{\partial x} = -[2x_1 \quad 2x_2] \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} = -4x_1^2 - 4x_2^2$, then the H_∞ control is achieved.

To end this section, we add a few remarks on our above results.

Remark 3.4. Although in this paper W is assumed to be one-dimensional, all our results can be extended to the system with multiplicative noise as follows:

$$\begin{cases} dx = (f(x) + g(x)u + k(x)d)dt + \sum_{i=1}^N (h_i(x) + l_i(x)d)dW_i, \\ z = \begin{bmatrix} m(x) \\ u \end{bmatrix}. \end{cases}$$

In this case, HJE (20) becomes

$$\begin{cases} \frac{\partial V'}{\partial x} f + \frac{1}{2}(\frac{\partial V'}{\partial x} k + \sum_{i=1}^N h_i' \frac{\partial^2 V}{\partial x^2} l_i)(\gamma^2 I - \sum_{i=1}^N l_i' \frac{\partial^2 V}{\partial x^2} l_i)^{-1}(k' \frac{\partial V}{\partial x} + \sum_{i=1}^N l_i' \frac{\partial^2 V}{\partial x^2} h_i) \\ \quad -\frac{1}{2} \frac{\partial V'}{\partial x} g g' \frac{\partial V}{\partial x} + \frac{1}{2} m'm + \frac{1}{2} \sum_{i=1}^N h_i' \frac{\partial^2 V}{\partial x^2} h_i = 0, \\ \gamma^2 I - \sum_{i=1}^N l_i' \frac{\partial^2 V}{\partial x^2} l_i > 0, V(0) = 0. \end{cases}$$

Following [33], for the linear part, all our results can also be extended to the system with W being a multi-dimensional Wiener process.

Remark 3.5. In (16), if $l(x)d$ is replaced by $l(x)u$, then by the same discussion as in Theorem 3.1, we can show that Theorem 3.1 still holds if we substitute

$$\begin{cases} \frac{\partial V'}{\partial x} f + \frac{1}{2\gamma^2} \frac{\partial V'}{\partial x} k k' \frac{\partial V}{\partial x} - \frac{1}{2}(\frac{\partial V'}{\partial x} g + h' \frac{\partial^2 V}{\partial x^2} l)(I + l' \frac{\partial^2 V}{\partial x^2} l)^{-1} \\ \qquad \cdot (g' \frac{\partial V}{\partial x} + l' \frac{\partial^2 V}{\partial x^2} h) + \frac{1}{2} m' m + \frac{1}{2} h' \frac{\partial^2 V}{\partial x^2} h = 0, \\ I + l' \frac{\partial^2 V}{\partial x^2} l > 0, V(0) = 0 \end{cases}$$

and

$$u_\infty^*(x) = \tilde{k}(x) = -\left(I + l' \frac{\partial^2 V}{\partial x^2} l\right)^{-1} \left(g' \frac{\partial V}{\partial x} + l' \frac{\partial^2 V}{\partial x^2} h\right)$$

for (20) and (21), respectively.

4. Finite horizon nonlinear H_∞ control. In this section, we study the finite horizon H_∞ control problem. Suppose the system is governed by the following stochastic time-varying equation:

$$\begin{cases} dx = (f(t,x) + g(t,x)u + k(t,x)d)dt + (h(t,x) + l(t,x)d)dW, \\ \qquad f(t,0) = 0, h(t,0) = 0, \\ z(t) = \begin{bmatrix} m(t,x) \\ u \end{bmatrix}, \ m(t,0) = 0, \end{cases} \tag{35}$$

where $d \in \mathcal{L}_{\mathcal{F}}^2([0,T], \mathcal{R}^{n_d})$ represents the exogenous disturbance, and f, g, k, h, and l are vector-valued functions, jointly continuous in all arguments. The so-called finite horizon H_∞ control is to find, if existing, an $u_T^* \in \mathcal{L}_{\mathcal{F}}^2([0,T], \mathcal{R}^{n_u})$, such that for any given $\gamma > 0$, and all $d \neq 0 \in \mathcal{L}_{\mathcal{F}}^2([0,T], \mathcal{R}^{n_d})$, $x(0) = 0$, the closed-loop system satisfies

$$\|z\|_{\mathcal{L}^2([0,T])} \leq \gamma \|d\|_{\mathcal{L}^2([0,T])}. \tag{36}$$

Similar to the definition of $\tilde{\mathcal{L}}_{zd}$, a perturbation operator $\mathcal{L}_{zd}^T : \mathcal{L}_{\mathcal{F}}^2([0,T], \mathcal{R}^{n_d}) \mapsto \mathcal{L}_{\mathcal{F}}^2([0,T], \mathcal{R}^{n_z})$ can be defined, and (36) is equivalent to

$$\|\tilde{\mathcal{L}}_{zd}^T\|_{[0,T]} = \sup_{d \in \mathcal{L}_{\mathcal{F}}^2([0,T], \mathcal{R}^{n_d}), d \neq 0, x(0)=0} \frac{||z||_{\mathcal{L}^2([0,T])}}{\|d\|_{\mathcal{L}^2([0,T])}} \leq \gamma.$$

In particular, if we let $u \equiv 0$ and take d as a control variable in (35), then when $\|\mathcal{L}^T\|_{[0,T]} := \|\mathcal{L}_{zd}^T\|_{[0,T]} \leq \gamma$ for any given $\gamma > 0$, the system is said to have $\mathcal{L}_2$-gain less than or equal to γ.

In analogy with the proof of Theorem 3.1, the following result is easily obtained.

THEOREM 4.1. *Assume $V_T(t,x) \in C_0^2([0,T] \times \mathcal{R}^n)$ satisfies the HJE*

$$\begin{cases} \mathcal{H}_T^1(t,x) := \frac{\partial V_T}{\partial t} + \frac{\partial V_T'}{\partial x} f + \frac{1}{2}(\frac{\partial V_T'}{\partial x} k + h' \frac{\partial^2 V_T}{\partial x^2} l)(\gamma^2 I - l' \frac{\partial^2 V_T}{\partial x^2} l)^{-1}(k' \frac{\partial V_T}{\partial x} + l' \frac{\partial^2 V_T}{\partial x^2} h) \\ \qquad - \frac{1}{2} \frac{\partial V_T'}{\partial x} g g' \frac{\partial V_T}{\partial x} + \frac{1}{2} m' m + \frac{1}{2} h' \frac{\partial^2 V_T}{\partial x^2} h = 0, \\ \gamma^2 I - l' \frac{\partial^2 V_T}{\partial x^2} l > 0, V_T(T,x) = 0, V_T(t,0) = 0 \ \forall (t,x) \in [0,T] \times \mathcal{R}^n. \end{cases} \tag{37}$$

Then (u_T^*, d_T^*) *is a saddle point for the following stochastic game problem:*

$$\min_{u\in\mathcal{L}_{\mathcal{F}}^2([0,T],\mathcal{R}^{n_u})}\max_{d\in\mathcal{L}_{\mathcal{F}}^2([0,T],\mathcal{R}^{n_d})} E\int_0^T (\|z\|^2-\gamma^2\|d\|^2)\,dt,$$

where u_T^* *and* d_T^* *are defined, respectively, as*

$$u_T^* = -g'\frac{\partial V_T}{\partial x}$$

and

$$d_T^* = \left(\gamma^2 I - l'\frac{\partial^2 V_T}{\partial x^2}l\right)^{-1}\left(k'\frac{\partial V_T}{\partial x}+l'\frac{\partial^2 V_T}{\partial x^2}h\right).$$

Moreover, u_T^* *is an* H_∞ *control for system* (35)*, and* d_T^* *is the corresponding worst case disturbance.*

Below, we point out the relationship on the solutions between finite and infinite horizon HJEs. Let $V(x)$ and $V_T(t,x,Q(x))$ stand for the solutions of (20) and (37) with terminal condition $V_T(T,x)=Q(x)\ge 0$ for all $x\in\mathcal{R}^n$, respectively. A generalized version of Lemma 2.6 of [16] is as follows.

PROPOSITION 4.1. *If* $V(x)\ge Q(x)$ *for all* $x\in\mathcal{R}^n$*, then* $V(x)\ge V_T(t,x,Q(x))\ge V_T(t,x,0)=V_T(t,x)\ge 0$ *for all* $(t,x)\in[0,T]\times\mathcal{R}^n$.

Proof. For any initial time $t\ge 0$ and state $x(t):=x\in\mathcal{R}^n$, one only needs to note the following identities:

$$\begin{aligned}\frac{1}{2}E\int_t^T(\|z\|^2-\gamma^2\|d\|^2)\,dt &= V(x)-EV(x(T))\\ &\quad+\frac{1}{2}E\int_t^T(\|u-u_\infty^*\|^2-\|d-d_\infty^*\|_{\gamma,l,V}^2+2\mathcal{H}_\infty^1(V(x)))\,dt \qquad (38)\end{aligned}$$

and

$$\begin{aligned}\frac{1}{2}E\int_t^T(\|z\|^2-\gamma^2\|d\|^2)\,dt &= V_T(t,x,Q(x))-EV_T(T,x(T),Q(x))\\ &\quad+\frac{1}{2}E\int_t^T(\|u-u_T^*\|^2-\|d-d_T^*\|_{\gamma,l,V_T(t,x,Q(x))}^2+2\mathcal{H}_T^1(t,x))\,dt. \qquad (39)\end{aligned}$$

The rest is similar to the proof of Lemma 2.6 of [16] and is omitted. □

PROPOSITION 4.2. *There is at most one solution to* (37).

Proof. Otherwise, let $V_T^{(1)}(\cdot,\cdot)$ and $V_T^{(2)}(\cdot,\cdot)$ be two solutions of (37). Set

$$J_T(x,u,d,x(t_0),t_0)=\frac{1}{2}E\int_{t_0}^T(\|z\|^2-\gamma^2\|d\|^2)\,dt,$$

$$u_{i,T}^*=-g'\frac{\partial V_T^{(i)}}{\partial x},$$

and

$$d_{i,T}^*=\left(\gamma^2 I-l'\frac{\partial^2 V_T^{(i)}}{\partial x^2}l\right)^{-1}\left(k'\frac{\partial V_T^{(i)}}{\partial x}+l'\frac{\partial^2 V_T^{(i)}}{\partial x^2}h\right),\qquad i=1,2.$$

For any $x(s) = y$, $(s, y) \in [0, T) \times \mathcal{R}^n$, from (39), we have

$$\begin{aligned} J_T(x, u^*_{1,T}, d^*_{1,T}, y, s) = V_T^{(1)}(s, y) &\le J_T(x, u^*_{2,T}, d^*_{1,T}, y, s) \\ &\le J_T(x, u^*_{2,T}, d^*_{2,T}, y, s) = V_T^{(2)}(s, y). \end{aligned}$$

Also, we have $V_T^{(2)}(s, y) \le V_T^{(1)}(s, y)$, so $V_T^{(2)}(s, y) = V_T^{(1)}(s, y)$. □

PROPOSITION 4.3. *$V_T(\cdot, \cdot)$ is monotonically increasing with respect to $T > 0$.*

Proof. For any $0 \le s \le T_0 \le T_1 < \infty$, $x(s) = y \in \mathcal{R}^n$, still by (39), we have

$$\begin{aligned} J_{T_0}(x, u^*_{T_0}, d^*_{T_0}, y, s) = V_{T_0}(s, y) &\le J_{T_0}(x, u^*_{T_1}, d^*_{T_0}, y, s) \\ &\le J_{T_1}(x, u^*_{T_1}, d^*_{T_0}, y, s) \le J_{T_1}(x, u^*_{T_1}, d^*_{T_1}, y, s), \\ &= V_{T_1}(s, y). \end{aligned}$$

This proposition is complete. □

If system (35) is time invariant, and

$$\bar{V}(t, x) := \lim_{T \to \infty} V_T(t, x, Q(x))$$

exists, then $\bar{V}$ depends only on x and is a solution of (20). We refer the reader to the proof of Corollary 2.7 of [16]. In particular, if there exists $(u^*_\infty, d^*_\infty) \in \mathcal{L}^2_{\mathcal{F}}(\mathcal{R}_+, \mathcal{R}^{n_u}) \times \mathcal{L}^2_{\mathcal{F}}(\mathcal{R}_+, \mathcal{R}^{n_d})$, such that $J_\infty(x, u^*_\infty, d^*_\infty, y, s) < \infty$, then by making use of Propositions 4.1, 4.2, and 4.3, $\bar{V}(x)$ exists due to the monotonicity and uniform boundedness of $V_T(t, x)$.

In general, the inverse of Theorem 4.1 is not true; i.e., $\|\mathcal{L}^T_{zd}\|_{[0,T]} \le \gamma$ does not necessarily imply that HJE (37) has a solution. An inverse result will be presented in the following together with some other conditions. To this end, assume $u = \tilde{k}(t, x)$ is an H_∞ control law of (35). We define $\tilde{V}_{T,\tilde{k}}(s, x) : [0, T] \times \mathcal{R}^n \mapsto \mathcal{R}^+$ as

$$\begin{aligned} \tilde{V}_{T,\tilde{k}}(s, x) &= -\frac{1}{2} \inf_{d \in \mathcal{L}^2_{\mathcal{F}}([s,T], \mathcal{R}^{n_d}), u = \tilde{k}, x(s) = x} E \int_s^T (\gamma^2 \|d\|^2 - \|z\|^2)\, dt \\ &= \sup_{d \in \mathcal{L}^2_{\mathcal{F}}([s,T], \mathcal{R}^{n_d}), u = \tilde{k}, x(s) = x} -\frac{1}{2} E \int_s^T (\gamma^2 \|d\|^2 - \|z\|^2)\, dt. \end{aligned}$$

It is easy to test the following properties of $\tilde{V}_{T,\tilde{k}}$: (i) $\tilde{V}_{T,\tilde{k}} \ge 0$; (ii) $\tilde{V}_{T,\tilde{k}}(T, x) = 0$ for all $x \in \mathcal{R}^n$. The following proposition can also be shown in the same way as [9] and [26].

PROPOSITION 4.4. (i) *$\|\mathcal{L}^T_{zd}\|_{[0,T]} \le \gamma$ implies $\tilde{V}_{T,\tilde{k}}(s, 0) = 0$ for all $s \in [0, T]$.*

(ii) *$\tilde{V}_{T,\tilde{k}}$ is finite on $[0, T] \times \mathcal{R}^n$ if and only if there exists a nonnegative function $V(s, x) : [0, T] \times \mathcal{R}^n \mapsto \mathcal{R}^+$ satisfying the following integral dissipation inequality (IDI):*

$$EV(T, x(T)) - V(s, x) \le \frac{1}{2} E \int_s^T (\gamma^2 \|d\|^2 - \|z\|^2)\, dt. \tag{40}$$

Moreover, when $\tilde{V}_{T,\tilde{k}}(s, x)$ is finite, $\tilde{V}_{T,\tilde{k}}$ is itself a solution of (40).

In the literature, such as [5] and [26], to guarantee the finiteness of $\tilde{V}_{T,\tilde{k}}(s, x)$, an essential concept on the system theory called "reachability" is introduced. Checking

the proof of Lemma 2.3 of [26], it can be found that even under the assumption of reachability, it is not necessary to have $\tilde{V}_{T,\tilde{k}}(0,x) < \infty$ for all $x \in \mathcal{R}^n$.

LEMMA 4.1. *If $\tilde{V}_{T,\tilde{k}}(s,x) \in C_0^2([0,T]\times\mathcal{R}^n)$ is finite with $\gamma^2 I - l'\frac{\partial^2 \tilde{V}_{T,\tilde{k}}}{\partial x^2} l > 0$ for some $\gamma > 0$, and $\|\mathcal{L}_{zd}^T\|_{[0,T]} \le \gamma$, then $\tilde{V}_{T,\tilde{k}}$ solves the HJE*

$$\begin{cases} \mathcal{H}(V_{T,\tilde{k}}) := \frac{\partial V_{T,\tilde{k}}}{\partial t} + \frac{\partial V'_{T,\tilde{k}}}{\partial x}(f+g\tilde{k}) + \frac{1}{2}\left(\frac{\partial V'_{T,\tilde{k}}}{\partial x}k + h'\frac{\partial^2 V_{T,\tilde{k}}}{\partial x^2}l\right)\left(\gamma^2 I - l'\frac{\partial^2 V_{T,\tilde{k}}}{\partial x^2}l\right)^{-1} \\ \qquad\qquad \times\left(k'\frac{\partial V_{T,\tilde{k}}}{\partial x} + l'\frac{\partial^2 V_{T,\tilde{k}}}{\partial x^2}h\right) + \frac{1}{2}(m'm + \tilde{k}'\tilde{k}) + \frac{1}{2}h'\frac{\partial^2 V_{T,\tilde{k}}}{\partial x^2}h = 0, \\ \gamma^2 I - l'\frac{\partial^2 V_{T,\tilde{k}}}{\partial x^2}l > 0, V_{T,\tilde{k}}(T,x) = 0, V_{T,\tilde{k}}(t,0) = 0 \ \forall (t,x) \in [0,T]\times\mathcal{R}^n. \end{cases} \tag{41}$$

Proof. We have shown that $\tilde{V}_{T,\tilde{k}}$ satisfies the boundary conditions of (41) above. Now, let $\hat{V} = -\tilde{V}_{T,\tilde{k}}$; then by the dynamic programming principle, $\hat{V}$ solves the following HJE [32]:

$$-\frac{\partial \hat{V}}{\partial t} + \max_{d\in U} H\left(t,x,d,-\frac{\partial \hat{V}}{\partial x},-\frac{\partial^2 \hat{V}}{\partial x^2}\right) = 0, \tag{42}$$

where (U,ρ) is a Polish space, $U \subset \mathcal{R}^{n_d}$, and the generalized Hamiltonian function H is defined as

$$\begin{aligned} H\left(t,x,d,-\frac{\partial \hat{V}}{\partial x},-\frac{\partial^2 \hat{V}}{\partial x^2}\right) &:= -\frac{1}{2}\gamma^2\|d\|^2 + \frac{1}{2}\|z\|^2 \\ &\quad -\frac{\partial \hat{V}'}{\partial x}(f+g\tilde{k}+kd) - \frac{1}{2}(h+ld)'\frac{\partial^2 \hat{V}}{\partial x^2}(h+ld) \\ &= \mathcal{H}(\tilde{V}_{T,\tilde{k}}) - \frac{\partial \tilde{V}_{T,\tilde{k}}}{\partial t} - \frac{1}{2}\|d-\hat{d}_T\|^2_{\gamma,l,\tilde{V}_{T,\tilde{k}}} \end{aligned}$$

with $\hat{d}_T = (\gamma^2 I - l'\frac{\partial^2 \tilde{V}_{T,\tilde{k}}}{\partial x^2}l)^{-1}(k'\frac{\partial \tilde{V}_{T,\tilde{k}}}{\partial x} + l'\frac{\partial^2 \tilde{V}_{T,\tilde{k}}}{\partial x^2}h)$. Obviously,

$$\begin{aligned} \max_{d\in U} H\left(t,x,d,-\frac{\partial \hat{V}}{\partial x},-\frac{\partial^2 \hat{V}}{\partial x^2}\right) &= H\left(t,x,\hat{d}_T,\frac{\partial \tilde{V}_{T,\tilde{k}}}{\partial x},\frac{\partial^2 \tilde{V}_{T,\tilde{k}}}{\partial x^2}\right) \\ &= \mathcal{H}(\tilde{V}_{T,\tilde{k}}) - \frac{\partial \tilde{V}_{T,\tilde{k}}}{\partial t}. \end{aligned}$$

Therefore, (42) is equivalent to $\mathcal{H}(\tilde{V}_{T,\tilde{k}}) = 0$. The proof of this lemma is complete. □

THEOREM 4.2. *If there exists an H_∞ control $u = \tilde{k}(t,x)$ for system (35), such that the conditions of Lemma 4.1 hold, then HJE (37) admits a unique solution.*

Proof. Apply Lemma 4.1 and note identity (39). We have

$$\begin{aligned} E\int_t^T (\|z\|^2 - \gamma^2\|d\|^2)\,dt &= E\int_t^T \left(\left\|\tilde{k} + g'\frac{\partial \tilde{V}_{T,\tilde{k}}}{\partial x}\right\|^2 - \|d-\hat{d}_T\|^2_{\gamma,l,\tilde{V}_{T,\tilde{k}}} + 2\mathcal{H}(\tilde{V}_{T,\tilde{k}})\right)dt \\ &= E\int_t^T \left(\left\|\tilde{k} + g'\frac{\partial \tilde{V}_{T,\tilde{k}}}{\partial x}\right\|^2 - \|d-\hat{d}_T\|^2_{\gamma,l,\tilde{V}_{T,\tilde{k}}}\right)dt. \end{aligned}$$

Obviously, in order to have $\|\mathcal{L}_{zd}^T\|_{[0,T]} \leq \gamma$, we must take $\tilde{k} = -g' \frac{\partial \tilde{V}_{T,\tilde{k}}}{\partial x}$ as an H_∞. Substituting $\tilde{k}$ into (41), (37) is derived. Uniqueness is followed from Proposition 4.2. □

5. Concluding remarks. This paper has discussed the stochastic H_∞ control for nonlinear systems with both state- and disturbance-dependent noise, including finite and infinite horizon cases. It has been shown that both finite and infinite horizon stochastic H_∞ designs are associated with two kinds of HJEs. Some results of nonlinear H_∞ control [5] for deterministic systems are generalized to the stochastic case. There are several interesting problems which merit further study, for instance, the relation between HJEs and invariant manifolds of Hamiltonian vector fields, and the relation between the primal nonlinear stochastic system and its linearization. In addition, when the state is not completely available, we must consider the output feedback H_∞ control, as done in [18] and [19]. However, since the nonlinear stochastic H_∞ filtering has been investigated in [31], it is easier to treat the aforementioned issue using our nonlinear SBRL (Lemma 3.2). Finally, when the control u enters the diffusion term in (16) and (35), which form would the HJE take? This is a very interesting topic, but the results in this paper would not shed much light on this extension. One reason for this is that u and d are no longer separable in the HJE. If the diffusion term depends only on u (and not also on d), then the problem can be tractable.

Acknowledgments. The authors would like to thank the Associate Editor and the two reviewers for their constructive comments that lead to a significant improvement of the paper.

REFERENCES

[1] R. Z. Has'minskii, *Stochastic Stability of Differential Equations*, Sijthoff and Noordhoff, Alphen aan den Rijn–Germantown, MD, 1980.

[2] D. Hinrichsen and A. J. Pritchard, *Stochastic H^∞*, SIAM J. Control Optim., 36 (1998), pp. 1504–1538.

[3] X. Chen and K. Zhou, *Multiobjective H_2/H_∞ control design*, SIAM J. Control Optim., 40 (2001), pp. 628–660.

[4] E. Gershon, D. J. N. Limebeer, U. Shaked, and I. Yaesh, *Robust H_∞ filtering of stationary continuous-time linear systems with stochastic uncertainties*, IEEE Trans. Automat. Control, 46 (2001), pp. 1788–1793.

[5] A. J. Van Der Schaft, *L_2-gain analysis of nonlinear systems and nonlinear state feedback H_∞ control*, IEEE Trans. Automat. Control, 37 (1992), pp. 770–784.

[6] D. J. Hill and P. J. Moylan, *Stability results for nonlinear feedback systems*, Automatica, 13 (1977), pp. 377–382.

[7] D. J. Hill and P. J. Moylan, *The stability of nonlinear dissipative systems*, IEEE Trans. Automat. Control, 21 (1976), pp 708–711.

[8] H. J. Kushner, *Stochastic stability*, in Stability of Stochastic Dynamical Systems, R. Curtain, ed., Lecture Notes in Math. 294, Springer, Berlin, 1972, pp. 97–124.

[9] C. I. Byrnes, A. Isidori, and J. C. Willems, *Passivity, feedback equivalence, and the global stabilization of minimum phase nonlinear systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 1228–1240.

[10] J. C. Willems, *Dissipative dynamic systems Part* I: *General theory*, Arch. Rational Mech., 45 (1972), pp. 321–393.

[11] T. C. Lee, B. S. Chen, and T. S. Lee, *Singular H_∞ control in nonlinear systems: Positive semidefinite storage functions and separation principle*, Int. J. Robust Nonlinear Control, 7 (1997), pp. 881–897.

[12] C. S. Wu and B. S. Chen, *Unified design for H_2, H_∞, and mixed control of spacecraft*, J. Guidance Control Dyn., 22 (1999), pp. 884–896.

[13] C. S. Wu and B. S. Chen, *Adaptive attitude control of spacecraft: Mixed H_2/H_∞ approach*, J. Guidance Control Dyn., 24 (2001), pp. 755–766.

[14] B. S. Chen, C. S. Tseng, and H. J. Uang, *Mixed H_2/H_∞ fuzzy output feedback control design for nonlinear dynamic systems: An LMI approach*, IEEE Trans. Fuzzy Systems, 8 (2000), pp. 249–265.

[15] D. J. N. Limebeer, B. D. O. Anderson, and B. Hendel, *A Nash game approach to mixed H_2/H_∞ control*, IEEE Trans. Automat. Control, 39 (1994), pp. 69–82.

[16] D. J. N. Limebeer, B. D. O. Anderson, P. P. Khargonekar, and M. Green, *A game theoretic approach to H^∞ control for time-varying systems*, SIAM J. Control Optim., 30 (1992), pp. 262–283.

[17] J. C. Doyle, K. Glover, P. P. Khargonekar, and B. Francis, *State-space solutions to standard H_2 and H_∞ control problems*, IEEE Trans. Automat. Control, 34 (1989), pp. 831–847.

[18] J. A. Ball, J. W. Helton, and M. L. Walker, *H^∞ control for nonlinear systems with output feedback*, IEEE Trans. Automat. Control, 38 (1993), pp. 546–559.

[19] A. Isidori and A. Astolfi, *Disturbance attenuation and H_∞-control via measurement feedback in nonlinear systems*, IEEE Trans. Automat. Control, 37 (1992), pp. 1283–1293.

[20] P. P. Khargonekar, K. M. Nagpal, and K. R. Poolla, *H_∞ control with transients*, SIAM J. Control Optim., 29 (1991), pp. 1373–1393.

[21] R. Ravi, K. M. Nagpal, and P. P. Khargonekar, *H^∞ control of linear time-varying systems: A state-space approach*, SIAM J. Control Optim., 29 (1991), pp. 1394–1413.

[22] M. D. S. Aliyu and E. K. Boukas, *H_∞ control for Markovian jump nonlinear systems*, in Proceedings of the 37th IEEE Conference on Decision and Control, Florida, 1998, pp. 766–771.

[23] E. K. Boukas and Z. K. Liu, *Robust H_∞ control of discrete time Markovian jump linear systems with mode-dependent time-delays*, IEEE Trans. Automat. Control, 46 (2001), pp. 1918–1924.

[24] D. P. de Farias, J. C. Geromel, J. B. R. do Val, and O. L. Costa, *Output feedback control of Markovian jump linear systems in continuous time*, IEEE Trans. Automat. Control, 45 (2000), pp. 944–949.

[25] T. Basar and P. Bernhard, *H^∞-Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*, Birkhäuser Boston, Boston, MA, 1995.

[26] W. M. Lu, *H_∞-control of nonlinear time-varying systems with finite time horizon*, Int. J. Control, 64 (1996), pp. 241–262.

[27] P. Florchinger, *A passive system approach to feedback stabilization of nonlinear control stochastic systems*, SIAM J. Control Optim., 37 (1999), pp. 1848–1864.

[28] T. Damm, *State-feedback H^∞-type control of linear systems with time-varying parameter uncertainty*, Linear Algebra Appl., 351/352 (2002), pp. 185–210.

[29] V. Borkar and S. Mitter, *A note on stochastic dissipativeness*, in Directions in Mathematical Systems Theory and Optimization, Lecture Notes in Control Inform. 286, Springer, Berlin, 2003, pp. 41–49.

[30] B. S. Chen and W. Zhang, *Stochastic H_2/H_∞ control with state-dependent noise*, IEEE Trans. Automat. Control, 49 (2004), pp. 45–57.

[31] W. Zhang, B. S. Chen, and C. S. Tseng, *Robust H_∞ filtering for nonlinear stochastic systems*, IEEE Trans. Signal Process., 53 (2005), pp. 589–598.

[32] J. Yong and X. Y. Zhou, *Stochastic Control: Hamiltonian Systems and HJB Equations*, Springer-Verlag, New York, 1999.

[33] V. A. Ugrinovskii and I. R. Petersen, *Absolute stabilization and minimax optimal control of uncertain systems with stochastic uncertainty*, SIAM J. Control Optim., 37 (1999), pp. 1089–1122.

[34] V. A. Ugrinovskii, *Robust H_∞ control in the presence of stochastic uncertainty*, Int. J. Control, 71 (1998), pp. 219–237.

[35] V. Dragan, A. Halanay, and A. Stoica, *The γ-attenuation problem for systems with state dependent noise*, Stochast. Anal. Appl., 17 (1999), pp. 395–404.

[36] V. Dragan and T. Morozan, *Global solutions to game-theoretic Riccati equation of stochastic control*, J. Differential Equations, 138 (1997), pp. 328–350.

[37] U. H. Thygesen, *On dissipation in stochastic systems*, in Proceedings of the American Control Conference, San Diego, CA, IEEE, Los Alamitos, CA, 1999, pp. 1430–1434.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 6, pp. 1992–2005

A PARAMETRIZATION OF SOLUTIONS OF THE DISCRETE-TIME ALGEBRAIC RICCATI EQUATION BASED ON PAIRS OF OPPOSITE UNMIXED SOLUTIONS*

HARALD K. WIMMER†

Abstract. The paper describes the set of solutions of the discrete-time algebraic Riccati equation. It is shown that each solution is a combination of a pair of opposite unmixed solutions. There is a one-to-one correspondence between solutions and invariant subspaces of the closed loop matrix of an unmixed solution. The results of the paper provide an extended counterpart of the parametrization theory of continuous-time algebraic Riccati equations by Willems, Coppel, and Shayman.

Key words. discrete-time algebraic Riccati equation, unmixed solutions, opposite solutions, parametrization by invariant subspaces, shorted operators, Schur complements

AMS subject classifications. 15A24, 93C55, 47A64

DOI. 10.1137/S0363012904441362

1. Introduction.

To a large extent the groundwork for the investigation of algebraic Riccati equations by geometric methods was laid by J. C. Willems. In his study of least squares stationary optimal control [24], he gave a complete description of the set of symmetric solutions of the real continuous-time algebraic Riccati equation (CARE)

$$-F^TX - XF + XGG^TX - Q = 0. \tag{1.1}$$

Assuming that (F, G) is controllable and that the associated Hamiltonian matrix

$$H = \begin{bmatrix} F & -GG^T \\ -Q & -F^T \end{bmatrix}$$

has no eigenvalues on the imaginary axis, Willems showed that (1.1) has a greatest solution X^+ and a least solution X^-. The eigenvalues of the corresponding closed loop matrix $F_{X^+} = F - GG^TX^+$ lie in the left half-plane and those of F_{X^-} are in the right half-plane. Moreover, according to [24] there is a one-to-one correspondence of solutions of (1.1) and invariant subspaces of F_{X^+} such that every solution of (1.1) can be expressed as a combination of X^+ and X^-. That result was refined by Coppel [5] and extended further by Shayman [23].

In this paper we consider the complex discrete-time algebraic Riccati equation (DARE)

$$X - F^*XF + (G^*XF + S)^*(R + G^*XG)^{-1}(G^*XF + S) - Q = 0. \tag{1.2}$$

We shall obtain a classification of solutions of (1.2) that corresponds to the Willems–Coppel–Shayman parametrization for (1.1).

Let us first recapitulate the parametrization result for the CARE (1.1). It will be assumed that elementary divisors of H have even degree when they belong to pure

*Received by the editors February 24, 2004; accepted for publication (in revised form) June 3, 2005; published electronically January 6, 2006.
http://www.siam.org/journals/sicon/44-6/44136.html
†Mathematisches Institut, Universität Würzburg, Am Hubland, D-97074 Würzburg, Germany (wimmer@mathematik.uni-wuerzburg.de).

imaginary eigenvalues. A solution X of (1.1) is called *unmixed* [23] if the spectrum of $F_X = F - GG^TX$ satisfies

$$\sigma(F_X) \cap \sigma(-F_X^T) \subseteq i\mathbb{R}. \tag{1.3}$$

It is known (see, e.g., [12]) that there exists an unmixed solution of (1.1) if and only if the pair $(F,G) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times p}$ is *sign-controllable*, i.e., if $\lambda \in \sigma(F)$ and $\operatorname{rank}[\lambda I - F, G] < n$ imply $\operatorname{rank}\left[-\overline{\lambda} I - F, G\right] = n$. Let X_1 be an unmixed solution of (1.1). Then there exists a unique (unmixed) solution X_2 such that F_{X_1} and F_{X_2} have, at most, pure imaginary eigenvalues in common if and only if (F,G) is controllable. In that case (X_1, X_2) is called a *pair of opposite unmixed solutions* [23]. The extremal solutions (X^-, X^+) are an example of such a pair.

Let $\operatorname{Inv}(F)$ denote the set of F-invariant subspaces of $\mathbb{R}^n$. If $i\alpha$ is an eigenvalue of F on the imaginary axis, we set $E_{\pm i\alpha} = \operatorname{Ker}(F^2 + \alpha^2 I)^n$ and we define

$$E_{i\mathbb{R}}(F) = \bigoplus_{\alpha\in\mathbb{R}} E_{\pm i\alpha}.$$

The following theorem can be regarded as an updated version of the Willems–Coppel–Shayman theory. It combines results of [24], [5], and [23] with an approach by Scherer [22] and observations on shorted operators in section 3.

THEOREM 1.1. *Let Γ be the set of real symmetric solutions of* (1.1). *Assume that (F,G) is controllable. Let (X_1, X_2) be a pair of opposite unmixed solutions of*

$$-F^TX - XF + XGG^TX - Q = 0, \tag{1.1}$$

and let $F_{X_2} = F - GG^TX_2$ be the closed loop matrix corresponding to X_2. Set $\Delta = X_1 - X_2$. Define

$$\mathcal{N} = \{N \in \operatorname{Inv}(F_{X_2}) \mid E_{i\mathbb{R}}(F_{X_2}) \subseteq N\}.$$

If $N \in \mathcal{N}$, then

$$(\Delta\, N)^\perp \oplus (N \cap \operatorname{Im}\Delta) = \mathbb{R}^n.$$

Let P_N be the projection on $(\Delta\, N)^\perp$ along $(N \cap \operatorname{Im}\Delta)$. Define

$$\kappa(N) = X_1P_N + X_2(I - P_N).$$

Then $\kappa : \mathcal{N} \to \Gamma$ is a bijection. If $X \in \Gamma$, then

$$\kappa^{-1}(X) = \operatorname{Ker}(X - X_2).$$

It is not difficult to explain (see also [18]) why direct sum decompositions of the form

$$(\Delta\, N)^\perp \oplus N = \mathbb{R}^n \tag{1.4}$$

and associated projections P_N should play a role in the theory of the CARE. Consider (1.1) with $Q = 0$, i.e.,

$$\mathcal{C}(X) = -F^TX - XF + XGG^TX = 0. \tag{1.5}$$

If (F, G) is controllable and $\sigma(F) \cap \sigma(-F^T) = \emptyset$, then (1.5) has a unique nonsingular solution Δ (see, e.g., [22, p. 102]). Let $N = \operatorname{Im}(I, O)^T$ be invariant under F and let Δ be partitioned according to F such that

$$F = \begin{bmatrix} F_1 & F_{12} \\ 0 & F_2 \end{bmatrix} \text{ and } \Delta = \begin{bmatrix} \Delta_1 & \Delta_{12} \\ \Delta_{12}^T & \Delta_2 \end{bmatrix}.$$

Then Δ_1 is nonsingular [22]. Let $\widetilde{\Delta}_2 = \Delta_2 - \Delta_{12}^T \Delta_1^{-1} \Delta_{12}$ be the Schur complement of Δ with respect to N. Set

$$S = \begin{bmatrix} I & -\Delta_1^{-1}\Delta_{12} \\ 0 & I \end{bmatrix}.$$

Then $S^T \Delta S = \operatorname{diag}(\Delta_1, \widetilde{\Delta}_2)$ and

$$S^{-1} F S = \begin{bmatrix} F_1 & * \\ 0 & F_2 \end{bmatrix}.$$

It is obvious that

$$\widetilde{\Delta} = \begin{bmatrix} 0 & 0 \\ 0 & \widetilde{\Delta}_2 \end{bmatrix}$$

is a solution of (1.5). In the terminology of [2], [16], or [6], the matrix $\widetilde{\Delta}$ is a *shorted matrix* of Δ. In section 3 we shall give a basis-free description of $\widetilde{\Delta}$ and we shall see that $\widetilde{\Delta} = \Delta P_N$, where P_N is the projection corresponding to (1.4).

Notation: Let $\mathbb{D}$ be the open unit disk and $\partial\mathbb{D} = \{z \in \mathbb{C};\, |z| = 1\}$ be the unit circle. Let Λ be a set of complex numbers and define

$$\Lambda^{\triangledown} = \{\bar{\lambda}^{-1} \mid \lambda \in \Lambda,\, \lambda \neq 0\}.$$

To a matrix of complex rational functions of the form $W(s) = D + C(sI - A)^{-1}B$, we associate $W^{\triangledown}(s) = D^* + B^*(s^{-1}I - A^*)C^*$. For $\lambda \in \mathbb{C}$ put

$$E_\lambda(F) = \operatorname{Ker}(\lambda I - F)^n.$$

Then $E_\lambda(F)$ is a generalized eigenspace of F if $\lambda \in \sigma(F)$. In particular, $E_0(F) = \operatorname{Ker} F^n$. We set

$$E_{\partial\mathbb{D}} = \oplus\{E_\eta(F),\, \eta \in \partial\mathbb{D}\}.$$

A subspace V of $\mathbb{C}^n$ is a *spectral subspace* of F if $V = \oplus\{E_\lambda(F);\, \lambda \in T\}$ for some $T \subseteq \sigma(F)$. If K is a hermitian matrix, we write $K > 0$ ($K \geq 0$) when K is positive (semi)definite. If F is nonsingular, we set $F^{-*} = (F^*)^{-1}$.

The main result of this paper is Theorem 5.3. It describes a parametrization of the set of hermitian solutions of the DARE (1.2). The proof of that theorem will proceed in several stages. In section 2 we review basic facts on the DARE. We recall that the difference of two solutions of (1.2) satisfies an associated DARE of the form

$$X - F^*XF + F^*XG(R + G^*XG)^{-1}G^*XF = 0.$$

Such equations, where $X = 0$ is a solution, will be considered in section 4. We have indicated before that (1.4) is related to Schur complements and shorted operators. That subject will be touched upon in section 3.

The geometric theory of Willems [24] and Coppel [5] was carried over to the DARE (1.2) first by G. Ruckebusch [20], [21, p. 129] and then by Ran and Trentelman [19]. Those papers give a geometric characterization of the set of hermitian solutions in terms of the pair of extremal solutions.

2. Basic facts of the DARE: Definitions. There is a wide class of problems in systems and control theory that require solutions of the DARE

$$\mathcal{D}(X) = X - F^*XF + (G^*XF + S)^*(R + G^*XG)^{-1}(G^*XF + S) - Q = 0. \tag{2.1}$$

An important example is the discrete-time linear quadratic problem of optimal control. Let the system

$$x(t+1) = Fx(t) + Gu(t),\ x(0) = x_0$$

be stabilizable, and let

$$J(x_0, u) = \sum_{t=0}^{\infty} [x^*(t)\ \ u^*(t)] \begin{bmatrix} Q & S^* \\ S & R \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}$$

be a positive semidefinite performance index. Then (see, e.g., [9]) there exists an optimal control $u(t)$ which minimizes $J(x_0, u)$. The optimal cost is given by $J_{\text{opt}} = x_0^* X_{\text{opt}} x_0$, where X_{opt} is the smallest positive semidefinite solution of (2.1), and the optimal control is

$$u_{\text{opt}}(t) = -(R + G^*X_{\text{opt}}G)^{-1}(S + G^*X_{\text{opt}}F)x(t). \tag{2.2}$$

Thus (2.2) gives rise to the closed loop system

$$x(t+1) = [F - G(R + G^*X_{\text{opt}}G)^{-1}(S + G^*X_{\text{opt}}F)]x(t). \tag{2.3}$$

The matrices in (2.1) are assumed to be complex, $F \in \mathbb{C}^{n\times n}$, $G \in \mathbb{C}^{n\times m}$, $S \in \mathbb{C}^{m\times n}$, $R = R^* \in \mathbb{C}^{m\times m}$, and $Q = Q^* \in \mathbb{C}^{n\times n}$. We are concerned with hermitian solutions X of (2.1). In this section we assemble basic facts and concepts related to (2.1). With regard to (2.3) we define

$$F_X = F - G(R + G^*XG)^{-1}(G^*XF + S)$$

as the closed loop matrix corresponding to a solution X. We say that a solution X of the DARE (2.1) is *unmixed* if F_X has the property

$$\sigma(F_X) \cap \sigma(F_X)^{\nabla} \subseteq \partial\mathbb{D}, \tag{2.4}$$

i.e., if $\lambda \in \sigma(F_X)$ and $\lambda \neq 0$ and $|\lambda| \neq 1$, then $\overline{\lambda}^{-1} \notin \sigma(F_X)$. We say that (X_1, X_2) is a *pair of opposite unmixed solutions* if the corresponding closed loop matrices satisfy

$$\sigma(F_{X_1}) \cap \sigma(F_{X_2}) \subseteq \partial\mathbb{D} \cup \{0\}. \tag{2.5}$$

Let

$$M - sL = \begin{bmatrix} F & 0 & G \\ Q & I & S^* \\ S & 0 & R \end{bmatrix} - s \begin{bmatrix} I & 0 & 0 \\ 0 & F^* & 0 \\ 0 & G^* & 0 \end{bmatrix} \tag{2.6}$$

be the extended symplectic pencil associated with (2.1), and let

$$\Psi(s) = \begin{bmatrix} G^*(s^{-1}I - F^*)^{-1} & I \end{bmatrix} \begin{bmatrix} Q & S^* \\ S & R \end{bmatrix} \begin{bmatrix} (sI - F)^{-1}G \\ I \end{bmatrix} \tag{2.7}$$

be the associated Popov matrix. The following identities and facts can be found in [11], [13], and [17]. If X is a solution of (2.1), then

$$\det(M - sL) = \det(R + G^*XG)\det(sI - F_X)\det(I - sF_X^*). \tag{2.8}$$

It follows from (2.8) that there exists a solution of (2.1) only if $M - sL$ is nonsingular, i.e., if $\det(M - sL)$ is not the zero polynomial. In that case we call

$$\sigma(M - sL) = \{\lambda \in \mathbb{C} \mid \det(M - \lambda L) = 0\}$$

the set of *characteristic roots* of the pencil (2.6). It is obvious from (2.8) that $\sigma(F_X) \subseteq \sigma(M - sL)$. Hence, if 0 is a characteristic root of $M - sL$, then we have $0 \in \sigma(F_X)$ for all solutions X, which accounts for the singleton $\{0\}$ in (2.5). Let X_1 and X_2 be solutions of (2.1) such that (2.5) holds. Then it follows from (2.8) that both X_1 and X_2 are unmixed. Each solution X of (2.1) gives rise to a factorization of the Popov matrix $\Psi(s)$, namely

$$\Psi(s) = \Phi_X^\nabla(s)(R + G^*XG)\Phi_X(s) \tag{2.9}$$

with

$$\Phi_X(s) = I + (R + G^*XG)^{-1}(G^*XF + S)(sI - F)^{-1}G.$$

If $\Psi(s)$ satisfies

$$\Psi(\eta) > 0 \text{ for some } \eta \in \partial\mathbb{D}, \tag{Ψ}$$

then (2.9) implies

$$R + G^*XG > 0 \tag{2.10}$$

for all solutions X.

Arguments in section 4 show that the conditions (Ψ) and (2.10) are essential for the derivation of Theorem 5.3. Hence it is not within the scope of this paper to deal with DAREs of the form (1.2) where $(R + G^*XG)^{-1}$ is replaced by a generalized inverse $(R + G^*XG)^\#$.

Existence of unmixed solutions was studied in [4], [25], [3]. We note the following result of [3]. Suppose that (Ψ) holds, $\Psi(\eta) \geq 0$ for almost all $\eta \in \partial\mathbb{D}$, and $\operatorname{rank}(F - \lambda I) = n$ for all $\lambda \neq 0$. Then each unmixed set Λ gives rise to a unique solution X such that $\sigma(F_X) \subseteq \Lambda$. In particular there exists a unique pair (X_-, X_+) such that

$$\sigma(F_{X_-}) \subseteq \{0\} \cup \{\lambda \in \mathbb{C}\,;\, |\lambda| \geq 1\} \text{ and } \sigma(F_{X_+}) \subseteq \overline{\mathbb{D}}. \tag{2.11}$$

It is known that Cayley transformations allow a passage from continuous-time to discrete-time algebraic Riccati equations [15], [1]. The use of such transformations requires invertibility assumptions (see, e.g., [1, p. 81]) which are not met by the general hypotheses in the present paper. We note that in the case of $0 \in \sigma(M - sL)$ a computational procedure is available [7] to obtain an equivalent DARE of smaller order such that the associated closed loop matrices are nonsingular.

3. Shorted operators and oblique projections. Let N be a subspace of $\mathbb{C}^n$ and Δ be a hermitian $n \times n$ matrix.

LEMMA 3.1. (i) *If*

$$N = \operatorname{Im} \begin{bmatrix} I_t \\ 0 \end{bmatrix}, \quad \text{and if} \quad \Delta = \begin{bmatrix} \Delta_1 & \Delta_{21}^* \\ \Delta_{21} & \Delta_2 \end{bmatrix} \tag{3.1}$$

is partitioned conformably, then we have

$$\mathbb{C}^n = (\Delta N)^\perp \oplus N \tag{3.2}$$

if and only if Δ_1 is nonsingular.

(ii) *Assume* (3.2) *and let P_N be the projection on $(\Delta N)^\perp$ along N. Then the matrix ΔP_N is hermitian. If N and Δ are given as in* (3.1)*, then*

$$\Delta P_N = \operatorname{diag}(0, \tilde{\Delta}_2), \qquad \tilde{\Delta}_2 = \Delta_2 - \Delta_{21}^* \Delta_1^{-1} \Delta_{21}. \tag{3.3}$$

Proof. (i) Suppose $\det \Delta_1 = 0$. If $u_1 \in \operatorname{Ker}\ \Delta_1$, $u_1 \neq 0$, then

$$u = \begin{bmatrix} u_1 \\ 0 \end{bmatrix} \in N$$

and

$$u^* \Delta N = [u_1^* \ \ 0] \begin{bmatrix} \Delta_1 \\ \Delta_{21} \end{bmatrix} = 0.$$

Hence $(\Delta N)^\perp \cap N \neq 0$. Conversely, if Δ_1 is nonsingular, then

$$\Delta N = \operatorname{Im} \begin{bmatrix} \Delta_1 \\ \Delta_{21} \end{bmatrix} = \operatorname{Im} \begin{bmatrix} I_t \\ \Delta_{21} \Delta_1^{-1} \end{bmatrix}$$

and

$$(\Delta N)^\perp = \operatorname{Im} \begin{bmatrix} -\Delta_1^{-1} \Delta_{21}^* \\ I_{n-t} \end{bmatrix} \tag{3.4}$$

imply (3.2). (ii) From $\operatorname{Im} P_N = (\Delta N)^\perp$ and $\Delta \operatorname{Im} (I - P_N) = \Delta N$ follows $(I - P_N)^* \Delta P_N = 0$. Hence, $\Delta P_N = P_N^* \Delta P_N$ is hermitian. Because of (3.4) the projection matrix P_N is given by

$$P_N = \begin{bmatrix} 0 & -\Delta_1^{-1} \Delta_{21}^* \\ 0 & I_{n-t} \end{bmatrix},$$

which yields ΔP_N as a shorted operator in block diagonal form (3.3). ∎

The following observation is a special case of a formula for the inverse of a block matrix [10, p. 18]. It will be needed in the proof of Theorem 4.5, which deals with a nonsingular solution Δ of a special DARE and its inverse.

LEMMA 3.2. *Assume that Δ is nonsingular, set $W = \Delta^{-1}$, and let*

$$\Delta = \begin{bmatrix} \Delta_1 & \Delta_{21}^* \\ \Delta_{21} & \Delta_2 \end{bmatrix} \ \text{and} \ W = \begin{bmatrix} * & * \\ * & W_2 \end{bmatrix}$$

be partitioned accordingly. (i) *Then Δ_1 is nonsingular if and only if W_2 is nonsingular.* (ii) *If Δ_1 is nonsingular, then $\widetilde{\Delta}_2 = \Delta_2 - \Delta_{21}^* \Delta_1^{-1} \Delta_{21}$ is nonsingular and $\widetilde{\Delta}_2^{-1} = W_2$.*

Proof. We include a proof which is an application of the preceding lemma. Set $N = \operatorname{Im}[I_t \ O]^T$ and $M = N^\perp$. (i) Assume that Δ_1 is nonsingular, or equivalently (3.2). Because of $(\Delta N)^\perp = \Delta^{-1}N^\perp$ we can write (3.2) as $WM \oplus M^\perp = \mathbb{C}^n$, which in turn is equivalent to $(WM)^\perp \oplus M = \mathbb{C}^n$. Because of $M = \operatorname{Im}[O \ I_{n-t}]^T$ such a decomposition exists if and only if W_2 is nonsingular. (ii) From $\Delta P_N = \operatorname{diag}(0, \widetilde{\Delta}_2)$ follows

$$\Delta P_N \Delta^{-1} = \begin{bmatrix} 0 & 0 \\ * & \widetilde{\Delta}_2 W_2 \end{bmatrix}.$$

Obviously $\Delta P_N \Delta^{-1}$ is the projection on $N^\perp$ along ΔN. Thus we obtain $\widetilde{\Delta}_2 W_2 = I$. ∎

Let $K = \operatorname{Ker}\Delta$ and $V = \operatorname{Im}\Delta$. We assume $0 < n_c = \dim V < n$ such that the decomposition $\mathbb{C}^n = K \oplus V$ is nontrivial. Let $U \subseteq V$. Then $U^{\perp_c}$ shall denote the orthogonal complement of U with respect to V. Set $\Delta_c = \Delta_{|V}$. Then $\Delta_c : V \to V$ is hermitian and nonsingular.

LEMMA 3.3. *Let N be a subspace of $\mathbb{C}^n$ with $K \subseteq N$. Set $N_c = N \cap V$. Then*

$$(\Delta_c N_c)^{\perp_c} \oplus N_c = V \tag{3.5}$$

if and only if

$$(\Delta N)^\perp \oplus (N \cap \operatorname{Im}\Delta) = \mathbb{C}^n. \tag{3.6}$$

Proof. Note that $(\Delta N)^\perp = K \oplus (\Delta_c N_c)^{\perp_c}$ and $N \cap \operatorname{Im}\Delta = N_c$. ∎

4. A DARE with solution $X = 0$. In this section we deal with the equation

$$\mathcal{H}(Y) = Y - F^*YF + F^*YG(R + G^*YG)^{-1}G^*YF = 0, \tag{4.1}$$

where R is nonsingular. Set $\Gamma = GR^{-1}G^*$. Then

$$F_Y = F - G(R + G^*YG)^{-1}G^*YF = (I + \Gamma Y)^{-1}F,$$

and we have $\mathcal{H}(Y) = Y - F^*YF_Y = 0$. We first note two results on nonsingular solutions of (4.1).

LEMMA 4.1 (see [26, p. 931]). *Let Y be a hermitian nonsingular $n \times n$ matrix. Assume that F is nonsingular. Then Y is a solution of* (4.1) *if and only if $W = Y^{-1}$ satisfies the discrete-time Lyapunov equation*

$$W - FWF^* = -\Gamma. \tag{4.2}$$

LEMMA 4.2 (see [26, p. 932]). *Assume that F is nonsingular, (F, G) is controllable, $R > 0$, and $\sigma(F) \cap \sigma(F^{-*}) = \emptyset$. Then* (4.1) *has a unique nonsingular solution.*

The subsequent result deals with eigenvalues of F in $\{0\} \cup \partial\mathbb{D}$ and corresponding generalized eigenspaces.

LEMMA 4.3 (see [26, p. 933]). *Assume $R > 0$ and $\operatorname{rank}(F - \eta I, G) = n$ for all $\eta \in \partial\mathbb{D}$. If Y is a solution of* (4.1)*, then $E_0(F) + E_{\partial\mathbb{D}}(F) \subseteq \operatorname{Ker} Y$.*

COROLLARY 4.4 (see [26, pp. 923–924]). *Assume $R > 0$, $\operatorname{rank}[F - \lambda I, \ G] = n$ if $\lambda \neq 0$, and $\sigma(F) \cap \sigma(F)^\nabla \subseteq \partial\mathbb{D}$. Then* (4.1) *has a unique solution Δ with $\operatorname{Ker}\Delta = E_0(F) + E_{\partial\mathbb{D}}(F)$.*

It is easy to show (see also Proposition 5.2) that $(0, \Delta)$ is a pair of opposite unmixed solutions of $\mathcal{H}(Y) = 0$. Therefore, the following result can already be viewed as a special case of the main theorem.

THEOREM 4.5. *The assumptions on the DARE*

$$\mathcal{H}(Y) = Y - F^*YF + F^*YG(R + G^*YG)^{-1}G^*YF = 0 \tag{4.1}$$

are the following: $R > 0$, $\sigma(F) \cap \sigma(F)^{\nabla} \subseteq \partial\mathbb{D}$, *and* $\operatorname{rank}[F - \lambda I,\ G] = n$ *if* $\lambda \neq 0$. *Define*

$$K = E_0(F) + E_{\partial\mathbb{D}}(F),$$

and

$$\mathcal{T} = \{Y \mid \mathcal{H}(Y) = 0\},\quad \mathcal{N} = \{N \in \operatorname{Inv} F \mid K \subseteq N\}. \tag{4.3}$$

Let Δ *be the solution of* (4.1) *with* $\operatorname{Ker} \Delta = K$. *Then the following holds.*

(i) *If* Y *is a solution of* (4.1), *then* $\operatorname{Ker} Y \in \mathcal{N}$.

(ii) *If* $N \in \mathcal{N}$, *then*

$$\mathbb{C}^n = (\Delta N)^{\perp} \oplus (N \cap \operatorname{Im} \Delta), \tag{4.4}$$

and if P_N *is the projection of* $\mathbb{C}^n$ *on* $(\Delta N)^{\perp}$ *along* $N \cap \operatorname{Im} \Delta$, *then* $\tilde{Y} = \Delta P_N$ *is the unique solution of* (4.1) *with* $\operatorname{Ker} \tilde{Y} = N$.

(iii) *For* $N \in \mathcal{N}$ *set* $\tilde{\kappa}(N) = \Delta P_N$. *Then the map* $\tilde{\kappa} : \mathcal{N} \to \mathcal{T}$ *is a bijection, and for* $Y \in \mathcal{T}$ *we have* $\tilde{\kappa}^{-1}(Y) = \operatorname{Ker} Y$.

Proof. Let us first prove the theorem under the stronger assumption that F is nonsingular and $\sigma(F) \cap \sigma(F^{-*}) = \emptyset$. In that case we have $K = 0$ and the solution Δ is nonsingular such that $N \cap \operatorname{Im} \Delta = N$ and $\mathcal{N} = \operatorname{Inv} F$.

(i) This is obvious since $\mathcal{H}(Y) = Y - F^*YF_Y = 0$ is equivalent to $F_Y^{-*}Y = YF$.

(ii) Assume $N = \operatorname{Im}[I\ O]^T$ such that

$$F = \begin{bmatrix} F_1 & * \\ 0 & F_2 \end{bmatrix}. \tag{4.5}$$

It is known from Lemma 4.1 that $W = \Delta^{-1}$ satisfies

$$W - FWF^* = -GR^{-1}G^*. \tag{4.6}$$

Let

$$\Delta = \begin{bmatrix} \Delta_1 & \Delta_{21}^* \\ \Delta_{21} & \Delta_2 \end{bmatrix},\ G = \begin{bmatrix} * \\ G_2 \end{bmatrix},\ \text{and}\ \Delta^{-1} = W = \begin{bmatrix} * & * \\ * & W_2 \end{bmatrix}$$

be partitioned according to (4.5). To establish a decomposition

$$\mathbb{C}^n = (\Delta N)^{\perp} \oplus N, \tag{4.7}$$

we recall Lemma 3.1 and the fact that (4.7) holds if and only if Δ_1 is nonsingular. Thus, according to Lemma 3.2 we have to show that W_2 is nonsingular. From (4.6) we obtain

$$W_2 - F_2W_2F_2^* = -G_2R^{-1}G_2^*. \tag{4.8}$$

In the preceding Stein equation the pair (F_2, G_2) is controllable, $\sigma(F_2) \cap \sigma(F_2^{-*}) = \emptyset$, and $R > 0$. Therefore (see, e.g., [14, p. 453]) the matrix W_2 is nonsingular.

Now let us show that $\tilde{Y} = \Delta P_N$ is a solution of (4.1). Since Δ is assumed to be nonsingular we have $\operatorname{Ker} \tilde{Y} = \operatorname{Ker} P_N = N$. Hence it follows from Lemmas 3.1 and 3.2 that $\tilde{Y} = P_N^* \Delta P_N = \operatorname{diag}(0, \widetilde{\Delta}_2)$, and $\tilde{\Delta}_2 = \Delta_2 - \Delta_{21}\Delta_1^{-1}\Delta_{21}^*$ is nonsingular and $\widetilde{\Delta}_2^{-1} = W_2$. Consider the equation

$$\mathcal{H}_2(Y_2) = Y_2 - F_2^* Y_2 F_2 + F_2^* Y_2 G_2 (R + G_2^* Y_2 G_2)^{-1} G_2^* Y_2 F_2. \tag{4.9}$$

It is easy to check that $\mathcal{H}(\tilde{Y}) = \operatorname{diag}\bigl(0, \mathcal{H}_2(\tilde{\Delta}_2)\bigr)$. Hence $\tilde{Y}$ is a solution of (4.1) if and only if $\mathcal{H}_2(\widetilde{\Delta}_2) = 0$. To show that $\widetilde{\Delta}_2$ is a solution of (4.9) we use Lemma 4.1 again, which says that $\mathcal{H}_2(\widetilde{\Delta}_2) = 0$ is equivalent to

$$\widetilde{\Delta}_2^{-1} - F_2 \widetilde{\Delta}_2^{-1} F_2^* = -G_2 R^{-1} G_2^*. \tag{4.10}$$

We have seen that the solution Δ gives rise to (4.8). Therefore, because $W_2 = \widetilde{\Delta}_2^{-1}$, (4.10) is satisfied.

To prove uniqueness, take a solution Y of (4.1) with $\operatorname{Ker} Y = N$. Then $Y = \operatorname{diag}(0, Y_2)$, and Y_2 is nonsingular satisfying (4.9). According to Lemma 4.2 there exists a unique nonsingular solution of (4.9), namely, $\widetilde{\Delta}_2$. Hence $Y = \operatorname{diag}(0, \widetilde{\Delta}_2) = \tilde{Y}$. Part (iii) is an obvious consequence of (i) and (ii).

We now discard the assumptions about F made at the beginning and consider the case $K = E_0(F) + E_{\partial\mathbb{D}}(F) \neq 0$. Let $V = \oplus\{E_\lambda(F) \mid \lambda \neq 0, \lambda \notin \partial\mathbb{D}\}$ be the F-invariant complement of K such that $\mathbb{C}^n = K \oplus V$. Suppose

$$K = \operatorname{Im} \begin{bmatrix} I_{n_a} \\ 0 \end{bmatrix} \text{ and } V = \operatorname{Im} \begin{bmatrix} 0 \\ I_{n_c} \end{bmatrix} \tag{4.11}$$

such that

$$F = \begin{bmatrix} F_a & 0 \\ 0 & F_c \end{bmatrix}, \; G = \begin{bmatrix} * \\ G_c \end{bmatrix}.$$

Then $\sigma(F_a) \subseteq \{0\} \cup \partial\mathbb{D}$, F_c is nonsingular, and $\sigma(F_c) \cap \sigma(F_c^{-*}) = \emptyset$. It follows from Lemma 4.3 that Y is a solution of (4.1) if and only if $Y = \operatorname{diag}(0, Y_c)$ and Y_c is a solution of

$$\mathcal{H}_c(Y_c) = Y_c - F_c^* Y_c F_c + F_c^* Y_c G_c (R + G_c^* Y_c G_c)^{-1} G_c^* Y_c F_c = 0. \tag{4.12}$$

Set $\mathcal{T}_c = \{Y_c \mid \mathcal{H}_c(Y_c) = 0\}$ and let $\iota : \mathcal{T}_c \to \mathcal{T}$ be the bijection given by $\iota(Y_c) = \operatorname{diag}(0, Y_c)$. Let Δ_c be the unique nonsingular solution of (4.12). Then $\Delta = \operatorname{diag}(0, \Delta_c)$ is the unique solution of (4.1) with $\operatorname{Ker} \Delta = K$.

Let $\hat{N}_c \in \operatorname{Inv} F_c$. We embed $\hat{N}_c$ into the space

$$N_c = \left\{ \begin{bmatrix} 0 \\ x_c \end{bmatrix} \in \mathbb{C}^n, \, x_c \in \hat{N}_c \right\}$$

and define $\tau(\hat{N}_c) = K \oplus N_c$. Then $\tau : \operatorname{Inv} F_c \to \mathcal{N}$ is a bijection. In accordance with (4.7) we have

$$\mathbb{C}^{n_c} = (\Delta_c \hat{N}_c)^\perp \oplus \hat{N}_c, \tag{4.13}$$

which is equivalent to $V = (\Delta_c N_c)^{\perp_c} \oplus N_c$. Hence, it follows from Lemma 3.3 that $N = \tau(\hat{N}_c) \in \mathcal{N}$ satisfies (4.4). Let $P_{\hat{N}_c}$ be the projection corresponding to (4.13) and set $\tilde{\kappa}_c(\hat{N}_c) = \Delta_c P_{\hat{N}_c}$. We know that $\tilde{\kappa}_c : \text{Inv } F_c \to \mathcal{T}_c$ is a bijection. Define $\mu_c = \tilde{\kappa}_c^{-1}$, i.e., $\mu_c(Y_c) = \text{Ker } Y_c$, $Y_c \in \mathcal{T}_c$. The maps $\tilde{\kappa}_c$ and μ_c can now be extended to bijections between $\mathcal{T}$ and $\mathcal{N}$. For $Y \in \mathcal{T}$ define $\mu(Y) = \text{Ker } Y$. Then $\mu = \tau\mu_c\iota^{-1} : \mathcal{T} \to \mathcal{N}$, and μ is bijective. Let P_N be the projection arising from (4.4). For $N \in \mathcal{N}$ define $\tilde{\kappa}(N) = \Delta P_N$. In the setting of (4.11) we have $P_N = \text{diag}(I, P_{\hat{N}_c})$. Therefore

$$\tilde{\kappa}(N) = \text{diag}(0, \Delta_c P_{\hat{N}_c}) = \iota(\Delta_c P_{\hat{N}_c}) = \iota\tilde{\kappa}_c(\hat{N}_c) = \iota\tilde{\kappa}_c\tau^{-1}(N).$$

Hence $\tilde{\kappa} : \mathcal{N} \to \mathcal{T}$ and $\tilde{\kappa} = \mu^{-1}$. □

In order to characterize unmixed solutions of $\mathcal{D}(X) = 0$, we need a discrete-time counterpart of a result of Scherer [22, p. 106]. Let $A \in \mathbb{C}^{n\times n}$ and $M \in \text{Inv } A$; let $\bar{A}$ denote the endomorphism of $\mathbb{C}^n/M$ induced by A; and let $A_{|M}$ be the restriction of A to M. We define $\sigma_i(A, M) = \sigma(A_{|M})$ and $\sigma_o(A, M) = \sigma(\bar{A})$. Assumptions and definitions in the lemma below are those of Theorem 4.5 and its proof.

LEMMA 4.6. *For $N \in \mathcal{N}$ let $Y = \Delta P_N$ be the corresponding solution of* (4.1). (i) *Then $F = F_Y$ on N and $N \in \text{Inv } F_Y$, and we have*

$$\sigma_i(F_Y, N) = \sigma_i(F, N) \text{ and } \sigma_o(F_Y, N) = \sigma_o(F, N)^\nabla. \tag{4.14}$$

(ii) *The solution Y is unmixed if and only if $N \in \mathcal{N}$ is a spectral subspace of F.*

Proof. Let us assume for simplicity that F is nonsingular and $\sigma(F) \cap \sigma(F^{-*}) = \emptyset$.

(i) As before, suppose $N = \text{Im}\,[I\ O]^T$ and let F be given by (4.5). Then $\sigma_i(F, N) = \sigma(F_1)$ and $\sigma_o(F, N) = \sigma(F_2)$. From $Y = \Delta P_N = \text{diag}(0, \widetilde{\Delta}_2)$ follows

$$F_Y = (I + \Gamma Y)^{-1}F = \begin{bmatrix} F_1 & * \\ 0 & A_2 \end{bmatrix}$$

with $A_2 = (I + \Gamma_2\widetilde{\Delta}_2)^{-1}F_2$. Thus $\sigma_i(F_Y, N) = \sigma(F_1)$. On the other hand, $Y - F^*YF_Y = 0$ implies $A_2 = \widetilde{\Delta}_2^{-1}F_2^{-*}\widetilde{\Delta}_2$, and we obtain $\sigma(A_2) = \sigma(F_2)^\nabla$.

(ii) N is a spectral subspace of F if and only if

$$\sigma(F_1) \cap \sigma(F_2) = \emptyset. \tag{4.15}$$

The assumption $\sigma(F) \cap \sigma(F^{-*}) = \emptyset$ together with (4.14) imply

$$\sigma(F_Y) \cap \sigma(F_Y^{-*}) = [\sigma(F_1) \cup \sigma(F_2^{-*})] \cap [\sigma(F_1^{-*}) \cup \sigma(F_2)] =$$
$$[\sigma(F_2^{-*}) \cap \sigma(F_1^{-*})] \cup [\sigma(F_1) \cap \sigma(F_2)] = [\sigma(F_1) \cap \sigma(F_2)]^\nabla \cup [\sigma(F_1) \cap \sigma(F_2)].$$

Hence, the property that Y is unmixed, i.e., $\sigma(F_Y) \cap \sigma(F_Y^{-*}) = \emptyset$, is equivalent to (4.15). □

5. The main result. The passage from the general DARE

$$\mathcal{D}(X) = X - F^*XF + (G^*XF + S)^*(R + G^*XG)^{-1}(G^*XF + S) - Q = 0 \tag{5.1}$$

to an equation of the form $\mathcal{H}(Y) = 0$ in (4.1) is a crucial step in the derivation of our main theorem. It is based on the following lemma for which we refer to [19] and [8, Lemma 5.2] or [1, Lemma 6.8.9].

LEMMA 5.1. (i) *Let X_2 be a solution of $\mathcal{D}(X) = 0$. Then X is a solution of $\mathcal{D}(X) = 0$ if and only if $Y = X - X_2$ is a solution of*

$$\mathcal{H}_2(Y) = Y - F_{X_2}^* Y F_{X_2} + F_{X_2}^* YG[(R + G^* X_2 G) + G^* YG]^{-1} G^* Y F_{X_2} = 0. \tag{5.2}$$

(ii) *Let X_2 and X be solutions of $\mathcal{D}(X) = 0$. Set $Y = X - X_2$. Then*

$$F_X = F_{X_2} - G(R + G^* XG)^{-1} G^* Y F_{X_2}. \tag{5.3}$$

Thus, when a solution X_2 of (5.1) is at our disposal we can pass from (5.1) to (5.2) and apply the results of section 4.

PROPOSITION 5.2. *Assume* (Ψ) *and*

$$\operatorname{rank}[F - \lambda I,\ G] = n \ \text{ if } \lambda \neq 0. \tag{5.4}$$

Let X_2 be an unmixed solution of $\mathcal{D}(X) = 0$. Set $K = E_0(F_{X_2}) + E_{\partial\mathbb{D}}(F_{X_2})$. Then there exists a unique solution Δ of (5.2) with $\operatorname{Ker} \Delta = K$. *Moreover (X, X_2) is a pair of opposite unmixed solutions of $\mathcal{D}(X) = 0$ if and only if $X = X_2 + \Delta$.*

Proof. The assumption (Ψ) implies $R + G^* X_2 G > 0$ in (5.2). Clearly (5.4) implies the corresponding condition for $F = F_{X_2}$. The assumption that X_2 should be an unmixed solution of (5.1) is equivalent to

$$\sigma(F_{X_2}) \cap \sigma(F_{X_2})^{\triangledown} \subseteq \partial\mathbb{D}. \tag{5.5}$$

Hence it follows from Corollary 4.4 that (5.2) has a unique solution Δ with $\operatorname{Ker} \Delta = K$. Again it is no loss of generality to assume

$$K = \operatorname{Im} \begin{bmatrix} I_{n-n_c} \\ 0 \end{bmatrix}$$

and $F_{X_2} = \operatorname{diag}(F_a, F_c)$. Then $\Delta = \operatorname{diag}(0,\ \Delta_c)$, $\det \Delta_c \neq 0$. Obviously (5.5) implies that F_c is nonsingular and

$$\sigma(F_c) \cap \sigma(F_c^{-*}) = \emptyset. \tag{5.6}$$

Let X be a solution of (5.1). Set $Y = X - X_2$. Then $\mathcal{H}_2(Y) = 0$, or equivalently

$$Y - F_{X_2}^* Y F_X = 0. \tag{5.7}$$

Because of Lemma 4.3 we have $K \subseteq \operatorname{Ker} Y$. Hence $Y = \operatorname{diag}(0, Y_c)$. The identity (5.3) yields

$$F_X = \begin{bmatrix} F_a & * \\ 0 & B \end{bmatrix}. \tag{5.8}$$

From (5.7) we obtain

$$Y_c - F_c^* Y_c B = 0. \tag{5.9}$$

Let us consider the solution $X = X_2 + \Delta$ and let F_X be as in (5.8). In this case we have $Y = \Delta = \operatorname{diag}(0,\ \Delta_c)$. Thus (5.9) becomes $\Delta_c - F_c^* \Delta_c B = 0$, where F_c and Δ_c are nonsingular. Hence $B = \Delta_c^{-1} F_c^{-*} \Delta_c$, $\sigma(B) = \sigma(F_c^{-*})$, and

$$\sigma(F_X) = \sigma(F_a) \cup \sigma(B) = \sigma(F_a) \cup \sigma(F_c^{-*}).$$

Thus (5.6) implies $\sigma(F_X) \cap \sigma(F_{X_2}) = \sigma(F_a) \subseteq \{0\} \cup \partial\mathbb{D}$, and therefore (X, X_2) is a pair of opposite unmixed solutions.

Now consider a solution X such that $Y = X - X_2 \neq \Delta$. Then $Y = \operatorname{diag}(0, Y_c)$ and Y_c is singular. Take $w \in \operatorname{Ker} Y_c$, $w \neq 0$. Recall that F_c in (5.9) is nonsingular. Hence $F_c B w = 0$, and $\operatorname{Ker} Y_c$ is invariant under B. Therefore we can assume that w is an eigenvector of B, say $Bw = \lambda w$. Interchanging the role of X and X_2 in (5.3), we obtain

$$F_{X_2} = \operatorname{diag}(F_a, F_c) = F_X - G(R + G^* X_2 G)^{-1} G^*(-Y) F_X.$$

Hence $F_c = B - W Y_c B$ with some $n_c \times n_c$ matrix W. Thus $F_c w = Bw$. Therefore $\lambda \in \sigma(F_c)$, and $\lambda \notin (\{0\} \cup \partial\mathbb{D})$. On the other hand $\lambda \in \sigma(F_X) \cap \sigma(F_{X_2})$. Hence (X, X_2) is not a pair of opposite unmixed solutions. ∎

At this point the pieces can be put together.

THEOREM 5.3. *Let (X_1, X_2) be a pair of opposite unmixed solutions of the DARE $\mathcal{D}(X) = 0$ in (5.1). Set $\Delta = X_1 - X_2$ and*

$$F_{X_2} = F - G(R + G^* X_2 G)^{-1}(G^* X_2 F + S).$$

Let $\mathcal{S}$ be the set of hermitian solutions of $\mathcal{D}(X) = 0$, and define

$$\mathcal{N} = \{N \in \operatorname{Inv} F_{X_2} \mid E_0(F_{X_2}) + E_{\partial\mathbb{D}}(F_{X_2}) \subseteq N\}. \tag{5.10}$$

Assume (Ψ) and $\operatorname{rank}[F - \lambda I, G] = n$ if $\lambda \neq 0$. If $N \in \mathcal{N}$, then

$$(\Delta\, N)^\perp \oplus (N \cap \operatorname{Im} \Delta) = \mathbb{C}^n. \tag{5.11}$$

Let P_N be the projection on $(\Delta\, N)^\perp$ along $N \cap \operatorname{Im} \Delta$. Define

$$\kappa(N) = X_1 P_N + X_2(I - P_N). \tag{5.12}$$

Then $\kappa : \mathcal{N} \to \mathcal{S}$ is a bijection. If $X \in \mathcal{S}$, then

$$\kappa^{-1}(X) = \operatorname{Ker}(X - X_2).$$

Proof. Let $\mathcal{T}_2$ denote the set of hermitian solutions of $\mathcal{H}_2(Y) = 0$ in (5.2). By Lemma 5.1 we have $X \in \mathcal{S}$, if and only if $Y = X - X_2 \in \mathcal{T}_2$. Hence $\mathcal{S} = X_2 + \mathcal{T}_2$. If (X_1, X_2) is a pair of opposite unmixed solutions, then it follows from Proposition 5.2 that $\Delta = X_1 - X_2$ is the solution of (5.2) with $\operatorname{Ker} \Delta = E_0(F_{X_2}) + E_{\partial\mathbb{D}}(F_{X_2})$. Hence we can parametrize the set $\mathcal{T}_2$ according to Theorem 4.5. If we set $\tilde{\kappa}(N) = \Delta P_N$, then $\tilde{\kappa} : \mathcal{N} \to \mathcal{T}_2$ is a bijection with $\tilde{\kappa}^{-1}(Y) = \operatorname{Ker} Y$. Thus $\mathcal{S} = X_2 + \mathcal{T}_2$ gives rise to a bijection $\kappa : \mathcal{N} \to \mathcal{S}$ defined by $\kappa(N) = X_2 + \tilde{\kappa}(N)$. Then $\kappa(N) = X_2 + (X_1 - X_2)P_N$ yields (5.12). Finally, for $X \in \mathcal{S}$ we have $\kappa^{-1}(X) = \tilde{\kappa}^{-1}(X - X_2) = \operatorname{Ker}(X - X_2)$. ∎

COROLLARY 5.4. *A solution X of (5.1) is unmixed if and only if $\operatorname{Ker}(X - X_2)$ is a spectral subspace of F_{X_2}.*

Proof. Let X be a solution of (5.1) and set $Y = X - X_2$ such that Y is a solution of $\mathcal{H}_2(Y) = 0$ in (5.2). We claim that X is an unmixed solution of (5.1) if and only if the corresponding matrix Y is an unmixed solution of (5.2). Let A be the closed loop matrix associated with Y (with respect to $\mathcal{H}_2 = 0$), i.e.,

$$A = F_{X_2} - G[(R + G^* X_2 G) + G^* Y G]^{-1} G^* Y F_{X_2}.$$

Now the identity (5.3) shows that $A = F_X$. We can apply Lemma 4.6, which tells us that Y is an unmixed solution of (5.2) if and only if $N = \tilde{\kappa}^{-1}(Y) = \operatorname{Ker} Y$ is a spectral subspace of F_{X_2}. □

The pair of opposite solutions (X_-, X_+) satisfying (2.11) consists of the smallest and the greatest solution of (5.1), i.e., for each $X \in \mathcal{S}$ we have $X_- \leq X \leq X_+$. In that particular case, Theorem 5.3 can be obtained from a result on intervals of solutions in [27].

REFERENCES

[1] H. Abou-Kandil, G. Freiling, V. Ionescu, and G. Jank, *Matrix Riccati Equations in Control and Systems Theory*, Birkhäuser, Basel, 2003.

[2] W. N. Anderson, Jr., *Shorted operators*, SIAM J. Appl. Math., 20 (1971), pp. 520–525.

[3] D. J. Clements and H. K. Wimmer, *Existence and uniqueness of unmixed solutions of the discrete-time algebraic Riccati equation*, Systems Control Lett., 50 (2003), pp. 343–346.

[4] A. N. Churilov, *On solutions of a quadratic matrix equation encountered in the investigation of discrete control systems*, Sov. Math. (Iz. VUZ), 30 (1986), pp. 81–89.

[5] W. A. Coppel, *Matrix quadratic equations*, Bull. Austral. Math. Soc., 10 (1974), pp. 377–401.

[6] G. Corach, A. Maestripieri, and D. Stojanoff, *Oblique projections and Schur complements*, Acta Sci. Math. (Szeged), 67 (2001), pp. 337–356.

[7] A. Ferrante and H. K. Wimmer, *Order reduction of discrete-time algebraic Riccati equations with singular closed loop matrix*, Eur. J. Control, submitted.

[8] G. Freiling and A. Hochhaus, *Properties of the solutions of rational matrix difference equations*, Comput. Math. Appl., 45 (2003), pp. 1137–1154.

[9] A. H. W. Geerts, *The algebraic Riccati equation and singular optimal control: The discrete-time case*, in Systems and Networks: Mathematical Theory and Applications, Proceedings of the International Symposium MTNS 93, Regensburg, 1993, Vol. II, U. Helmke et al., eds., Akademie Verlag, Berlin, 1994, pp. 129–134.

[10] R. A. Horn and Ch. R. Johnson, *Matrix Analysis*, Cambridge University Press, New York, 1985.

[11] V. Ionescu and C. Oară, *Generalized discrete-time Riccati theory*, SIAM J. Control Optim., 34 (1996), pp. 601–619.

[12] V. Kučera, *Algebraic Riccati equations: Hermitian and definite solutions*, in The Riccati Equation, Commun. Control Engrg., S. Bittanti et al., eds., Springer-Verlag, Berlin, 1991, pp. 53–88.

[13] P. Lancaster and L. Rodman, *Algebraic Riccati Equations*, Clarendon Press, Oxford, UK, 1995.

[14] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, Academic Press, San Diego, 1985.

[15] V. Mehrmann, *A step towards a unified treatment of continuous and discrete time control problems*, Linear Algebra Appl., 241–243 (1996), pp.749–779.

[16] S. K. Mitra and M. L. Puri, *Shorted matrices—an extended concept and some applications*, Linear Algebra Appl., 42 (1982), pp. 57–79.

[17] B. P. Molinari, *The stabilizing solution of the discrete algebraic Riccati equation*, IEEE Trans. Automat. Control, 20 (1975), pp. 396–399.

[18] M. Pavon and H. K. Wimmer, *Suboptimal Markovian smoothing estimates based on continuous curves of solutions of the algebraic Riccati inequality*, Automatica J. IFAC, 38 (2002), pp. 1017–1025.

[19] A. C. M. Ran and H. L. Trentelman, *Linear quadratic problems with indefinite cost for discrete time systems*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 776–797.

[20] G. Ruckebusch, *Représentations Markoviennes de processus Gaussiens stationaires*, C.R. Acad. Sci. Paris Ser. A, 282 (1976), pp. 649–651.

[21] G. Ruckebusch, *Représentations Markoviennes de processus Gaussiens stationaires et applications statistiques*, in Journées de Statistique des Processus Stochastiques, Proceedings, Grenoble 1977, Lecture Notes in Math. 636, D. Dacunha-Castelle and B. Van Cutsem, eds., Springer, Berlin, 1978, pp. 115–139.

[22] C. Scherer, *The solution set of the algebraic Riccati equation and the algebraic Riccati inequality*, Linear Algebra Appl., 153 (1991), pp. 99–122.

[23] M. A. Shayman, *Geometry of the algebraic Riccati equation, Part* I, SIAM J. Control Optim., 21 (1983), pp. 375–394.

[24] J. C. Willems, *Least-squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, 16 (1971), pp. 621–634.
[25] H. K. Wimmer, *Unmixed solutions of the discrete-time algebraic Riccati equation*, SIAM J. Control Optim., 30 (1992), pp. 867–878.
[26] H. K. Wimmer, *Hermitian solutions of the discrete-time algebraic Riccati equation*, Internat. J. Control, 63 (1996), pp. 921–936.
[27] H. K. Wimmer, *Intervals of solutions of the discrete-time algebraic Riccati equation*, Systems Control Lett., 36 (1999), pp. 207–212.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 6, pp. 2006–2037

LINEAR PROGRAMMING APPROACH TO DETERMINISTIC LONG RUN AVERAGE PROBLEMS OF OPTIMAL CONTROL*

VLADIMIR GAITSGORY† AND SERGEY ROSSOMAKHINE†

Abstract. We establish that deterministic long run average problems of optimal control are "asymptotically equivalent" to infinite-dimensional linear programming problems (LPPs) and we establish that these LPPs can be approximated by finite-dimensional LPPs, the solutions of which can be used for construction of the optimal controls. General results are illustrated with numerical examples.

Key words. long run average optimal control, singularly perturbed control systems, occupational measures, averaging method, linear programming

AMS subject classifications. 34E15, 34C29, 34A60, 93C70

DOI. 10.1137/040616802

1. Introduction and description of the problems. In this paper we show that, under some conditions, deterministic long run average problems of optimal control are "asymptotically equivalent" to infinite-dimensional linear programming problems (LPPs) and we establish that these LPPs can be approximated by finite-dimensional LPPs, the solutions of which can be used for numerical construction of the optimal controls.

Infinite horizon problems of optimal control have been studied intensively in both deterministic and stochastic settings (see Anderson and Kokotovic [3], Arisawa, Ishii, and Lions [5], Bardi and Capuzzo-Dolcetta [10], Bensoussan [12], Carlson, Haurie, and Leizarowitz [14], Colonius and Kliemann [17], Fleming and Soner [21], Grüne [29], Kushner [34], Kushner and Dupuis [35], Vigodner [46], and references therein). In the stochastic setting, the linear programming formulation is a common tool for treating the problems (see, e.g., Basak, Borkar, and Ghosh [11], Borkar [13], Hernandez-Lerma and Lasserre [31], Stockbridge [44], Yin and Zhang [48]). Finite-dimensional approximations of LPPs arising in stochastic optimal control problems were considered by Helmes and Stockbridge [30] and by Mendiondo and Stockbridge [38]. A linear programming approach to long run average optimal control problems in the deterministic setting appears to be new and, to the best of our knowledge, there are no publications devoted to this topic (under different assumptions and for a different problem, a linear programming formulation was discussed in Evans and Gomes [20]). A linear programming approach to deterministic optimal control problems on a finite time interval has been studied in Rubio [42].

Let us introduce the problems that we will be dealing with. Consider the control system

$$\dot{y}(\tau) \;=\; f(u(\tau), y(\tau)), \quad \tau \in [0, S], \tag{1}$$

*Received by the editors October 12, 2004; accepted for publication (in revised form) June 8, 2005; published electronically January 6, 2006.

http://www.siam.org/journals/sicon/44-6/61680.html

†Centre for Industrial and Applied Mathematics, University of South Australia, Mawson Lakes, SA 5095, Australia (v.gaitsgory@unisa.edu.au, s.rossomakhine@unisa.edu.au). The work of the first author was supported by Australian Research Council Discovery-Project grant DP0346099, IREX grant X00106494, and Linkage International grant LX0560049. The work of the second author was supported by Australian Research Council Discovery-Project grant DP0346099.

where the function $f(u,y) : U \times \mathbb{R}^m \to \mathbb{R}^m$ is continuous in (u,y) and satisfies Lipschitz conditions in y; the controls are Lebesgue measurable functions $u(\tau) : [0,S] \to U$ and U is a compact metric space.

A pair $(u(\tau), y(\tau))$ is called *admissible* on the interval $[0,S]$ if (1) is satisfied for almost all $\tau \in [0,S]$ and $y(\tau) \in Y \ \ \forall \tau \in [0,S]$, where Y is a given compact subset of $\mathbb{R}^m$. The pair is called admissible on $[0,\infty)$ if it is admissible on any interval $[0,S]$, $S > 0$.

Let $g(u,y) : U \times \mathbb{R}^m \to \mathbb{R}^1$ be a continuous function. We will be considering the asymptotics of the optimal control problem

$$\frac{1}{S} \inf_{(u(\cdot),y(\cdot))} \int_0^S g(u(\tau), y(\tau)) d\tau \stackrel{\text{def}}{=} G(S), \tag{2}$$

where inf is over all admissible pairs on the interval $[0,S]$. Along with (2), we will be referring to the infinite time horizon optimal control problem

$$\inf_{(u(\cdot),y(\cdot))} \lim_{S\to\infty} \frac{1}{S} \int_0^S g(u(\tau), y(\tau)) d\tau \stackrel{\text{def}}{=} G_\infty, \tag{3}$$

where inf is over all admissible pairs on the interval $[0,\infty)$ such that the limit in the above expression exists. If this inf is sought over the periodic admissible pairs only, that is, over the pairs such that

$$(u(\tau), y(\tau)) = (u(\tau + T), y(\tau + T)) \ \forall \tau \geq 0 \tag{4}$$

for some $T > 0$, then (3) becomes equivalent to a so-called periodic optimization problem (see, e.g., Colonius [15])

$$\inf_{(u(\cdot),y(\cdot))} \frac{1}{T} \int_0^T g(u(\tau), y(\tau)) d\tau \stackrel{\text{def}}{=} G_{per}, \tag{5}$$

where inf is over the length of the time interval T and over the admissible pairs defined on $[0,T]$ which satisfy the periodicity condition $y(0) = y(T)$.

A very special family of admissible pairs on $[0,\infty)$ is that consisting of constant valued controls and corresponding steady state solutions of (1):

$$(u(\tau), y(\tau)) \ = \ (u,y) \ \in \ M \stackrel{\text{def}}{=} \{(u,y) \mid (u,y) \in U \times Y, \ f(u,y) = 0\}. \tag{6}$$

If inf is sought over the admissible pairs from this family, the problem (3) is reduced to

$$\inf_{(u,y)\in M} g(u,y) \stackrel{\text{def}}{=} G_{ss}, \tag{7}$$

which is called a steady state optimization problem. It is easy to see that the optimal values of the above introduced problems satisfy the inequalities

$$\overline{\lim}_{S\to\infty} G(S) \ \leq \ G_\infty \ \leq G_{per} \ \leq \ G_{ss}. \tag{8}$$

The approach that we are developing in the paper is based on a reformulation of problem (2) as the problem of minimization over the set of occupational measures generated on the interval $[0,S]$ by the admissible pairs of (1) and on the fact that this set is proven to converge (as $S \to \infty$) to a set of probability measures characterized by

linear constraints (as has been recently established in [26]). Note that it is the presence of this convergence that constitutes the main difference between our approach and a linear programming approach to deterministic optimal control problems on a finite time interval developed by Rubio [42].

Note also in conclusion that results obtained in the paper have a potential for applications in asymptotic and numerical analysis of singularly perturbed control systems (**SPCS**), which have been the focus of many researchers (see Alvarez and Bardi [1, 2], Artstein [6, 7], Colonius and Fabbri [16], Donchev and Dontchev [19], Gaitsgory [26], Grammel [28], Kabanov and Pergamenshchikov [32], Leizarowitz [36], Naidu [39], and Quincampoix and Watbled [41] for the most recent developments and also for references to earlier results in the area). One such application follows directly from the fact that tending S to infinity in problem (2) is equivalent to tending ϵ to zero in the problem

$$\inf_{(u^\epsilon(\cdot),y^\epsilon(\cdot))} \int_0^1 g(u^\epsilon(t),y^\epsilon(t))dt \stackrel{\text{def}}{=} G(\epsilon) \tag{9}$$

considered on the admissible pairs $(u^\epsilon(\cdot),y^\epsilon(\cdot)) \in U \times Y$ of the SPCS

$$\epsilon\frac{dy^\epsilon(t)}{dt} \;=\; f(u^\epsilon(t),y^\epsilon(t)), \tag{10}$$

where (9) and (10) are obtained from (2) and (1) with $\epsilon \stackrel{\text{def}}{=} \frac{1}{S}$ and with $t = \tau\epsilon, u^\epsilon(t) = u(\frac{t}{\epsilon}), y^\epsilon(t) = y(\frac{t}{\epsilon})$. By formally taking $\epsilon = 0$ in (9)–(10), one obtains the so-called reduced problem, which proves to be equivalent to the steady state optimization problem (7). This implies that the statement about the validity of the equality $\lim_{\epsilon\to 0} G(\epsilon) = G(0)$, which can be interpreted as a weak version of Tichonov's theorem for the SPCS under consideration, is true only if the "less than or equal to" inequalities in (8) are satisfied as exact equalities. More elaborate connections between SPCS and long run average problems of optimal control implying the applicability of results of this paper in dealing with SPCS have been established in Alvarez and Bardi [1, 2], Artstein and Gaitsgory [8], and Gaitsgory [24, 25]; different, Tichonov-theorem-type results can be found in Kokotovic, Khalil, and O'Reilly [33], O'Malley [40], and Veliov [45].

The paper is organized as follows. In section 2, we give the occupational measures formulation of problem (2). In section 3, we show that, as S tends to infinity, the set of occupational measures converges to the set of probability measures with linear constraints, and we introduce the infinite-dimensional LPP defined on this set, which determines the asymptotics of problem (2) (Propositions 2 and 5, Corollaries 3 and 6). In section 4, we establish that the infinite-dimensional LPP can be approximated by a finite-dimensional LPP (Propositions 7 and 9). In section 5, we discuss the possibility of using the solution of the latter to construct an approximation to the solution of the periodic optimization problem (5) and we also illustrate the idea of the construction with two numerical examples. The proofs for sections 3, 4, and 5 are given in section 6.

2. Occupational measures formulation. Let $\mathcal{P}(U \times Y)$ stand for the space of probability measures defined on the Borel subsets of $U \times Y$. Given an arbitrary admissible (on the interval $[0, S]$) pair $(u(\tau), y(\tau))$, one can define a probability measure $\gamma^{(u(\cdot),y(\cdot))} \in \mathcal{P}(U \times Y)$ by taking

$$\gamma^{(u(\cdot),y(\cdot))}(Q) \stackrel{\text{def}}{=} \frac{1}{S}\text{meas}\left\{\tau \;\middle|\; (u(\tau),y(\tau)) \in Q\right\} \tag{11}$$

for any Borel $Q \subset U \times Y$, where meas $\{\cdot\}$ stands for the Lebesgue measure on $[0, S]$. Such a probability measure is called the *occupational measure* generated by the pair $(u(\tau), y(\tau))$. Note that the occupational measure generated by a steady state admissible pair $(u(\tau), y(\tau)) = (u, y) \in M$ (as in (6)) is just the Dirac measure at (u, y).

It is easy to see that (11) is equivalent to the equality

$$\int_{U\times Y} \chi_Q(u,y)\gamma^{(u(\cdot),y(\cdot))}(du,dy) \;=\; \frac{1}{S}\int_0^S \chi_Q(u(\tau),y(\tau))d\tau, \tag{12}$$

where $\chi_Q(\cdot)$ is the indicator function of the set Q: $\chi_Q(u, y) = 1 \;\; \forall (u, y) \in Q$ and $\chi_Q(u, y) = 0 \;\; \forall (u, y) \notin Q$. The validity of (12) for any indicator function leads to the validity of a similar equality for the simple functions (that is, linear combinations of the indicator functions) and, thus, with the help of a standard approximation argument, leads to the validity of the equality

$$\int_{U\times Y} q(u,y)\gamma^{(u(\cdot),y(\cdot))}(du,dy) \;=\; \frac{1}{S}\int_0^S q(u(\tau),y(\tau))d\tau \tag{13}$$

for any continuous function $q(u, y) : U \times \mathbb{R}^m \to \mathbb{R}^1$.

Denote by $\Gamma(S) \subset \mathcal{P}(U \times Y)$ the set of all occupational measures generated by the admissible pairs on the interval $[0, S]$. Using this notation and (13), one can rewrite problem (2) in the equivalent form

$$\inf_{\gamma\in\Gamma(S)} \int_{U\times Y} g(u,y)\gamma(du,dy) \;=\; G(S). \tag{14}$$

In what follows, the convergence properties of $G(S)$ (as S tends to infinity) are established on the basis of the corresponding convergence properties of $\Gamma(S)$. To describe these convergence properties, let us introduce a metric ρ on $\mathcal{P}(U \times Y)$ as follows:

$$\rho(\gamma',\gamma'') \stackrel{\text{def}}{=} \sum_{j=1}^{\infty} \frac{1}{2^j} \left| \int_{U\times Y} q_j(u,y)\gamma'(du,dy) - \int_{U\times Y} q_j(u,y)\gamma''(du,dy) \right| \tag{15}$$

$\forall \gamma', \gamma'' \in \mathcal{P}(U \times Y)$, where $q_j(\cdot), j = 1, 2, \dots$, is a sequence of Lipschitz continuous functions which is dense in the unit ball of $C(U\times Y)$ (the space of continuous functions on $U \times Y$). Note that this metric is consistent with the weak convergence topology of $\mathcal{P}(U \times Y)$. Namely, a sequence $\gamma^k \in \mathcal{P}(U \times Y)$ converges to $\gamma \in \mathcal{P}(U \times Y)$ in this metric if and only if

$$\lim_{k\to\infty} \int_{U\times Y} q(u,y)\gamma^k(du,dy) \;=\; \int_{U\times Y} q(u,y)\gamma(du,dy) \tag{16}$$

for any continuous $q(u, y) : U \times Y \to \mathbb{R}^1$. Note also that, the space $\mathcal{P}(U\times Y)$ is known to be compact in its weak convergence topology and, hence, being equipped with the metric (15), it becomes a compact metric space.

Using the metric ρ, one can define the "distance" $\rho(\gamma, \Gamma)$ between $\gamma \in \mathcal{P}(U \times Y)$ and $\Gamma \subset \mathcal{P}(U\times Y)$ and define the Hausdorff metric $\rho_H(\Gamma_1, \Gamma_2)$ between $\Gamma_1 \subset \mathcal{P}(U\times Y)$ and $\Gamma_2 \subset \mathcal{P}(U \times Y)$ as follows:

$$\rho(\gamma,\Gamma) \stackrel{\text{def}}{=} \inf_{\gamma'\in\Gamma} \rho(\gamma,\gamma'), \qquad \rho_H(\Gamma_1,\Gamma_2) \stackrel{def}{=} \max\left\{ \sup_{\gamma\in\Gamma_1} \rho(\gamma,\Gamma_2), \sup_{\gamma\in\Gamma_2} \rho(\gamma,\Gamma_1) \right\}. \tag{17}$$

The following simple lemma is implied by the definitions above.

LEMMA 1. *Let Γ be a subset of $\mathcal{P}(U \times Y)$.*

(i) *If $\lim_{S\to\infty} \sup_{\gamma\in\Gamma(S)} \rho(\gamma, \Gamma) = 0$, then, for any continuous $q(u, y) : U \times Y \to \mathbb{R}^1$,*

$$\underline{\lim}_{S\to\infty} \inf_{\gamma\in\Gamma(S)} \int_{U\times Y} q(u, y)\gamma(du, dy) \;\geq\; \inf_{\gamma\in\Gamma} \int_{U\times Y} q(u, y)\gamma(du, dy).$$

(ii) *If $\lim_{S\to\infty} \rho_H(\Gamma(S), \Gamma) = 0$, then*

$$\lim_{S\to\infty} \inf_{\gamma\in\Gamma(S)} \int_{U\times Y} q(u, y)\gamma(du, dy) \;=\; \inf_{\gamma\in\Gamma} \int_{U\times Y} q(u, y)\gamma(du, dy).$$

Proof. The proof is obvious. □

3. Infinite-dimensional LPPs. Define the set $W \subset \mathcal{P}(U \times Y)$ by the equation

$$W \stackrel{\text{def}}{=} \{\gamma :\; \gamma \in \mathcal{P}(U \times Y);\; \int_{U\times Y} (\phi'(y))^T f(u, y)\gamma(du, dy) = 0 \;\; \forall \phi(\cdot) \in C^1\}, \tag{18}$$

where C^1 is the space of continuously differentiable functions $\phi(y) : \mathbb{R}^m \to \mathbb{R}^1$ and $\phi'(y)$ is a vector column of partial derivatives (the gradient) of $\phi(y)$.

Note that the set W can be empty. It is easy to see, for example, that W is empty if there exists a continuously differentiable function $\phi(\cdot) \in C^1$ such that

$$\max_{(u,y)\in U\times Y} (\phi'(y))^T f(u, y) \;< 0. \tag{19}$$

The set W is not empty if the set of steady state or periodic admissible pairs is not empty since the occupational measure generated by each such pair is contained in W. In fact, let $(u(\cdot), y(\cdot))$ be a periodic admissible pair (that is, (4) is satisfied with some positive T) and let $\gamma^{(u(\cdot),y(\cdot))}$ be the occupational measure generated by this pair on the interval $[0, T]$. Then, by (13),

$$\int_{U\times Y} (\phi'(y))^T f(u, y)\gamma^{(u(\cdot),y(\cdot))}(du, dy) \;=\; \frac{1}{T}\int_0^T (\phi'(y(\tau)))^T f(u(\tau), y(\tau))d\tau$$

$$= \;\; \frac{\phi(y(T)) - \phi(y(0))}{T} \;=\; 0 \quad \forall \phi(\cdot) \in C^1 \qquad \Rightarrow \qquad \gamma^{(u(\cdot),y(\cdot))} \in W.$$

PROPOSITION 2. *If the set W is empty, then there exists $S_0 > 0$ such that $\Gamma(S)$ is empty for $S \geq S_0$. If $\Gamma(S)$ is not empty for $S > 0$, then W is not empty and*

$$\lim_{S\to\infty} \sup_{\gamma\in\Gamma(S)} \rho(\gamma, W) = 0. \tag{20}$$

Proof. The proof is similar to the corresponding part of the proof of Theorem 2.1(i) in [26]. For the sake of completeness, we have displayed it in section 6. □

Assume that W is not empty and consider the problem

$$\min_{\gamma\in W} \int_{U\times Y} g(u, y)\gamma(du, dy) \stackrel{\text{def}}{=} G^*, \tag{21}$$

where $g(\cdot)$ is the same as in (14) (and the same as in (2)–(7)). It can be easily seen that the set W is convex and compact. Moreover, since both the objective function and the constraints defining W are linear in γ, problem (21) is that of infinite-dimensional linear programming (see, e.g., [4]).

COROLLARY 3. *The lower limit of the optimal values of* (2) *satisfies the inequality*

$$\underline{\lim}_{S\to\infty} G(S) \;=\; \underline{\lim}_{S\to\infty} \inf_{\gamma\in\Gamma(S)} \int_{U\times Y} g(u,y)\gamma(du,dy) \;\geq\; G^*.$$

Proof. The proof follows from Lemma 1(i), Proposition 2, and the validity of the representation (14). □

COROLLARY 4 (criteria of optimality). (i) *If an admissible pair* $(u(\cdot),y(\cdot)) : [0,\infty) \to U\times Y$ *is such that*

$$\lim_{S\to\infty} \frac{1}{S}\int_0^S g(u(\tau),y(\tau))d\tau \;=\; G^*,$$

then this pair is a solution of problem (3) *and* $G_\infty = G^*$.

(ii) *If a periodic (with a period* T*) admissible pair* $(u(\cdot),y(\cdot))$ *is such that*

$$\frac{1}{T}\int_0^T g(u(\tau),y(\tau))d\tau \;=\; G^*,$$

then this pair is a solution of problems (3) *and* (5)*, and also* $G_\infty = G_{per} = G^*$.

(iii) *If a steady state admissible pair* $(u(\tau),y(\tau)=(u,y)\in M$ *(as defined in* (6)*) is such that*

$$g(u,y) \;=\; G^*,$$

then this pair is a solution of problems (3), (5), *and* (7)*, and also* $G_\infty = G_{per} = G_{ss} = G^*$.

Proof. The proof follows from inequalities (8) and Corollary 3. □

Denote by $\mathcal{P}(U)$ the space of probability measures defined on the Borel subsets of U and consider the system

$$\dot y(\tau) \;=\; \bar f(\nu(\tau),y(\tau)), \quad \tau\in[0,S], \tag{22}$$

where $\nu(\tau)\in\mathcal{P}(U)$ are relaxed controls (see [47]) and $\bar f(\nu,u) \stackrel{\text{def}}{=} \int_U f(u,y)\nu(du)$.

A pair $(\nu(\tau),y(\tau))$ will be called *relaxed admissible* on the interval $[0,S]$ if (22) is satisfied for almost all $\tau\in[0,S]$ and $y(\tau)\in Y \;\; \forall\tau\in[0,S]$.

Assumption 1. For any Lipschitz continuous function $q(u,y) : U\times\mathbb{R}^m \to \mathbb{R}^1$,

$$\left|\frac{1}{S}\sup_{(u(\cdot),y(\cdot))}\int_0^S q(u(\tau),y(\tau))d\tau - \frac{1}{S}\sup_{(\nu(\cdot),y(\cdot))}\int_0^S \bar q(\nu(\tau),y(\tau))d\tau\right| \;\stackrel{\text{def}}{=}\; \alpha_q(S) \;\to\; 0 \tag{23}$$

as $S\to\infty$, where $\bar q(\nu,u) \stackrel{\text{def}}{=} \int_U q(u,y)\nu(du)$, with the first sup being over all admissible pairs and the second being over all relaxed admissible pairs.

Remark 1. The fulfillment of Assumption 1 is related to the applicability of Filippov–Wazewski type theorems on Y (see Frankowska and Rampazo [22]). In particular, it is satisfied with $\alpha_q(S)\equiv 0$ if Y is forward invariant with respect to the

solutions of system (1), that is, if for an arbitrary control $u(\tau)$, any solution $y(\tau)$ of (1) with the initial conditions in Y does not leave Y (see, e.g., Theorem 10.4.4 in Aubin and Frankowska [9]). Assumption 1 is also satisfied with $\alpha_q(S) \equiv 0$ if $f(u,y) \equiv f(y)$ (the case of uncontrolled dynamics, with inf's in (2)–(5) being over the admissible trajectories having different initial conditions). In this case U can be formally defined to consist of only one point and systems (1) and (22) are identical.

Assumption 1 is not satisfied if, for example, the set of admissible pairs is empty, while the set of relaxed admissible pairs is not, as in the case when $m = 1$, $f(u,y) = -y + u$, with U consisting of two points $U = \{-1, 1\}$, and Y consisting of one point $Y = \{0\}$.

PROPOSITION 5. *Let $\Gamma(S)$ be nonempty and Assumption 1 be satisfied. Then*

$$\lim_{S\to\infty} \rho_H(co\Gamma(S), W) \;=\; 0, \tag{24}$$

where $co\Gamma(S)$ stands for the convex hull of $\Gamma(S)$.

Proof. The proof of (24) is similar to the proof of Theorem 1(i) in [26], which was established under a stronger assumption implying the validity of Assumption 1. The necessary adjustments for the case under consideration are made in section 6. □

COROLLARY 6. *If Assumption 1 is satisfied, then the limit of the optimal value of (2) exists and is equal to G^*,*

$$\lim_{S\to\infty} G(S) \;=\; G^*. \tag{25}$$

Also, if the solution γ^* of problem (21) is unique, then, for any $\gamma^S \in \Gamma(S)$ such that $\lim_{S\to\infty} \int_{U\times Y} g(u,y)\gamma^S(du,dy) = G^*$,

$$\lim_{S\to\infty} \rho(\gamma^S, \gamma^*) \;=\; 0. \tag{26}$$

Proof. Since

$$\inf_{\gamma\in co\Gamma(S)} \int_{U\times Y} g(u,y)\gamma(du,dy) \;=\; \inf_{\gamma\in\Gamma(S)} \int_{U\times Y} g(u,y)\gamma(du,dy),$$

then, by (14),

$$\inf_{\gamma\in co\Gamma(S)} \int_{U\times Y} g(u,y)\gamma(du,dy) \;=\; G(S).$$

The validity of (25) follows now from Lemma 1(ii) and Proposition 5. The validity of (26) is, in turn, implied by (25) and Proposition 2. □

Note that the solution γ^* of problem (21) can be unique only if it is an extreme point of W (since (21) is an LPP) and that, using (24), one can show (although not shown here) that, for any extreme point γ of W, there exists $\gamma^S \in \Gamma(S)$ such that $\lim_{S\to\infty} \rho(\gamma^S, \gamma) \;=\; 0$.

Let γ^* be a solution of problem (21) which is an extreme point of W and let $\gamma^S \in \Gamma(S)$ satisfy (26). Assume that there exists an admissible pair $(u^{\gamma^*}(\cdot), y^{\gamma^*}(\cdot))$: $[0,\infty) \to U\times Y$ that generates γ^S on any interval $[0,S]$ (we will say that γ^* is generated by the pair on $[0,\infty)$ in this case). Then, for any continuous $q(u,y) : U \times Y \to \mathbb{R}^1$,

$$\lim_{S\to\infty} \frac{1}{S} \int_0^S q(u^{\gamma^*}(\tau), y^{\gamma^*}(\tau))d\tau = \int_{U\times Y} q(u,y)\gamma^*(du,dy) \tag{27}$$

and, in particular, for $q(u,y) = g(u,y)$,

$$\lim_{S\to\infty} \frac{1}{S}\int_0^S g(u^{\gamma^*}(\tau), y^{\gamma^*}(\tau))d\tau = \int_{U\times Y} g(u,y)\gamma^*(du,dy) = G^*.$$

Thus, by Corollary 4(i), this pair will be a solution of problem (3). Also, by Corollary 4(ii), (iii), this pair will be a solution of the periodic optimization problem (5) (and the steady state problem (7)) if it proves to be periodic (and, respectively, steady state).

4. Finite-dimensional approximations. Let $\{\phi_i(\cdot)\ , i = 1,2,\dots\}$ be a sequence of continuously differentiable functions such that any function $\phi(\cdot) \in C^1$ and its gradient $\phi'(\cdot)$ can be simultaneously approximated on Y by linear combinations of functions from $\{\phi_i(\cdot), i = 1,2,\dots\}$ and their corresponding gradients. That is, for any $\phi(\cdot) \in C^1$ and any $\delta > 0$, there exist $\beta_1,\dots,\beta_k$ (real numbers) such that

$$\max_{y\in Y}\left\{\left|\phi(y) - \sum_1^k \beta_i\phi_i(y)\right| + \left\|\phi'(y) - \sum_1^k \beta_i\phi_i'(y)\right\|\right\} \le\ \delta,$$

with $||\cdot||$ being a norm in $\mathbb{R}^m$. An example of such an approximating sequence is the sequence of monomials $y_1^{i_1}\dots y_m^{i_m}$, $i_1,\dots,i_m = 0,1,\dots$, where $y_j (j = 1,\dots,m)$ stands for the jth component of y (see, e.g., [37]).

Using the system $\{\phi_i(\cdot), i = 1,2,\dots\}$, one can represent the set W in the form of a countable system of equations:

$$W = \left\{\gamma \mid \gamma \in \mathcal{P}(U\times Y); \int_{U\times Y} (\phi_i'(y))^T f(u,y)\gamma(du,dy) = 0,\ i = 1,2,\dots\right\}. \tag{28}$$

Let us assume that the gradients $\phi_i'(\cdot), i = 1,\dots,N$, are linearly independent on any open ball B in $\mathbb{R}^m$ (that is, the equality $\sum_{i=1}^N v_i\phi_i'(y) = 0\ \forall y \in B$ can be valid only with $v_i = 0,\ i = 1,\dots,N$) and let us define the set W_N by truncation of the system of equations in (28):

$$W_N \stackrel{\text{def}}{=} \left\{\gamma \mid \gamma \in \mathcal{P}(U\times Y); \int_{U\times Y} (\phi_i'(y))^T f(u,y)\gamma(du,dy) = 0,\ i = 1,2,\dots,N\right\}. \tag{29}$$

Consider the LPP

$$\min_{\gamma\in W_N} \int_{U\times Y} q(u,y)\gamma(du,dy) \stackrel{\text{def}}{=} G_N. \tag{30}$$

Note that W_N is a convex and compact subset of $\mathcal{P}(U\times Y)$ and that $W \subset W_N$, which implies

$$G^* \ge G_N. \tag{31}$$

Note also that the set W_N is empty if (19) is true with $\phi(y) \stackrel{\text{def}}{=} \sum_{i=1}^N v_i\phi_i(y)$, where v_i are real numbers.

PROPOSITION 7. *The set W is not empty if and only if there exists $N_0 \ge 1$ such that W_N is not empty for $N \ge N_0$. If W is not empty, then*

$$\lim_{N\to\infty} \rho_H(W_N, W) = 0 \tag{32}$$

and

$$\lim_{N\to\infty} G_N = G^*. \tag{33}$$

Also, if γ_N is a solution of problem (30) *and $\lim_{N'\to\infty}\rho(\gamma_{N'},\gamma)=0$ for some subsequence of integers N' tending to infinity, then γ is a solution of* (21). *If the solution γ^* of problem* (21) *is unique, then, for any solution γ_N of* (30), $\lim_{N\to\infty}\rho(\gamma_N,\gamma^*)=0$.

Proof. By Lemma 1(ii), the validity of (33) follows from the validity of (32). The other statements included in the proposition readily follow from (32) and (33). The validity of (32) is established in section 6. □

Let us introduce another assumption which we need to consider.

Assumption 2. The inequality

$$\sum_{i=1}^{N} v_i(\phi_i'(y))^T f(u,y) \le 0 \quad \forall (u,y)\in U\times Y \tag{34}$$

is valid only with $v_i = 0 \ \ \forall\ i = 1,\dots,N$.

This assumption is satisfied if there exists a closed subset Y^* of Y with a nonempty interior such that from the validity of (34) it follows that $\sum_{i=1}^{N} v_i\phi_i'(y) = 0\ \forall y\in Y^*$ (the equality of v_i to zero is implied in this case by linear independence of $\phi_i'(\cdot)$). The existence of such Y^* can be guaranteed, for instance, in two cases specified in the statement below.

Proposition 8. *A closed set $Y^*\subset Y$ with a nonempty interior such that from the fact that*

$$(\phi'(y))^T f(u,y)\le 0 \quad \forall (u,y)\in U\times Y \tag{35}$$

it follows that $\phi'(y)=0\ \forall y\in Y^*$ exists if one of the following conditions is satisfied:

(i) *The set $f(y,U)\stackrel{\text{def}}{=}\{x\in\mathbb{R}^m : x = f(y,u),\ u\in U\}$ is convex for $y\in Y$, and there exists $\bar y\in \operatorname{int} Y$ such that*

$$0\in \operatorname{int} f(\bar y, U), \tag{36}$$

where "int" stands for the interior of the corresponding set.

(ii) *There exists $Y^0\subset Y$ such that the closure of Y^0 has a nonempty interior and such that any two points in Y^0 are connected by an admissible trajectory. That is, for any $y',y''\in Y^0$, there exists an admissible pair $(u(\tau),y(\tau))$ defined on some interval $[0,S]$ such that $y(0)=y'$ and $y(S)=y''$.*

Proof. The proof is in section 6. □

Remark 2. Note that Y^0 in Proposition 8(ii) can be equal to Y in which case Y is a subset of complete controllability of system (1) (see [29]). Note also that both Assumptions 1 and 2 can be easily verified if there exist positive definite matrices A_1 and A_2 such that

$$\begin{aligned}&(f(u,y')-f(u,y''))^T A_1(y'-y'')\\ &\quad\le -(y'-y'')^T A_2(y'-y'')\quad \forall y',y''\in\mathbb{R}^m,\ \forall u\in U.\end{aligned} \tag{37}$$

The latter is a Liapunov-type stability condition that implies the validity of Assumption 3.1 in [24] and, thus, guarantees the existence of a compact set $Y^*\subset\mathbb{R}^m$, which is forward invariant with respect to the solutions of system (1) and which is the global attractor for the solutions of this system starting outside Y^* (Theorem 3.1(ii)

in [24]). The existence of such Y^* leads to the fulfillment of Assumption 1 in case $Y^* \subset Y$. Also, the set Y^* contains all periodic and steady state solutions of the system (Lemma 3.1 in [24]) and, moreover, it can be shown that Y^* is equal to the closure of the set of all points belonging to the periodic orbits. The latter implies the validity of Proposition 8(ii) (and, hence, the validity of Assumption 2) if the interior of Y^* is not empty. Condition (37) is satisfied, for example, if system (1) is linear (that is, $f(u,y) = Ay + Du$, with $u \in U \subset \mathbb{R}^n$ and A, D being matrices of the corresponding dimensions) and if the eigenvalues of A have negative real parts. The nonemptiness of the interior of Y^* can be guaranteed in this case if U has a nonempty interior and if the Kalman controllability matrix $\{D, AD, \dots, A^{m-1}D\}$ has the rank m.

Assume that, for any $\Delta > 0$, Borel sets $Q^{\Delta}_{l,k} \subset U \times Y (l = 1, \dots, L^{\Delta}, k = 1, \dots, K^{\Delta})$ (called cells in what follows) are defined in such a way that two different cells do not intersect, the union of all cells is equal to $U \times Y$ and

$$\sup_{(u,y) \in Q^{\Delta}_{l,k}} ||(u,y) - (u_l, y_k)|| \le c\Delta, \quad c = \text{const}, \tag{38}$$

for some point $(u_l, y_k) \in Q^{\Delta}_{l,k}$, where, for simplicity of notation, it is assumed (from now on) that U is a compact subset of $\mathbb{R}^n$ and $||\cdot||$ stands for a norm in $\mathbb{R}^{n+m}$. Fix these points (u_l, y_k) $(l = 1, \dots, L^{\Delta},\ k = 1, \dots, K^{\Delta})$ and define a polyhedral set $W^{\Delta}_N \subset \mathbb{R}^{L^{\Delta} + K^{\Delta}}$ by the equation

$$W^{\Delta}_N \stackrel{\text{def}}{=} \left\{ \gamma = \{\gamma_{l,k}\} \ge 0 \ : \ \sum_{l,k} \gamma_{l,k} = 1, \right. \\ \left. \sum_{l,k} (\phi_i'(y_k))^T f(u_l, y_k) \gamma_{l,k} = 0,\ i = 1, 2, \dots, N \right\}, \tag{39}$$

where $\sum_{l,k} \stackrel{\text{def}}{=} \sum_{l=1}^{L^{\Delta}} \sum_{k=1}^{K^{\Delta}}$. Consider a finite-dimensional LPP

$$\min_{\gamma \in W^{\Delta}_N} \sum_{l,k} \gamma_{l,k} g(u_l, y_k) \stackrel{\text{def}}{=} G^{\Delta}_N. \tag{40}$$

Note that the set W^{Δ}_N is the set of probability measures on $U \times Y$ which assign nonzero probabilities only to the points (u_l, y_k), and, as such,

$$W^{\Delta}_N \subset W_N \quad \Rightarrow \quad G^{\Delta}_N \ge G_N. \tag{41}$$

PROPOSITION 9. *Let Assumption* 2 *be satisfied. Then the set W_N is not empty if and only if there exists $\Delta_0 > 0$ such that W^{Δ}_N is not empty for $\Delta \le \Delta_0$. If W_N is not empty, then*

$$\lim_{\Delta \to 0} \rho_H(W^{\Delta}_N, W_N) = 0 \tag{42}$$

and

$$\lim_{\Delta \to 0} G^{\Delta}_N = G_N. \tag{43}$$

Also, if γ^{Δ}_N is a solution of problem (40) *and $\lim_{\Delta' \to 0} \rho(\gamma^{\Delta'}_N, \gamma_N) = 0$ for some sequence of Δ' tending to zero, then γ_N is a solution of* (30). *If the solution γ_N of problem* (30) *is unique, then, for any solution γ^{Δ}_N of* (40), $\lim_{\Delta \to 0} \rho(\gamma^{\Delta}_N, \gamma_N) = 0$.

Proof. The proof is in section 6. □

5. Numerical solution of periodic optimization problems. Let us assume that a solution γ^* of problem (21) is unique and that it is generated by a T-periodic admissible pair $(u^{\gamma^*}(\cdot), y^{\gamma^*}(\cdot))$ (see Remark 3 about these assumptions below). Note that, due to Corollary 4(ii), this pair will be a solution of the periodic optimization problem (5). Let

$$\Theta \stackrel{\text{def}}{=} \{(u,y) \ : \ (u,y) = (u^{\gamma^*}(\tau), y^{\gamma^*}(\tau)) \text{ for some } \tau \in [0,T]\}. \tag{44}$$

This set can be considered as the graph of the optimal feedback control function, which is defined on the optimal state trajectory $\mathcal{Y} \stackrel{\text{def}}{=} \{y \ : \ (u,y) \in \Theta\}$ by the equation $\psi(y) \stackrel{\text{def}}{=} u \ \ \forall (u,y) \in \Theta$. For the definition of $\psi(\cdot)$ to make sense, it is assumed that the set Θ is such that from the fact that $(u',y) \in \Theta$ and $(u'',y) \in \Theta$ it follows that $u' = u''$ (this assumption is satisfied if the closed curve defined by $y^{\gamma^*}(\tau),\ \tau \in [0,T]$, does not intersect itself).

Let $\gamma_N^\Delta \stackrel{\text{def}}{=} \{\gamma_{l,k}^\Delta\}$ be a basic solution of the finite-dimensional LPP (40), that is, a solution of (40) which is an extreme point of W_N^Δ. Let

$$\Theta_N^\Delta \stackrel{\text{def}}{=} \{(u_l, y_k) \ : \ \gamma_{l,k}^\Delta > 0\}, \quad \mathcal{Y}_N^\Delta \stackrel{\text{def}}{=} \{y \ : \ (u,y) \in \Theta_N^\Delta\}, \quad \psi_N^\Delta(y) \stackrel{\text{def}}{=} u \ \forall (u,y) \in \Theta_N^\Delta, \tag{45}$$

where again it is assumed that from the fact that $(u',y) \in \Theta_N^\Delta$ and $(u'',y) \in \Theta_N^\Delta$ it follows that $u' = u''$. Note that the set Θ_N^Δ (and the set $\mathcal{Y}_N^\Delta$) can contain no more than $N+1$ elements since γ_N^Δ, being a basic solution of the LPP (40), has no more than $N+1$ positive elements (see, e.g., [18, p. 81]).

The two propositions below establish that the set Θ_N^Δ converges (in the specified sense) to the set Θ, thus leading to the corresponding convergences of $\mathcal{Y}_N^\Delta$ to $\mathcal{Y}$ and of $\psi_N^\Delta(y)$ to $\psi(y)$.

We will be using the following notation. B will stand for the open unit ball in $\mathbb{R}^{n+m}$: $B \stackrel{\text{def}}{=} \{(u,y) \ : \ ||(u,y)|| < 1\}$ and, for any $Q \subset U \times Y$, $\gamma_N^\Delta(Q)$ will denote the γ_N^Δ measure of Q: $\gamma_N^\Delta(Q) \stackrel{\text{def}}{=} \sum_{(u_l,y_k) \in Q \cap \Theta_N^\Delta} \gamma_{l,k}^\Delta$.

PROPOSITION 10. *Let Assumptions* 1 *and* 2 *be satisfied and let* γ^* *be the unique solution of* (21). *Then, corresponding to an arbitrary small* $r > 0$ *and arbitrary small* $\delta > 0$, *there exists* N_0 *such that, for* $N \geq N_0$ *and* $\Delta \leq \Delta_N$ *(*Δ_N *is positive and may depend on* N*),*

$$\gamma_N^\Delta(\Theta_N^\Delta / (\Theta + rB)) \ < \ \delta, \tag{46}$$

$$\Theta_N^{\Delta,\delta} \ \subset \ \Theta + rB, \tag{47}$$

where $\Theta_N^{\Delta,\delta} \stackrel{\text{def}}{=} \{(u_l, y_k) \ : \ \gamma_{l,k}^\Delta \geq \delta\}$.

Proof. The proof is in section 6. □

Assumption 3. For any $(u,y) \in cl\Theta$ (the closure of Θ) and any $r > 0$, the set $B_r(u,y) \stackrel{\text{def}}{=} ((u,y) + rB) \cap (U \times Y)$ has a nonzero γ^*-measure: $\gamma^*(B_r(u,y)) > 0$.

Note that this assumption is satisfied if the optimal control function $u^{\gamma^*}(\cdot) : [0,T] \to U$ is piecewise continuous and at every discontinuity point τ the value of $u^{\gamma^*}(\tau)$ is equal to either the limit from the left ($u^{\gamma^*}(\tau) = \lim_{\tau' \to \tau -} u^{\gamma^*}(\tau')$) or the limit from the right ($u^{\gamma^*}(\tau) = \lim_{\tau' \to \tau +} u^{\gamma^*}(\tau')$).

PROPOSITION 11. *Let the conditions of Proposition* 10 *and Assumption* 3 *be satisfied. Then, corresponding to an arbitrary small* $r > 0$*, there exists* N_0 *such that, for* $N \geq N_0$ *and* $\Delta \leq \Delta_N$ *(*Δ_N *is positive and may depend on* N*),*

$$\Theta \subset \Theta_N^\Delta + rB. \tag{48}$$

Proof. The proof is in section 6. □

Based on the consideration above, one can propose the following steps to construct an approximate solution to the periodic optimization problem (5):

(1) Find a basic solution γ_N^Δ and the optimal value G_N^Δ of the LPP (40) for N large and Δ small enough; the values of N and Δ can be identified as being, respectively, *large enough* and *small enough* if a further increase of N and a reduction of Δ lead only to insignificant changes of the optimal value G_N^Δ and, thus, the latter can be considered to be approximately equal to G^* (see Propositions 7 and 9).

(2) Define $\Theta_N^\Delta, \mathcal{Y}_N^\Delta, \psi_N^\Delta(y)$ as in (45). Note that, as follows from Propositions 10 and 11, if γ^* is the unique solution of (21) and it is generated by a periodic admissible pair, then one can expect that the points of $\mathcal{Y}_N^\Delta$ will be concentrated around a closed curve being the optimal state trajectory, while $\psi_N^\Delta(y)$ will give a pointwise approximation to the optimal feedback control.

(3) Extrapolate the function $\psi_N^\Delta(y)$ to some neighborhood of $\mathcal{Y}_N^\Delta$ and integrate system (1) starting from an initial point $y(0) \in \mathcal{Y}_N^\Delta$ and using the extrapolation of $\psi_N^\Delta(y)$ as the feedback control. One can expect that, thus, the obtained solution of the system will return to a small vicinity of the starting point $y(0)$ and it will be possible to identify the end point of the integration period, T^Δ, as the moment the solution enters this vicinity.

(4) Adjust the initial condition and/or control to obtain a periodic admissible pair $(u^\Delta(\tau), y^\Delta(\tau))$ defined on the interval $[0, T^\Delta]$. Find the integral

$$\frac{1}{T^\Delta} \int_0^{T^\Delta} g(u^\Delta(\tau), y^\Delta(\tau)) d\tau$$

and compare its value with G_N^Δ. If this value proves to be close to G_N^Δ, then, by Corollary 4(ii), the constructed admissible pair is a "good" approximation to the solution of the periodic optimization problem (5).

Remark 3. Under certain conditions (e.g., under the conditions mentioned in Remark 2), the set of occupational measures generated by periodic regimes is dense in W and $G^* = G_\infty = G_{per}$ (compare with Corollary 4). If this is the case, then the assumption that there exists a solution γ^* of problem (21), which is generated by a periodic admissible pair, is equivalent to the assumption that there exists a solution of the periodic optimization problem (5), and the assumption that γ^* is a unique solution of problem (21) implies that all solutions of (5) generate the same occupational measure (namely, γ^*). Note that these assumptions are difficult to verify and one may attempt to use the above steps to find an approximate solution of (5) without such a verification. If, as the result of executing these steps, a periodic admissible pair that gives the value of the objective function close to G_N^Δ is constructed, then one can consider this pair as an approximate solution to problem (5) and use it, if necessary, for further analysis of the existence and structure of the "exact" solution.

Let us illustrate the construction with the following two examples.

Example 1. Let k and ω be positive parameters such that

$$\omega > 1, \qquad k\omega < 1. \tag{49}$$

Consider a differential equation

$$\ddot{x}(\tau) + k\dot{x}(\tau) + \omega^2 x(\tau) = u(\tau), \tag{50}$$

where $x(\tau)$ and $u(\tau)$ are scalars and $u(\tau) \in [-1, 1]$. Via a standard replacement of variables (i.e., $x(\tau) = y_1(\tau)$ and $\dot{x}(\tau) = y_2(\tau)$), equation (50) is reduced to the system of the form (1) with

$$y \stackrel{\text{def}}{=} (y_1, y_2), \qquad f(u, y) \stackrel{\text{def}}{=} (\ y_2,\ -\omega^2 y_1 - k y_2 + u\ ), \qquad U = [-1, 1].$$

Since this system is linear and stable, condition (37) is satisfied and, hence, the system has a forward invariant set $Y^* \subset \mathbb{R}^2$ that is a global attractor for its solutions. The Kalman controllability matrix of the system has rank 2 and, consequently, the interior of Y^* is not empty. Thus, as follows from Remark 2, both Assumptions 1 and 2 are satisfied if Y is such that it contains Y^* (in what follows this is achieved by choosing Y large enough). Also, all periodic and steady state solutions of the system are contained in Y^*, which means that all such solutions are admissible, with the set of steady state admissible pairs (see (6)) being, in this case, equal to

$$M = \left\{(u, y)\ :\ u \in [-1, 1];\ y = (y_1, y_2),\ y_1 = \frac{u}{\omega^2},\ y_2 = 0\right\}.$$

Take

$$g(u, y) \stackrel{\text{def}}{=} u^2 - y_1^2 \tag{51}$$

and consider the steady state optimization problem (7). By the first inequality in (49), its solution and the optimal value are $u = 0, y_1 = y_2 = 0, G_{ss} = 0$. It is easy to verify that this steady state solution is not optimal in the corresponding periodic optimization problem (5). To see this, it is enough to consider the $\frac{2\pi}{\omega}$-periodic admissible pair $(u(\tau), y(\tau))$, with $u(\tau) = \cos(\omega\tau)$ and $y(\tau) = (\frac{1}{\omega k}\sin(\omega\tau), \frac{1}{k}\cos(\omega\tau))$. The value of the objective function obtained on this pair is

$$\tilde{G} \stackrel{\text{def}}{=} \frac{\omega}{2\pi}\int_0^{\frac{2\pi}{\omega}} \left(\cos^2(\omega\tau) - \frac{1}{\omega^2 k^2}\sin^2(\omega\tau)\right) d\tau \ = \ \frac{1}{2}\left(1 - \frac{1}{\omega^2 k^2}\right) \ < 0. \tag{52}$$

The last inequality follows from the second inequality in (49), which postulates smallness of the "friction coefficient" k compared to the "proper frequency" ω and, thus, makes it possible to diminish the value of the objective function via the resonance oscillations of the state variables (such an interpretation of the example was given by Pervozvanskii; see Examples 3.2 and 4.2 in [24]).

Let us demonstrate numerical results obtained with the use of the proposed linear programming approach for the case when $k = 0.3$ and $\omega = 2$. Note that, for these values of the parameters, $\tilde{G} = \frac{1}{2}(1 - \frac{1}{0.36}) \approx -0.889$.

Let us take $Y \stackrel{\text{def}}{=} \{(y_1, y_2) \mid y_i \in [-5, 5], i = 1, 2\}$ (it is straightforward to verify that $Y^* \subset Y$ in this case) and define

$$u_i \stackrel{\text{def}}{=} -1 + i\Delta, \quad y_{1,j} \stackrel{\text{def}}{=} -5 + j\Delta, \quad y_{2,k} \stackrel{\text{def}}{=} -5 + k\Delta, \tag{53}$$

where $i = 0, 1, \ldots, \frac{2}{\Delta}$ and $j, k = 0, 1, \ldots, \frac{10}{\Delta}$ (Δ being chosen in such a way that $\frac{2}{\Delta}$ is integer). Using a slightly different system of notation (adjusted to the case under

consideration and to the grid defined by (53)) and using monomials as the functions defining the constraints in (39), one can rewrite the LPP (40) in the form

$$\min_{\gamma\in W_N^\Delta}\sum_{i,j,k}((u_i)^2-(y_{1,j})^2)\gamma_{i,j,k}\stackrel{\text{def}}{=}G_N^\Delta, \tag{54}$$

with

(55)

$$W_N^\Delta\stackrel{\text{def}}{=}\left\{\gamma=\{\gamma_{i,j,k}\}\ge 0\colon \sum_{i,j,k}\gamma_{i,j,k}=1,\sum_{i,j,k}(\phi'_{l_1,l_2}(y_{1,j},y_{2,k}))^T f(u_i,y_{1,j},y_{2,k})\gamma_{i,j,k}=0,\right.$$
$$\left. l_1,l_2=0,1,\dots,N\right\},$$

where $\phi_{l_1,l_2}(y_1,y_2)\stackrel{\text{def}}{=}y_1^{l_1}y_2^{l_2}$. Problem (54) was solved for $N=7$ and $N=10$ and for $\Delta=0.2,0.1,0.05,0.025,0.0125$ (note that in both Examples 1 and 2 that follow we used ILOG CPLEX 8.0. (http://www.ilog.com) as a linear programming solver). The optimal values of the LPPs obtained with these values of the parameters are (respectively) $G_7^{0.2}\approx -1.312$, $G_7^{0.1}\approx -1.329$, $G_7^{0.05}\approx -1.331$, $G_7^{0.025}\approx -1.331$, $G_7^{0.0125}\approx -1.331$, and $G_{10}^{0.2}\approx -1.280$; $G_{10}^{0.1}\approx -1.325$, $G_{10}^{0.05}\approx -1.326$, $G_{10}^{0.025}\approx -1.327$, $G_{10}^{0.0125}\approx -1.327$. On the basis of these results and Proposition 9 (see (42)) one may conclude that $G_{10}=\lim_{\Delta\to 0}G_{10}^\Delta\approx -1.327$. Hence, by (31), $G^*\ge -1.327$ (within the given proximity). From Corollary 4(ii) (see also (8) and Corollary 3) it now follows that, if for some admissible T-periodical pair $(u(\tau),y(\tau))$,

$$\frac{1}{T}\int_0^T(u^2(\tau)-y_1^2(\tau))d\tau\ \approx -1.327, \tag{56}$$

then this pair is an approximate solution of problems (3) and (5).

Let $\{\gamma_{i,j,k}^{N,\Delta}\}$ stand for the solution of problem (54). The sets Θ_N^Δ and $\mathcal{Y}_N^\Delta$ can then be represented in the form

$$\Theta_N^\Delta=\{(u_i,y_{1,j},y_{2,k})\ :\ \gamma_{i,j,k}^{N,\Delta}\ne 0\},\qquad \mathcal{Y}_N^\Delta=\left\{(y_{1,j},y_{2,k})\ :\ \sum_i\gamma_{i,j,k}^{N,\Delta}\ \ne 0\right\}.$$

Let us mark with dots the points on the plane (y_1,y_2) which belong to $\mathcal{Y}_N^\Delta$ for $N=10$ and $\Delta=0.0125$. The result of such marking is depicted in Figure 1.

Figure 1 clearly identifies a closed curve, which one can expect to be an approximation to the optimal state trajectory. As can be seen from this figure, the value of y_1 is uniquely determined by the value of y_2 for $(y_1,y_2)\in\mathcal{Y}_N^\Delta$ with $y_1\ge 0$ and for $(y_1,y_2)\in\mathcal{Y}_N^\Delta$ with $y_1<0$. Having this in mind, let us mark with dots the points $(u_i,y_{2,k})$ on the plane (u,y_2) for which $\gamma_{i,j,k}^{N,\Delta}\ne 0$ and $y_{1,j}\ge 0$ (Figure 2) and, also, the points for which $\gamma_{i,j,k}^{N,\Delta}\ne 0$ and $y_{1,j}<0$ (Figure 3).

The points marked with the dots in Figure 2 define $\psi_N^\Delta(y)\stackrel{\text{def}}{=}\psi_N^\Delta(y_1,y_2)$ as a function of y_2 (denoted as $\bar\psi_N^\Delta(y_2)$) for $y_1\ge 0$, and the points marked with the dots in Figure 3 define $\psi_N^\Delta(y)\stackrel{\text{def}}{=}\psi_N^\Delta(y_1,y_2)$ as another function of y_2 (denoted as $\bar{\bar\psi}_N^\Delta(y_2)$) for $y_1<0$. Note that in both cases $(y_1,y_2)\in\mathcal{Y}_N^\Delta$. Let us extend the definition of

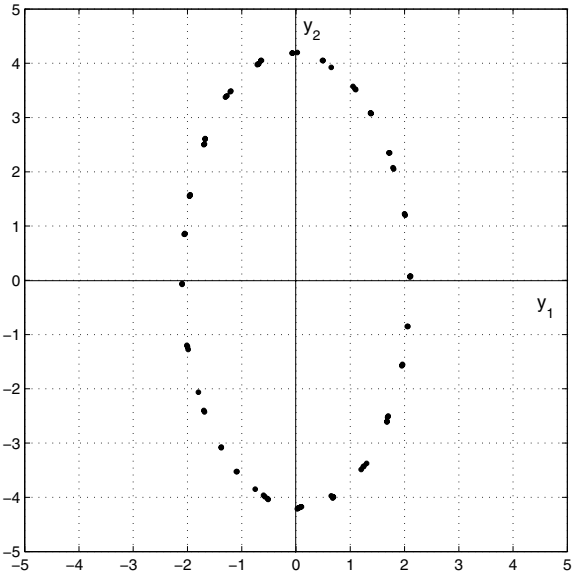


FIG. 1. $\mathcal{Y}_N^\Delta$.

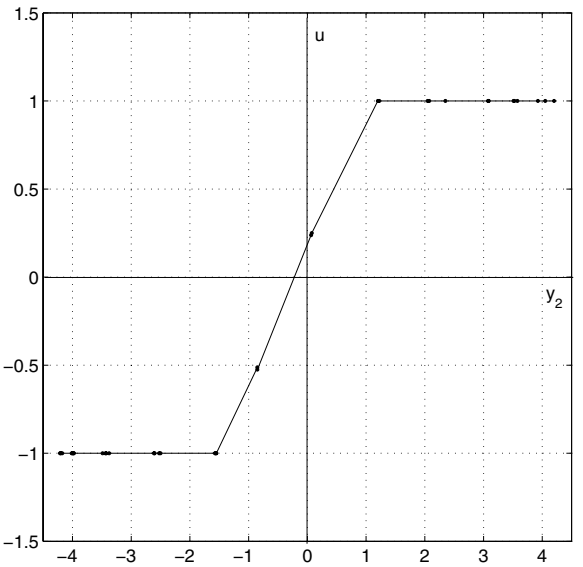


FIG. 2. $\psi_N^\Delta(y_1, y_2) = \bar{\psi}_N^\Delta(y_2)$ *for* $y_1 \geq 0$ *and* $(y_1, y_2) \in \mathcal{Y}_N^\Delta$.

$\psi_N^\Delta(y_1, y_2)$ by connecting the points in Figures 2 and 3 with piecewise linear functions (we will denote this extension also as $\psi_N^\Delta(y_1, y_2)$) and integrate the system with the feedback control thus obtained and with the initial condition being at one of the points marked in Figure 1. Denote by $\tilde{y}^\Delta(\tau) = (\tilde{y}_1^\Delta(\tau), \tilde{y}_2^\Delta(\tau))$ the resulting solution of the system and by $u^\Delta(\tau) = \psi_N^\Delta(\tilde{y}_1^\Delta(\tau), \tilde{y}_2^\Delta(\tau))$ the corresponding open loop control. The function $\tilde{y}^\Delta(\tau)$ proves to be nonperiodic but it returns to a small vicinity of $\tilde{y}^\Delta(0)$

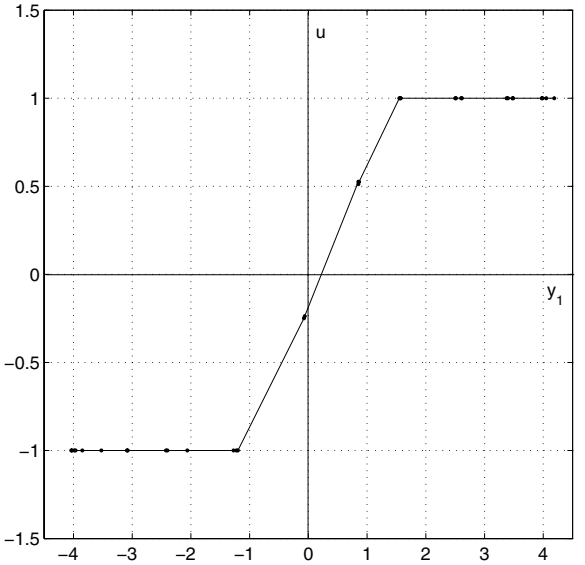


FIG. 3. $\psi_N^\Delta(y_1, y_2) = \bar{\psi}_N^\Delta(y_2)$ *for* $y_1 < 0$ *and* $(y_1, y_2) \in \mathcal{Y}_N^\Delta$.

with $\tau \approx 3.16$. Take $T^\Delta \stackrel{\text{def}}{=} 3.16$ and denote by $y^\Delta(\tau)$ the solution of the system which is obtained when applying the control $u^\Delta(\tau)$ on the interval $[0, T^\Delta]$ and which satisfies the periodicity condition $y^\Delta(0) = y^\Delta(T^\Delta)$. Note that such a solution exists, it is unique and, for the system under consideration, it can be easily found numerically (see, e.g., [23, p. 39]). The periodic admissible pair $(u^\Delta(\tau), y^\Delta(\tau))$ that has been constructed by following the indicated steps is shown in Figures 4 and 5.

The value of the objective function calculated on this pair is approximately equal to -1.324. Comparing it with (56), one can see that it is close to the optimal one and, hence, the pair $(u^\Delta(\tau), y^\Delta(\tau))$ can be considered to be an approximate solution to problem (5). Let us emphasize that we have not verified the assumptions that the solution of the periodic optimization problem (5) exists and that it is unique. Based on the form of the obtained approximate solution, one may conjecture that these assumptions are satisfied in the given example.

Example 2. Consider system (1) with

$$y \stackrel{\text{def}}{=} (y_1, y_2), \qquad u \stackrel{\text{def}}{=} (u_1, u_2), \qquad f(u, y) \stackrel{\text{def}}{=} (\, -y_1 + u_1,\ -y_2 + u_2)$$

(that is, $n = 2$ and $m = 2$) and with

$$U \stackrel{\text{def}}{=} [-1, 1] \times [-1, 1], \qquad Y \stackrel{\text{def}}{=} [-1, 1] \times [-1, 1].$$

As in Example 1, the system under consideration is linear and stable, and, hence, it has a forward invariant set Y^* which is also a global attractor of its solutions. Moreover, it can be easily verified that Y^* coincides with Y introduced above. This implies that both Assumptions 1 and 2 are satisfied (see Remarks 1 and 2).

Let the function $g(u, y)$ be defined by the equation

$$g(u, y) \stackrel{\text{def}}{=} -y_1 u_2 + y_2 u_1 \tag{57}$$

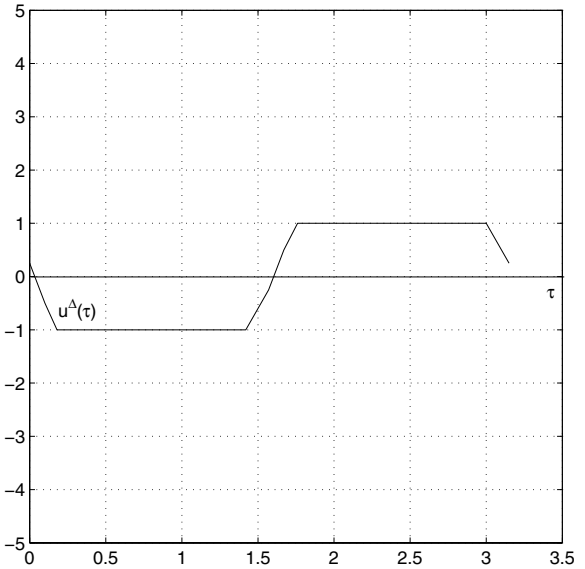


FIG. 4. $u^{\Delta}(\tau)$.

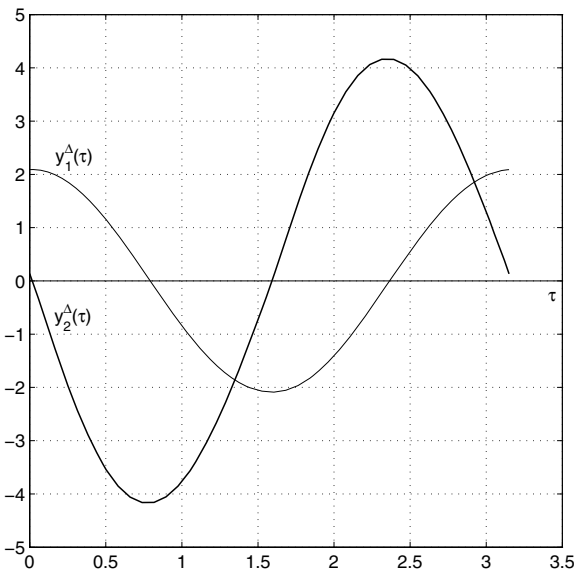


FIG. 5. $y_1^{\Delta}(\tau)$ *and* $y_2^{\Delta}(\tau)$.

and let

$$u_{1,i}^{\Delta} \stackrel{\text{def}}{=} -1+i\Delta, \quad u_{2,l}^{\Delta} \stackrel{\text{def}}{=} -1+l\Delta, \quad y_{1,j}^{\Delta} \stackrel{\text{def}}{=} -1+j\Delta, \quad y_{2,k}^{\Delta} \stackrel{\text{def}}{=} -1+k\Delta, \tag{58}$$

where $i,l,j,k = 0,1,\dots,\frac{2}{\Delta}$ (Δ is such that $\frac{2}{\Delta}$ is integer). The LPP (40) takes the

form

$$\min_{\gamma\in W_N^\Delta}\sum_{i,l,j,k}(-y_{1,j}^\Delta u_{2,l}^\Delta + y_{2,k}^\Delta u_{1,i}^\Delta)\gamma_{i,l,j,k} \stackrel{\text{def}}{=} G_N^\Delta, \tag{59}$$

where

(60)

$$W_N^\Delta \stackrel{\text{def}}{=} \left\{ \gamma = \{\gamma_{i,l,j,k}\} \geq 0 \ : \ \sum_{i,l,j,k}\gamma_{i,l,j,k} = 1, \right.$$
$$\left. \sum_{i,l,j,k}(\phi'_{l_1,l_2}(y_{1,j},y_{2,k}))^T f(u_{1,i},u_{2,l},y_{1,j},y_{2,k})\gamma_{i,l,j,k} = 0, l_1,l_2 = 0,1,\dots,N \right\},$$

with $\phi_{l_1,l_2}(y_1,y_2) \stackrel{\text{def}}{=} y_1^{l_1}y_2^{l_2}$.

Problem (59) was solved for $N = 10$ and for $\Delta = 0.2, 0.1, 0.05, 0.025$, 0.0125, 0.00625, and 0.003125. The optimal values of the LPPs obtained with these parameters are $G_{10}^{0.2} \approx -0.7035$, $G_{10}^{0.1} \approx -0.7579$, $G_{10}^{0.05} \approx -0.7671$, $G_{10}^{0.025} \approx -0.7678$, $G_{10}^{0.0125} \approx -0.7679$, $G_{10}^{0.00625} \approx -0.7679$, $G_{10}^{0.003125} \approx -0.7679$.

Similarly to Example 1, one may conclude that $G^* \geq G_{10} = \lim_{\Delta\to 0} G_{10}^\Delta \approx -0.7679$. Consequently, by Corollary 4(ii), if for some admissible T-periodic pair $(u(\tau), y(\tau))$,

$$\frac{1}{T}\int_0^T (-y_1(\tau)u_2(\tau) + y_2(\tau)u_1(\tau))d\tau \ \approx -0.7679, \tag{61}$$

then this pair is an approximate solution of the periodic optimization problem under consideration.

Let $\{\gamma_{i,l,j,k}^{N,\Delta}\}$ be the solution of problem (59). Then

$$\Theta_N^\Delta = \{(u_{1,i}^\Delta, u_{2,l}^\Delta, y_{1,j}^\Delta, y_{2,k}^\Delta) \ : \ \gamma_{i,l,j,k}^{N,\Delta} \neq 0\}, \qquad \mathcal{Y}_N^\Delta = \left\{ (y_{1,j}^\Delta, y_{2,k}^\Delta) \ : \ \sum_{i,l}\gamma_{i,l,j,k}^{N,\Delta} \ \neq 0 \right\}.$$

Figure 6 represents the result of marking with dots the points on the plane (y_1, y_2) which belong to $\mathcal{Y}_N^\Delta$ for $N = 10$ and $\Delta = 0.003125$.

The image created by the points marked in Figure 6 reminds a square. The analysis of the results of the linear programming solution showed that the function $\psi_N^\Delta(y) \stackrel{\text{def}}{=} \psi_N^\Delta(y_1, y_2)$ is equal to $(-1, 1)$ at every point belonging to the upper side of the "square," and it is equal to $(-1, -1)$, $(1, -1)$, and $(1, 1)$ at the points belonging to, respectively, left, bottom, and right sides of the square. Let us extend the definition of $\psi_N^\Delta(y)$ as follows:

$u_1 = -1,\ u_2 = 1$ for $-0.5 < y_1 \leq 0.9,\ 0.5 \leq y_2 \leq 0.9$;
$u_1 = -1,\ u_2 = -1$ for $-0.9 \leq y_1 \leq -0.5,\ -0.5 < y_2 \leq 0.9$;
$u_1 = 1,\ u_2 = -1$ for $-0.9 \leq y_1 < 0.5,\ -0.9 \leq y_2 \leq 0.5$;
$u_1 = 1,\ u_2 = 1$ for $0.5 \leq y_1 \leq 0.9,\ -0.9 \leq y_2 < 0.5$.

Proceeding as in Example 1, we integrate the system with thus defined feedback control and with the initial condition being at one of the points marked in Figure 6. The resulting solution of the system $\tilde{y}^\Delta(\tau) = (\tilde{y}_1^\Delta(\tau), \tilde{y}_2^\Delta(\tau))$ remains in the area of definition of the feedback control and it returns to a small vicinity of $\tilde{y}^\Delta(0)$ with $\tau \approx 6.1$.

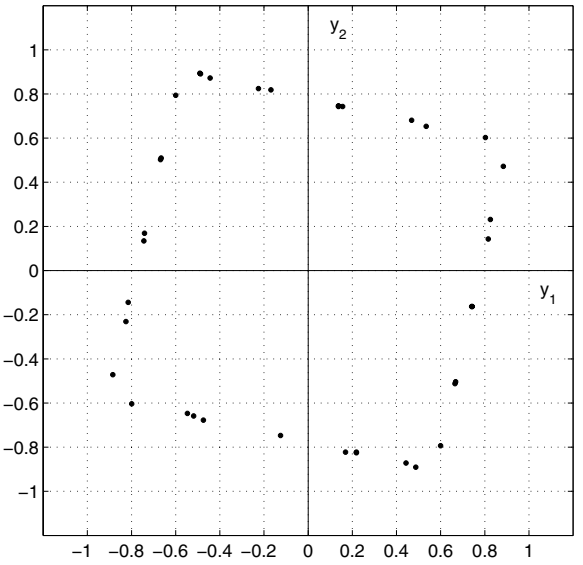


FIG. 6. $\mathcal{Y}_N^\Delta$.

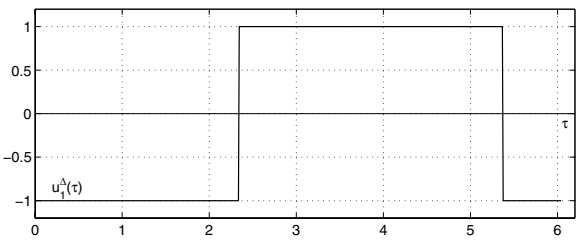


FIG. 7. $u_1^\Delta(\tau)$.

Take $T^\Delta \stackrel{\text{def}}{=} 6.1$. Let $u^\Delta(\tau) = (u_1^\Delta(\tau), u_2^\Delta(\tau))$ be the open loop control defined by the feedback control on the trajectory specified above and let $y^\Delta(\tau) = (y_1^\Delta(\tau), y_2^\Delta(\tau))$ be the solution of the system obtained with this open loop control which satisfies the periodicity condition: $y^\Delta(0) = y^\Delta(T^\Delta)$. The components of the constructed periodic admissible pair $(u^\Delta(\tau), y^\Delta(\tau))$ are shown in Figures 7, 8, and 9.

The value of the objective function calculated on the pair $(u^\Delta(\tau), y^\Delta(\tau))$ is approximately equal to -0.7679 and, hence, this pair is an approximate solution to problem (5). Note that in this example too the assumptions that the solution of the periodic optimization problem (5) exists and that it is unique have not been verified. However, again, based on the form of the obtained approximate solution, one may conjecture that the solution exists and that it has a form similar to that of the obtained approximate solution.

Example 2 (continued). The set of steady state admissible pairs in Example 2 is equal to

$$M = \{(u, y) \ : \ u = (u_1, u_2), \ \ y = (y_1, y_2), \quad y_i = u_i \in [0, 1], \ \ i = 1, 2\}.$$

One can see that, for every $(u, y) \in M$, $g(u, y) = 0$. Hence, $G_{ss} = 0 \ < G_{per} = G^* \approx -0.7679$. Consider the periodic optimization problem (5) in which $g(u, y)$ is replaced

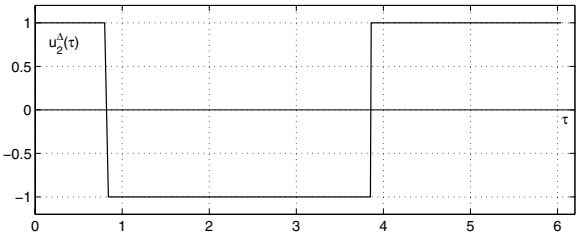


FIG. 8. $u_2^\Delta(\tau)$.

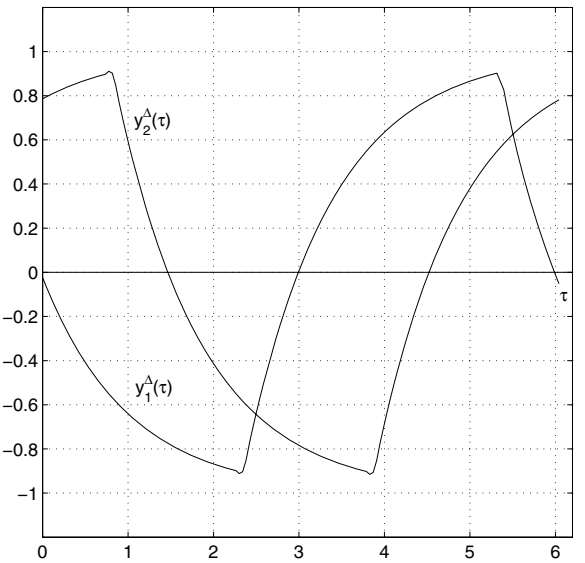


FIG. 9. $y_1^\Delta(\tau)$ *and* $y_2^\Delta(\tau)$.

by $g^\lambda(u,y)$,

$$g^\lambda(u,y) \stackrel{\text{def}}{=} -y_1u_2 + y_2u_1 + \lambda(u_1^2+u_2^2+y_1^2+y_2^2), \tag{62}$$

where $\lambda \geq 0$. For every $(u,y) \in M$, $g^\lambda(u,y) = \lambda(u_1^2+u_2^2+y_1^2+y_2^2)$. Consequently, the optimal value of the corresponding steady state optimization problem (denoted G_{ss}^λ) is equal to zero too: $G_{ss}^\lambda = 0 \ \forall\lambda \geq 0$. It is well known (see, e.g., [23]) that, if U and Y are convex, the system is linear, and the integrand in the objective function is convex, then the optimal values of the periodic and steady state optimization problems coincide. One can verify (by direct calculation) that the Hessian of the function $g^\lambda(u,y)$ has nonnegative eigenvalues for $\lambda \geq 0.5$. That is, this function is convex and, hence, $G_p^\lambda = G_{ss}^\lambda = 0 \ \forall\lambda \geq 0.5$, where G_p^λ stands for the optimal value of the periodic optimization problem (5) considered with $g^\lambda(u,y)$ instead of $g(u,y)$. Consider the LPP

$$\min_{\gamma\in W_N^\Delta} \sum_{i,l,j,k} (-y_{1,j}^\Delta u_{2,l}^\Delta + y_{2,k}^\Delta u_{1,i}^\Delta + \lambda((u_{1,i}^\Delta)^2 + (u_{2,l}^\Delta)^2 + (y_{1,j}^\Delta)^2 + (y_{2,k}^\Delta)^2))\gamma_{i,l,j,k} \stackrel{\text{def}}{=} G_N^{\Delta,\lambda}, \tag{63}$$

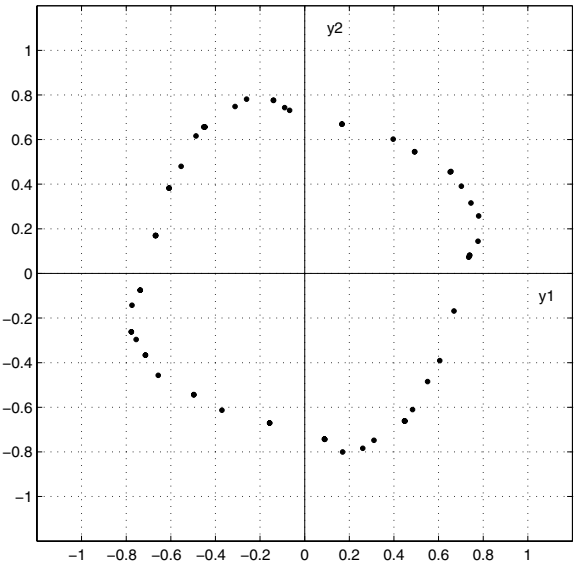


FIG. 10. $\mathcal{Y}_N^\Delta$ *for* $\lambda = 0.2$.

where W_N^Δ is as in (60). As above, the solution of this problem with $N = 10$ and $\Delta = 0.003125$ allows one to find an approximation to the solution of the corresponding periodic optimization problem for different λ. In particular, for $\lambda = 0.2, 0.33, 0.35$, the approximations to the optimal values of the latter are, respectively, $G_{per}^{0.2} \approx -0.2767$, $G_{per}^{0.33} \approx -0.0358$, $G_{per}^{0.35} \approx -0.0050$. For $\lambda \geq 0.36$, $G_{per}^{\lambda} \approx G_{ss}^{\lambda} = 0$ (for $\lambda \geq 0.5$, this being due to the convexity of the function $g^\lambda(u, y)$). Figures 10–13, represent the results of marking with dots the points on the plane (y_1, y_2) which belong to $\mathcal{Y}_N^\Delta$ for $\lambda = 0.2, 0.33, 0.35, 0.36$ ($N = 10$, $\Delta = 0.003125$).

6. Proofs for sections 3, 4, and 5.

Proof of Proposition 2. To prove the validity of (20), let us define $\kappa(S)$ by the equation

$$\kappa(S) \stackrel{\text{def}}{=} \sup_{\gamma \in \Gamma(S)} \rho(\gamma, W) \tag{64}$$

and show that $\kappa(S)$ tends to zero as S tends to infinity. Assume it is not the case. Then there exist a positive number δ and sequences $S^k \to \infty$, $\gamma^k \in \Gamma(S^k)$, such that $\rho(\gamma^k, W) \geq \delta$ for $k = 1, 2, \ldots$. Without loss of generality one may assume that there exists $\lim_{k\to\infty} \gamma^k \stackrel{\text{def}}{=} \gamma \in \mathcal{P}(U \times Y)$ (since $\mathcal{P}(U \times Y)$ is compact). From the continuity of the metric it follows that

$$\rho(\gamma, W) \geq \delta. \tag{65}$$

By the definition of the convergence in $\mathcal{P}(U \times Y)$ (see (16)),

$$\lim_{k\to\infty} \int_{U\times Y} (\phi'(y))^T f(u, y)\gamma^k(du, dy) \;=\; \int_{U\times Y} (\phi'(y))^T f(u, y)\gamma(du, dy) \tag{66}$$

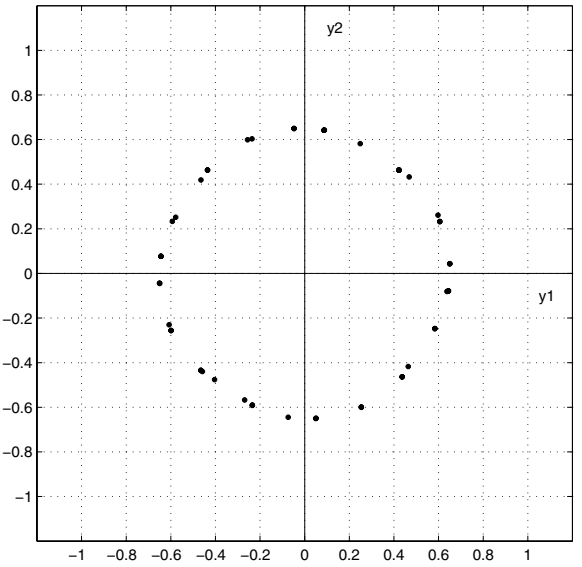


FIG. 11. $\mathcal{Y}_N^{\Delta}$ *for* $\lambda = 0.33$.

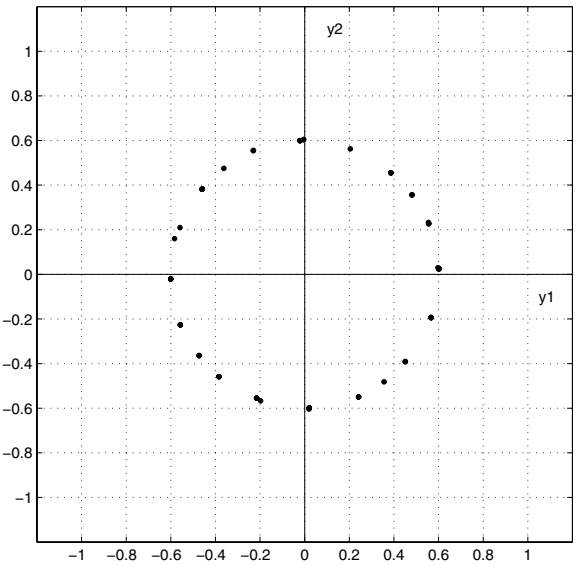


FIG. 12. $\mathcal{Y}_N^{\Delta}$ *for* $\lambda = 0.35$.

for any $\phi \in C^1$. Also, from the fact that $\gamma^k \in \Gamma(S^k)$ it follows that there exists an admissible pair $(u^k(\tau), y^k(\tau))$ defined on the interval $[0, S^k]$ such that

$$\int_{U\times Y} (\phi'(y))^T f(u,y)\gamma^k(du,dy) \;=\; \frac{1}{S^k}\int_0^{S^k} (\phi'(y^k(\tau)))^T f(u^k(\tau), y^k(\tau))d\tau.$$

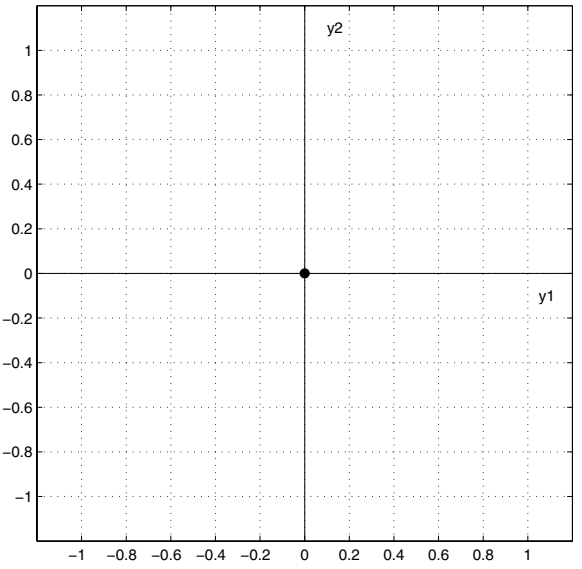


FIG. 13. $\mathcal{Y}_N^{\Delta}$ *for* $\lambda = 0.36$.

The second integral is apparently equal to

$$\frac{\phi(y^k(S^k)) - \phi(y^k(0))}{S^k}$$

and tends to zero as S^k tends to infinity (since $y^k(\tau) \in Y \ \forall \ \tau \in [0, S^k]$ and Y is a compact set). This and (66) imply that

$$\int_{U\times Y} (\phi'(y))^T f(u,y)\gamma(du,dy) \ = \ 0 \ \forall \phi \in C^1 \quad \Rightarrow \quad \gamma \ \in \ W.$$

The latter contradicts (65) and, hence, $\kappa(S)$ defined in (64) tends to zero as S tends to infinity. This proves (20).

From the consideration above it follows that, if there exists a sequence of S^k tending to infinity such that $\Gamma(S^k) \neq \emptyset$, then the set W is not empty. Hence, if the latter is empty, then $\Gamma(S) = \emptyset$ for all S large enough. □

Proof of Proposition 5. Let $\hat{Y}$ be a compact set which contains Y in its interior and let $q_l(u,y) : U \times \hat{Y} \to \mathbb{R}^1$, $l = 1, 2, \dots$, be a sequence of Lipschitz continuous functions which is dense in $C(U \times \hat{Y})$ (the space of continuous functions on $U \times \hat{Y}$). Let

(67)
$$h(u,y) = (q_1(u,y), \dots, q_j(u,y)), \qquad \bar{h}(\nu,y) = (\bar{q}_1(\nu,y), \dots, \bar{q}_j(\nu,y)), \quad j = 1, 2, \dots,$$

where $\bar{q}_j(\nu,u) \stackrel{\text{def}}{=} \int_U q_j(u,y)\nu(du)$. Define the sets $V_h(S,y)$ and $\bar{V}_h(S,y)$ by the equations

$$V_h(S) \stackrel{\text{def}}{=} \bigcup_{\big(u(\cdot),\ y(\cdot)\big)} \left\{ \frac{1}{S} \int_0^S h\big(u(\tau), y(\tau)\big) d\tau \right\},$$

$$\bar{V}_h(S) \stackrel{\text{def}}{=} \bigcup_{\big(\nu(\cdot),\ y(\cdot)\big)} \left\{ \frac{1}{S} \int_0^S \bar{h}\big(\nu(\tau), y(\tau)\big) d\tau \right\},$$

where the unions are over all admissible and relaxed admissible pairs, respectively.

Let d_H stand for the Hausdorff metric defined on compact subsets of $\mathbb{R}^j$ by the Euclidean norm. Using a standard argument based on the separability of convex sets (see, e.g., Lemma 4.2 in [26]), one can verify that Assumption 1 is equivalent to the statement that, for any $h(\cdot)$ and $\bar{h}(\cdot)$ as in (67),

$$d_H(\bar{co}V_h(S), \bar{co}\bar{V}_h(S)) \stackrel{\text{def}}{=} \kappa_h(S) \ \to \ 0 \tag{68}$$

as $S \to \infty$, where $\bar{co}$ stands for the closed convex hull of the corresponding set.

From the definition of the metric ρ in the form (15) and the convexity of W it follows that $\sup_{\gamma \in \Gamma(S)} \rho(\gamma, W) = \sup_{\gamma \in co\Gamma(S)} \rho(\gamma, W)$. Hence, by (20),

$$\lim_{S \to \infty} \sup_{\gamma \in co\Gamma(S)} \rho(\gamma, W) = 0$$

and, to prove (24), it is enough to show that

$$\sup_{\gamma \in W} \rho(\gamma, co\Gamma(S)) \le \kappa(S), \qquad \lim_{S \to \infty} \kappa(S) = 0. \tag{69}$$

Let us take an arbitrary $\gamma \in W$. From Lemma 5.1 in [26] and Theorem 4.1 in [43] it follows (see [26]) that there exist a probability space $(\Omega, \mathcal{F}, P)$ and a $(\mathcal{P}(U) \times \mathbb{R}^m)$-valued random process $(\nu(\tau), y(\tau)) = (\nu(\tau, \omega), y(\tau, \omega))$ such that (i) for any $h(\cdot)$ and $\bar{h}(\cdot)$ defined in (67),

$$E[\bar{h}(\nu(\tau, \omega), y(\tau, \omega))] \ = \ \int_{U \times Y} h(u, y) \gamma(du, dy) \ \ \forall \tau \ge 0; \tag{70}$$

and (ii) for some $\Omega' \subset \Omega$ with $P(\Omega') = 1$ and for any $\omega \in \Omega'$, the pair $(\nu(\cdot, \omega), y(\cdot, \omega))$ is relaxed admissible on any interval $[0, S]$.

From (ii) it follows that

$$\frac{1}{S} \int_0^S \bar{h}\big(\nu(\tau, \omega), y(\tau, \omega)\big) d\tau \ \in \ \bar{V}_h(S) \ \ \forall \omega \in \Omega',$$

while (i) implies that

$$\int_{U \times Y} h(u, y) \gamma(du, dy) \ = \ E \frac{1}{S} \int_0^S [\bar{h}\big(\nu(\tau, \omega), y(\tau, \omega)\big)] d\tau \ \in \ \bar{co}\bar{V}_h(S).$$

Using the above inclusion and taking into account (68) (as well as the fact that γ is an arbitrary element of W), one can conclude that

$$\bigcup_{\gamma \in W} \left\{ \int_{U \times Y} h(u, y) \gamma(du, dy) \right\} \ \subset \ \bar{co}V_h(S) \ + \ \kappa_h(S) B_j, \tag{71}$$

where B_j is the closed unit ball in $\mathbb{R}^j$ (j is the number of components of $h(\cdot)$; see (67)). Applying now Lemma 3.5 from [27] in exactly the same way as it is done

on page 335 in [26], one can prove the validity of (69). This completes the proof of (24). □

Proof of Proposition 7 *(continued).* Since $W \subset W_N$, to prove that (32) is valid, it is enough to show that

$$\lim_{N\to\infty} \sup_{\gamma\in W_N} \rho(\gamma, W) = 0. \tag{72}$$

Assume it is not true. Then there exist a positive number δ, a subsequence of positive integers $N' \to \infty$, and a sequence of probability measures $\gamma_{N'} \in W_{N'}$ such that $\rho(\gamma_{N'}, W) \geq \delta$. Due to the compactness of $\mathcal{P}(U \times Y)$, one may assume (without loss of generality) that there exists $\bar{\gamma} \in \mathcal{P}(U \times Y)$ such that

$$\lim_{N'\to\infty} \rho(\gamma_{N'}, \bar{\gamma}) = 0 \qquad \Rightarrow \qquad \rho(\bar{\gamma}, W) \geq \delta. \tag{73}$$

From the fact that $\gamma_{N'} \in W_{N'}$ it follows that, for any integer i and $N' \geq i$,

$$\int_{U\times Y} (\phi_i'(y))^T f(u,y)\gamma_{N'}(du,dy) = 0 \quad \Rightarrow \quad \int_{U\times Y} (\phi_i'(y))^T f(u,y)\bar{\gamma}(du,dy) = 0.$$

Since the latter is valid for any $i = 1, 2, \dots$, one can conclude that $\bar{\gamma} \in W$, which contradicts (73). This proves (32). □

Proof of Proposition 8. In case (i), from (36) it follows that there exist an open ball $B \subset Y$ centered at $\bar{y}$ and a number $r > 0$ such that the closed ball $\bar{B}_r$ centered at 0 and having the radius $r > 0$ will satisfy the inclusion

$$\bar{B}_r \subset f(y, U) \quad \forall y \in B. \tag{74}$$

Let us show that, if $\phi(\cdot)$ satisfies (35), then $\phi'(y) = 0\ \forall y \in B$ and, thus, Y^* can be taken to be equal to the closure of B. Assume that $\phi'(y) \neq 0$ for some $y \in B$. By (74), there exist $u \in U$ such that $\frac{r}{||\phi'(y)||}\phi'(y) = f(y,u)$. Hence, by (35),

$$(\phi'(y))^T\phi'(y) = \frac{||\phi'(y)||}{r}(\phi'(y))^T f(y,u) \leq 0 \quad \Rightarrow \quad \phi'(y) = 0.$$

The obtained contradiction proves the statement.

In case (ii), let us show that Y^* can be taken to be equal to the closure of Y^0. To show this, it is enough to establish that from (35) it follows that $\phi(y) = \text{ const } \forall y \in Y^0$ (which leads to that $\phi(y) = \text{ const } \forall y \in Y^*$ and, hence, to that $\phi'(y) = 0\ \forall y \in \text{int } Y^*$).

Let $y', y'' \in Y^0$ and $(u(\tau), y(\tau))$ be an admissible pair such that $y(0) = y'$ and $y(S) = y''$. Then, by (35),

$$\phi(y'') - \phi(y') = \int_0^S (\phi'(y(\tau)))^T f(u(\tau), y(\tau))d\tau \leq 0 \quad \Rightarrow \quad \phi(y'') \leq \phi(y').$$

Since y', y'' are arbitrary points in Y^0, the latter implies that

$$\phi(y) = \text{ const } \forall y \in Y^0. \qquad \square$$

Proof of Proposition 9. First, note that, by (41), the set W_N is not empty if W_N^Δ is not empty.

Let us assume that the set W_N is not empty and show that W_N^Δ is not empty and that (42) is valid (the validity of (43) follows from (42) on the basis of Lemma 1(ii);

the other statements included in the proposition are immediate consequences of (42) and (43)).

From (38) and the fact that the functions $(\phi_i'(y))^T f(u,y)$ are continuous it follows that

$$\sup_{(u,y)\in Q_{l,k}^{\Delta}} |(\phi_i'(y))^T f(u,y) - (\phi_i'(y_k))^T f(u_l,y_k)| \le \kappa(\Delta), \quad i=1,\dots,N, \tag{75}$$

for some $\kappa(\Delta)$ such that $\lim_{\Delta\to 0}\kappa(\Delta)=0$. Define the set $Z_N^{\Delta}\subset \mathbb{R}^{L^{\Delta}+K^{\Delta}}$ by the equation

$$Z_N^{\Delta} \stackrel{\text{def}}{=} \left\{ \gamma=\{\gamma_{l,k}\}\ge 0 \ : \ \sum_{l,k}\gamma_{l,k}=1, \ \left|\sum_{l,k}(\phi_i'(y_k))^T f(u_l,y_k)\gamma_{l,k}\right| \le \kappa(\Delta), \right. \\ \left. i=1,2,\dots,N \right\}. \tag{76}$$

For any Δ, let $\gamma^{\Delta}\in W_N$ be such that $\rho(\gamma^{\Delta},Z_N^{\Delta}) = \max_{\gamma\in W_N}\rho(\gamma,Z_N^{\Delta})$ (γ^{Δ} exists since W_N is compact) and show that

$$\lim_{\Delta\to 0}\max_{\gamma\in W_N}\rho(\gamma,Z_N^{\Delta}) = \lim_{\Delta\to 0}\rho(\gamma^{\Delta},Z_N^{\Delta}) = 0. \tag{77}$$

Let $\gamma_{l,k}^{\Delta}\stackrel{\text{def}}{=}\int_{Q_{l,k}^{\Delta}}\gamma^{\Delta}(du,dy)$. By (75),

$$\left|\sum_{l,k}(\phi_i'(y_k))^T f(u_l,y_k)\gamma_{l,k}^{\Delta}\right| \\ = \left|\sum_{l,k}(\phi_i'(y_k))^T f(u_l,y_k)\gamma_{l,k}^{\Delta} - \int_{U\times Y}(\phi_i'(y))^T f(u,y)\gamma^{\Delta}(du,dy)\right| \\ \le \sum_{l,k}\int_{Q_{l,k}^{\Delta}}|(\phi_i'(y_k))^T f(u_l,y_k)-(\phi_i'(y))^T f(u,y)|\gamma^{\Delta}(du,dy) \le \kappa(\Delta), \quad i=1,2,\dots,N.$$

Hence, denoting $\tilde{\gamma}^{\Delta}\stackrel{\text{def}}{=}(\gamma_{l,k}^{\Delta})$, one obtains that $\tilde{\gamma}^{\Delta}\in Z_N^{\Delta}$ and, consequently,

$$\rho(\tilde{\gamma}^{\Delta},Z_N^{\Delta})=0. \tag{78}$$

Let $q(u,y):U\times Y\to\mathbb{R}^1$ be an arbitrary continuous function and let $\kappa_q(\Delta)$ be such that

$$\sup_{(u,y)\in Q_{l,k}^{\Delta}}|q(u,y)-q(u_l,y_k)| \le \kappa_q(\Delta), \qquad \lim_{\Delta\to 0}\kappa_q(\Delta)=0.$$

Then

$$\left|\int_{U\times Y}q(u,y)\gamma^{\Delta}(du,dy) - \sum_{l,k}q(u_l,y_k)\gamma_{l,k}^{\Delta}\right| \\ = \left|\sum_{l,k}\int_{Q_{l,k}}q(u,y)\gamma^{\Delta}(du,dy) - \sum_{l,k}\int_{Q_{l,k}}q(u_l,y_k)\gamma^{\Delta}(du,dy)\right| \le \kappa_q(\Delta).$$

The fact that the latter inequality is valid for an arbitrary continuous $q(u,y)$ implies that $\lim_{\Delta\to 0}\rho(\gamma^\Delta,\tilde\gamma^\Delta)=0$, which, along with (78), implies the validity of (77).

By (41), $\max_{\gamma\in W_N^\Delta}\rho(\gamma,W_N)=0$. Hence, to prove (42), it is enough to establish that

$$\lim_{\Delta\to 0}\max_{\gamma\in W_N}\rho(\gamma,W_N^\Delta)=0. \tag{79}$$

Since (as can be easily verified using the triangle inequality),

$$\max_{\gamma\in W_N}\rho(\gamma,W_N^\Delta)\le\max_{\gamma\in W_N}\rho(\gamma,Z_N^\Delta)+\max_{\gamma\in Z_N^\Delta}\rho(\gamma,W_N^\Delta)$$

and since (77) has been already verified, equality (79) will be established if one shows that

$$\lim_{\Delta\to 0}\max_{\gamma\in Z_N^\Delta}\rho(\gamma,W_N^\Delta)\ =\ \lim_{\Delta\to 0}\rho(\bar\gamma^\Delta,W_N^\Delta)\ =0, \tag{80}$$

where $\bar\gamma^\Delta=\{\bar\gamma^\Delta_{l,k}\}\in Z_N^\Delta$ is such that $\rho(\bar\gamma^\Delta,W_N^\Delta)=\max_{\gamma\in Z_N^\Delta}\rho(\gamma,W_N^\Delta)$ for any $\Delta>0$.

Let $q_j(\cdot)$ be the same as in definition (15) of the metric ρ. Consider the following finite-dimensional linear program:

$$F_J(\Delta)\stackrel{\text{def}}{=}\min_{\gamma=\{\gamma_{l,k}\}\in W_N^\Delta}\sum_{j=1}^{J}\frac{1}{2^j}\left|\sum_{l,k}q_j(u_l,y_k)\gamma_{l,k}-\sum_{l,k}q_j(u_l,y_k)\bar\gamma^\Delta_{l,k}\right|. \tag{81}$$

To prove that (80) is valid, it is enough to show that

$$\lim_{\Delta\to 0}F_J(\Delta)\ =\ 0,\qquad J=1,2,\dots. \tag{82}$$

Below it is shown that the optimal value of the problem dual to (81) tends to zero as Δ tends to zero. Since the latter coincides with $F_J(\Delta)$, this will prove (82). Also, from (82) it follows that $F_J(\Delta)$ is bounded and, hence, W_N^Δ is not empty for Δ small enough (see, e.g., Theorem 2 on page 129 in [18]).

Let us rewrite problem (81) in the equivalent form:

$$F_J(\Delta)\ =\ \min_{\gamma=\{\gamma_{l,k}\}\in W_N^\Delta}\sum_{j=1}^{J}\frac{1}{2^j}\theta_j, \tag{83}$$

where

$$-\sum_{l,k}q_j(u_l,y_k)\gamma_{l,k}+\theta_j\ge-\sum_{l,k}q_j(u_l,y_k)\bar\gamma^\Delta_{l,k}, \tag{84}$$

$$\sum_{l,k}q_j(u_l,y_k)\gamma_{l,k}+\theta_j\ge\sum_{l,k}q_j(u_l,y_k)\bar\gamma^\Delta_{l,k}. \tag{85}$$

The problem dual to (83)–(85) is

$$F_J(\Delta)\ =\ \max_{\lambda_i,\mu_j,\eta_j,\zeta}\sum_{j=1}^{J}(-\mu_j+\eta_j)\left(\sum_{l,k}q_j(u_l,y_k)\bar\gamma^\Delta_{l,k}\right)+\zeta, \tag{86}$$

where $\lambda_i, i = 1, \dots, N$; μ_j, η_j, $j = 1, \dots J$, and ζ satisfy the following relationships:

$$\sum_{i=1}^{N} \lambda_i (\phi_i'(y^k))^T f(u_l, y_k) + \sum_{j=1}^{J} (-\mu_j + \eta_j) q_j(u_l, y_k) + \zeta \le 0, \tag{87}$$

$l = 1, 2, \dots, L^\Delta$, $k = 1, 2, \dots, K^\Delta$, and

$$\mu_j + \eta_j = \frac{1}{2^j}, \qquad \mu_j \ge 0, \ \eta_j \ge 0, j = 1, , 2, \dots, J. \tag{88}$$

Before proving (82), let us verify that $F_J(\Delta)$ is bounded for Δ small enough (which, by (81), is equivalent to that W_N^Δ is not empty). Assume it is not. Then there exist a sequence $\Delta^r, r = 1, 2, \dots, \lim_{r\to\infty} \Delta^r = 0$, and sequences λ_i^r, μ_j^r, η_j^r, ζ^r, satisfying (87)–(88) with $\Delta = \Delta^r$, $r = 1, 2, \dots$, such that $\lim_{r\to\infty}(\zeta^r + \sum_{i=1}^N |\lambda_i^r|) = \infty$ and

$$\lim_{r\to\infty} \frac{\zeta^r}{\zeta^r + \sum_{i=1}^N |\lambda_i^r|} \stackrel{\text{def}}{=} a \ge 0, \qquad \lim_{r\to\infty} \frac{\lambda_i^r}{\zeta^r + \sum_{i=1}^N |\lambda_i^r|} \stackrel{\text{def}}{=} v_i,$$

where

$$a + \sum_{i=1}^{N} |v_i| = 1. \tag{89}$$

Dividing (87) by $\zeta^r + \sum_{i=1}^N |\lambda_i^r|$ and passing to the limit as $r \to \infty$, one can obtain

$$\sum_{i=1}^{N} v_i (\phi_i'(y))^T f(u, y) + a \le 0 \quad \forall (u, y) \in U \times Y, \tag{90}$$

where it is taken into account that every point $(u, y) \in U \times Y$ can be presented as the limit of (u_l, y_k) belonging to the sequence of cells $Q_{l,k}^{\Delta^r}$ such that $(u, y) \in Q_{l,k}^{\Delta^r}$.

Two cases are possible: $a > 0$ and $a = 0$. If $a > 0$, then the validity of (90) implies that the function $\phi(y) \stackrel{\text{def}}{=} \sum_{i=1}^N v_i \phi_i(y)$ satisfies (19) which would lead to W_N being empty. The set W_N, however, is not empty (by our assumption) and, hence, the only case to consider is $a = 0$. In this case, (90) becomes

$$\sum_{i=1}^{N} v_i (\phi_i'(y))^T f(u, y) \le 0 \quad \forall (u, y) \in U \times Y. \tag{91}$$

By Assumption 2, (91) can be valid only with all v_i being equal to zero. This contradicts (89) and, thus, proves that $F_J(\Delta)$ is bounded for Δ small enough (and that W_N^Δ is not empty).

From the fact that $F_J(\Delta)$ is bounded it follows that a solution $\lambda_i^\Delta, i = 1, \dots, N$; μ_j^Δ, η_j^Δ, $j = 1, \dots, J$, and ζ^Δ of the problem (86)–(88) exists. Using this solution,

one can obtain the following estimates:

$$\begin{aligned}
0 \le F_J(\Delta) &= \sum_{j=1}^{J}(-\mu_j^\Delta + \eta_j^\Delta)\left(\sum_{l,k} q_j(u_l, y_k)\bar\gamma_{l,k}^\Delta\right) + \zeta^\Delta \\
&= \sum_{l,k}\bar\gamma_{l,k}^\Delta\left(\sum_{j=1}^{J}(-\mu_j^\Delta + \eta_j^\Delta)q_j(u_l, y_k)\right) + \zeta^\Delta \\
&\le \sum_{l,k}\bar\gamma_{l,k}^\Delta\left(-\sum_{i=1}^{N}\lambda_i^\Delta(\phi_i'(y^k))^T f(u_l, y_k) - \zeta^\Delta\right) + \zeta^\Delta \\
&= -\sum_{i=1}^{N}\lambda_i^\Delta\left(\sum_{l,k}(\phi_i'(y^k))^T f(u_l, y_k)\bar\gamma_{l,k}^\Delta\right) \le \sum_{i=1}^{N}|\lambda_i^\Delta|\kappa(\Delta),
\end{aligned}$$

where the last two relationships are implied by the fact that $\bar\gamma^\Delta = \{\bar\gamma_{l,k}^\Delta\} \in Z_N^\Delta$ (see (76)).

To prove (82), it is now sufficient to show that $\sum_{i=1}^N |\lambda_i^\Delta|$ remains bounded as $\Delta \to 0$. Assume it is not. Then there exists a sequence $\Delta^r, r = 1, 2, \dots, \lim_{r\to\infty}\Delta^r = 0$, and sequences $\lambda_i^r, \mu_j^r, \eta_j^r, \zeta^r$, satisfying (87)–(88) with $\Delta = \Delta^r$, $r = 1, 2, \dots$, such that

(92)
$$\lim_{r\to\infty}\sum_{i=1}^{N}|\lambda_i^r| = \infty, \qquad \lim_{r\to\infty}\frac{\zeta^r}{\sum_{i=1}^N|\lambda_i^r|} = 0, \qquad \lim_{r\to\infty}\frac{\lambda_i^r}{\sum_{i=1}^N|\lambda_i^r|} \stackrel{\text{def}}{=} v_i, \qquad \sum_{i=1}^{N}|v_i| = 1.$$

Dividing (87) by $\sum_{i=1}^N|\lambda_i^r|$ and passing to the limit as $r \to \infty$, one obtains that the inequality (91) is valid, which, by Assumption 2, implies that $v_i = 0$, $i = 1, \dots, N$. This contradicts the last equality in (92) and, thus, proves (82). □

Proof of Proposition 10. Assume that (46) is not true. Then there exist a number $r > 0$ and a sequence N_i tending to infinity as i tends to infinity and there exist sequences $\Delta_{i,j}$, with each $\Delta_{i,j}$ tending to zero as j tends to infinity (with i being fixed) such that

$$\begin{aligned}
\gamma_{N_i}^{\Delta_{i,j}}((U\times Y)/(\Theta + rB)) = \gamma_{N_i}^{\Delta_{i,j}}(\Theta_{N_i}^{\Delta_{i,j}}/(\Theta + rB)) \ &\ge \ \delta \\
\Rightarrow \gamma_{N_i}^{\Delta_{i,j}}(\Theta + rB) \ &< \ 1 - \delta.
\end{aligned} \tag{93}$$

Due to compactness of $\mathcal{P}(U \times Y)$, one may assume (without loss of generality) that there exists $\gamma_{N_i} \in \mathcal{P}(U \times Y)$ such that $\lim_{j\to\infty}\rho(\gamma_{N_i}^{\Delta_{i,j}}, \gamma_{N_i}) = 0$. Hence, since $\Theta + rB$ is an open set,

$$\underline{\lim}_{j\to\infty}\gamma_{N_i}^{\Delta_{i,j}}(\Theta + rB) \ \ge \ \gamma_{N_i}(\Theta + rB). \tag{94}$$

By Proposition 9, γ_{N_i} is a solution of problem (30). Consequently, by Proposition 7 and the fact that γ^* is the unique solution of (21),

$$\lim_{i\to\infty}\rho(\gamma_{N_i}, \gamma^*) = 0 \qquad \Rightarrow \qquad \underline{\lim}_{i\to\infty}\gamma_{N_i}(\Theta + rB) \ \ge \ \gamma^*(\Theta + rB) \ = \ 1. \tag{95}$$

The relationships (94) and (95) imply that, for i and j large enough, $\gamma_{N_i}^{\Delta_{i,j}}(\Theta + rB) \ge 1 - \delta$. This contradicts (93) and, thus, proves (46). The inclusion (47) follows from (46). □

Proof of Proposition 11. Assume that the proposition is not true. Then there exist a number $r > 0$ and sequence: $(u_i, y_i) \in \Theta,\ N_i,\ \Delta_{i,j},\ i = 1, 2, \dots,\ j = 1, 2, \dots,$ with

$$\lim_{i\to\infty}(u_i, y_i) = (\bar{u}, \bar{y}) \in cl\Theta, \qquad \lim_{i\to\infty} N_i = \infty, \qquad \lim_{j\to\infty} \Delta_{i,j} = 0$$

such that

$$d((u_i, y_i), \Theta_{N_i}^{\Delta_{i,j}}) \geq r \qquad \Rightarrow \qquad d((\bar{u}, \bar{y}), \Theta_{N_i}^{\Delta_{i,j}}) \geq \frac{r}{2}, \tag{96}$$

where $d((u, y), Q)$ stands for the distance between a point $(u, y) \in U \times Y$ and a set $Q \subset U \times Y$: $d((u, y), Q) \stackrel{\text{def}}{=} \inf_{(u',y')\in Q}\{||(u, y) - (u', y')||\}$. The second inequality in (96) implies that

$$\left((\bar{u}, \bar{y}) + \frac{r}{2}B\right) \cap \Theta_{N_i}^{\Delta_{i,j}} = \emptyset \qquad \Rightarrow \qquad \gamma_{N_i}^{\Delta_{i,j}}\left((\bar{u}, \bar{y}) + \frac{r}{2}B\right) = 0.$$

Similarly to the proof of Proposition 10, one may assume, without loss of generality, that there exists $\gamma_{N_i} \in \mathcal{P}(U \times Y)$ such that $\lim_{j\to\infty} \rho(\gamma_{N_i}^{\Delta_{i,j}}, \gamma_{N_i}) = 0$. Hence, since the set $(\bar{u}, \bar{y}) + \frac{r}{2}B$ is open,

$$0 \ = \ \lim_{j\to\infty} \gamma_{N_i}^{\Delta_{i,j}}\left((\bar{u}, \bar{y}) + \frac{r}{2}B\right) \ \geq \ \gamma_{N_i}\left((\bar{u}, \bar{y}) + \frac{r}{2}B\right).$$

As in the proof of Proposition 10, γ_{N_i} is a solution of problem (30) (see Proposition 9). Consequently, from Proposition 7 and from the fact that γ^* is the unique solution of (21) it follows that

$$\lim_{i\to\infty} \rho(\gamma_{N_i}, \gamma^*) = 0 \qquad \Rightarrow \qquad 0 \ = \ \lim_{i\to\infty} \gamma_{N_i}\left((\bar{u}, \bar{y}) + \frac{r}{2}B\right) \ \geq \ \gamma^*\left((\bar{u}, \bar{y}) + \frac{r}{2}B\right).$$

The latter contradicts Assumption 3 and, thus, proves the proposition. □

Acknowledgments. We thank J.-P. Aubin for inspiring discussions (during V. Gaitsgory's visits to Universities of Dauphine and Paris 1 in 2002–2004). We also express our gratitude to O. Alvarez, M. Bardi, and V. Veliov for supplying us with important references on the subject matter.

REFERENCES

[1] O. Alvarez and M. Bardi, *Viscosity solutions methods for singular perturbations in deterministic and stochastic control*, SIAM J. Control Optim., 40 (2001), pp. 1159–1188.

[2] O. Alvarez and M. Bardi, *Singular perturbations of nonlinear degenerate parabolic PDEs: A general convergence result*, Arch. Ration. Mech. Anal., 170 (2003), pp. 17–61.

[3] B. D. O. Anderson and P. V. Kokotovic, *Optimal control problems over large time intervals*, Automatica, 23 (1987), pp. 355–363.

[4] E. J. Anderson and P. Nash, *Linear Programming in Infinite-Dimensional Spaces*, John Wiley, Chichester, 1987.

[5] M. Arisawa, H. Ishii, and P.-L. Lions, *A characterization of the existence of solutions for Hamilton-Jacobi equations in ergodic control problems with applications*, Appl. Math. Optim., 42 (2000), pp. 35–50.

[6] Z. Artstein, *An occupational measure solution to a singularly perturbed optimal control problem*, Control Cybernet., 31 (2002), pp. 623–642.

[7] Z. Artstein, *Invariant measures and their projections in nonautonomous dynamical systems*, Stoch. Dyn., 4 (2004), no. 3, pp. 439–459.

[8] Z. Artstein and V. Gaitsgory, *The value function of singularly perturbed control systems*, Appl. Math. Optim., 41 (2000), pp. 425–445.

[9] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser Boston, Boston, MA, 1990.

[10] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser Boston, Boston, MA, 1997.

[11] G. K. Basak, V. S. Borkar, and M. K. Ghosh, *Ergodic control of degenerate diffusions*, Stoch. Anal. Appl., 15 (1997), pp. 1–17.

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

[13] A. Arapostathis, V. S. Borkar, and M. Ghosh, *Ergodic Control of Diffusion Processes*, Springer, Berlin, to appear.

[14] D. A. Carlson, A. B. Haurie, and A. Leizarowitz, *Optimal Control on Infinite Time Horizon*, 2nd ed., Springer-Verlag, Berlin, 1991.

[15] F. Colonius, *Optimal Periodic Control*, Lecture Notes in Math. 1313, Springer-Verlag, Berlin, 1988.

[16] F. Colonius and R. Fabbri, *Controllability for systems with slowly varying parameters*, ESAIM Control Optim. Calc. Var., 9 (2003), pp. 207–216.

[17] F. Colonius and W. Kliemann, *Infinite time optimal control and periodicity*, Appl. Math. Optim., 20 (1989), pp. 113–130.

[18] G. B. Dantzig, *Linear Programming and Extensions*, Princeton University Press, Princeton, NJ, 1963.

[19] T. D. Donchev and A. L. Dontchev, *Singular perturbations in infinite-dimensional control systems*, SIAM J. Control Optim., 42 (2003), pp. 1795–1812.

[20] L. C. Evans and D. Gomes, *Linear programming interpretations of Mather's variational principle*, ESAIM Optim. Calc. Var., 8 (2002), pp. 693–702.

[21] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1991.

[22] H. Frankowska and F. Rampazo, *Filippov and Filippov-Wazewski's theorems on closed domains*, J. Differential Equations, 161 (2000), pp. 449–478.

[23] V. Gaitsgory, *Control of Systems with Fast and Slow Motions*, Nauka, Moscow, 1991 (in Russian).

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

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

[26] V. Gaitsgory, *On representation of the limit occupational measures set of a control system with applications to singularly perturbed control systems*, SIAM J. Control Optim., 43 (2004), pp. 325–340.

[27] V. Gaitsgory and M. T. Nguyen, *Multiscale singularly perturbed control systems: Limit occupational measures sets and averaging*, SIAM J. Control Optim., 41 (2002), pp. 954–974.

[28] G. Grammel, *On nonlinear control systems with multiple time scales*, J. Dyn. Control Systems, 10 (2004), pp. 11–28.

[29] L. Grüne, *On the relation between discounted and average optimal value functions*, J. Differential Equations, 148 (1998), pp. 65–99.

[30] K. Helmes and R. H. Stockbridge, *Numerical comparison of controls and verification of optimality for stochastic control problems*, J. Optim. Theory Appl., 106 (2000), pp. 107–127.

[31] O. Hernandez-Lerma and J. B. Lasserre, *Markov Chains and Invariant Probabilities*, Birkhäuser-Verlag, Basel, 2003.

[32] Y. Kabanov and S. Pergamenshchikov, *Two-Scale Stochastic Systems*, Springer-Verlag, Berlin, Heidelberg, 2003.

[33] P. V. Kokotovic, H. K. Khalil, and J. O'Reilly, *Singular Perturbation Methods in Control: Analysis and Design*, SIAM Classics in Appl. Math. 25, SIAM, Philadelphia, 1999.

[34] H. J. Kushner, *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*, Birkhäuser Boston, Boston, MA, 1990.

[35] H. J. Kushner and P. G. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, 2nd ed., Springer-Verlag, New York, 2001.

[36] A. Leizarowitz, *Order reduction is invalid for singularly perturbed control problems with vector fast variables*, Math. Control Signals Systems, 15 (2002), pp. 101–119.

[37] J. G. Llavona, *Approximation of Continuously Differentiable Functions*, Math. Stud. 130, North-Holland, Amsterdam, 1986.

[38] M. S. Mendiondo and R. H. Stockbridge, *Approximation of infinite-dimensional linear programming problems which arise in stochastic control*, SIAM J. Control Optim., 36 (1998), pp. 1448–1472.

[39] S. D. Naidu, *Singular perturbations and time scales in control theory and applications: An overview*, Dyn. Contin. Discrete Impuls. Syst., Ser. B Appl. Algorithms, 9 (2002), pp. 233–278.

[40] R. E. O'Malley, Jr., *Singular perturbations and optimal control*, in Mathematical Control Theory, W. A. Copel, ed., Lecture Notes in Math. 680, Springer-Verlag, Berlin, 1978, pp. 170–218.

[41] M. Quincampoix and F. Watbled, *Averaging method for discontinuous Mayer's problem of singularly perturbed control systems*, Nonlinear Anal., 54 (2003), pp. 819–837.

[42] J. E. Rubio, *Control and Optimization. The Linear Treatment of Nonlinear Problems*, Manchester University Press, Manchester, 1985.

[43] R. H. Stockbridge, *Time-average control of a Martingale problem. Existence of a stationary solution*, Ann. Prob., 18 (1990), pp. 190–205.

[44] R. H. Stockbridge, *Time-average control of a Martingale problem: A linear programming formulation*, Ann. Prob., 18 (1990), pp. 206–217.

[45] V. Veliov, *A generalization of Tichonov theorem for singularly perturbed differential inclusions*, J. Dyn. Control Systems, 3 (1997), pp. 1–28.

[46] A. Vigodner, *Limits of Singularly Perturbed Control Problems: Dynamical Systems Approach*, Ph.D. Thesis, The Weizmann Institute of Science, Rehovot, Israel, 1995.

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

[48] G. G. Yin and Q. Zhang, *Continuous-Time Markov Chains and Applications. A Singular Perturbation Approach*, Springer, New York, 1997.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 6, pp. 2038–2062

BIFURCATIONS OF 1-PARAMETER FAMILIES OF CONTROL-AFFINE SYSTEMS IN THE PLANE*

BRONISLAW JAKUBCZYK† AND WITOLD RESPONDEK‡

Abstract. We define bifurcations of control-affine systems in the plane and classify all generic 1-parameter bifurcations at control-regular points. More precisely, we classify topological bifurcations of invariants of usual feedback equivalence. Such bifurcations form six different classes: two bifurcations of equilibrium sets, two bifurcations of critical sets, and two bifurcations of pairs of invariants. We also classify all generic 1-parameter families of control-affine systems with respect to orbital feedback equivalence.

Key words. control system, bifurcation, invariants, classification, orbital feedback equivalence, 1-parameter families

AMS subject classifications. Primary: 93B10; Secondary: 34H05, 37G05, 93C15

DOI. 10.1137/S0363012903431165

1. Introduction.

In this paper we define and study bifurcations of 1-parameter families of smooth, control-affine systems

$$\Sigma \ : \ \dot{\xi} = f(\xi, \epsilon) + g(\xi, \epsilon)u,$$

where ξ lies in an open subset $X \subset \mathbb{R}^2$ or in a differential manifold $X = M^2$ and $u \in \mathbb{R}$. We classify the generic bifurcations at control-regular points (i.e., with $g(\xi, \epsilon) \neq 0$).

A local bifurcation of a parameter dependent dynamical system $\dot{\xi} = f(\xi, \epsilon)$ occurs at an equilibrium if there is a change, when the parameter ϵ varies, of topological character of the solution curves nearby the equilibrium (see, e.g., [IJ], [Ku]). Understanding bifurcations of such equations is important from several points of view, and the already known classification is rather rich (see, e.g., [AAIS]).

Analogous definition of bifurcation applied to a control system is not suitable since the set of trajectories of Σ is rich. (Local invariants of the feedback group include functional invariants, already for generic systems on $\mathbb{R}^2$; see [JR1], [JR2], [Zh].)

Therefore we consider only the most characteristic trajectories: constant trajectories, time-critical trajectories, and so-called, fast (quasi) trajectories. Thus, we attach to our system three basic invariants (equivariants) of feedback transformations. Namely, the *equilibria set* and the *critical set* are defined, respectively, by

$$E_\epsilon = \{p \in X \mid f(p, \epsilon) \text{ and } g(p, \epsilon) \text{ are linearly dependent}\},$$

$$C_\epsilon = \{p \in X \mid [g, f](p, \epsilon) \text{ and } g(p, \epsilon) \text{ are linearly dependent}\},$$

*Received by the editors July 10, 2003; accepted for publication (in revised form) June 21, 2005; published electronically January 6, 2006. The paper was completed during the stay of both authors at the Mittag-Leffler Institute in Stockholm.

http://www.siam.org/journals/sicon/44-6/43116.html

†Institute of Applied Mathematics and Mechanics, Warsaw University, ul. Banacha 2, 02-097 Warsaw, Poland (B.Jakubczyk@impan.gov.pl). On leave from the Institute of Mathematics, Polish Academy of Sciences. This author was partially supported by KBN grant 2 P03A 001 24.

‡Institut National des Sciences Appliquées de Rouen, Laboratoire de Mathématiques de l'INSA, Pl. Emile Blondel, 76 131 Mont Saint Aignan, Cedex, France (wresp@insa-rouen.fr). This author was partially supported by the Excellence Centre of IMPAN-BC.

where $[g, f] = Df\, g - Dg\, f$ is the Lie bracket of g and f. The *canonical foliation* $\mathcal{G}_\epsilon$ (foliation of *fast trajectories*) consists of orbits (nonparameterized integral curves) of the control vector field $g(\cdot, \epsilon)$. Note that the fast trajectories are not true trajectories of Σ but only "asymptotic trajectories," corresponding to "arbitrarily large" control.

The choice of these invariants is additionally justified by the following remarks. Consider a generic system $\Sigma : \dot{\xi} = f(\xi) + ug(\xi)$, $\xi \in X$, $u \in \mathbb{R}$, and piecewise continuous control. (a) The pair of invariants $(E, \mathcal{G})$ determines the set of unparameterized trajectories of Σ in the region X, where $g(\xi) \neq 0$. Namely, the trajectories of Σ are exactly the piecewise C^1 curves in X intersecting the leaves of $\mathcal{G}$ transversally at points $p \in X \setminus E$ and tangent to the leaves at points in E. If we add to the invariants the "drift direction transversal to the leaves of $\mathcal{G}$," then the same remark is valid for oriented, unparameterized trajectories. (b) It follows from (a) that all controllability properties of Σ, expressed in terms of oriented unparameterized trajectories, are determined by the pair $(E, \mathcal{G})$ and the "drift direction." (c) Locally time-minimal and time-maximal trajectories are contained in C.

We will study bifurcations of these invariants. Roughly speaking (see Definition 2.1 for a rigorous statement), a bifurcation occurs if the triplet of basic invariants $(E_\epsilon, C_\epsilon, \mathcal{G}_\epsilon)$ at ϵ_0 is not topologically conjugated to the triplets at nearby values of ϵ. The same definition can be used for any subset of the triplet $(E_\epsilon, C_\epsilon, \mathcal{G}_\epsilon)$. In particular, we define bifurcations of the equilibrium set E_ϵ (E-bifurcations), of the critical set C_ϵ (C-bifurcations), as well as bifurcations of the pairs (E_ϵ, C_ϵ), $(E_\epsilon, \mathcal{G}_\epsilon)$, and $(C_\epsilon, \mathcal{G}_\epsilon)$.

In our considerations we will use the following local prenormal form of Σ:

$$\Sigma_{pre} \ : \quad \dot{x} = f_1(x, y, \epsilon), \quad \dot{y} = u,$$

where (x, y) are in a neighborhood of $0 \in \mathbb{R}^2$. Namely, every system Σ can be transformed to Σ_{pre} by a family of local diffeomorphisms $(x, y) = \phi_\epsilon(\xi)$, which rectify the vector fields $g(\xi, \epsilon)$ to $\partial/\partial y$, and by a static feedback transformation. For Σ_{pre},

$$E_\epsilon = \{(x, y) : f_1(x, y, \epsilon) = 0\}, \qquad C_\epsilon = \left\{(x, y) : \frac{\partial f_1}{\partial y}(x, y, \epsilon) = 0\right\},$$

and

$$\mathcal{G}_\epsilon = \{x = \text{const}\}.$$

From the last expression for C_ϵ, it is clear that C_ϵ consists of locally time-critical curves (time-minimal, if $f_1 \partial^2 f_1/\partial y^2 < 0$, and time-maximal, if $f_1 \partial^2 f_1/\partial y^2 > 0$).

Our first main result says that, generically, there are only six nonequivalent bifurcations of planar systems Σ at control-regular points. Throughout the paper, by a generic system Σ we mean a 1-parameter family of pairs (f, g) of vector fields that belongs to a dense set $\mathcal{GS}$, which is a countable intersection of open and dense subsets in the C^∞ Whitney topology of the space of all pairs (f, g) defined on $X \times I$ (see [Hi] for properties of the Whitney topology). The set of generic systems $\mathcal{GS}$ is given by the transversality conditions (G1)–(G6) in Theorem 3.3 and Lemma 4.4.

THEOREM 1.1. *Let Σ be a smooth, generic, 1-parameter family of control-affine systems. If $g(p, \epsilon_0) \neq 0$ and Σ bifurcates locally at (p, ϵ_0), then the bifurcation is equivalent to one of the following:*

(i) *an E-bifurcation, which can be of two types described in section 2.1;*
(ii) *a C-bifurcation, which can be of two types described in section 2.2;*
(iii) *an $E\mathcal{G}$-bifurcation (or an EC-bifurcation), described in section 2.3;*
(iv) *a $C\mathcal{G}$-bifurcation described in section 2.4.*

Here equivalence of bifurcations is understood as equivalence of the triples of invariants $(E_\epsilon, C_\epsilon, \mathcal{G}_\epsilon)$ under smooth, invertible local transformations of the form

$$\tilde{\xi} = \phi(\xi, \epsilon), \qquad \tilde{\epsilon} = \eta(\epsilon). \tag{Eq}$$

The above theorem is a consequence of our second main result, Theorem 3.3, which locally classifies generic families Σ under smooth orbital feedback equivalence. This C^∞ classification holds at control-regular points.

Using weaker, topological equivalence (Eq), a classification of generic bifurcations of $(E_\epsilon, C_\epsilon, \mathcal{G}_\epsilon)$ at points where g vanishes is also possible but requires different techniques. As the results in [Ru] show, there are four additional bifurcations in that case. A generic local topological classification of 1-parameter families, around points where g vanishes, gives seven nonequivalent families (four that bifurcate and three that do not).

Clearly, the set of equilibria, the set of (time) critical trajectories, and the set of fast trajectories have already proved their importance in several problems concerning control-affine planar systems. In constructing the time-optimal synthesis on $\mathbb{R}^2$ for a system $\dot{x} = f(x) + ug(x)$ with constraints $|u| \leq 1$, both the equilibria set and the critical set play an important role, while the set of fast trajectories is not significant; see [Ba1], [Ba2], [BsP], [BP], [Su1], [Su2]. In studying generic controllability problems and singularities of the boundary of the reachable set for such systems [Da1], [Da2], the equilibrium set is important, while the other two invariants do not appear. The main role is played by two vector fields, X_+ and X_-, $X_\pm = f \pm g$, and by two foliations of oriented orbits of these vector fields, called limiting directions [Da2]. In that case the problem is reduced to local classification of two generic, oriented foliations. See [Da2] and [BsP], which summarize deep studies of planar systems with constraints from controllability and optimal control points of view.

The study of bifurcations of control systems was initiated by Abed and Fu [AF1], [AF2], in a different setting, for systems of the form $\dot{\xi} = f(\xi, u, \epsilon)$. They assumed that the uncontrolled system, defined by taking $u = 0$, undergoes a bifurcation at $\epsilon = \epsilon_0$, and they studied stabilizability of the system by quadratic and cubic feedbacks.

Our approach is close in spirit to that of Kang [Ka1], who studied bifurcations of the set of equilibria and of the linear approximation of the system at an equilibrium.

A control system does not need a parameter to bifurcate—the control can play the same role. This point of view is presented by Krener, Kang, and Chang [KKC]. They consider systems $\dot{\xi} = f(\xi, u)$, for which the set of equilibria is conveniently parameterized by the control u. According to their definition, a control bifurcation takes place at an equilibrium if the linear approximation of the system looses stabilizability.

2. Bifurcations. We will use the following definition of bifurcation. For a subset $\Omega \subset X \times I$ and a fixed parameter $\epsilon \in I = (a, b)$ we denote $\Omega_\epsilon = \{\xi \in X : (\xi, \epsilon) \in \Omega\}$. Assume $0 \in I$. We denote the system Σ with a fixed value ϵ of parameter by Σ_ϵ.

DEFINITION 2.1. *We say that the family Σ does not bifurcate, locally at $(\xi_0, \epsilon_0) = (p, 0)$, if there exits a neighborhood $\Omega \subset X \times I$ of $(p, 0)$ and a family of homeomorphisms $\chi_\epsilon : \Omega_\epsilon \to \Omega_0$, continuous with respect to (ξ, ϵ), such that for Σ_ϵ restricted to Ω_ϵ we have*

$$\chi_\epsilon(E_\epsilon) = E_0, \quad \chi_\epsilon(C_\epsilon) = C_0, \quad \text{and} \quad \chi_\epsilon(\mathcal{G}_\epsilon) = \mathcal{G}_0$$

for all $\epsilon \in I$ close enough to 0. Otherwise we say that Σ bifurcates locally *or has a* local bifurcation *at $(\xi, \epsilon) = (p, 0)$.*

Analogous definition applies to bifurcations at arbitrary (ξ_0, ϵ_0). Strictly speaking, in our definition we should say that the triple $(E_\epsilon, C_\epsilon, \mathcal{G}_\epsilon)$ bifurcates or that Σ bifurcates with respect to $(E_\epsilon, C_\epsilon, \mathcal{G}_\epsilon)$. The same definition will be used for any subset of the triplet $(E_\epsilon, C_\epsilon, \mathcal{G}_\epsilon)$. In particular, we define bifurcations of the equilibrium set E_ϵ, of the critical set C_ϵ, and of the pairs (E_ϵ, C_ϵ), $(E_\epsilon, \mathcal{G}_\epsilon)$, and $(C_\epsilon, \mathcal{G}_\epsilon)$.

In the above definition we use a family of homeomorphisms χ_ϵ. Another possibility is to use, instead, a family of smooth diffeomorphisms ϕ_ϵ. We will call the corresponding bifurcations *topological* and *differential* bifurcations. In turns out, however, that in both topological and smooth categories the bifurcations of generic 1-parameter families of control-affine planar systems are the same at control-regular points.

DEFINITION 2.2. *We say that two (local) bifurcations of Σ_ϵ and $\tilde{\Sigma}_\epsilon$ are locally equivalent if there is a local, smooth, invertible transformation $(\tilde{\xi}, \tilde{\epsilon}) = (\phi(\xi, \epsilon), \eta(\epsilon))$ which transforms the triple $(E_\epsilon, C_\epsilon, \mathcal{G}_\epsilon)$ into the triple $(\tilde{E}_{\tilde{\epsilon}}, \tilde{C}_{\tilde{\epsilon}}, \tilde{\mathcal{G}}_{\tilde{\epsilon}})$.*

Assuming that the families of vector fields $f(\xi, \epsilon)$ and $g(\xi, \epsilon)$ are expressed in coordinates on $X \subset \mathbb{R}^2$ as column vectors, we define the functions of (ξ, ϵ):

$$e = \det(f, g), \qquad c = \det([g, f], g).$$

For Σ in the prenormal form we have $e(x, y, \epsilon) = f_1(x, y, \epsilon)$ and $c(x, y, \epsilon) = \frac{\partial f_1}{\partial y}(x, y, \epsilon)$.

Below we will describe in detail all generic planar bifurcations. To simplify the notation, we will denote the state of the transformed system $\tilde{\Sigma}_{\tilde{\epsilon}}$ by $\tilde{\xi} = (x, y)$, its control by $\tilde{u} = v$, and its parameter by $\tilde{\epsilon} = \epsilon$. This system will be in the prenormal form Σ_{pre}. Then its canonical foliation will be in the form

$$\mathcal{G}_\epsilon = \{x = \text{const}\}.$$

Let $\varphi_\epsilon(x, y) = \varphi(x, y, \epsilon)$ be a smooth function. We will denote by $d\varphi$ the differential of φ in the (x, y)-space and by $D\varphi$ the differential in the (x, y, ϵ)-space, that is,

$$d\varphi = \left(\frac{\partial \varphi}{\partial x}, \frac{\partial \varphi}{\partial y}\right), \qquad D\varphi = \left(\frac{\partial \varphi}{\partial x}, \frac{\partial \varphi}{\partial y}, \frac{\partial \varphi}{\partial \epsilon}\right).$$

By $\operatorname{hess}(\varphi)$ we denote the matrix of second order partial derivatives of φ with respect to x, y, and $L_g\varphi$ denotes the derivative of φ along g.

2.1. *E*-bifurcations. Consider Σ around a point $(p, \epsilon_0) \in X \times I$ such that $g(p, \epsilon_0) \neq 0$ and assume that the equilibrium set E_ϵ locally bifurcates while the invariants C_ϵ and $\mathcal{G}_\epsilon$ do not bifurcate at ϵ_0. We will call such a phenomenon an *equilibria bifurcation* or an *E-bifurcation.*

PROPOSITION 2.3. *For a generic Σ, an E-bifurcation appears at a point $(\xi, \epsilon) = (p, \epsilon_0)$ if and only if the following conditions are fulfilled at this point:*

$$e = 0, \quad c = 0, \quad L_g c \neq 0, \quad de = 0, \quad \det \operatorname{hess}(e) \neq 0, \quad \text{and} \quad De \neq 0.$$

Such a bifurcation is equivalent to one of the two which are described by the following families of invariants: $C_\epsilon = \{y = 0\}$, $\mathcal{G}_\epsilon = \{x = \text{const}\}$ and

$$E_\epsilon = \{x^2 + y^2 = \epsilon\} \quad \text{or} \quad E_\epsilon = \{x^2 - y^2 = \epsilon\}.$$

Moreover, $E_\epsilon = \{x^2 + y^2 = \epsilon\}$ corresponds to $\det \operatorname{hess}(e) > 0$, while $E_\epsilon = \{x^2 - y^2 = \epsilon\}$ corresponds to $\det \operatorname{hess}(e) < 0$.

Description of the *E*-bifurcations. In the first E-bifurcation, called *birth of equilibria*, the equilibrium set E_ϵ is empty, for $\epsilon < 0$, it consists of a single point

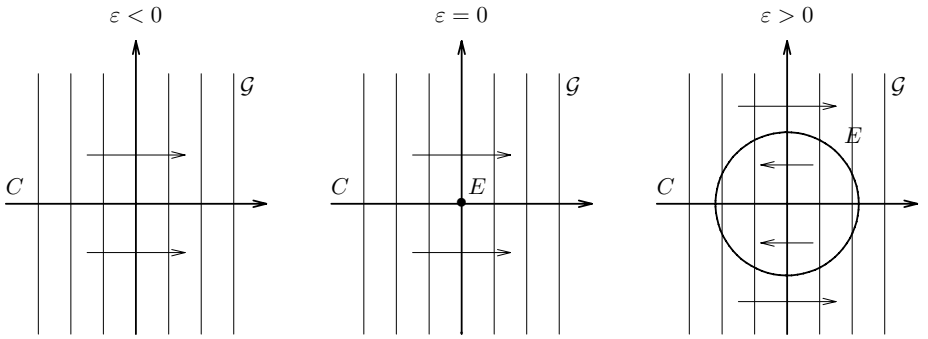


FIG. 2.1. *E-bifurcation (birth of equilibria).*

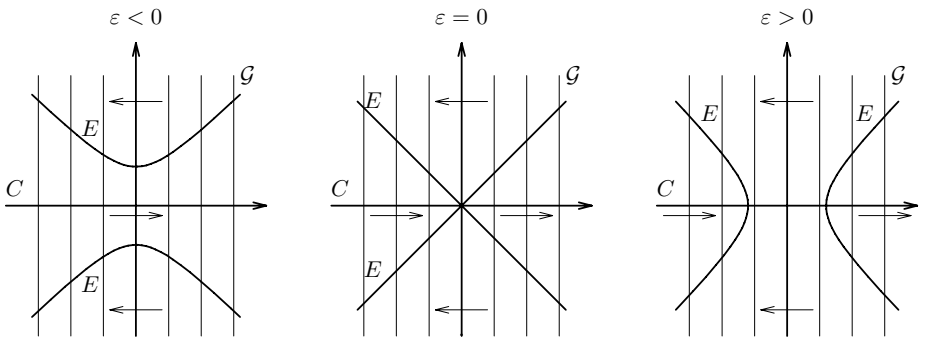


FIG. 2.2. *E-bifurcation (cross of equilibria).*

$(0,0) \in \mathbb{R}^2$, for $\epsilon = 0$, and of a circle of equilibria, for $\epsilon > 0$ (see Figure 2.1). The arrows in the figures will symbolically denote the direction of the drift f, transverse to $\mathcal{G}_\epsilon$.

In the second E-bifurcation, called *cross of equilibria* (Figure 2.2), the equilibrium set E_ϵ consists of two curves that do not intersect, for $\epsilon \neq 0$. The two curves approach each other when ϵ tends to zero and, for $\epsilon = 0$, the set E_0 is formed by two curves which intersect (transversally, if we consider differential bifurcations).

Notice that the fact that E_ϵ bifurcates implies that the pairs (E_ϵ, C_ϵ) and $(E_\epsilon, \mathcal{G}_\epsilon)$ bifurcate as well (although neither C_ϵ nor $\mathcal{G}_\epsilon$, considered separately, bifurcates). For instance, in the case of the birth of equilibria bifurcation, the curves E_ϵ and C_ϵ do not intersect for $\epsilon < 0$ (since E_ϵ is empty), their intersection is one point, for $\epsilon = 0$, and consists of two points for $\epsilon > 0$.

2.2. *C*-bifurcations. Consider a system Σ around a point (p, ϵ_0) such that $g(p, \epsilon_0) \neq 0$ and assume that the critical set C_ϵ locally bifurcates while the equilibrium set E_ϵ and the canonical foliation $\mathcal{G}_\epsilon$ do not bifurcate at $\epsilon = \epsilon_0$. We will call such a bifurcation *critical set bifurcation* or *C-bifurcation*. Generically, only two such bifurcations are possible.

PROPOSITION 2.4. *A generic family Σ has a C-bifurcation at a point $(\xi, \epsilon) = (p, \epsilon_0)$ if and only if the following conditions are fulfilled at this point:*

$$e \neq 0, \quad c = 0, \quad dc = 0, \quad L_g^2 c \neq 0, \quad \det \operatorname{hess}(c) \neq 0, \quad Dc \neq 0.$$

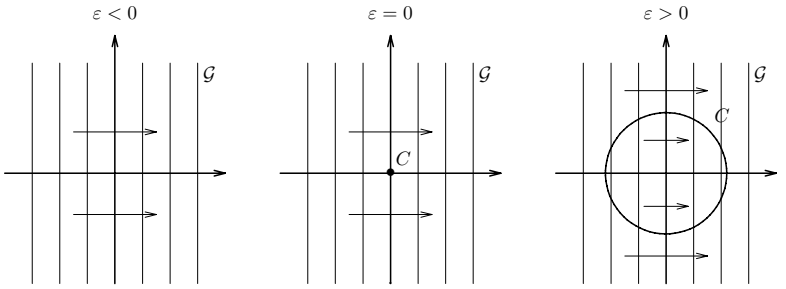


FIG. 2.3. *C-bifurcation (birth of critical curve).*

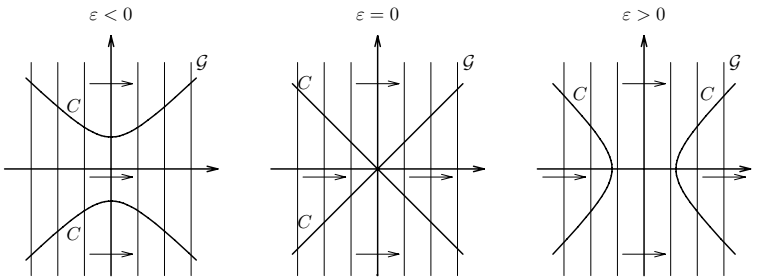


FIG. 2.4. *C-bifurcation (cross of critical curves).*

Such a bifurcation is equivalent to one of the two which are described by the following invariants: $E_\epsilon = \emptyset$ *(empty set),* $\mathcal{G}_\epsilon = \{x = \text{const}\}$ *and*

$$C_\epsilon = \{x^2 + y^2 = \epsilon\} \quad \text{or} \quad C_\epsilon = \{x^2 - y^2 = \epsilon\}.$$

In these bifurcations, the critical curve bifurcates in exactly the same way as the equilibria curve does in the E-bifurcations, which is illustrated in Figures 2.3 and 2.4 (the first corresponds to $\det \operatorname{hess}(c) > 0$, while the second corresponds to $\det \operatorname{hess}(c) < 0$).

2.3. *EG*-bifurcations or *EC*-bifurcations. Consider a generic family Σ such that neither of the invariants E_ϵ, C_ϵ, $\mathcal{G}_\epsilon$ bifurcates, while the pair $(E_\epsilon, \mathcal{G}_\epsilon)$ bifurcates. We call this *EG-bifurcation*. Only one such bifurcation is possible.

PROPOSITION 2.5. *A generic family* Σ *has a local EG-bifurcation at a point* (p, ϵ_0) *if and only if the conditions*

$$e = 0, \quad c = 0, \quad L_g c = 0, \quad L_g^2 c \neq 0, \quad de \neq 0, \quad dc \neq 0$$

are fulfilled at this point and the differentials De and Dc are linearly independent. Such a bifurcation is locally equivalent to one with the invariants $\mathcal{G}_\epsilon = \{x = \text{const}\}$ *and*

$$E_\epsilon = \{y^3 + (x - \epsilon)y + x = 0\},$$

$$C_\epsilon = \{3y^2 + x - \epsilon = 0\}.$$

Description of the *EG*-bifurcation. Both the equilibria set $E_\epsilon = \{e_\epsilon = 0\}$ and the critical set $C_\epsilon = \{c_\epsilon = 0\}$ are smooth curves and, considered separately, they do not bifurcate. The equilibria curve E_ϵ passes, for any value of ϵ, through the

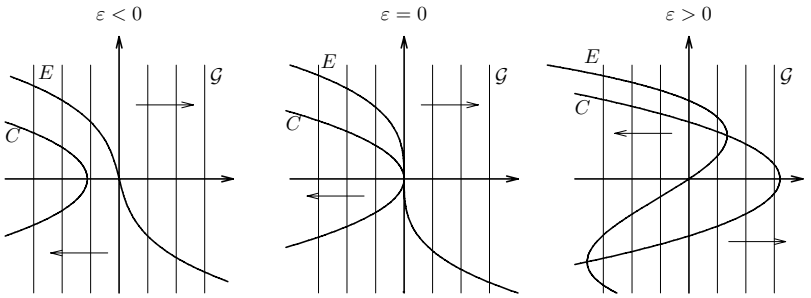


FIG. 2.5. *EG-bifurcation,* $\gamma = 1$.

origin $(0,0)$ (Figure 2.5, with the drift direction chosen so that $\gamma = 1$ in Theorem 3.3). We can equivalently write $E_\epsilon = \{x = y(\epsilon - y^2)(y+1)^{-1}\}$. Thus, the curve E_ϵ intersects each curve $\{x = c = \text{const}\}$ of $\mathcal{G}_\epsilon$ at zeros of the function $f(y) = y(\epsilon - y^2)(y+1)^{-1} - c$, i.e., it intersects it once if $\epsilon \leq 0$ and three times if $\epsilon > 0$ (for $x = c$ small enough).

Observe that the pair (E_ϵ, C_ϵ) also has a bifurcation. Eliminating x from the equations for E_ϵ and C_ϵ we get $2y^3 + 3y^2 = \epsilon$. For small y, the function on the left behaves like $3y^2$, which gives no solutions, if $\epsilon < 0$, and two solutions, if $\epsilon > 0$. Thus, for $\epsilon < 0$, the curves E_ϵ and C_ϵ do not intersect, for $\epsilon = 0$ they just intersect at $(0,0) \in \mathbb{R}^2$ (and are tangent), for $\epsilon > 0$ they intersect at two different points close to $(0,0)$. At the points of intersection, the curve E_ϵ is tangent to the curves of $\mathcal{G}_\epsilon$. Thus the above bifurcation can be called either an $E\mathcal{G}$-bifurcation or an EC-bifurcation.

2.4. $C\mathcal{G}$-bifurcations. We will analyze families Σ for which neither of the invariants E_ϵ, C_ϵ, $\mathcal{G}_\epsilon$, considered separately, bifurcates while the pair $(C_\epsilon, \mathcal{G}_\epsilon)$ bifurcates. We will call this a *$C\mathcal{G}$-bifurcation*. For generic Σ only one such bifurcation appears.

PROPOSITION 2.6. *For a generic family Σ a local $C\mathcal{G}$-bifurcation appears at a point (p, ϵ_0) if and only if the conditions*

$$e \neq 0, \quad c = 0, \quad L_g c = 0, \quad L_g^2 c = 0, \quad L_g^3 c \neq 0, \quad dc \neq 0, \quad d(L_g c) - (L_g^4 c / 7 L_g^3 c)\, dc \neq 0$$

are fulfilled at this point and the differentials Dc and $D(L_g c)$ are linearly independent. Such a $C\mathcal{G}$-bifurcation is locally equivalent to one having $E_\epsilon = \emptyset$, $\mathcal{G}_\epsilon = \{x = \text{const}\}$, and

$$C_\epsilon = \{y^3 + (x - \epsilon)y + x = 0\}.$$

Description of the $C\mathcal{G}$-bifurcation. Here, for any ϵ the critical curve C_ϵ passes through the origin in $\mathbb{R}^2$. We can equivalently write $C_\epsilon = \{x = y(\epsilon - y^2)(y+1)^{-1}\}$. Thus, the critical curve C_ϵ intersects each curve $\{x = c = \text{const}\}$ of $\mathcal{G}_\epsilon$ at zeros of the function $f(y) = y(\epsilon - y^2)(y+1)^{-1} - c$, i.e., it intersects it once if $\epsilon \leq 0$ and three times if $\epsilon > 0$ for $x = c$ small enough (Figure 2.6, with $\theta = 1$ and with the drift direction corresponding to $a(0,0) > 0$ in Theorem 3.3).

Observe that neither of the invariants E_ϵ, C_ϵ, $\mathcal{G}_\epsilon$, considered separately, bifurcates. Since $E = \emptyset$, there is neither $E\mathcal{G}$-bifurcation nor EC-bifurcation in this case.

When considering differential bifurcations, we can give an alternative description of the above $C\mathcal{G}$-bifurcation. The critical curve $C_\epsilon = \{c_\epsilon = 0\}$ is a regular curve passing through the origin for any value of ϵ. If $\epsilon < 0$, the critical curve intersects transversally all integral curves $\{x = \text{const}\}$ of $\mathcal{G}_\epsilon$. For $\epsilon = 0$ the critical curve C_0 is tangent at $(0,0) \in \mathbb{R}^2$ to the integral curve $\{x = 0\}$ of $\mathcal{G}_0$. Finally, for $\epsilon > 0$ it becomes tangent to exactly two integral curves of $\mathcal{G}_\epsilon$ at two points close to $(0,0)$.

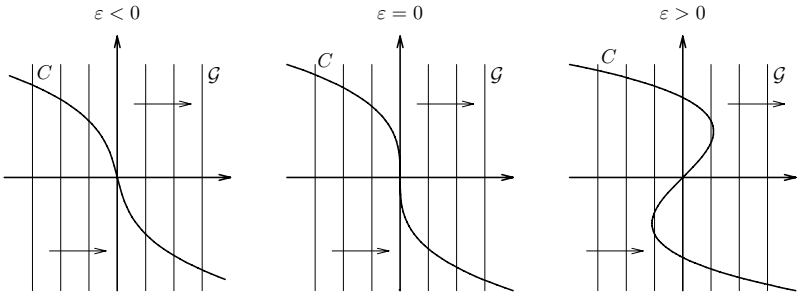


FIG. 2.6. *$C\mathcal{G}$-bifurcation, $a(0,0) > 0$.*

3. Orbital classification of 1-parameter families.

3.1. Equivalence of families. Consider a family of control systems on $X \subset \mathbb{R}^2$

$$\Sigma \ : \quad \dot{\xi} = f(\xi, \epsilon) + g(\xi, \epsilon)u, \tag{3.1}$$

where $u \in \mathbb{R}$ and $\epsilon \in I$ (open interval). Here $f(\xi, \epsilon) = f_\epsilon(\xi)$ and $g(\xi, \epsilon) = g_\epsilon(\xi)$ are families of vector fields on X, C^∞-smooth with respect to (ξ, ϵ).

Consider C^∞ local invertible transformations $X \times \mathbb{R} \times I \to X \times \mathbb{R} \times I$ of the form

$$\Gamma \ : \ \begin{array}{rcl} \tilde{\xi} & = & \phi(\xi, \epsilon) = \phi_\epsilon(\xi), \\ \tilde{u} & = & \psi(\xi, u, \epsilon) = \psi_\epsilon(\xi, u), \\ \tilde{\epsilon} & = & \eta(\epsilon), \end{array}$$

where ψ is affine with respect to u so that $u = \psi^{-1}(\xi, \tilde{u}, \epsilon) = \alpha(\xi, \epsilon) + \beta(\xi, \epsilon)\tilde{u}$. Invertibility at (ξ_0, ϵ_0) means that $d\phi_{\epsilon_0}(\xi_0)$ is of rank 2 and $\beta(\xi_0, \epsilon_0) \neq 0$, $\eta'(\epsilon_0) \neq 0$.

A *feedback transformation* $\Gamma = (\phi, \psi, \eta)$ brings Σ into $\tilde{\Sigma}$: $\dot{\tilde{\xi}} = \tilde{f}(\tilde{\xi}, \tilde{\epsilon}) + \tilde{g}(\tilde{\xi}, \tilde{\epsilon})\tilde{u}$, with $\tilde{f}_{\tilde{\epsilon}} = \tilde{f}(\cdot, \tilde{\epsilon})$ and $\tilde{g}_{\tilde{\epsilon}} = \tilde{g}(\cdot, \tilde{\epsilon})$ given by

$$\tilde{f}_{\tilde{\epsilon}} = (\phi_\epsilon)_*(f_\epsilon + \alpha_\epsilon g_\epsilon), \qquad \tilde{g}_{\tilde{\epsilon}} = (\phi_\epsilon)_*(\beta_\epsilon g_\epsilon),$$

where $\tilde{\epsilon} = \eta(\epsilon)$, $\alpha_\epsilon = \alpha(\cdot, \epsilon)$, and $\beta_\epsilon = \beta(\cdot, \epsilon)$. Throughout, for any vector field f and any diffeomorphism $\tilde{\xi} = \phi(\xi)$ we denote $(\phi_* f)(\tilde{\xi}) = d\phi(\xi) \cdot f(\xi)$, where $\xi = \phi^{-1}(\tilde{\xi})$.

An *orbital feedback transformation* $\Gamma_{orb} = (\phi, \psi, h, \eta)$ contains additionally a family of positive valued smooth functions $h_\epsilon(\xi) = h(\xi, \epsilon)$ which change the time scale of the system according to $dt/d\tau = h(\xi, \epsilon)$. Thus, by definition, Γ_{orb} transforms Σ into $\tilde{\Sigma}$ with $\tilde{f}_{\tilde{\epsilon}} = (\phi_\epsilon)_*(h_\epsilon f_\epsilon + h_\epsilon \alpha_\epsilon g_\epsilon)$ and $\tilde{g}_{\tilde{\epsilon}} = (\phi_\epsilon)_*(h_\epsilon \beta_\epsilon g_\epsilon)$. We can incorporate the action of h_ϵ on g_ϵ by choosing $\hat{\alpha}_\epsilon = h_\epsilon \alpha_\epsilon$ and $\hat{\beta}_\epsilon = h_\epsilon \beta_\epsilon$. The transformation becomes

$$\tilde{f}_{\tilde{\epsilon}} = (\phi_\epsilon)_*(h_\epsilon f_\epsilon + \hat{\alpha}_\epsilon g_\epsilon), \qquad \tilde{g}_{\tilde{\epsilon}} = (\phi_\epsilon)_*(\hat{\beta}_\epsilon g_\epsilon), \tag{3.2}$$

where $\tilde{\epsilon} = \eta(\epsilon)$. Throughout the paper we assume that the functions h_ϵ are positive valued and constant on the trajectories of g_ϵ, i.e., the derivative of h_ϵ along g_ϵ vanishes:

$$L_{g_\epsilon} h_\epsilon = 0. \tag{3.3}$$

We shall suppress the dependence on ϵ in our notation, writing the relations (3.2), (3.3) in the form $\tilde{f} = \phi_*(hf + \hat{\alpha}g)$, $\tilde{g} = \phi_*(\hat{\beta}g)$, and $L_g h = 0$.

DEFINITION 3.1. *Two families Σ and $\tilde{\Sigma}$ are called* locally orbitally feedback equivalent *(or* orbitally equivalent*) at (ξ_0, ϵ_0) and $(\tilde{\xi}_0, \tilde{e}_0)$, respectively, if there exists a local, invertible at (ξ_0, ϵ_0), C^∞-transformation $\Gamma_{orb} = (\phi, \psi, h, \eta)$, satisfying*

$(\phi,\eta)(\xi_0,\epsilon_0)=(\tilde\xi_0,\tilde\epsilon_0)$, *with* $h(\xi,\epsilon)$ *positive valued and* $L_gh=0$, *transforming* Σ *into* $\tilde\Sigma$. *If* $h\equiv 1$, *then we say that* Σ *and* $\tilde\Sigma$ *are locally feedback equivalent.*

Orbital feedback equivalence, defined by relations (3.2) and (3.3) (introduced in [J1] as "mild feedback equivalence"), allows one to get a simple classification of generic families. The equivalence does not change basic features of Σ. Namely, consider again the functions

$$e=\det(f,g),\qquad c=\det([g,f],g).$$

Replacing f and g by the orbitally equivalent pair

$$\tilde f=h\,f+\alpha\,g,\qquad \tilde g=\beta\,g\qquad\text{with}\quad L_gh=0$$

gives $[\tilde g,\tilde f]=h\beta[g,f]+\varphi g$, where $\varphi=\beta L_g(\alpha)-\alpha L_g(\beta)-hL_f(\beta)$, and so

$$\tilde e=h\beta\,e,\qquad \tilde c=h\beta^2c. \tag{3.4}$$

The following fact follows then immediately from (3.4) and from the equalities $E_\epsilon=\{p\in X\ :\ e(p,\epsilon)=0\}$ and $C_\epsilon=\{p\in X\ :\ c(p,\epsilon)=0\}$.

PROPOSITION 3.2. *If* Σ *and* $\tilde\Sigma$ *are orbitally feedback equivalent, via the transformation* $\Gamma_{orb}=(\phi,\psi,h,\eta)$, *then the ideals* (e) *and* (c) *generated by the functions* $e(\xi,\epsilon)$ *and* $c(\xi,\epsilon)$, *respectively, are transformed by the map* (ϕ,η) *(remain unchanged, if* $(\phi,\eta)=id$*). In particular, with* $\tilde\epsilon=\eta(\epsilon)$, *we have*

$$\phi_\epsilon(E_\epsilon)=\tilde E_{\tilde\epsilon},\quad \phi_\epsilon(C_\epsilon)=\tilde C_{\tilde\epsilon},\quad\text{and}\quad \phi_\epsilon(\mathcal{G}_\epsilon)=\tilde{\mathcal{G}}_{\tilde\epsilon}.$$

The above property of E, C, $\mathcal{G}$ and of the ideals (e), (c) is called *equivariance*. Thus E, C, $\mathcal{G}$ are said to be *equivariant* or, by abuse of language, *invariant* with respect to orbital feedback equivalence.

3.2. Classification theorem. Let Σ be a smooth family of systems (3.1). Consider the functions $e=\det(f,g)$ and $c=\det([g,f],g)$ depending on (ξ,ϵ). We introduce the sequence of functions of (ξ,ϵ):

$$c^0=e,\quad c^1=c,\qquad c^k=L_g^{k-1}c,\quad k\ge 2.$$

Denote by $j(c^i,c^{i+1})$ and $J(c^i,c^{i+1},c^{i+2})$ the Jacobians with respect to ξ and (ξ,ϵ),

$$j(c^i,c^{i+1})=\det\left(dc^i,dc^{i+1}\right),$$

$$J(c^i,c^{i+1},c^{i+2})=\det\left(Dc^i,Dc^{i+1},Dc^{i+2}\right),$$

where $dh=(\partial h/\partial\xi_1,\partial h/\partial\xi_2)$ and $Dh=(\partial h/\partial\xi_1,\partial h/\partial\xi_2,\partial h/\partial\epsilon)$.

The following theorem classifies generic Σ under the orbital feedback equivalence. To simplify the exposition, we denote the state of the transformed system by $\tilde\xi=(x,y)\in\mathbb{R}^2$ and its control by $\tilde u=v$. In the last column below we list codes of invariants (they are defined after the theorem) describing the equivalence class. The diagrams in Figures 3.1 and 3.2 explain the structure of the classification as well as relations between the codes and the corresponding invariant conditions expressed in terms of e, $c^1,c^2,\dots$, and their differentials.

THEOREM 3.3. *A generic family Σ, given by* (3.1)*, is locally orbitally feedback equivalent, around any point at which g does not vanish, to one of the following canonical forms at $0 \in \mathbb{R}^2$ and $\epsilon = 0$, where the code appearing in the last column characterizes the systems equivalent to the normal form:*

$$
\begin{array}{lllll}
(O) & \dot{x} = y + 1, & \dot{y} = v, & (11), \\
(E) & \dot{x} = y, & \dot{y} = v, & (01), \\
(C) & \dot{x} = \tau y^2 + 1, & \dot{y} = v, & (101), \\
(EC) & \dot{x} = y^2 + \gamma x, & \dot{y} = v, & (0_1 01), \\
(C\mathcal{G}) & \dot{x} = \delta y^3 + xy + 1, & \dot{y} = v, & (10_1 01), \\
(E_{bif}) & \dot{x} = \sigma_e y^2 + x^2 - \epsilon, & \dot{y} = v, & (\overline{0_{01}0}1), \\
(C_{bif}) & \dot{x} = \sigma_c y^3 + (x^2 - \epsilon)y + 1, & \dot{y} = v, & (1\overline{0_{01}0}1), \\
(E\mathcal{G}_{bif}) & \dot{x} = y^3 + (x - \epsilon)y + \gamma x, & \dot{y} = v, & (\overline{0_1 0_1 0}1). \\
(C\mathcal{G}_{bif}) & \dot{x} = y^4 + (\theta x - \epsilon)y^2 + xy + a(x, \epsilon), & \dot{y} = v, & (1\overline{0_1 0_1 0}1)_{mod}.
\end{array}
$$

Above a is a smooth function of (x, ϵ) satisfying $a(0,0) \neq 0$ and $\operatorname{sgn} a(0,0) = \kappa$. The integers τ, γ, δ, σ_e, σ_c, θ, and κ take values ± 1 and are orbitally feedback invariant.

The class of families (3.1)*, which are locally orbitally equivalent at (p, ϵ_0) to one of the above canonical forms, is characterized by the following conditions at (p, ϵ_0):*

$$
\text{(G1)} \qquad (c^i, c^{i+1}, c^{i+2}, c^{i+3}) \neq (0,0,0,0) \qquad \text{for} \quad i = 0, 1,
$$

$$
\text{(G2)} \qquad (c^i, c^{i+1}, c^{i+2}, dc^i) \neq (0,0,0,0) \qquad \text{for} \quad i = 0, 1,
$$

$$
\text{(G3)} \qquad (c^i, Dc^i) \neq (0,0) \qquad \text{for} \quad i = 0, 1,
$$

$$
\text{(G4)} \qquad (c^i, dc^i, \det \operatorname{hess}(c^i)) \neq (0,0,0) \qquad \text{for} \quad i = 0, 1,
$$

$$
\text{(G5)} \qquad (c^i, c^{i+1}, c^{i+2}, J(c^i, c^{i+1}, c^{i+2})) \neq (0,0,0,0) \qquad \text{for} \quad i = 0, 1,
$$

$$
\text{(G6)} \qquad (c^i, c^{i+1}, c^{i+2}, dc^{i+1}_{mod}) \neq (0,0,0,0) \qquad \text{for} \quad i = 0, 1.
$$

Above, in condition (G6)*, we use the functions $c^1_{mod} = c^1$ and $c^2_{mod} = c^2 - (c^5/7c^4)c^1$.*

The conditions (G1)–(G5) *are equivalent to the following condition:*

$$
\text{(G)} \qquad (c^i, c^{i+1}, j(c^i, c^{i+1}), J(c^i, c^{i+1}, j(c^i, c^{i+1}))) \neq (0,0,0,0) \qquad \text{for} \quad i = 0, 1.
$$

Remark 3.1. Notice that any orbital feedback transformation preserving the prenormal form $\Sigma_{pre} : \dot{x} = f_1(x, y, \epsilon)$, $\dot{y} = u$ satisfies $\tilde{x} = \phi(x, \epsilon)$. Therefore the orbital feedback classification of families Σ of control-affine systems in the plane reduces to the classification of families of systems $\dot{x} = f_1(x, v, \epsilon)$ on the line under the invertible transformations $\tilde{x} = \phi(x, \epsilon)$, $\tilde{v} = \psi(x, v, \epsilon)$, $\tilde{e} = \eta(\epsilon)$ and time rescaling $dt = h(x, \epsilon)d\tau$. The above theorem can be reformulated as local classification of such systems.

Remark 3.2. For generic families Σ, the sets $E = \{(p, \epsilon) : e(p, \epsilon) = 0\}$ and $C = \{(p, \epsilon) : c(p, \epsilon) = 0\}$ are submanifolds in $X \times I$. Their singularities of contact with the foliations $\{\epsilon = \text{const}\}$ and $\{\epsilon = \text{const}, x = \text{const}\}$ (in the notation of the above theorem) are very simple. For more degenerated families this aspect of our

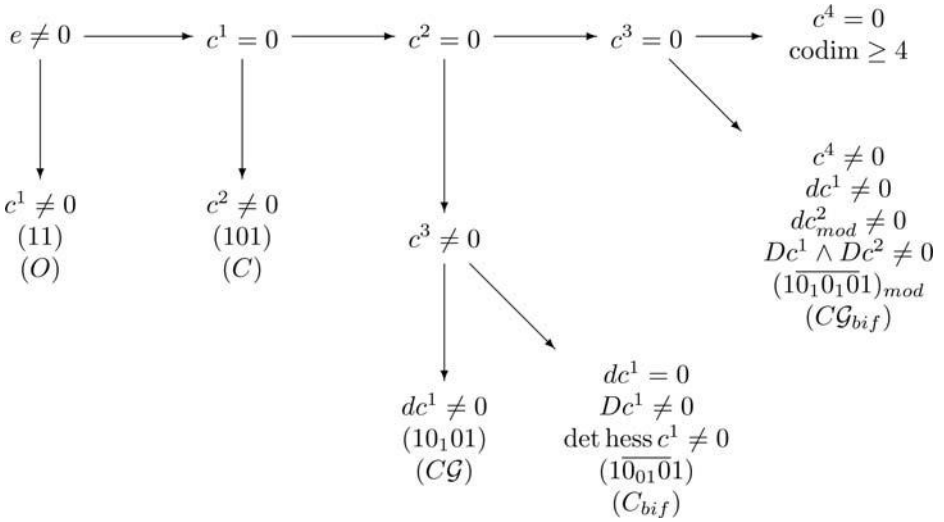


FIG. 3.1. *Bifurcation diagram out of equilibria.*

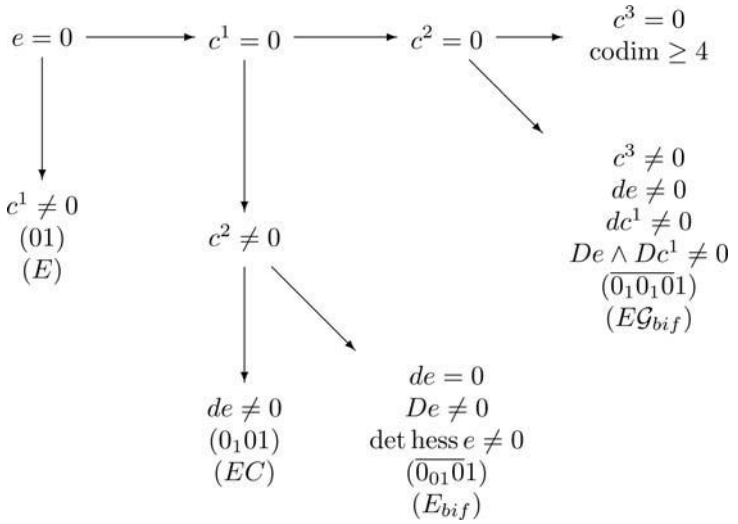


FIG. 3.2. *Bifurcation diagram at equilibria.*

problem is nontrivial. In understanding general singularities of contact of the sets E and C the results in [Go] and [Za] should be helpful.

To characterize the equivalence classes we encode our invariants (equivariants) in a compact way by introducing the following notation. We define the *code of the first level* of Σ at the point (p, ϵ_0) to be equal to $(\chi^0\ \chi^1\ \cdots)$, where $\chi^i = 0$ if $c^i(p, \epsilon_0) = 0$ and $\chi^i = 1$ if $c^i(p, \epsilon_0) \neq 0$. We will need only sequences of finite length, namely, we cut the sequence χ^i after 1 appears for the first time, for $i \geq 1$ (further χ^i are not invariant). Thus codes of the first level are of minimal length 2, of the form

$$(0 \cdots 01) \quad \text{or} \quad (10 \cdots 01).$$

Note that if a function $\varphi : X \to \mathbb{R}$ vanishes at p, then the next invariant fact about the ideal generated by φ is vanishing or not of $d\varphi$ at p. This will be encoded by 0_0 (meaning $\varphi(p) = 0$, $d\varphi(p) = 0$) and 0_1 (meaning $\varphi(p) = 0$, $d\varphi(p) \neq 0$). If φ depends also on ϵ, we still mean the differential $d\varphi$ taken with respect to $\xi \in X$, only.

We define the *code of the second level* of the family Σ, at a given point, as its code of the first level with each zero in it having the subscript 0 or 1, depending on whether the differential dc^i of the corresponding c^i vanishes. Since for each 0 which is followed by 1 in the code of the first level we have $dc^i \neq 0$ (because $L_g c^i = c^{i+1}$ does not vanish), we should always place the subscript 1. This subscript will be omitted. Thus, for example, instead of the code $(10_1 0_1 1)$ we write $(10_1 01)$ and this means that, at the given point, $e \neq 0$, $L_g c = L_g^2 c = 0$, $L_g^3 c \neq 0$, and $dc \neq 0$.

If $\varphi : X \to \mathbb{R}$ and $\varphi(p) = 0$, $d\varphi(p) = 0$, then the next invariant fact about the ideal generated by φ is whether the Hessian $\operatorname{hess}(\varphi)$ has full rank at p. We write

$$0_{01} \quad \text{if} \quad \varphi(p) = 0,\ d\varphi(p) = 0,\ \text{and}\ \det \operatorname{hess}(\varphi)(p) \neq 0,$$

$$0_{00} \quad \text{if} \quad \varphi(p) = 0,\ d\varphi(p) = 0,\ \text{and}\ \det \operatorname{hess}(\varphi)(p) = 0.$$

(The point p is replaced by (p, ϵ_0), if φ depends on ϵ.)

The *code of the third level* of the system at a given point is defined as its code of the second level, completed with the information about vanishing or not of the determinant of the Hessian. We put the subscript 01 to any of the first two or three functions in the sequence $c^0, c^1, \ldots$, if it vanishes at the given point, its differential also vanishes, and the determinant of the Hessian is nonzero. For example, the system has the code $(10_{01}01)$ if and only if, at the given point, $e \neq 0$, $c^1 = c^2 = 0$, $c^3 \neq 0$, $dc^1 = 0$, and $\det \operatorname{hess}(c) \neq 0$. The subscript 00 will not appear in our classification since it describes singularities of higher codimension.

Finally, the *complete code*, or simply the *code*, is the code of appropriate level completed with the information about linear independence of the differentials Dc^i of the functions c^i, which is marked by overlying the corresponding zeros. We add this information only in the cases where it does not follow from the basic code. For example, the code $(10_1 01)$ is complete because it follows from the code that $dc^1 \neq 0$, $L_g c^1 = 0$, and $L_g c^2 \neq 0$, which means that dc^1 and dc^2 are linearly independent and thus so are Dc^1 and Dc^2. The complete code $(\overline{0_1 0_1 0}1)$ is equivalent to the conditions $c^0 = e = 0$, $c^1 = c = 0$, $c^2 = 0$, $c^3 \neq 0$, $dc^0 \neq 0$, $dc^1 \neq 0$ and to linear independence of Dc^0, Dc^1, and Dc^2. They are equivalent to the conditions stated in Proposition 2.5 (linear independence of Dc^0, Dc^1 is equivalent to linear independence of Dc^0, Dc^1, and Dc^2, since $L_g c^0 = 0$, $L_g c^1 = 0$, and $L_g c^2 \neq 0$).

The complete code $(1\overline{0_1 0_1 0}1)$ means that, at the given point, $e \neq 0$, $c^1 = c^2 = c^3 = 0$, $c^4 \neq 0$, $dc^1 \neq 0$, $dc^2 \neq 0$, and Dc^1, Dc^2, Dc^3 are linearly independent. The condition $dc^2 \neq 0$ here is not feedback invariant; therefore in the theorem we use a modified version of this code $(1\overline{0_1 0_1 0}1)_{mod}$, which means that we replace the condition $dc^2 \neq 0$ with the condition $dc^2_{mod} = dc^2 - (c^5/7c^4)dc^1 \neq 0$ (which is orbitally feedback invariant; cf. Lemma 3.9). The set of conditions defined by $(1\overline{0_1 0_1 0}1)_{mod}$ is equivalent to the condition stated in Proposition 2.6.

We will show in section 3.3 that the following holds.

PROPOSITION 3.4. (i) *The conditions* (G1)–(G6)*, as well as* (G) *and the conditions defined by the codes listed in Theorem* 3.3*, are invariant under orbital feedback equivalence.* (ii) *The conditions* (G1)–(G5) *are equivalent to the condition* (G).

3.3. More about invariants. We will systematize the already introduced invariants and describe some new ones which appeared in the classification theorem. We define them for the system Σ given by $\dot{\xi} = f(\xi) + ug(\xi)$ and assume that $g(p) \neq 0$. (Analogous definitions are valid for the family $\dot{\xi} = f(\xi, \epsilon) + ug(\xi, \epsilon)$, with p replaced

by (p, ϵ_0) and $g(p, \epsilon_0) \neq 0$.) At nonequilibrium points it is convenient to replace the system with the 1-form ω uniquely defined by the equations (cf., e.g., [Su1])

$$\omega(g) \equiv 0, \qquad \omega(f) \equiv 1.$$

Replacing f, g by an equivalent pair $\tilde{f} = hf + \alpha g,\ \tilde{g} = \beta g$ gives new $\tilde{\omega} = h^{-1}\omega$.

The *multiplicity* of Σ at p is defined by

$$\mu(p) = \min\{k \geq 0 : (L_g^k c)(p) \neq 0\}.$$

Here and below we use the functions $e = \det(f, g)$ and $c = \det([g, f], g)$. By definition $\mu(p) = \infty$, if such k does not exist, and $\mu(p) = 0$ if $p \notin C$. At nonequilibrium points, the multiplicity μ can be equivalently defined using the 1-form ω:

$$\mu(p) = \min\{k \geq 0 : \omega(ad_g^{k+1} f)(p) \neq 0\}, \quad p \notin E.$$

If $p \in C = \{c = 0\}$ and $dc(p) \neq 0$, then, geometrically, multiplicity $\mu(p)$ is the order of tangency (plus one) of the critical curve to the trajectory of g at p.

Recall the functions $c^0 = e,\ c^1 = c$, and generally $c^k = L_g^{k-1} c$, for $k \geq 2$. At any $p \in X$ we define two sequences of ideals of function germs at p generated by $c^1, \ldots, c^k$ and, respectively, by $c^0, \ldots, c^k$:

$$I_p^k = I_p(c^1, \ldots, c^k), \quad J_p^k = I_p(c^0, c^1, \ldots, c^k).$$

At nonequilibrium points $p \notin E$, we similarly define

$$c_1 = \omega([g, f]), \quad c_2 = \omega([g, [g, f]]), \quad c_k = \omega(ad_g^k f)$$

and

$$I_{k,p} = I_p(c_1, \ldots, c_k).$$

PROPOSITION 3.5. *The ideals I_p^k, J_p^k, $I_{k,p}$ and the multiplicity $\mu(p)$ are invariant (more exactly, equivariant) under orbital feedback equivalence and $I_p^k = I_{k,p}$, if $p \notin E$.*

PROPOSITION 3.6. *If $p \notin E$ and the multiplicity $\mu(p) = \mu$ is odd, then the index*

$$\sigma_{\mu+1}(p) := \operatorname{sgn} c_{\mu+1}(p) = \operatorname{sgn} (e\, c^{\mu+1})(p)$$

is well defined and invariant under orbital feedback equivalence.

We first prove the following useful lemma (we define $I_p^0 = I_{0,p} = \{0\}$ and $c_0 = 1$).

LEMMA 3.7. *If $\tilde{f} = h\, f + \alpha\, g$ and $\tilde{g} = \beta\, g$, where $L_g h = 0$, then we have for $k \geq 1$*

$$\tilde{c}^k = h\beta^{k+1} c^k \mod I_p^{k-1}, \qquad \tilde{c}_k = \beta^k c_k \mod I_{k-1,p},$$

$$\tilde{c}^k = h\beta^{k+1} c^k + m_k h \beta^k L_g \beta\, c^{k-1}, \qquad \mod c^1, c^2, \ldots, c^{k-2},$$

$$\tilde{c}_k = \beta^k c_k + n_k \beta^{k-1} L_g \beta\, c_{k-1}, \qquad \mod c_1, c_2, \ldots, c_{k-2},$$

where $m_1 = 0$, $m_k = m_{k-1} + k$, and $n_1 = 0$, $n_k = n_{k-1} + k - 1$, $k \geq 2$.

Proof. Recall that (formula (3.4)) replacing f and g by the orbitally equivalent pair $\tilde{f} = h\, f + \alpha\, g$, $\tilde{g} = \beta\, g$, gives $\tilde{c} = h\beta^2 c$, $\tilde{e} = h\beta\, e$, and $\tilde{\omega} = h^{-1}\omega$.

We will use induction with respect to k. For $k=1$ we have $\tilde{c}^1 = \tilde{c} = h\beta^2 c = h\beta^2 c^1$. Let the first relation hold for $k-1$. Since $\tilde{c}^k = L_{\tilde{g}}\tilde{c}^{k-1}$ and $\tilde{c}^{k-1} = h\beta^k c^{k-1} + \sum_{i\le k-2}\varphi_i c^i$ (the induction assumption), we get from the Leibniz formula $L_g(\varphi\psi) = \psi L_g(\varphi) + \varphi L_g\psi$ that $L_{\tilde{g}}\tilde{c}^{k-1} = h\beta^k L_{\tilde{g}} c^{k-1} = h\beta^{k+1}c^k \mod I_p^{k-1}$. This implies the first formula. The third formula can be proved analogously.

The second and the fourth formulas follow from $ad_{\tilde{g}}\tilde{f} = h\beta\, ad_g f \mod g$ and the relation

$$ad_{\tilde{g}}^k \tilde{f} = h\beta^k ad_g^k f + n_k h\beta^{k-1} L_g\beta\, ad_g^{k-1} f \mod g,\ ad_g f, \dots,\ ad_g^{k-2} f,$$

which can be easily proved by induction, for $k \ge 2$, using the general property of Lie bracket $[\varphi f, \psi g] = \varphi\psi[f,g] + \varphi(L_f\psi)\, g - \psi(L_g\varphi)\, f$ and the property $L_g h = 0$. □

Proof of Proposition 3.5. It is enough to show invariance when the state transformation is $\phi = id$. Then the equality of ideals $I_p^k = \tilde{I}_p^k$, $J_p^k = \tilde{J}_p^k$, and $I_{k,p} = \tilde{I}_{k,p}$ follows, by induction with respect to k, from the lemma and nonvanishing of h and β.

Finally, the equality $I_p^k = I_{k,p}$ follows from equivariance of both ideals and the fact that for the system in the prenormal form Σ_{pre} we have

$$c^k = \frac{\partial^k f_1}{\partial y^k}, \qquad c_k = \frac{1}{f_1}\frac{\partial^k f_1}{\partial y^k}.$$

(f_1 does not vanish at $p \notin E$.) Invariance of $\mu(p)$ is a consequence of $I_p^k = \tilde{I}_p^k$ and the fact that $\mu(p)$ is the minimal k such that I_p^k contains functions nonvanishing at p. □

Proof of Proposition 3.6. Invariance of $\operatorname{sgn} c_k(p)$ and of $\operatorname{sgn}(e\, c^k)(p)$ in Proposition 3.6 follows from the lemma, the fact that the functions in I_p^{k-1} and $I_{k-1,p}$ vanish at p, if $\mu(p) = k$, and from $\beta(p) \ne 0$, $h(p) > 0$ (note that $\tilde{e} = h\beta e$). The equality $\operatorname{sgn} c_k(p) = \operatorname{sgn}(e\, c^k)(p)$ follows from the formulas for c_k and c^k in the proof of Proposition 3.5, for the system Σ_{pre}, and from $e = f_1$. □

We introduce new invariants. The *signature indices* of e and c is defined by

$$\sigma_e(p) = \operatorname{sgn}(\det \operatorname{hess}(e)(p)) \quad \text{if} \quad p \in E, \quad de(p) = 0,$$

$$\sigma_c(p) = \operatorname{sgn}(\det \operatorname{hess}(c)(p)) \quad \text{if} \quad p \in C, \quad dc(p) = 0.$$

The *stability index* is defined as

$$\gamma(p) = \operatorname{sgn}\lambda(p) \qquad \text{if} \quad p \in C \cap E, \quad de(p) \ne 0,$$

where λ is the eigenvalue of the uncontrollable mode of the linear part of Σ, at p.

Moreover, the following discrete invariants can be defined at nonequilibrium points of a given multiplicity (assuming, additionally, $dc(p) \ne 0$ for $\delta(p)$):

$$\tau(p) = \operatorname{sgn}(\omega(ad_g^2 f))(p) = \operatorname{sgn}(e\, L_g c)(p) = \sigma_2(p), \qquad \mu(p) = 1,$$

$$\delta(p) = -\operatorname{sgn}(\omega(ad_f^2 g)\, \omega(ad_g^3 f))(p) = \operatorname{sgn}(L_f c\, L_g^2 c)(p), \qquad \mu(p) = 2,$$

$$\kappa(p) = \operatorname{sgn}(\omega(ad_g^4 f))(p) = \operatorname{sgn}(e\, L_g^3 c)(p) = \sigma_4(p), \qquad \mu(p) = 3.$$

PROPOSITION 3.8. *All the indices τ, δ, κ, σ_e, σ_c, γ are well defined and are invariants of orbital feedback equivalence.*

Proof. Invariance of τ and κ follows from Proposition 3.6.

To show invariance of δ we take $\tilde f = h\,f+\alpha\,g$ and $\tilde g = \beta\,g$. Then $[\tilde f,\tilde g] = h\beta[f,g]+\varphi g$ and from our assumption $L_g h = 0$ we see that $[\tilde f,[\tilde f,\tilde g]] = [hf+\alpha g, h\beta[f,g]+\varphi g] = h^2\beta[f,[f,g]]+\alpha\beta h[g,[f,g]]-h\beta(L_{[f,g]}h)f \mod g, ad_f g$. Since $\mu(p)=2$ and so $p\in C$, the vectors g, $[f,g]$, and $[g,[f,g]]$ are linearly dependent at p and so $L_{[f,g]}h(p)=0$, by $L_g h = 0$ and $\omega([f,g])(p)=0$. This means that the second term in the expression for $[\tilde f,[\tilde f,\tilde g]](p)$ vanishes at p. Thus $\tilde\omega(ad^2_{\tilde f}\tilde g)(p) = (h\beta\omega(ad^2_f g))(p)$. From Lemma 3.7 we have $\tilde\omega(ad^3_{\tilde g}\tilde f)(p) = (\beta^3\omega(ad^3_g f))(p)$, which means that the factor $(h\beta^4)(p)$ appears in the expression for δ, after the transformation. This factor is positive, and thus the first formula for δ gives orbital feedback invariant. The invariance of the second formula for δ follows analogously. Equality of both formulas is easy to see for the system Σ_{pre}.

Invariance of σ_e and σ_c follows from $\det \operatorname{hess}(\varphi\psi))(p) = \varphi^2(p)\det\operatorname{hess}(\psi)(p)$, which holds if $\psi(p)=0$ and $d\psi(p)=0$.

Finally, invariance of γ is a consequence of independence of the eigenvalue λ of the feedback transformations and of the positivity of h. The proof is complete. □

For characterizing the last normal form in Theorem 3.3 and defining the invariant θ appearing there we need the condition $dc^2 \neq 0$ which, however, is not feedback invariant (by Lemma 3.7 we have $\tilde c^1 = h\beta^2 c^1$, $\tilde c^2 = h\beta^3 c^2 + 2h\beta^2 L_g\beta\, c^1$, and $d\tilde c^2|_p = h\beta^3 dc^2|_p + 2h\beta^2 L_g\beta\, dc^1|_p$, if $c^1(p)=c^2(p)=0$). Thus, we introduce modified versions of the functions c^2 and c_2:

$$c^2_{mod} = c^2 - \frac{c^5}{7c^4}c^1, \qquad c_{2\,mod} = c_2 - \frac{c_5}{10c_4}c_1.$$

They are defined at p if $c^4(p)\neq 0$ (equivalently, $c_4(p)\neq 0$) and $e(p)\neq 0$, for $c_{2\,mod}$.

LEMMA 3.9. *If $\mu(p)=3$ for Σ given by (f,g), and $\tilde f = hf+\alpha g$, $\tilde g = \beta g$, $L_g h = 0$, then*

$$d\tilde c^2_{mod}|_p = h\beta^3 dc^2_{mod}|_p, \qquad d\tilde c_{2\,mod}|_p = \beta^2 dc_{2\,mod}|_p.$$

Thus, with $\beta(p)\neq 0$, the conditions $dc^2_{mod}(p)\neq 0$ and $dc_{2\,mod}(p)\neq 0$ are invariant.

Proof. We prove only the first equality, using the third formula in Lemma 3.7. (The proof of the second equality uses the fourth formula in Lemma 3.7 and is analogous.) Denote $a = L_g\beta/\beta$. From $\mu(p)=3$ we have $c^1(p)=c^2(p)=c^3(p)=0$. Thus, by Lemma 3.7,

$$\begin{aligned}
\tilde c^2 &= h\beta^3(c^2+2ac^1),\\
\tilde c^4|_p &= h\beta^5 c^4|_p,\\
\tilde c^5|_p &= h\beta^6(c^5+14ac^4)|_p,\\
d\tilde c^1|_p &= h\beta^2 dc^1|_p,\\
d\tilde c^2|_p &= h\beta^3(dc^2+2adc^1)|_p.
\end{aligned}$$

Using also $\tilde c^1(p)=0$, we get

$$d\tilde c^2_{mod}|_p = d\tilde c^2|_p - \frac{\tilde c^5}{7\tilde c^4}d\tilde c^1|_p = h\beta^3(dc^2+2a\,dc^1)|_p - h\beta^3\frac{c^5+14ac^4}{7c^4}dc^1|_p,$$

thus $d\tilde c^2_{mod}|_p = h\beta^3(dc^2-(c^5/7c^4)dc^1)|_p = h\beta^3 dc^2_{mod}|_p$, by $c^1(p)=0$. □

If $e(p)\neq 0$, then the condition $dc^2_{mod}(p)\neq 0$ is equivalent to $(L_f c^2_{mod}, L_g c^2_{mod})|_p \neq (0,0)$. If, additionally, $\mu(p)=3$, then $c^3(p) = L_g c^2(p) = 0$ and $dc^2_{mod}(p)\neq 0$ is

equivalent to $L_f c^2_{mod}(p) \neq 0$. Similarly, $dc_{2\,mod}(p) \neq 0$ is equivalent to $L_f c_{2\,mod}(p) \neq 0$. The following definition is then correct if $\mu(p) = 3$, $e(p) \neq 0$, and $dc^2_{mod}(p) \neq 0$:

$$\theta(p) = \operatorname{sgn}(L_f c_{2\,mod})(p) = \operatorname{sgn}(e\, L_f c^2_{mod})(p).$$

PROPOSITION 3.10. *θ is invariant under orbital feedback equivalence.*

Proof. With $\tilde f = h\, f + \alpha\, g$, $\tilde g = \beta g$, we have $\tilde e = h\beta e$ and, by Lemma 3.9 and $dc^2_{mod}(g)|_p = 0$ (since $c^2(p) = c^3(p) = 0$), the expression $\tilde e\, L_{\tilde f} \tilde c^2_{mod}|_p = \tilde e\, d\tilde c^2_{mod}(\tilde f)|_p$ is equal to $h^3\beta^4 e\, dc^2_{mod}(f)|_p = h^3\beta^4 e\, L_f c^2_{mod}|_p$. Since h is positive, $\operatorname{sgn}(e\, L_f c^2_{mod})(p)$ does not change. Similarly we check that $\operatorname{sgn}(L_f c_{2\,mod})(p)$ does not change. Equality of both expressions for $\theta(p)$ can be checked directly for the prenormal form Σ_{pre}. □

To prove Proposition 3.4, note that the ideals I^k, J^k introduced earlier are well defined for the family $\Sigma:\ \dot\xi = f(\xi,\epsilon) + u g(\xi,\epsilon)$ and consist of functions of (ξ,ϵ). We denote by I^k_{p,ϵ_0} and J^k_{p,ϵ_0} the corresponding ideals of function germs at (p,ϵ_0). They have the same invariant properties as I^k_p and J^k_p.

Proof of Proposition 3.4. (i) To show invariance of conditions (G1), (G2), (G6) we reformulate them in invariant terms, using the ideals I^k_{p,ϵ_0} and J^k_{p,ϵ_0}, and use Proposition 3.5. Denote by C^∞_{p,ϵ_0} the ideal of all smooth function germs at (p,ϵ_0). From the definition of the ideals I^k_{p,ϵ_0} and J^k_{p,ϵ_0} we easily see that these conditions are equivalent to

$$\text{(G1)} \qquad J^3_{p,\epsilon_0} = C^\infty_{p,\epsilon_0}, \qquad I^4_{p,\epsilon_0} = C^\infty_{p,\epsilon_0},$$

$$\text{(G2)} \qquad J^2_{p,\epsilon_0} \neq C^\infty_{p,\epsilon_0} \Longrightarrow dc^0(p,\epsilon_0) \neq 0, \qquad I^3_{p,\epsilon_0} \neq C^\infty_{p,\epsilon_0} \Longrightarrow dc^1(p,\epsilon_0) \neq 0,$$

$$\text{(G3)} \qquad J^2_{p,\epsilon_0} \neq C^\infty_{p,\epsilon_0} \Longrightarrow dc^1(p,\epsilon_0) \neq 0, \qquad I^3_{p,\epsilon_0} \neq C^\infty_{p,\epsilon_0} \Longrightarrow dc^2_{mod}(p,\epsilon_0) \neq 0,$$

and, thus, they are invariant. (In the last implication we use the invariance of the condition $dc^2_{mod}(p,\epsilon_0) \neq 0$, implied by Lemma 3.9.) Invariance of each of the conditions (G3) and (G4) is a consequence of the fact that the ideals generated by c^i, $i = 0$, 1, are invariant. Invariance of (G5) follows from the invariance of the ideals J^2_{p,ϵ_0} and I^3_{p,ϵ_0}. Invariance of the ideal generated by c^i and c^{i+1} implies invariance of (G).

Finally, invariance of the conditions defined by each of the codes listed in Theorem 3.3 is easy to show using Proposition 3.5 and Lemma 3.7. In the case of the code $(1\overline{0_1 0_1 0}1)_{mod}$, the condition $dc^2_{mod}(p,\epsilon_0) \neq 0$ is invariant by Lemma 3.9.

(ii) Throughout the proof all functions are assumed to be evaluated at (p,ϵ_0). The symbol $*$ stands for nonzero numbers. For a function $h = h(x,y,\epsilon)$ we will denote its partial derivatives by h_x, h_y, h_ϵ. We assume Σ in the prenormal form Σ_{pre}, then $c^{i+1} = c^i_y$, $i \geq 0$.

(G1)–(G5) $\Longrightarrow$ (G). Assume that (G1)–(G5) hold but (G) does not hold, that is, for $i = 0$ or 1,

$$(c^i, c^{i+1}, j, J(c^i, c^{i+1}, j)) = (0,0,0,0), \quad \text{where} \quad j = j(c^i, c^{i+1}) = c^i_x c^{i+1}_y - c^i_y c^{i+1}_x.$$

We have $c^i = 0$, $c^i_y = c^{i+1} = 0$ and $j = c^i_x \cdot c^{i+1}_y = 0$. In the subcase $c^i_x = 0$, we have $dc^i = 0$, $j_y = 0$, $j_x = \operatorname{hess}(c^i)$, and then $J(c^i, c^{i+1}, j) = c^i_\epsilon \cdot c^{i+2} \cdot \operatorname{hess}(c^i) = 0$. Thus either (G2) or (G3) or (G4) is violated. In the subcase $c^{i+1}_y = c^{i+2} = 0$ we have $j_y = c^i_x c^{i+2}_y = c^i_x c^{i+3}$ and $J(c^i, c^{i+1}, j) = -(c^i_x \cdot c^{i+1}_\epsilon - c^{i+1}_x \cdot c^i_\epsilon) \cdot c^i_x \cdot c^{i+3} = 0$. Thus either (G1) or (G2) or (G5) is violated.

(G) $\Longrightarrow$ (G1)–(G5). Note that if $j(c^i, c^{i+1}) \neq 0$, then we have $dc^i \neq 0 \neq dc^{i+1}$ and $(c^{i+1}, c^{i+2}) = (c^i_y, c^{i+1}_y) \neq (0,0)$. Thus, if (G) holds for $i = 0$ or 1, with $(c^i, c^{i+1}, j(c^i, c^{i+1})) \neq (0,0,0)$, then it is easy to see that (G1)–(G5) hold for i. So we can assume that, for $i = 0$ or 1,

$$(c^i, c^{i+1}, j(c^i, c^{i+1}), J(c^i, c^{i+1}, j(c^i, c^{i+1}))) = (0,0,0,*).$$

We have $c^i_y = c^{i+1} = 0$ and thus $j(c^i, c^{i+1}) = c^i_x \cdot c^{i+1}_y = 0$. In the subcase $c^i_x = 0$, we have $J(c^i, c^{i+1}, j(c^i, c^{i+1})) = c^i_\epsilon \cdot c^{i+1}_y \cdot \operatorname{hess}(c^i) \neq 0$, in particular, $c^{i+1}_y = c^{i+2} \neq 0$. It follows that (G1)–(G5) hold for i. In the subcase $c^{i+1}_y = c^i_{yy} = 0$ we have $J(c^i, c^{i+1}, j(c^i, c^{i+1})) = -(c^i_x \cdot c^{i+1}_\epsilon - c^{i+1}_x \cdot c^i_\epsilon) \cdot c^i_x \cdot c^{i+2}_y \neq 0$, in particular, $c^{i+2}_y = c^{i+3} \neq 0$. It follows that (G1)–(G5) hold for i. □

4. Main proofs. In this section we will prove our main results, Theorems 1.1 and 3.3. We will also prove Propositions 2.3, 2.4, 2.5, and 2.6. When proving these results we will use the following corollary of the Mather theorem on universal unfoldings (cf., e.g., Theorem 14.8 in [BL] or Chapters IV and XI in [Ma]).

THEOREM 4.1 (Mather theorem). *Let $\varphi = \varphi(z, w)$ be a C^∞-function from a neighborhood of $(0,0) \in \mathbb{R} \times \mathbb{R}^p$ into $\mathbb{R}$. Assume that $k \geq 2$ and*

$$\frac{\partial^i \varphi}{\partial z^i}(0,0) = 0$$

for $1 \leq i \leq k-1$ and

$$\frac{\partial^k \varphi}{\partial z^k}(0,0) \neq 0.$$

Then there exists a local transformation $z = \psi(\bar{z}, w)$ invertible with respect to $\bar{z}$, such that

$$\varphi(\psi(\bar{z}, w), w) = \sigma \bar{z}^k + \sum_{i=0}^{k-2} a_i(w) \bar{z}^i,$$

where $\sigma = 1$ if k is odd, $\sigma = \pm 1$ if k is even, and $a_i(0) = 0$ for $1 \leq i \leq k-2$.

We consider points where $g(p, \epsilon) \neq 0$; thus we assume Σ in the prenormal form

$$\Sigma_{pre}: \quad \dot{x} = f_1(x, y, \epsilon), \quad \dot{y} = u.$$

The proof of the first part of Theorem 3.3 will be based on the following lemma.

LEMMA 4.2. *If Σ has finite multiplicity $\mu = k - 1 \geq 1$ at $(0,0) \in \mathbb{R}^2 \times \mathbb{R}$, then it is locally feedback equivalent to*

$$\Sigma_{\epsilon\, spe}: \quad \dot{x} = y^k + \sum_{i=0}^{k-2} a_i(x, \epsilon) y^i, \quad \dot{y} = u,$$

where $a_i(x, \epsilon)$ are smooth functions and $a_1(0,0) = \cdots = a_{k-2}(0,0) = 0$. Moreover, the orbital feedback transformation $\Gamma_\epsilon = (\phi, \psi, \eta, h)$ of the form

$$\tilde{x} = \phi(x, \epsilon), \quad \tilde{y} = \bar{\psi}(x, \epsilon) y, \quad \tilde{\epsilon} = \eta(\epsilon), \quad h = h(x, \epsilon)$$

satisfying $h\,(\partial\phi/\partial x)^{-1} \bar{\psi}^k = 1$, $h(x, \epsilon) > 0$, transforms $\tilde{\Sigma}_{\tilde{\epsilon}\, spe}$ into $\Sigma_{\epsilon\, spe}$ with

$$\text{(TF)} \qquad a_i = \bar{h}\, \bar{\psi}^i\, \tilde{a}_i(\phi, \eta), \qquad i = 0, \ldots, k-2,$$

where $\bar{h} = h\,(\partial\phi/\partial x)^{-1}$.

Further normalization of the special normal form $\Sigma_{\epsilon\, spe}$ in Lemma 4.2 to the normal forms in Theorem 3.3 will be done case by case, using the transformation formula (TF) in the lemma.

The proof of the second part of Theorem 3.3 follows from two lemmas.

LEMMA 4.3. *Any 1-parameter family Σ of systems which satisfies the conditions* (G1)–(G6) *of Theorem* 3.3 *at a point (ξ, ϵ) has, at this point, one of the nine codes of invariants listed in this theorem.*

LEMMA 4.4. *Conditions* (G1), (G2), ..., (G6) *are generic. More precisely, the set of 1-parameter families of pairs (f, g) of vector fields satisfying the conditions* (G1)–(G6) *in Theorem* 3.3, *at every point (ξ, ϵ) where $g(\xi, \epsilon) \neq 0$, is a countable intersection of open and dense subsets in the C^∞ Whitney topology of the space of pairs (f, g) defined on $X \times I$ (in particular, it is dense).*

Proof of Lemma 4.2. Finite multiplicity at $(0, 0)$ implies that $g(0, 0) \neq 0$. Thus, we can bring Σ to the prenormal form $\dot{x} = f_1(x, y, \epsilon)$, $\dot{y} = u$. The fact that multiplicity is equal to $\mu = k - 1$ is equivalent to

$$\frac{\partial^k f_1}{\partial y^k}(0, 0, 0) \neq 0, \qquad \frac{\partial^i f_1}{\partial y^i}(0, 0, 0) = 0, \qquad i = 1, \ldots, k-1$$

(cf. section 3.3). Using Theorem 4.1 for the function $\varphi = f_1$, with $z = y$, $w = (x, \epsilon)$, we find an invertible with respect to $\tilde{y}$ transformation $y = \psi(x, \tilde{y}, e)$ which brings the function $f_1 = f_1(x, y, \epsilon)$ into the polynomial form, with respect to $\tilde{y}$. This transformation applied to $\Sigma_{\epsilon\, pre}$ gives $\tilde{\Sigma}_\epsilon$ of the form $\dot{x} = \sigma\tilde{y}^k + \sum_{0 \leq i \leq k-2} a_i \tilde{y}^i$, $\dot{\tilde{y}} = (\partial\psi/\partial y)^{-1} u$. Since $\partial\psi/\partial y$ does not vanish (by invertibility of ψ), we can define new $\tilde{u} = (\partial\psi/\partial y)^{-1} u$. If $\sigma = -1$, then we additionally change $x \mapsto -x$ and obtain the system equations in the form $\Sigma_{\epsilon\, spe}$. □

Proof of Lemma 4.3. We fix a point of consideration $(\xi, \epsilon) = (p, \epsilon_0)$ (all equalities will be meant at this point, only). Condition (G1) says that there cannot be more then three consecutive zeros in a code (of the first level) at (p, ϵ_0) of a family Σ which satisfies condition (G1). Thus the codes of the first level are (11), (01), (101), (001), (1001), (0001), and (10001). We will call the number of zeros in the code of the first level "deficiency of the code." Thus, (G1) admits codes of deficiency 0, 1, 2 and 3. Below we determine the corresponding complete codes admitted by (G1)–(G6).

The possible codes of the second level are (11), (01), (101) (deficiency 0 or 1), $(0_1 01)$, $(0_0 01)$, $(1_0 1_1 01)$, $(1_0 0_0 01)$ (deficiency 2), and the codes of deficiency 3: $(0_1 0_1 01)$, $(0_1 0_0 01)$, $(0_0 0_1 01)$, $(0_0 0_0 01)$, $(1_0 1_1 01)$, $(1_0 1_0 0_0 01)$, $(1_0 0_0 1_1 01)$, and $(1_0 0_0 0_0 01)$. Note that the codes of deficiency 0 and 1 are complete.

For the codes $(0_0 01)$ and $(1_0 0_0 01)$ of deficiency 2 we apply condition (G4), which says that if c^i and dc^i vanish, then we have $\det \operatorname{hess}(c^i) \neq 0$ This gives the codes $(0_{01} 01)$, $(1_0 1_{01} 01)$ of level 3. Thus the codes of level 3, with deficiency 2, which are admissible are $(0_1 01)$, $(1_0 1_1 01)$, $(0_{01} 01)$, and $(1_0 1_{01} 01)$. Condition (G3) implies that if c^i and dc^i vanish, we have $\frac{\partial c^i}{\partial \epsilon} \neq 0$. Thus, among codes of deficiency 2, the complete admissible codes are $(0_1 01)$, $(1_0 1_1 01)$, $(\overline{0_{01}} 01)$ and $(1_0 \overline{1_{01}} 01)$ (by $dc^i = 0$ and $c^{i+2} \neq 0$, the conditions $\frac{\partial c^i}{\partial \epsilon} \neq 0$, and "$Dc^i$, Dc^{i+1} linearly independent", are equivalent).

Condition (G2) implies that if there are three zeros in the code of first level, i.e., $(c^i, c^{i+1}, c^{i+2}) = (0, 0, 0)$, then $dc^i \neq 0$. Thus all the codes of deficiency 3 satisfying condition (G2) have the subscript 1 at the first zero. Similarly, condition (G6) implies that if $(c^i, c^{i+1}, c^{i+2}) = (0, 0, 0)$, then $dc^{i+1} \neq 0$ ($dc^2_{mod} \neq 0$, if $i = 1$). Thus, among all possible codes of the second level, of deficiency 3, only the codes $(0_1 0_1 01)$ and $(1_0 1_0 1_1 01)_{mod}$ are admitted by conditions (G1), (G2), and (G6).

Finally, condition (G5) says that for codes of deficiency 3, the Jacobian is nonzero. Thus, the admissible complete codes of deficiency 3 are $(\overline{0_1 0_1 0}1)$ and $(1\overline{0_1 0_1 0}1)_{mod}$. In this way we conclude that the codes characterized by conditions (G1)–(G6) are (11), (01), (101), $(0_1 01)$, $(10_1 01)$, $(\overline{0_{01}0}1)$, $(1\overline{0_{01}0}1)$, $(\overline{0_1 0_1 0}1)$, and $(1\overline{0_1 0_1 0}1)_{mod}$. □

Proof of Lemma 4.4. We shall use the Thom transversality theorem (cf. [Hi, Chapter 3, Theorem 2.8]). Consider the space $J^k\Sigma$ of k-jets of pairs of families of vector fields (f, g) defined on $X \times I$. Let $S_1, \ldots, S_N$ be a family of submanifolds in $J^k\Sigma$ and $\operatorname{codim} S_i > \dim(X \times I) = 3$. Thom's theorem says, in particular, that the set of pairs (f, g) of families of vector fields, whose k-jet extensions do not intersect the submanifolds $S_1, \ldots, S_N$, is a countable intersection of open and dense subsets in the space of C^∞ pairs (f, g), with the C^∞ Whitney topology. Thus, it is enough to show that the set of k-jets of (f, g) which do not satisfy one of the conditions (G1)–(G6) consists of a finite number of submanifolds of codimension not less then 4.

Proposition 3.4 says that each of the conditions (G1)–(G6) is orbitally feedback invariant. Thus, it is enough to consider the 5-jets of (f, g) for the system in prenormal form Σ_{pre} for which $g = \partial/\partial y$, $f = f_1\partial/\partial x$, and

$$c^i = \frac{\partial^i f_1}{\partial y^i}.$$

Condition (G1) is not satisfied at some point if either $f_1 = 0$, $\partial f_1/\partial y = 0$, $\partial^2 f_1/\partial y^2 = 0$, and $\partial^3 f_1/\partial y^3 = 0$, or $\partial f_1/\partial y = 0$, $\partial^2 f_1/\partial y^2 = 0$, $\partial^3 f_1/\partial y^3 = 0$, and $\partial^4 f_1/\partial^4 y = 0$. Clearly, both conditions define submanifolds in $J^4\Sigma$ of codimension 4.

Condition (G2) is violated at some point if $f_1 = 0$, $\partial f_1/\partial y = 0$, $\partial^2 f_1/\partial y^2 = 0$, and $\partial f_1/\partial x = 0$, or $\partial f_1/\partial y = 0$, $\partial^2 f_1/\partial y^2 = 0$, $\partial^3 f_1/\partial y^3 = 0$, and $\partial^2 f_1/\partial x \partial y = 0$. Again, both conditions define submanifolds in $J^4\Sigma$ of codimension 4. Similarly negation of condition (G3) defines two submanifolds in $J^4\Sigma$ of codimension 4.

Negation of condition (G6) gives two submanifolds of codimension 4 in $J^5\Sigma$: $f_1 = 0$, $\partial f_1/\partial y = 0$, $\partial^2 f_1/\partial y^2 = 0$, $\partial^2 f_1/\partial y \partial x = 0$, and $\partial f_1/\partial y = 0$, $\partial^2 f_1/\partial y^2 = 0$, $\partial^3 f_1/\partial y^3 = 0$, $\partial^3 f_1/\partial y^2 \partial x - (f_1^{(5)}/7f_1^{(4)})\partial^2 f_1/\partial y \partial x = 0$, $f_1^{(4)} \neq 0$, where $f_1^{(i)}$ is the ith partial derivative of f_1 with respect to y. ($f_1^{(4)} \neq 0$ is implied by (G1).)

Negation of condition (G4) means $c^i = 0$, $dc^i = 0$, $\det \operatorname{hess}(c^i) = 0$ (for each $i = 0, 1$), where $c^0 = f_1$ and $c^1 = \partial f_1/\partial y$. The corresponding set in $J^4\Sigma$ is, for each $i = 0, 1$, the union of two submanifolds $c^i = 0$, $dc^i = 0$, $\operatorname{hess}(c^i) = 0$ and $c^i = 0$, $dc^i = 0$, $\operatorname{rank} \operatorname{hess}(c^i) = 1$, in the space of 2-jets of functions c^i (and so in $J^4\Sigma$). The first submanifold is given by 6 independent equations, so it is of codimension 6, and the second is given by 4 independent equations (it has codimension 4). Here we use the fact that the set of symmetric 2×2 matrices of rank 1 is a codimension 1 submanifold in the space of all symmetric 2×2 matrices.

Similarly, the negation of condition (G5) defines the set $c^i = 0$, $c^{i+1} = 0$, $c^{i+2} = 0$, $J(c^i, c^{i+1}, c^{i+2}) = 0$, for each $i = 0, 1$. Since, for the prenormal form, $c^{i+1} = \partial c^i/\partial y$ and $c^{i+2} = \partial^2 c^i/\partial y^2$, it follows that the Jacobian matrix has two zero elements and vanishing of the Jacobian is equivalent to one of the conditions $\partial^3 c^i/\partial y^3 = 0$ (which defines a submanifold S_1 of codimension 4) or $\det A = 0$, where A is the 2×2 matrix with entries $(\partial c^i/\partial x, \partial c^i/\partial \epsilon)$ in the first row and $(\partial^2 c^i/\partial y \partial x, \partial^2 c^i/\partial y \partial \epsilon)$ in the second row. Again, the set of 2×2 matrices with determinant zero consists of matrices of rank 1, which form a submanifold of codimension 1 in the set of all 2×2 matrices, and of the zero matrix (submanifold of codimension 4). Altogether, this means that negative of condition (G5) gives three submanifolds (for each $i = (0, 1)$) in $J^4\Sigma$: S_2 of codimensions $3 + 1 = 4$ and S_3 of codimension $3 + 4 = 7$, and the third one S_1, of codimension 4, which was defined earlier. This completes the proof. □

Proof of Theorem 3.3. From Lemma 4.4 it follows that the class of systems Σ characterized by the conditions (G1)–(G6), satisfied at every (p, ϵ), where $g(p, \epsilon) \neq 0$, is generic. Lemma 4.3 implies that, at every such (p, ϵ), the system satisfying (G1)–(G6) fulfils exactly one of the conditions defined by the codes listed in the first part of the theorem. It is easy to check that for any of the nine normal forms listed in the theorem the corresponding conditions defined by the codes are satisfied. Moreover, they are invariant under orbital feedback equivalence (Proposition 3.4). Thus the conditions defined by the codes are necessary for equivalence of Σ to a given canonical form.

In the remaining part of the proof we will show their sufficiency. We will call Σ generic if it satisfies (G1)–(G6). Consider a generic Σ around (p, ϵ_0). Without loosing generality we can choose $(p, \epsilon_0) = (0, 0)$. Since in each code (of the first level) in Theorem 3.3 we have at most three 0, followed by 1, in each case the multiplicity μ is finite and equal to 0, 1, 2, or 3. We shall proceed in order of increasing multiplicity.

$\boldsymbol{\mu = 0}$. This appears in the codes (11) and (01). We first bring the system to the prenormal form $\dot{x} = f_1(x, y, \epsilon)$, $\dot{y} = u$. For both codes (11) and (01), the 1 at the second place means that $c(0, 0) = \partial f_1/\partial y(0, 0, 0) \neq 0$. Thus we can choose the new coordinate $\tilde{y} = f_1(x, y, \epsilon) - f_1(0, 0, 0)$ and applying a feedback transformation $u \to \tilde{u} = \psi(x, y, u)$ we bring the system to the form $\dot{x} = y + a$, $\dot{y} = u$, where $a = f_1(0, 0, 0) = e(0, 0, 0)$. In the case of the code (01) we have $a = e(0, 0, 0) = 0$, that is, we get the canonical form (E). In the case (11) we have $a \neq 0$. Changing x, y for x/a and y/a gives the system equation $\dot{x} = y + 1$, i.e., the canonical form (O).

When the multiplicity is $\mu = 1$, 2, or 3 we can apply Lemma 4.2 and bring Σ to the special normal form $\Sigma_{\epsilon\, spe}$. Its further simplification will be done case by case using the transformation formula (TF) in Lemma 4.2. We will write only the first equation of the system, the second being always $\dot{y} = u$, after a suitable feedback modification.

$\boldsymbol{\mu = 1}$. We have $\mu = 1$ in the codes (101), $(0_1 01)$, $(\overline{0_{01}0}1)$. The special normal form in Lemma 4.2 is

$$\dot{x} = y^2 + a_0(x, \epsilon).$$

In the case of the code (101) we have e nonvanishing and so $a_0(0, 0) \neq 0$. We can choose the new coordinate $\tilde{x} = \phi(x, \epsilon) = \int_0^x 1/a_0(s, \epsilon) ds$, then a_0 becomes 1, i.e., the first system equation becomes $\dot{\tilde{x}} = y^2/a_0(x, \epsilon) + 1$. Choosing $\tau = \operatorname{sgn} a_0(0, 0)$ and $\tilde{y} = y\sqrt{\tau/a_0(x, \epsilon)}$ we bring the system to the canonical form (C): $\dot{x} = \tau y^2 + 1$.

In the case $(0_1 01)$ we have $a_0(0, 0) = 0$ and $\partial a_0/\partial x(0, 0) \neq 0$. Thus, we can choose as the new coordinate $\tilde{x} = \phi(x, \epsilon) = a_0(x, \epsilon)$ and change the time scale by $h = \pm(\partial\phi/\partial x)^{-1}$, with h positive. Then the first system equation becomes $\dot{x} = \pm y^2 \pm x$. Changing possibly x for $-x$ we get the canonical form (EC): $\dot{x} = y^2 \pm x$.

The code $(\overline{0_{01}0}1)$ implies $a_0(0, 0) = 0$, $da_0(0, 0) = 0$, and $\partial^2 a_0/\partial x^2(0, 0) \neq 0$. Applying, e.g., Theorem 4.1, we see that we can change the coordinate x for $\tilde{x} = \phi(x, \epsilon)$ so that, taking also $h = \pm(\partial\phi/\partial x)^{-1}$, we get the new coefficient a_0 in the form $a_0 = \pm x^2 + k(\epsilon)$, where k is a smooth function. Changing, if necessary, x for $-x$ and $k(\epsilon)$ for $-k(\epsilon)$ we bring the system to the form $\dot{x} = \pm y^2 + x^2 + k(\epsilon)$. Linear independence of De and Dc, implied by the code $(\overline{0_{01}0}1)$, gives $(\partial k/\partial\epsilon)(0) \neq 0$ and thus replacing $k(\epsilon)$ with ϵ brings the system to the canonical form (E_{bif}): $\dot{x} = \pm y^2 + x^2 + \epsilon$.

$\boldsymbol{\mu = 2}$. This concerns the codes $(10_1 01)$, $(1\overline{0_{01}0}1)$, and $(\overline{0_1 0_1 0}1)$. We will treat the first two cases together. They have the code of the first level (1001) so, in the

special normal form $\Sigma_{\epsilon\, spe}$ in Lemma 4.2, we have the first system equation

$$\dot{x} = y^3 + a_1(x,\epsilon)y + a_0(x,\epsilon)$$

and $a := e(0,0,0) = a_0(0,0) \neq 0$. We change the time scale by $h = a_0^{-1}\operatorname{sgn} a$ to get the new $a_0 = \pm 1$ and then change y for $\tilde{y} = y\,h^{1/3}$. We obtain the first system equation $\dot{x} = y^3 + a_1(x,\epsilon)y \pm 1$ (with a_1 changed). It remains to normalize a_1.

In the case (10_101) we have $dc(0,0,0) \neq 0$, so $\partial a_1/\partial x(0,0) \neq 0$. This allows us to introduce the new coordinate $\tilde{x} = a_1(x,\epsilon)$. Changing simultaneously the time scale with the positive valued function $h = \pm(\partial a_1/\partial x)^{-1}$ we get the system $\dot{x} = \pm y^3 \pm xy \pm 1$. Changing possibly x for $-x$ we get the last term $a_0 = 1$. Finally, replacing if necessary y by $-y$, we obtain $a_1 = x$ and the desired canonical form (C$\mathcal{G}$): $\dot{x} = \pm y^3 + xy + 1$.

In the case of the code $(1\overline{0_{01}0}1)$ we have $dc(0,0,0) = 0$ and $\det \operatorname{hess}(c) \neq 0$. Since in this case $c = 3y^2 + a_1(x,\epsilon)$, we have $\partial a_1/\partial x(0,0) = 0$ and $\partial^2 a_1/\partial x^2(0,0) \neq 0$. Applying, e.g., Theorem 4.1 we see that we can change the coordinate x for $\tilde{x} = \phi(x,\epsilon)$ so that the function a_1 becomes $a_1 = x^2 + k(\epsilon)$. Changing simultaneously the time scale with the function $h = \pm(\partial\phi/\partial x)^{-1}$ we bring the system to the form $\dot{x} = \pm y^3 \pm (x^2 \pm k(\epsilon))y \pm 1$. Linear independence of Dc^1 and Dc^2, implied by the code $(1\overline{0_{01}0}1)$, gives $\partial k/\partial\epsilon(0) \neq 0$. Thus, we can replace $\pm k(\epsilon)$ with new ϵ, which gives $\dot{x} = \pm y^3 \pm (x^2 - \epsilon)y \pm 1$. Finally, changing possibly x for $-x$, and y for $-y$ brings the system to the canonical form (C_{bif}): $\dot{x} = \pm y^3 + (x^2 - \epsilon)y + 1$.

The last two cases, with the codes $(\overline{0_10_10}1)$ (where $\mu = 2$) and $(1\overline{0_10_10}1)_{mod}$ (where $\mu = 3$), are more complicated and will be treated together.

The difficulty in reaching the canonical form is reduced to the following lemma (whose proof is given below after completing the proof of Theorem 3.3) on equivalence of ratios of smooth 1-parameter families of functions. For $\varphi = \varphi(x,y)$ denote $\varphi'_x = \partial\varphi/\partial x$ and $\nabla\varphi = (\partial\varphi/\partial x, \partial\varphi/\partial y)$.

LEMMA 4.5. *Let $A(x,y)$ and $B(x,y)$ be families of smooth functions defined for (x,y) in a neighborhood of $(0,0) \in \mathbb{R}^2$ and assume that* (i) *$A'_x(0,0) \neq 0$ and $B'_x(0,0) \neq 0$;* (ii) *$\nabla A(0,0)$ and $\nabla B(0,0)$ are linearly independent. Then there is a smooth, local, invertible transformation $(\tilde{x},\tilde{y}) = \chi(x,y)$ of the form $\tilde{x} = \phi(x,y)$, $\tilde{y} = \eta(y)$, with $\chi(0,0) = (0,0)$, such that*

$$\frac{A^3}{B^2} \circ \chi(x,y) = \frac{(x-y)^3}{x^2}.$$

We continue the proof of Theorem 3.3. Consider a family Σ with one of the codes $(\overline{0_10_10}1)$ and $(1\overline{0_10_10}1)_{mod}$. We transform the system to the special normal form $\Sigma_{\epsilon\, spe}$ in Lemma 4.2, where $k = 3$ and $k = 4$, respectively. Then, bringing the system to the corresponding canonical form will be done in two steps.

Step 1. We transform x and ϵ so that the new coefficients of the systems

$$\dot{x} = y^3 + Ay + B, \qquad \dot{y} = u, \qquad \text{and}$$
$$\dot{x} = y^4 + Ay^2 + By + a_0, \quad \dot{y} = u,$$

satisfy the condition

$$A^3B^{-2} = (x-\epsilon)^3x^{-2},$$

where A, B, and a_0 are functions of (x,ϵ). This is possible by Lemma 4.5 (where ϵ is changed for y). Indeed, it follows from the conditions defined by the codes $(\overline{0_10_10}1)$

and $(1\overline{0_1 0_1 0}1)_{mod}$ that the functions $A(x,\epsilon)$ and $B(x,\epsilon)$ satisfy the assumptions of the lemma, with y replaced by ϵ. Applying the transformation χ of the lemma and an appropriate change of the time scale we obtain again the systems in the special normal forms above, with the new coefficients $A = (x-\epsilon)\bar{A}$ and $B = x\bar{B}$ satisfying the condition $A^3B^{-2} = (x-\epsilon)^3x^{-2}$, i.e., $\bar{A}^3\bar{B}^{-2} = 1$. ($\bar{A}$ and $\bar{B}$ are nonvanishing functions of (x,ϵ).)

Step 2. In the first case we define the orbital transformation

$$x = \tilde{x}, \quad y = \bar{B}^{1/3}\tilde{y}, \quad \epsilon = \tilde{\epsilon}, \quad h = \pm\bar{B}^{-1}$$

and check easily that under the condition $\bar{A}^3\bar{B}^{-2} = 1$, implying $\bar{A} = \bar{B}^{2/3}$, the system in the special form is transformed to $\dot{x} = \pm(y^3 + (x-\epsilon)y + x)$. Changing, possibly, y for $-y$ we finally come to the desired canonical form (E$\mathcal{G}_{bif}$): $\dot{x} = y^3 + (x-\epsilon)y \pm x$.

In the second case the condition $\bar{A}^3\bar{B}^{-2} = 1$, which gives $\bar{A} = \bar{B}^{2/3}$, implies that the orbital transformation $x = \tilde{x},\ y = \bar{B}^{1/3}\tilde{y},\ \epsilon = \tilde{\epsilon},\ h = \bar{B}^{-4/3}$ brings the system in the special form to the last canonical form (C$\mathcal{G}_{bif}$): $\dot{x} = y^4 + (x-\epsilon)y^2 + xy + a(x,\epsilon)$. The proof of Theorem 3.3 is complete. □

Proof of Lemma 4.5. We will reduce the problem to a special case and apply the homotopy method. By (i) we can choose new coordinate $\tilde{x} = B(x,y)$ and then assume that $B = x$. The equation $A(x,y) = 0$ has a locally unique solution $x = k(y)$, by the implicit function theorem (use condition (i)). Moreover, we can write $A(x,y) = (x-k(y))\bar{A}(x,y)$, where $\bar{A}(0,0) \neq 0$. From condition (ii) and $B = x$ it follows that $k'(0) \neq 0$, so we can take $\tilde{y} = k(y)$ as new coordinate. Then

$$\frac{A^3}{B^2} = gb, \qquad \text{where} \qquad g = \frac{(x-y)^3}{x^2} = (x-y)\left(1-\frac{y}{x}\right)^2$$

and $b(x,y) = \bar{A}^3(x,y)$, $b(0,0) > 0$. (We can assume that $\bar{A}(0,0) > 0$ since, in the contrary, we can change $x = B \mapsto -x = -B$, and $y \mapsto -y$.)

We include the functions g and $g\,b$ into the family

$$F(x,y,t) = g(x,y)\,a(x,y,t),$$

where $t \in [0,1]$ and

$$a(x,y,t) = 1 + t(b(x,y) - 1).$$

Since $b(0,0) > 0$, we have $a(0,0,t) > 0$ for $t \in [0,1]$. We will find a vector field of the form

$$Y = \frac{\partial}{\partial t} + u(y,t)\frac{\partial}{\partial y} + v(x,y,t)\frac{\partial}{\partial x}$$

such that $u(0,t) = 0$, $v(0,0,t) = 0$, and $L_Y F = 0$. ($L_Y F$ denotes the directional derivative of F along Y.) This will mean that F is constant on the trajectories of Y. In particular, if $\psi_t(x,y,0)$ is the trajectory of Y after time t, starting from $(x,y,0)$, then

$$F(\psi_t(x,y,0)) = F(\psi_0(x,y,0)) = F(x,y,0).$$

(The trajectory from $(0,0,0)$ is well defined for $t \in [0,1]$, since $u(0,t) = v(0,0,t) = 0$; thus it is also defined for such t from $(x,y,0)$, for (x,y) small.) It follows from the definition of the function F that $F(x,y,0) = g(x,y)$ and $F(x,y,1) = g(x,y)\,b(x,y)$.

Thus, taking in the above equality $t = 1$ we get the assertion of the proposition with χ defined by

$$\psi_1(x, y, 0) = (\chi(x, y), 1).$$

It remains to solve the equation $L_Y F = 0$. It takes the form

$$gb + u(g'_y a + g a'_y) + v(g'_x a + g a'_x) = 0 \tag{LF}$$

with $u = u(y, t)$ and $v = v(x, y, t)$ unknown. Given that $g = (x - y)(1 - y/x)^2$ we compute

$$g'_y = -3\left(1 - \frac{y}{x}\right)^2, \quad g'_x = \left(1 - \frac{y}{x}\right)^2 \left(1 + 2\frac{y}{x}\right),$$

which allows us to divide (LF) over $(1 - y/x)^2$. We get the equivalent equation

$$(x - y)b + u(3a + (x - y)a'_y) + v(a + 2ay/x + (x - y)a'_x) = 0.$$

Since $a(0, 0, t) > 0$, the coefficient $C = -3a + (x - y)a'_y$ at u is invertible in a neighborhood of $(0, 0, t)$, $t \in [0, 1]$. Dividing the equation over C and replacing b by $\tilde{b} = b/C$ and v by $\tilde{v} = v/(xC)$ we get the simpler equation

$$(x - y)\tilde{b} + u + \tilde{v}(xa + 2ya + x(x - y)a'_x) = 0.$$

We note that the coefficient $E = xa + 2ya + x(x - y)a'_x$ at $\tilde{v}$ has the property $\partial E/\partial x(0, 0, t) = a(0, 0, t) > 0$. We can thus use it as a local coordinate and introduce the local coordinates $(\tilde{x}, y, t)$ in a neighborhood of the segment $(0, 0, t)$, $t \in [0, 1]$, where $\tilde{x} = E(x, y, t)$. Note that if $y = 0$, then $\tilde{x} = E = 0$ if and only if $x = 0$. The function $(x - y)\tilde{b}(x, y, t) =: c(\tilde{x}, y, t)$ can be decomposed into

$$c(\tilde{x}, y, t) = c_0(y, t) + \tilde{x}c_1(\tilde{x}, y, t) \quad \text{with} \quad c_0(0, t) = 0.$$

(The latter equality follows from the former by putting $\tilde{x} = y = 0$.) Then our equation becomes

$$c_0(y, t) + \tilde{x}c_1(\tilde{x}, y, t) + u(y, t) + \tilde{v}(\tilde{x}, y, t)\tilde{x} = 0$$

and it has a solution $u(y, t) = -c_0(y, t)$, $\tilde{v}(\tilde{x}, y, t) = -c_1(\tilde{x}, y, t)$. This solution satisfies

$$u(0, t) = -c_0(0, t) = 0, \quad v(0, 0, t) = 0.$$

(We recall that $v = \tilde{v}xC$ and, if $y = 0$, then $\tilde{x} = 0$ if and only if $x = 0$.) In this way we have constructed the vector field Y satisfying $L_Y F = 0$, and the above conditions guarantee that the trajectory starting from $(x, y, t) = (0, 0, 0)$ satisfies $\psi_t(0, 0, 0) = (0, 0, t)$, in particular $\psi_1(0, 0, 0) = (\chi(0, 0), 1) = (0, 0, 1)$, as required. □

Proof of Theorem 1.1 *and Propositions* 2.3, 2.4, 2.5, and 2.6. Theorem 1.1 follows from Propositions 2.3, 2.4, 2.5, and 2.6, thus we only prove the propositions. They are rather straightforward consequences of Theorem 3.3.

Consider a generic family Σ. By Lemma 4.4 we can assume that genericity means that Σ satisfies the conditions (G1)–(G6) of Theorem 3.3. Lemma 4.3 implies that Σ has one of the codes in Theorem 3.3, at a given point (p, ϵ_0). By Theorem 3.3, the

family Σ is locally orbitally feedback equivalent to one of the normal forms: (O), (E), (C), (EC), (C$\mathcal{G}$), (E_{bif}), (C_{bif}), ($E\mathcal{G}_{bif}$), or ($C\mathcal{G}_{bif}$).

Observe that the normal forms (O), (E), (C), (EC), and (C$\mathcal{G}$) do not depend on the parameter ϵ so neither do the equivariants $\mathcal{I}_\epsilon = (E_\epsilon, C_\epsilon, \mathcal{G}_\epsilon)$. Thus, if the family Σ is equivalent at (p, ϵ_0) to one of those forms, it does not bifurcate at (p, ϵ_0).

It is straightforward to check that for the remaining four normal forms (E_{bif}), (C_{bif}), ($E\mathcal{G}_{bif}$), and ($C\mathcal{G}_{bif}$) the invariant condition expressed in the corresponding code in Theorem 3.3 is equivalent to the condition given in the respective Proposition 2.3, 2.4, 2.5, or 2.6. Thus, it is enough to analyze the invariants $\mathcal{I}_\epsilon = (E_\epsilon, C_\epsilon, \mathcal{G}_\epsilon)$, for each of the four normal forms, and show two facts: (i) the invariants are of the form stated in the propositions; (ii) they undergo the corresponding bifurcation described in the proposition. Statement (ii) was proved in the description of bifurcations given after the statements of the propositions. Thus we only check statement (i).

Clearly, for all of the normal forms (E_{bif}), (C_{bif}), ($E\mathcal{G}_{bif}$), and ($C\mathcal{G}_{bif}$) we have $\mathcal{G}_\epsilon = \{x = \text{const}\}$. Moreover, for the first one

$$(E_{bif}) \qquad \dot{x} = \sigma_e y^2 + x^2 - \epsilon, \qquad \dot{y} = v,$$

with $\sigma_e = \pm 1$, the equilibrium and critical sets are given by

$$E_\epsilon = \{x^2 + \sigma_e y^2 = \epsilon\}, \qquad C_\epsilon = \{y = 0\},$$

respectively. This together with the description of the corresponding bifurcations of E_ϵ in section 2.1 proves Proposition 2.3.

In the normal form

$$(C_{bif}) \qquad \dot{x} = \sigma_c y^3 + (x^2 - \epsilon)y + 1, \qquad \dot{y} = v$$

(where $\sigma_c = \pm 1$), the equilibrium set of E_ϵ is empty in a neighborhood of $(0, 0)$ and the critical set is given (after replacing y by $3^{-1/2}y$) by $C_\epsilon = \{x^2 + \sigma_c y^2 = \epsilon\}$. Thus this family has one of C-bifurcations (section 2.2). This proves Proposition 2.4.

In the case of normal form

$$(E\mathcal{G}_{bif}) \qquad \dot{x} = y^3 + (x - \epsilon)y + \gamma x, \qquad \dot{y} = v,$$

at $(0, 0)$ (where $\gamma = \pm 1$) the equilibrium set E_ϵ and the critical set C_ϵ are given by

$$E_\epsilon = \{y^3 + (x - \epsilon)y + \gamma x = 0\} \qquad \text{and} \qquad C_\epsilon = \{3y^2 + x - \epsilon = 0\}.$$

Replacing y for γy we get the same equation for C_ϵ and $E_\epsilon = \{y^3 + (x - \epsilon)y + x = 0\}$. Thus, the family ($E\mathcal{G}_{bif}$) has an $E\mathcal{G}$-bifurcation at $(0, 0)$, as shown in section 2.3.

In the normal form

$$(C\mathcal{G}_{bif}) \qquad \dot{x} = y^4 + (\theta x - \epsilon)y^2 + xy + a(x, \epsilon), \qquad \dot{y} = v,$$

the equilibrium set E_ϵ is empty and the critical set is $C_\epsilon = \{4y^3 + 2(\theta x - \epsilon)y + x = 0\}$. Replacing x, y, and ϵ by $\theta x/2$, $\theta y/2$, and $\epsilon/2$ we get $C_\epsilon = \{y^3 + (x - \epsilon)y + x = 0\}$. This shows Proposition 2.6. The proof is complete. □

Acknowledgment. The support of the Mittag-Leffler Institute is gratefully acknowledged. The authors thank M. Zhitomirskii for a discussion which led to the birth of this paper and J. K. Kowalski for preparing the figures.

REFERENCES

[AF1] E. H. Abed and J.-H. Fu, *Local feedback stabilization and bifurcation control, Part* I, *Hopf bifurcations*, Systems Control Lett., 7 (1986), pp. 11–17.

[AF2] E. H. Abed and J.-H. Fu, *Local feedback stabilization and bifurcation control, Part* II, *Stationary bifurcations*, Systems Control Lett., 8 (1987), pp. 463–473.

[AAIS] V. I. Arnold, V. S. Afraimovitch, Yu. Ilyashenko, and L. P. Shilnikov, *Theory of Bifurcations*, Encyclopaedia Math. 5, Springer, New York, 1994.

[Ba1] M. Baitmann, *Controllability regions on the plane*, Differ. Equ., 14 (1978), pp. 407–417.

[Ba2] M. Baitmann, *Switching lines in the plane*, Differ. Equ., 14 (1978), pp. 1093–1101.

[BsP] U. Boscain and B. Piccoli, *Optimal Synthesis for Control Systems on* 2*-D Manifolds*, Math. Appl. 43, Springer, New York, 2004.

[BP] A. Bressan and B. Piccoli, *A generic classification of time optimal planar stabilizing feedbacks*, SIAM J. Control Optim., 36 (1998), pp. 12–32.

[BL] Th. Bröcker and L. Lander, *Differentiable Germs and Catastrophes*, London Math. Soc. Lecture Note Ser. 17, Cambridge University Press, Cambridge, UK, 1975.

[Da1] A. Davydov, *Singularities of limit direction fields of two-dimensional contol systems*, Math. USSR-Sb., 64 (1989), pp. 471–493.

[Da2] A. Davydov, *Qualitative Theory of Control Systems*, Transl. Math. Monogr. 141, AMS, Providence, RI, 1994.

[Go] V. V. Goryunov, *Geometry of bifurcation diagrams of simple projections onto the line*, Funct. Anal. Appl., 15 (1981), pp. 77–82.

[Hi] M. Hirsch, *Differential Topology*, Springer, New York, 1976.

[J1] B. Jakubczyk, *Equivalence and invariants of nonlinear control systems*, in Nonlinear Controllability and Optimal Control, H. J. Sussmann, ed., Marcel Dekker, New York, 1990, pp. 177–218.

[JR1] B. Jakubczyk and W. Respondek, *Feedback equivalence of planar systems and stabilizability*, in Robust Control of Linear Systems and Nonlinear Control, M. A. Kaashoek, J. H. van Schuppen, and A.C.M. Ran, eds., Birkhäuser, Boston, 1990, pp. 447–456.

[JR2] B. Jakubczyk and W. Respondek, *Feedback classification of analytic control systems in the plane*, in Analysis of Controlled Dynamical Systems, B. Bonnard et al., eds., Birkhäuser, Boston, 1991, pp. 262–273.

[IJ] G. Iooss and D. Joseph, *Elementary Stability and Bifurcation Theory*, Springer-Verlag, New York, Heidelberg, Berlin, 1980.

[Ka1] W. Kang, *Bifurcation and normal form of nonlinear control systems:- Part* I *and Part* II, SIAM J. Control Optim., 36 (1998), pp. 193–212, pp. 213–232.

[Ka2] W. Kang, *Normal Form, Invariants, and Bifurcations of Nonlinear Control Systems in the Particle Deflection Plane*, in Dynamics, Bifurcations and Control, F. Colonius and L. Grüne, eds., Lecture Notes in Control and Inf. Sci. 273, Springer, Berlin, Heidelberg, 2002, pp. 67–87.

[KKC] A. J. Krener, W. Kang, and D. E. Chang, *Control bifurcations*, IEEE Trans. Automat. Control, 49 (2004), pp. 1231–1246.

[Ku] Y. A. Kuznetsov, *Elements of Applied Bifurcation Theory*, Springer-Verlag, New York, 1998.

[Ma] J. Martinet, *Singularities of Smooth Functions and Maps*, London Math. Soc. Lecture Note Ser. 58, Cambridge University Press, Cambridge, UK, 1982.

[R] W. Respondek, *Feedback classification of nonlinear control systems in* $\mathbb{R}^2$ *and* $\mathbb{R}^3$, in Geometry of Feedback and Optimal Control, B. Jakubczyk and W. Respondek, eds., Marcel Dekker, New York, 1998, pp. 347–382.

[Ru] M. Rupniewski, *Local bifurcations of control affine systems in the plane* (submitted to Journal of Dynamical and Control Systems).

[Su1] H. J. Sussmann, *The structure of time optimal trajectories for single-input systems in the plane: The* C^∞ *nonsingular case*, SIAM J. Control Optim., 25 (1987), pp. 433–465.

[Su2] H. J. Sussmann, *The structure of time–optimal trajectories for single-input systems in the plane: The general real analytic case*, SIAM J. Control Optim., 25 (1987), pp. 868–904.

[Za] V. M. Zakalyukin, *Flag contact singularities*, in Real and Complex Singularities, J. W. Bruce and F. Tari, eds., Research Notes in Math. 412, Chapman and Hall/CRC, Boca Raton, FL, 2000, pp. 134–146.

[Zh] M. Zhitomirskii, *Finitely determined* 1*-forms* ω, $\omega|_0 \neq 0$ *are reduced to the models of Darboux and Martinet*, Funct. Anal. Appl., 19 (1985), pp. 71–72.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 6, pp. 2063–2078

OPTION PRICING WITH MARKOV-MODULATED DYNAMICS*

A. JOBERT† AND L. C. G. ROGERS†

Abstract. Markov-modulated models for equity prices have recently been extensively studied in the literature. In this paper, we apply some old results on the Wiener–Hopf factorization of Markov processes to a range of option-pricing problems for such models. The first example is the perpetual American put, where the exact (numerical) solution is obtained without discretizing any PDE. We then show how the methodology of Rogers and Stapleton [*Finance Stoch.*, 2 (1997), pp. 3–17] can be used to tackle finite-horizon problems and illustrate the methodology by pricing European, American, single barrier, and double barrier options under Markov-modulated dynamics.

Key words. Markov-modulated, Markov-chain, option, Black–Scholes model, Wiener–Hopf factorization, risk-sensitive control, regime switching

AMS subject classifications. 15A23, 15A24, 60J27, 60J70, 93E20

DOI. 10.1137/050623279

1. Introduction.

Though outstandingly successful as a leading-order model for an asset price, the familiar log-Brownian paradigm fails in various ways, such as the fact that implied volatility is not constant. Among the many attempted variations and extensions, one of the most natural is to allow the dynamics of the underlying process to be a log-Brownian motion whose volatility and rate of return are stochastic in some way. Allowing the volatility to be stochastic is the central theme of the extensive literature on *stochastic volatility modeling*, of which [24, 15, 13, 2, 14] make up a small sample. Allowing the rate of return to be stochastic is of relevance to portfolio optimization, but not to asset pricing,[1] and the literature on *risk-sensitive optimal control* develops this theme in various ways; see, for example [5, 6, 3, 20].

Perhaps the simplest way to introduce additional randomness into the standard log-Brownian model is to let the volatility and rate of return be functions of a finite-state Markov chain; we can imagine that such a model might describe regime-switching behavior of some kind, perhaps related to the business cycle, or other economic indicators. The terms *regime-switching* and *Markov-modulated dynamics* are used to describe such models, and there are already interesting contributions here, such as applications to option pricing [12, 11, 10, 9, 8, 4, 26], portfolio optimization [27, 25], and optimal trading strategies [28]. In applications, it is likely that the number of states of the Markov chain will be small (otherwise estimation becomes a problem), and it is then natural to think of such a model as "nearly" a log-Brownian motion, with occasional parameter shifts. Some explicit solutions can be found for a two-state Markov chain, but as the problems get harder we are soon led into PDE-related numerical methods (smoothed approximation of boundary conditions [26], two-point boundary value problems [28], discretization of associated dynamic programming equations [12]). The coupled PDEs which arise in these models will rarely be soluble in closed form, though finite-difference methods are still quite competitive.

*Received by the editors January 25, 2005; accepted for publication (in revised form) June 22, 2005; published electronically January 6, 2006.
http://www.siam.org/journals/sicon/44-6/62327.html
†Statistical Laboratory, University of Cambridge, Wilberforce Road, Cambridge CB3 0WB, UK (A.Jobert@statslab.cam.ac.uk, L.C.G.Rogers@statslab.cam.ac.uk).
[1]Of course, for asset pricing we change measure so that the rate of return becomes the riskless rate.

This paper approaches such models from a different direction: by viewing the Markov-modulated asset as nearly a Markov chain. This approach relies on some old work on Wiener–Hopf factorization of Markov processes (and in particular, Markov chains) dating back to the paper of Barlow, Rogers, and Williams [1] from 1980; the focus is on level-crossings of the asset price process. We find a quite different toolkit applies in this approach, namely, linear algebra; this leads to numerical schemes that are very efficient and quite able to handle moderate-sized problems, which we will illustrate by pricing a perpetual American put[2] on such an asset. Let us emphasize immediately a key difference between the present approach and the traditional PDE approach: here we shall be obtaining (numerically) *exact* solutions to the problem, *not* just approximations. There is no need to solve any dynamic programming equation or discretize any PDE.

Next, we move to a pricing framework with finite time horizon and Markov regimes. The approach here extends the methodology developed by Rogers and Stapleton [22] for the standard log-Brownian model and is illustrated by pricing the European call, the American put, and finally double barrier options.

2. General setup and noisy Wiener–Hopf factorization. The stock price is modeled as

$$dS_t = S_t[\sigma(\xi_t)dW_t + r(\xi_t)dt], \tag{1}$$

where W_t is a standard Brownian motion, r denotes as usual the risk-free interest rate, σ denotes the Markov-modulated volatility of the stock, and ξ is an irreducible Markov chain with values in the finite set I, $|I| = d$. Notice that the riskless rate may vary with the underlying Markov chain. The log price $X_t = \log(S_t)$ then satisfies

$$dX_t = \sigma(\xi_t)dW_t + \left[r(\xi_t) - \frac{1}{2}\sigma(\xi_t)^2\right]dt, \tag{2}$$

which can be rewritten as, say,

$$dX_t = \sigma(\xi_t)dW_t + v(\xi_t)dt. \tag{3}$$

The idea of the Wiener–Hopf factorization approach is to study the crossings back and forth over levels of X. To help in this, define for $t \geq 0$[3]

$$\tau_t^{\pm} \equiv \inf\{u : \pm X_u > t\}. \tag{4}$$

We aim to characterize the distribution of the times $\tau_t^{\pm}$ and the law of the chain at these times, and to do this we will seek martingales M_t^f of the following form:

$$M_t^f = \exp\left(-\int_0^t r(\xi_u)du\right) f(\xi_t, X_t) \tag{5}$$

for some function f. Itô's formula gives, up to a local martingale part,

$$dM_t^f \doteq \exp\left(-\int_0^t r(\xi_u)du\right)\left[(Q-R)f + \frac{1}{2}\Sigma f_{XX} + Vf_X\right]dt, \tag{6}$$

[2]This problem was solved for a two-state chain by Guo and Zhang [12].

[3]With the usual convention that $\inf(\emptyset) = \infty$.

where R is the diagonal matrix whose ith diagonal element is equal to $r(i)$, Σ is the diagonal matrix whose ith diagonal element is equal to $\sigma(i)^2$, and $V = R - \frac{1}{2}\Sigma$. We therefore require

$$(Q - R)f + \frac{1}{2}\Sigma f_{XX} + V f_X = 0. \tag{7}$$

Seeking separable f of the form $f(\xi_t, X_t) = g(\xi_t)\exp(-\lambda X_t)$ gives rise to the following equation to be solved in λ and g:

$$(Q - R)g + \frac{1}{2}\lambda^2\Sigma g - \lambda V g = 0, \tag{8}$$

Now this is just the "quadratic eigenvalue" problem considered by Kennedy and Williams [17], which can be reduced to a standard eigenvalue problem as follows. Premultiplying the above equation by $2\Sigma^{-1}$ gives

$$2\Sigma^{-1}(Q - R)g + \lambda^2 g - 2\lambda\Sigma^{-1}Vg = 0. \tag{9}$$

This can be reformulated as a system of equations

$$\left\{ \begin{array}{ll} \lambda g & = h, \\ \lambda h & = 2\Sigma^{-1}Vh - 2\Sigma^{-1}(Q - R)g, \end{array} \right. \tag{10}$$

which can be rewritten as the following (standard) eigenvalue problem:

$$A \left(\begin{array}{c} g \\ h \end{array} \right) \equiv \left(\begin{array}{cc} 0 & I \\ -2\Sigma^{-1}(Q - R) & 2\Sigma^{-1}R - I \end{array} \right) \left(\begin{array}{c} g \\ h \end{array} \right) = \lambda \left(\begin{array}{c} g \\ h \end{array} \right). \tag{11}$$

If (g, λ) solve (11), then

$$M_t^f = \exp\left(-\int_0^t r(\xi_u)du - \lambda X_t \right) g(\xi_t) \tag{12}$$

is a martingale. The argument given in [1] serves to show that there are exactly d eigenvalues of A in the left open half plane, and d in the right open half plane, a fact that will be needed later.

3. Markov-modulated perpetual American put. Our goal in this section is to compute the value

$$v(j, x) \equiv \sup_{\tau} E\left[\exp\left(-\int_0^{\tau} r(\xi_s)ds \right) (K - e^{X_\tau})^+ \mid \xi_0 = j, X_0 = x \right] \tag{13}$$

of the perpetual American put with Markov-modulated dynamics. The special case of the problem where there is no Markov modulation (that is, $|I| = 1$) is well known:[4] the optimal rule is to wait until the price of the asset falls below some critical boundary value L^*, and then immediately exercise. Standard first passage time calculations for Brownian motion lead to the following closed-form expression for the perpetual American put:

$$v(x) = \left\{ \begin{array}{ll} K - \exp(x) & \text{if } x \leq \log(L^*), \\ (K - L^*)(L^*)^\gamma \exp(-\gamma x) & \text{if } x > \log(L^*), \end{array} \right. \tag{14}$$

where $\gamma = 2r/\sigma^2$, $L^* = \gamma K/(\gamma + 1)$.

[4]See the original solution of McKean [19] and Karatzas [16] for a discussion in a more general setting.

When $|I| > 1$, the optimal rule is to exercise when the price of the asset falls below some critical level, which depends on the current state of the modulating Markov chain ξ. This intuitively obvious form of the solution follows immediately from the next simple result.

PROPOSITION 1. *If $\varphi(x) \equiv (K - e^x)^+$, then for each $j \in I$ the function*

$$x \mapsto v(j,x) - \varphi(x)$$

is nondecreasing in $(0, \log(K))$.

Proof. Pick $0 < x < x + \delta < \log(K)$ and let τ^* denote the optimal stopping time to be used if $X_0 = x$. We consider instead what would happen if we were to use the stopping rule τ^* but with initial log-price $x + \delta$. Using the elementary inequality $(a-b)^+ \geq a^+ - b^+$, we get[5]

$$\begin{aligned}
v(j, x+\delta) &\equiv \sup_\tau E[e^{-R(\tau)}\varphi(X_\tau)|\, \xi_0 = j, X_0 = x+\delta] \\
&\geq E[e^{-R(\tau^*)}\varphi(X_{\tau^*})|\, \xi_0 = j, X_0 = x+\delta] \\
&= E[e^{-R(\tau^*)}\varphi(X_{\tau^*} + \delta)|\, \xi_0 = j, X_0 = x] \\
&= E[e^{-R(\tau^*)}(K - e^{X(\tau^*)+\delta})^+|\, \xi_0 = j, X_0 = x] \\
&= E[e^{-R(\tau^*)}(K - e^{X(\tau^*)} - (e^\delta - 1)e^{X(\tau^*)})^+|\, \xi_0 = j, X_0 = x] \\
&\geq E[e^{-R(\tau^*)}\{\, (K - e^{X(\tau^*)})^+ - (e^\delta - 1)e^{X(\tau^*)}\,\}|\, \xi_0 = j, X_0 = x] \\
&= v(j,x) - (e^\delta - 1)E[e^{-R(\tau^*)}e^{X(\tau^*)}|\, \xi_0 = j, X_0 = x] \\
&\geq v(j,x) - (e^\delta - 1)e^x \\
&= v(j,x) - \varphi(x) + \varphi(x+\delta),
\end{aligned}$$

using the fact that $e^{-R(t)+X(t)}$ is a martingale, and therefore a supermartingale. ∎

Immediately from Proposition 1, the optimal stopping time is of the form

$$\tau = \inf\{t : X_t < b(\xi_t)\}, \tag{15}$$

where the constants $(b_i)_{i\in I}$ must be found.

This problem was solved by Guo and Zhang [12] in the simple case of two states, where a closed-form expression can be derived for the price. Note that -1 is always an eigenvalue of A, which is a key observation that makes the two-state problem tractable. However, the current methodology will work for any number of states. The time-0 value of the stopping rule (15) defined by the levels $(b_i)_{i\in I}$ is

$$v(j,x) = \mathbf{E}\left[\exp\left(-\int_0^\tau r(\xi_t)dt\right)(K - \exp(b(\xi_\tau)))^+ | S_0 = \exp(x); \xi_0 = j\right] \tag{16}$$

There are thus two problems:

(1) Given some thresholds b_i, derive the value function;

(2) find the optimal b_i.

PROBLEM 1. *Let us suppose given $(b_i)_{i\in I}$, where without loss of generality*[6] *$b_1 > b_2 > \cdots > b_d$; our goal is to compute the value function associated with this set of threshold levels.*

[5] We use the abbreviation $R(t) \equiv \int_0^t r(\xi_s)ds$.

[6] This assumption amounts to an inessential relabeling of the states and is merely for convenient discussion. When it comes in practice to identifying the thresholds, no assumption is made on the ordering, and all possible orderings are considered. We show in Proposition 2 that there is a unique solution for the thresholds, whose ordering is determined by the parameters of the problem.

Let us start with x in the interval $[b_1, \infty)$. Here, the value function is larger than the payoff function whatever the initial state. Recall that we are looking for a martingale M_t^f of the form of (5) for some function f which satisfies (7) and which will be represented as a weighted sum

$$f(\xi, x) = \Sigma_{i=1}^d w_i g_i(\xi) \exp(-\lambda_i x), \tag{17}$$

where for each i, (λ_i, g_i) satisfies (8), with $\lambda_i > 0$. We restrict our attention to the d eigenvalues with positive real part because this means that the martingale

$$M_t \equiv \exp\left(-\int_0^t r(\xi_u)du\right) \sum_i w_i g_i(\xi_t) \exp(-\lambda_i X_t)$$

is bounded on $[0, \tau_1]$, where $\tau_1 \equiv \inf\{t : X_t < b_1\}$. Therefore we may apply the optional sampling theorem to obtain

$$\begin{aligned} &\mathbf{E}\left[\exp\left(-\int_0^{\tau_1} r(\xi_u)du\right) \sum_i w_i g_i(\xi_{\tau_1}) \exp(-\lambda_i X_{\tau_1}) \;\middle|\; X_0 = x, \xi_0 = j\right] \\ &= \sum_i w_i g_i(j) \exp(-\lambda_i x). \end{aligned} \tag{18}$$

This is the expression for the value function over the interval $[b_1, \infty)$. In particular, for $j = 1$, this completes the determination of the time-0 price when the underlying chain is initially in state 1, provided we impose

$$(K - \exp(b_1))^+ = \sum_i w_i g_i(1) \exp(-\lambda_i b_1). \tag{19}$$

This gives us a first equation satisfied by the d unknown weights w, and

$$v(1, x) = \begin{cases} K - \exp(x) & \text{if } x \le b_1, \\ \sum_i w_i g_i(1) \exp(-\lambda_i x) & \text{if } x \ge b_1, \end{cases} \tag{20}$$

where w still needs to be determined. Continuity at b_1 in (19) restricts w to a $(d-1)$-dimensional subspace; to go further, we must look at the next interval $I_2 = [b_2, b_1]$ and match values and slopes of V across b_1.

When $x \in I_2$, ξ can jump to state 1, causing exercise to happen. So we now need to modify slightly the Wiener–Hopf argument and the equation for f. Let $\tilde{\Sigma}$, $\tilde{R}$, $\tilde{V}$ be the diagonal matrices defined in the following way: for every $i = 2, \ldots, d$, $\tilde{\Sigma}(i,i) = \sigma(i)^2$, $\tilde{R}(i,i) = r(i)$ and $\tilde{V} = \tilde{R} - \frac{1}{2}\tilde{\Sigma}$. Let $\tilde{Q}$ be the submatrix derived from Q by removing its first row and first column.

We still seek a martingale M_t^f of the form of (5) for some function f, which now satisfies the following modified equation:

$$(\tilde{Q} - \tilde{R})f + \frac{1}{2}\tilde{\Sigma} f_{XX} + \tilde{V} f_X + \tilde{K} = 0, \tag{21}$$

where $\tilde{K}$ is defined so that it accounts for jumps to the payoff function in state 1: $\tilde{K} = \tilde{q}(K - \exp(x))$, where $\tilde{q}$ denotes a $(d-1)$-dimensional vector, such that $\tilde{q}(i) = q_{i1}$ for every $i = 2, \ldots, d$. The value function over the interval $[b_2, b_1]$ is characterized by (21).

A particular solution to (21) is easily obtained and is of the form $B + C\exp(x)$. The homogeneous equation

$$(\tilde{Q} - \tilde{R})f + \frac{1}{2}\tilde{\Sigma}f_{XX} + \tilde{V}f_X = 0 \tag{22}$$

is structurally similar to (7) and is solved similarly. Let $\tilde{\lambda_i}$ and $\tilde{g_i}$ denote the $2(d-1)$ corresponding eigenvalues and eigenvectors for this new problem. For any scalars $\tilde{w_i}$,

$$\exp\left(-\int_0^t r(\xi_u)du\right)\left(\Sigma_{i=1}^{2(d-1)}\tilde{w_i}\exp(-\tilde{\lambda_i}X_t)\tilde{g_i}(\xi_t) + B + C\exp(X_t)\right) \tag{23}$$

is a martingale, at least if we stop at first exit from I_2, and is bounded up to that time. Provided that we ensure that $v(j,\cdot)$ joins in a C^1 fashion across b_1, for $j = 2,\dots,d$, we therefore have for any $x \in I_2$, and any $j = 2\dots d$,

(24)

$$\mathbf{E}^{x,j}\left[\exp\left(-\int_0^{\tau_2} r(\xi_u)du\right)\left(\sum_i \tilde{w_i}\tilde{g_i}(\xi_{\tau_2})\exp(-\tilde{\lambda_i}X_{\tau_2}) + B_j + C_j\exp(X_{\tau_2})\right)\right]$$
$$= \sum_i \tilde{w_i}\tilde{g_i}(j)\exp(-\tilde{\lambda_i}x) + B_j + C_j\exp(x),$$

where $\tau_2 = \min\{t : X_t \le b_2\}$, and $\mathbf{E}^{x,j}$ denotes the usual expectation conditional upon $X_0 = x$, $\xi_0 = j$. This is the expression for the value function in the interval I_2; in particular, for $j = 2$, this completes the determination of the time-0 price when the underlying chain is initially in state 2, provided we impose continuity at b_2.

Notice that at the end of the first step, we were left with $d-1$ degrees of freedom. Once we have solved the problem over the interval $[b_2, b_1]$, matching the values and the slopes of $v(j,\cdot)$, $j = 2,\dots,d$ across b_1, we have $2\times(d-1)$ new linear equations, for the $2(d-1)$ new unknowns $\tilde{w_i}$. Continuity across b_2 of $v(2,\cdot)$ provides us with another equation so that at the end of our second step, we are left with $d-2$ degrees of freedom.

From the above, it is now clear that we can proceed recursively, from b_1 to b_d, by solving d successive problems of this type and considering the standard eigenvalue problem associated with our modified setup and our updated generator for the underlying Markov chain. At the end of the dth problem over the interval $[b_d, b_{d-1}]$, we no longer have any degrees of freedom, once we have imposed the continuity of $v(d,x)$ across b_d. Finally, over $[0, b_d]$, we have: $v(1,x) = \cdots = v(d,x) = K - \exp(x)$. This deals with the first problem, namely, given thresholds to compute the value function.[7]

PROBLEM 2. *The method just presented shows how for any given sequence of threshold values we may compute the value. For optimality, we need to make $v(j,\cdot)$ be C^1 at b_j for $j = 1,\dots,d$. This gives us d nonlinear equations to be solved in d unknowns, which can be solved by standard numerical techniques; we used sequential quadratic programming. The latter optimization routine is converging efficiently toward a set of b values which make v to be C^1. It remains to check that*

$$(Q - R)v + \frac{1}{2}\Sigma v_{XX} + V v_X \le 0$$

[7]The above procedure leaves us in fact with a linear system to solve in order to determine the unknown weights on every subinterval: d weights on $[b_1,\infty)$, $2\times(d-1)$ weights on $[b_2,b_1]$, $2\times(d-2)$ weights on $[b_3,b_2]$, ... and finally 2 weights on $[b_d,b_{d-1}]$. This gives rise to a linear system with d^2 unknowns and d^2 equations.

everywhere; from this, it follows that the solution v found is in fact optimal. The following results deals with this point.

PROPOSITION 2. *Suppose that thresholds $(b_j) < \log K$ have been found such that the (unique) bounded solution f to the coupled system of ODEs*

$$\frac{1}{2}\sigma_i^2 f_{XX}(i,X) + V_i f_X(i,X) - r_i f(i,X) + \sum_j q_{ij} f(j,X) = 0 \qquad (X > b_i), \tag{25}$$

$$f(i,X) = \varphi(X) \qquad (X \le b_i) \tag{26}$$

is C^1 in X at each point (j, b_j). Then the (b_j) are uniquely determined, and f is the value of the problem.

Proof. The proof proceeds in a number of steps. Given the thresholds (b_j), we set $\tau^* = \inf\{t : X_t \le b(\xi_t)\}$, and we observe that

$$f(\xi_{t\wedge\tau^*}, X_{t\wedge\tau^*}) \exp\{-R(t\wedge\tau^*)\} \text{ is a bounded martingale,}$$

and so in particular

$$f(j,x) = E\big[\, \varphi(X_{\tau^*}) e^{-R(\tau^*)} \,\big|\, \xi_0 = j, X_0 = x \,\big].$$

Since $\varphi \ge 0$, it follows that $f > 0$.

(i) We claim that $f(j,x) > \varphi(x)$ whenever $x > b_j$. To see why, let $\tau_0 \equiv \inf\{t : f(\xi_t, X_t) \le \varphi(X_t)\} \le \tau^*$, and observe that

$$\begin{aligned} f(j,x) &= E\big[\, \varphi(X_{\tau^*}) e^{-R(\tau^*)} \,\big|\, \xi_0 = j, X_0 = x \,\big] \\ &= E\big[\, (K - e^{X(\tau^*)}) e^{-R(\tau^*)} \,\big|\, \xi_0 = j, X_0 = x \,\big] \\ &= E\big[\, \varphi(X_{\tau_0}) e^{-R(\tau_0)} \,\big|\, \xi_0 = j, X_0 = x \,\big] \\ &= E\big[\, (K - e^{X(\tau_0)}) e^{-R(\tau_0)} \,\big|\, \xi_0 = j, X_0 = x \,\big]. \end{aligned}$$

The fact that $\exp(X_t - R_t)$ is a martingale[8] tells us that

$$E\big[\, K e^{-R(\tau^*)} \,\big|\, \xi_0 = j, X_0 = x \,\big] = E\big[\, K e^{-R(\tau_0)} \,\big|\, \xi_0 = j, X_0 = x \,\big],$$

whence immediately $\tau^* = \tau_0$, and the claim is proved.

(ii) We claim next that $f(j,\cdot) - \varphi(\cdot)$ is nondecreasing in $(0, \log(K))$. The proof of this is in effect a reprise of the proof of Proposition 1. As there, we take two starting points $x, x+\delta \in (0, \log(K))$, and let τ denote the stopping time that would be used if we started from x. Using the fact that $f \ge \varphi$, we have

$$\begin{aligned} f(j,x+\delta) &= E[e^{-R(\tau)} f(\xi_\tau, X_\tau) |\, \xi_0 = j, X_0 = x+\delta] \\ &\ge E[e^{-R(\tau)} \varphi(X_\tau) |\, \xi_0 = j, X_0 = x+\delta] \\ &= E[e^{-R(\tau)} \varphi(X_\tau + \delta) |\, \xi_0 = j, X_0 = x] \\ &= E[e^{-R(\tau)} (K - e^{X(\tau)+\delta})^+ |\, \xi_0 = j, X_0 = x] \\ &= E[e^{-R(\tau)} (K - e^{X(\tau)} - (e^\delta - 1) e^{X(\tau)})^+ |\, \xi_0 = j, X_0 = x] \\ &\ge E[e^{-R(\tau)} \big\{\, (K - e^{X(\tau)})^+ - (e^\delta - 1) e^{X(\tau)} \,\big\} |\, \xi_0 = j, X_0 = x] \\ &= f(j,x) - (e^\delta - 1) E[e^{-R(\tau)} e^{X(\tau)} |\, \xi_0 = j, X_0 = x] \\ &\ge f(j,x) - (e^\delta - 1) e^x \\ &= f(j,x) - \varphi(x) + \varphi(x+\delta). \end{aligned}$$

[8] It is in fact the discounted stock price.

(iii) The final step is to prove that

$$\Phi(i,x) \equiv \frac{1}{2}\sigma_i^2 f_{XX}(i,X) + V_i f_X(i,X) - r_i f(i,X) + \sum_j q_{ij} f(j,X) \le 0 \tag{27}$$

in $X \le b_i$. From (26), we have that in fact

$$\Phi(i,x) = -r_i K + \sum_{j \ne i} (f(j,X) - \varphi(X)),$$

which is seen to be nondecreasing in $[0, b_i]$, in view of point (ii) proved above. It is therefore sufficient to prove that $\Phi(i, b_i-) \le 0$ to establish (27). By the C^1 property of the solution f, we note that all of the terms in $\Phi(i,\cdot)$ are continuous across b_i except perhaps the second derivative term; thus any discontinuity in Φ is entirely accounted for by the jump in this. But now consider the function $f(i,\cdot) - \varphi(\cdot)$. Its second derivative at b_i- is zero, and yet its second derivative at b_i+ must be nonnegative, since the function is nonnegative to the right of b_i, and the function and its first derivative both vanish there. We deduce that the change in the second derivative of $f(i,\cdot)$ at b_i is nonnegative, and the conclusion (27) follows.

(iv) The standard verification argument for optimal control now shows that stopping at τ^* is optimal and that f is the value function of the problem. □

As a check, take the case $d = 1$, which is the standard perpetual American put problem mentioned earlier; we have $R = r$, $\Sigma = \sigma^2$, $V = r - \frac{1}{2}\sigma^2$, $Q = 0$ and set $\gamma = 2r/\sigma^2$. Now the matrix A defined at (11) is simply

$$A = \begin{pmatrix} 0 & 1 \\ \gamma & \gamma - 1 \end{pmatrix}$$

with eigenvalues γ and -1, so the solution is of the form

$$v(x) = \begin{cases} K - \exp(x) & \text{if } x \le b, \\ w \exp(-\gamma x) & \text{if } x > b, \end{cases} \tag{28}$$

where the critical level b and the weight w are to be determined. The C^1 condition for v at b becomes

$$\begin{cases} K - \exp(b) &= w \exp(-\gamma b), \\ \exp(b) &= \gamma w \exp(-\gamma b), \end{cases}$$

from which we easily deduce the form given in (14).

3.1. Numerical results.

3.1.1. Two states. First, we check that we recover the results of Guo and Zhang [12] for the simple case of two states. The strike K is taken to be equal to 5,

$$R = \begin{pmatrix} 0.03 & 0 \\ 0 & 0.03 \end{pmatrix},$$

$$Q = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} 0.25 & 0 \\ 0 & 0.81 \end{pmatrix}.$$

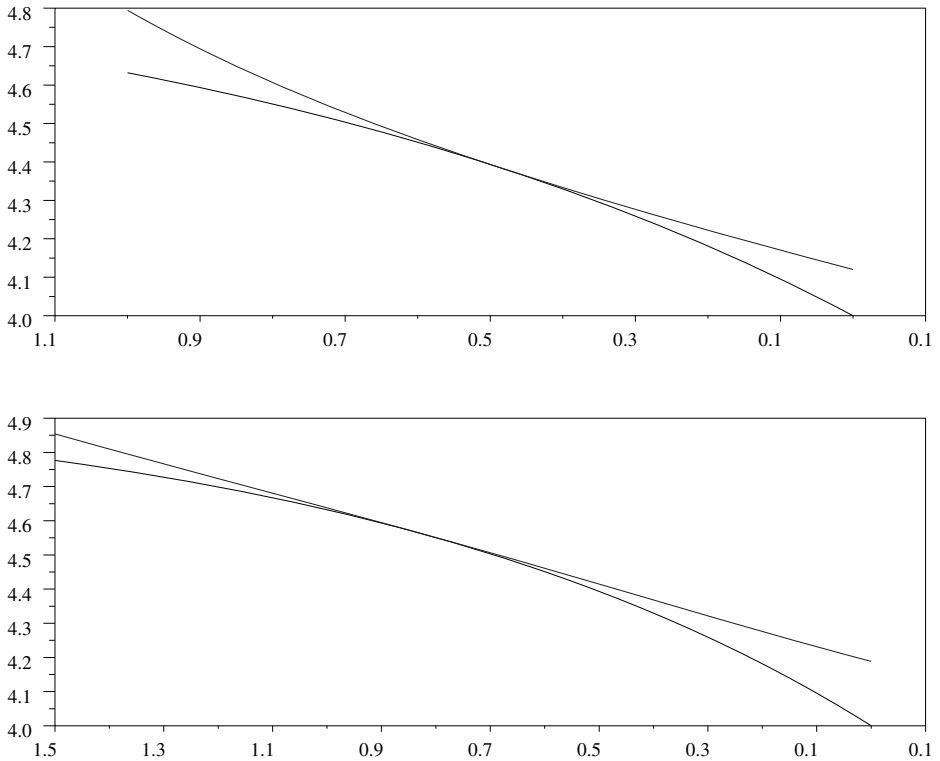


FIG. 1. *Perpetual American put with two states (value function against log price).*

This gives optimal thresholds: $\exp(b) = (0.612, 0.441)$, which is close to Guo and Zhang's solution $(0.614, 0.441)$. Figure 1 plots the price of the Markov-modulated perpetual American put in every state of the chain and enables us to visualize the corresponding smooth pasting conditions. Above the optimal thresholds, where smooth pasting occurs, the upper curve is the value function for the perpetual American put and the lower curve is the reward function. Below the optimal thresholds, the value function is equal to the reward function represented by the lower curve. We keep drawing the upper curve below the optimal thresholds for the sole purpose of assessing the quality of smooth pasting. All the plots below are drawn using a logarithmic scale for the stock price.

Decreasing the volatility in the second state,

$$\Sigma = \begin{pmatrix} 0.25 & 0 \\ 0 & 0.49 \end{pmatrix},$$

leads to higher optimal thresholds $\exp(b) = (0.801, 0.646)$. On the other hand, increasing the jump intensity from state 2 to state 1, where $\sigma_2^2 = 0.81$ and $\sigma_1^2 = 0.25$, decreases the average volatility and we expect therefore our optimal thresholds to be bigger, which turns out to be the case: $\exp(b) = (0.633, 0.455)$.

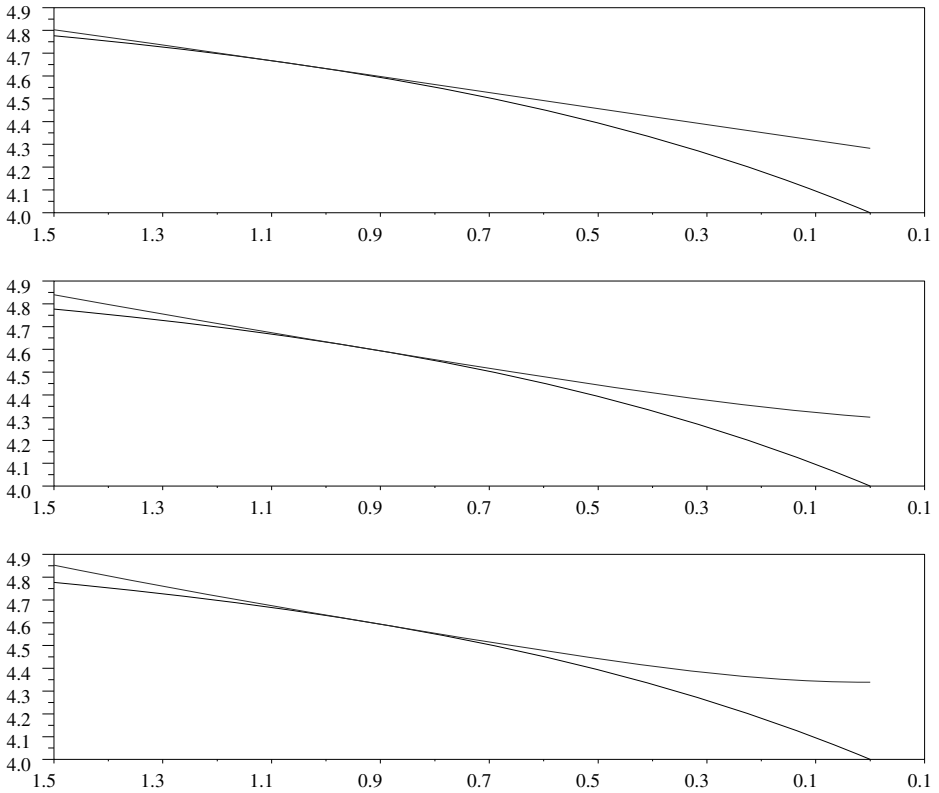


FIG. 2. *Perpetual American put with three states (value function against log price).*

3.1.2. Three states. Consider now the case of an underlying Markov chain with three states (high, low, and intermediate levels for the volatility):

$$R = \begin{pmatrix} 0.03 & 0 & 0 \\ 0 & 0.03 & 0 \\ 0 & 0 & 0.03 \end{pmatrix},$$

$$Q = \begin{pmatrix} -2 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 1 & -2 \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} 0.25 & 0 & 0 \\ 0 & 0.50 & 0 \\ 0 & 0 & 0.81 \end{pmatrix}.$$

This gives the following thresholds: $\exp(b) = (0.600, 0.544, 0.455)$. Figure 2 plots the results.

3.1.3. More states. The methodology specified above enables us to deal with a moderately large number of states; in this example, there are eight. In each of the states, r is taken to be equal to 0.03. The jump intensities from one state to another are taken to be equal to 1, the volatility matrix is given by a diagonal matrix with diagonal entries $(0.35, 0.4, 0.6, 0.7, 0.75, 0.8, 0.85, 0.9)$, and in Figure 3, we plot

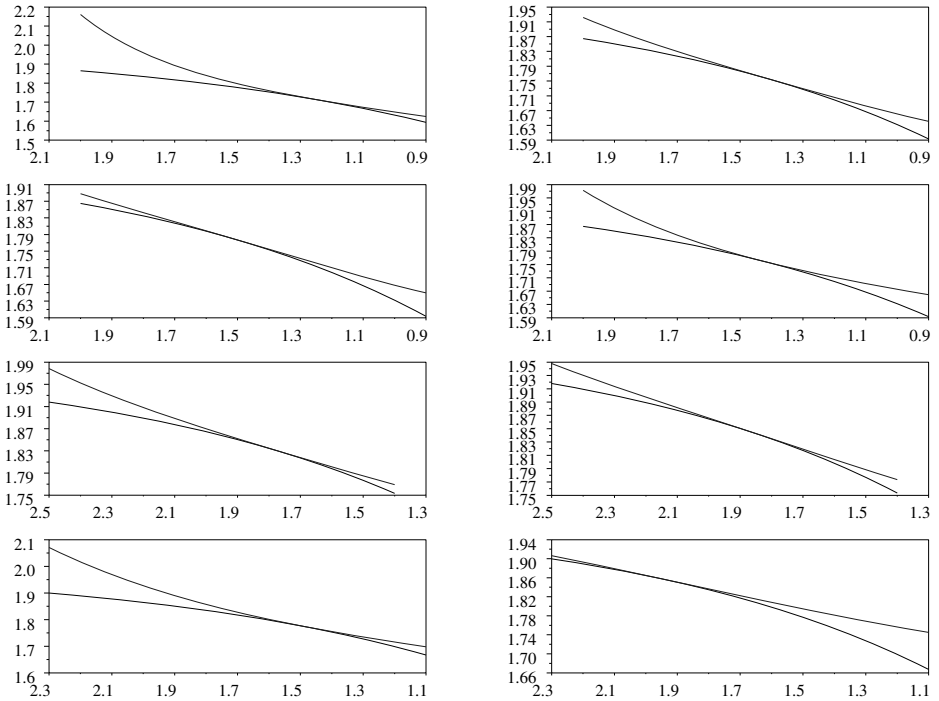


FIG. 3. *Perpetual American put with eight states (value function against log price).*

the value functions for the eight states of the chain, with the corresponding optimal thresholds,

$$\exp(b) = (0.290, 0.239, 0.238, 0.227, 0.220, 0.169, 0.155, 0.140).$$

As one would expect, the less volatile a state is, the bigger is the corresponding threshold.

4. Binomial pricing with Markov regimes. The standard binomial pricing method approximates the log-price process by a random walk, which jumps at the times $\Delta t, 2\Delta t, \ldots$ and, at each jump, moves either up or down. The probability of an up step and the size of the jump are chosen to match the drift and variance to the Black–Scholes asset. In this section, we will extend the alternative random walk approximation introduced by Rogers and Stapleton [22] to the case of Markov regimes.

With X still denoting the Markov-modulated log-price (2), the idea of [22] was to fix some $\Delta x > 0$ and view X only at the discrete set of times at which it has moved by Δx from where we last observed it. Formally, if

$$\begin{cases} \tau_0 = 0, \\ \tau_{n+1} \equiv \inf\{t > \tau_n : |X(t) - X(\tau_n)| > \Delta x\} \quad \text{if } n \geq 0, \end{cases} \tag{29}$$

then we take $(X(\tau_n))$ as the discrete approximation to X, observed for ν steps, where $\nu \equiv \sup\{n : \tau_n < T\}$ (T is the expiry of the option). We need to compute the distribution of $(X(\tau_1), \xi(\tau_1))$. Take $X_0 = 0$, $\tau \equiv \tau_1$ for notational simplicity.

Let λ_i and g_i denote the eigenvalues and the eigenvectors of the eigenvalue problem (11):

$$\begin{pmatrix} 0 & I \\ -2\Sigma^{-1}(Q-R) & 2\Sigma^{-1}R - I \end{pmatrix} \begin{pmatrix} g \\ h \end{pmatrix} = \lambda \begin{pmatrix} g \\ h \end{pmatrix}.$$

There are d negative and d positive eigenvalues. For any scalars w_i,

$$\exp\left(-\int_0^t r(\xi_u)du\right) \Sigma_{i=1}^{2d} w_i g_i(\xi_t) \exp(-\lambda_i X_t)$$

is a martingale, so by the optional sampling theorem,

(30)
$$\mathbf{E}\left[\exp\left(-\int_0^\tau r(\xi_u)du\right) \sum_i w_i g_i(\xi_\tau) \exp(-\lambda_i X_\tau) | X_0 = 0, \xi_0 = j\right] = \sum_i w_i g_i(j).$$

The discounted probability of an upwards step from state j to state k is given by

$$P^+_{j,k} = \mathbf{E}\left[\exp\left(-\int_0^\tau r(\xi_u)du\right) \mathbf{I}\{X_\tau = \Delta x, \xi_\tau = k\} | \xi_0 = j\right] \tag{31}$$

where $\mathbf{I}$ denotes the indicator function. Therefore, in order to find $P^+_{j,k}$ we need to solve the following system:

$$\begin{cases} \sum_i w_i g_i(\xi) \exp(-\lambda_i \Delta x) = \mathbf{I}\{\xi = k\} & \forall \xi = 1, \ldots, d, \\ \sum_i w_i g_i(\xi) \exp(+\lambda_i \Delta x) = 0 & \forall \xi = 1, \ldots, d. \end{cases} \tag{32}$$

This leaves us with $2d$ equations for the $2d$ unknown weights, from which we calculate the discounted probability of an upwards step. Similarly, we can compute the probability of a downwards step from state j to state k by replacing in the above system Δx with $-\Delta x$. When the initial logarithmic price is equal to x, the price of a standard European call option in this Markov-modulated framework is now computed using the following dynamic programming equation, written using vector notation:

$$\begin{cases} V_0(x) = (\exp(x) - K)^+, \\ V_{n+1}(x) = P^+ V_n(x + \Delta x) + P^- V_n(x - \Delta x), \end{cases} \tag{33}$$

where n is the number of time steps to go before expiry T. The matrices P^+ and P^- are defined above and denote, respectively, the up and down transition matrices for the underlying Markov chain.

Observe that (as in Rogers and Stapleton [22]) this approximation is well suited to pricing barrier options; we merely change appropriately the matrices $P^\pm$ at the vertices adjacent to the barrier(s).

Once we have computed the discounted probabilities of an upwards and a downwards step, it remains for us to deal with the fact that the number ν of time steps is random. One solution to this problem is to match bond prices so that

$$\mathbf{E}\left[\exp\left(-\int_0^T r(\xi_u)du\right) 1\right] \simeq \mathbf{E}\left[\exp\left(-\int_0^{\tau_\nu} r(\xi_u)du\right) 1\right]. \tag{34}$$

Let $\overline{P} = P^+ + P^-$. From the above, it is enough to find ν so that

$$\pi \overline{P}^{\nu} 1 = \pi \exp\left[T(Q-R)1\right], \tag{35}$$

where π denotes the invariant distribution of the underlying Markov chain. This simple approximation turns out to give very satisfactory results for the Markov-modulated setup.

Notice finally that for the case of the American put, the dynamic programming equation for the value function now becomes

$$\begin{cases} V_0(x) = (\exp(x) - K)^+, \\ V_{n+1}(x) = \max\{(K - \exp(x))^+, P^+ V_n(x + \Delta x) + P^- V_n(x - \Delta x)\}. \end{cases} \tag{36}$$

The case of the finite expiry Markov-modulated American put was tackled by Buffington and Elliott [4] but only in the case of a two-state Markov chain and by extending the Barone-Adesi–Whaley analytic approximation.

4.1. Numerical results.

4.1.1. Markov-modulated European call. One way of checking our results is to consider the case when the chain switches between two identical states for the volatility. The price in each state should then be equal approximately to the Black–Scholes price for this given volatility. The expiry time is taken to be equal to one year; the initial stock price is $S_0 = 95$. The strike is $K = 100$. Finally, the size of the space grid is taken to be $\Delta x = 0.022$. We take

$$R = \begin{pmatrix} 0.03 & 0 \\ 0 & 0.03 \end{pmatrix},$$

$$Q = \begin{pmatrix} -0.01 & 0.01 \\ 0.01 & -0.01 \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} 0.25 & 0 \\ 0 & 0.25 \end{pmatrix}.$$

The price in each state of the chain is found to be equal to 17.9667, compared to the Black–Scholes value of 17.9506 (relative error: 0.0009).

4.1.2. Markov-modulated American put. Here we compare with the prices tabulated in [21].

(i) Let us first consider the case of two identical states, with $T = 0.5$, $\Delta x = 0.022$, $X_0 = \log(85)$, $K = 100$:

$$R = \begin{pmatrix} 0.06 & 0 \\ 0 & 0.06 \end{pmatrix},$$

$$Q = \begin{pmatrix} -0.01 & 0.01 \\ 0.01 & -0.01 \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} 0.16 & 0 \\ 0 & 0.16 \end{pmatrix}.$$

The price in each of the two states is found to be equal to 18.0285, which needs to be compared with the value 18.0374 found by Rogers [21] (relative error: 0.0005).

(ii) Let us now decrease the volatility in the second state:

$$\Sigma = \begin{pmatrix} 0.16 & 0 \\ 0 & 0.10 \end{pmatrix}.$$

The prices in each of the two states are now found to be $(17.3070, 16.7677)$. Decreasing the volatility decreases the price in the two states, as expected. Correspondingly, increasing the volatility in one of the states increases the price in the two states, as shown by

$$\Sigma = \begin{pmatrix} 0.40 & 0 \\ 0 & 0.16 \end{pmatrix},$$

where the price is now given by $(20.3454, 19.1986)$.

(iii) Taking example (i) and changing just the start value X_0 to $\log(100)$ allows us to compare our values with other values in [22]. We find the price is $(9.9279, 9.9279)$, to be compared with 9.9458 (relative error: 0.00192). Taking $X_0 = \log(115)$, the price is $(5.1109, 5.1109)$, to be compared with 5.1265 (relative error: 0.00304).

It therefore turns out that the random walk approximation provides an accurate and very quick method for Markov-modulated asset dynamics.

The solution for the finite expiry American put should provide a way of checking the results of the preceding section for the perpetual American put, by letting T tend to ∞. For the example of Guo and Zhang [12], where $\exp(b) = (0.616, 0.441)$, the time-0 price of the Markov-modulated perpetual American putm, $(4.2239, 4.2758)$, compares well with our results for the finite expiry American put when $T = 40$: $(4.2180, 4.2692)$ (relative errors: 0.00139, 0.00155).

A similar check can be made for the perpetual American put example with three states, where the prices are $(4.2278, 4.2486, 4.2706)$, to be compared with $(4.2244, 4.2439, 4.2631)$ for the finite expiry case, where $T = 40$ (relative errors: 0.0008, 0.0011, 0.0017).

4.1.3. Markov-modulated barrier options. In this section, we price a number of double knockout barrier options in a Markov-modulated setup.

(i) Let us first consider the case of constant barriers, where we compare our results with those of Geman and Yor [7]. With two identical states, taking $T = 1$, $X_0 = \log(100)$, $K = 100$, $\Delta x = 0.022$, $b^* = \log(150)$, and $b_* = \log(75)$, and

$$R = \begin{pmatrix} 0.05 & 0 \\ 0 & 0.05 \end{pmatrix},$$

$$Q = \begin{pmatrix} -0.01 & 0.01 \\ 0.01 & -0.01 \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} 0.25 & 0 \\ 0 & 0.25 \end{pmatrix}.$$

the price of the double knockout is then found to be $(0.8994, 0.8994)$, to be compared with the value 0.89 computed by Geman and Yor [7] (relative error: 0.01061).

(ii) Changing K to 87.5 and b_* to $\log(50)$, the price becomes $(3.8274, 3.8274)$, to be compared with 3.8075 from Geman and Yor (relative error: 0.00519).

(iii) The next example is the same as the previous one, but now we have two different volatility levels:

$$\Sigma = \begin{pmatrix} 0.50 & 0 \\ 0 & 0.25 \end{pmatrix}.$$

The price of the double knockout is now equal to $(2.6055, 2.5882)$.

(iv) We finally consider the case of a double knockout with moving barriers, which are linear for the log-price $b^* = \log(U)+\Delta x_1 t$ and $b_* = \log(L)+\Delta x_2 t$. We compare our results with those of Kunitomo and Ikeda [18] Let us take: $T = 0.5$, $X_0 = \log(1000)$, $K = 1000$, $\Delta x_1 = 0.1$, $\Delta x_2 = -0.1$, $L = 500$, and $U = 1500$:

$$R = \begin{pmatrix} 0.05 & 0 \\ 0 & 0.05 \end{pmatrix},$$

$$Q = \begin{pmatrix} -0.01 & 0.01 \\ 0.01 & -0.01 \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} 0.04 & 0 \\ 0 & 0.04 \end{pmatrix}.$$

The price of the double knockout in this Markov-modulated setup is found to be $(67.2596, 67.2596)$, to be compared with 67.78 from Kunitomo and Ikeda [18] and 67.7834 from Rogers and Zane [23] (relative error: 0.00773).

5. Conclusions. We have shown how to use classical results from the Wiener–Hopf factorization of Markov processes to price options on a Markov-modulated asset. Such a model can accommodate "bull" and "bear" markets, as well as changes in interest rate and volatility. This method has been applied to the optimal stopping problem of the Markov-modulated perpetual American put. It yields a very efficient and accurate numerical method, which amounts to computing the eigenvalues and eigenvectors of some particular matrices. There is no dynamic programming nor discretization of any PDE. Finally, with a finite time horizon, the approach can be used to construct a modified binomial lattice methodology, which has been applied to the European call, the American put, and double barrier options in a Markov-modulated setup. This modified binomial method turns out to provide an efficient numerical scheme for Markov-modulated option pricing.

REFERENCES

[1] M. T. Barlow, L. C. G. Rogers, and D. Williams, *Wiener-Hopf factorization for matrices*, in Séminaire de Probabilités XIV, Lecture Notes in Math. 784, Springer-Verlag, Berlin, 1980, pp. 324–331.

[2] O. Barndorff-Nielsen and N. Shepard, *Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics*, J. Roy. Stat. Soc. Ser. B, 63 (2000), pp. 1–42.

[3] T. R. Bielecki and S. R. Pliska, *Risk-sensitive dynamic asset management*, Appl. Math. Optim., 39 (1999), pp. 337–360.

[4] J. Buffington and R. J. Elliott, *American options with regime switching*, Internat. J. Theoret. Appl. Finance, 5 (2002), pp. 497–514.

[5] W. H. Fleming and S. J. Sheu, *Risk-sensitive control and an optimal investment model*, Math. Finance, 10 (2000), pp. 197–213.

[6] W. H. Fleming and S. J. Sheu, *Risk-sensitive control and an optimal investment model* (ii), Ann. Appl. Probab., 12 (2000), pp. 730–767.

[7] H. Geman and M. Yor, *Pricing and hedging double-barrier options: A probabilistic approach*, Math. Finance, 6 (1996), 365–378.
[8] X. Guo, *Inside Information and Stock Fluctuations*, Ph.D. Thesis, Rutgers University, New Brunswick, NJ, 1999.
[9] X. Guo, *An explicit solution to an optimal stopping problem with regime switching*, J. Appl. Probab., 38 (2001), pp. 464–481.
[10] X. Guo, *Information and option pricing*, Quantitative Finance, 1 (2001), pp. 38–44.
[11] X. Guo and L. Shepp *Some optimal stopping problem with non-trivial boundaries for pricing exotic options*, J. Appl. Probab., 38 (2001), pp. 1–12.
[12] X. Guo and Q. Zhang, *Closed-form solutions for perpetual American put options with regime switching*, SIAM J. Appl. Math., 64 (2004), pp. 2034–2049.
[13] S. L. Heston, *A closed-form solution for options with stochastic volatility and applications to bond and currency options*, Rev. Financial Stud., 6 (1993), pp. 327–343.
[14] D. G. Hobson and L. C. G. Rogers, *Complete models with stochastic volatility*, Math. Finance, 8 (1998), pp. 27–48.
[15] J. Hull and A. White *The pricing of options on assets with stochastic volatility*, J. Finance, 42 (1987), pp. 281–299.
[16] I. Karatzas, *On the pricing of American options* Appl. Math. Optim., 17 (1988), pp. 37–60.
[17] J. E. Kennedy and D. Williams, *Probabilistic factorization of a quadratic matrix polynomial*, Math. Proc. Cambridge Philos. Soc., 107 (1990), pp. 591–600.
[18] N. Kunitomo and M. Ikeda, *Pricing options with curved boundaries*, Math. Finance, 2 (1992), pp. 275–298.
[19] H. P. McKean, *Appendix: A free boundary problem for the heat equation arising from a problem in mathematical economics*, Indust. Management Rev., 6 (1965), pp. 32–39.
[20] H. Nagai, *Optimal strategies for risk-sensitive portfolio optimization problems for general factor models*, SIAM J. Control Optim., 41 (2003), pp. 1779–1800.
[21] L. C. G. Rogers, *Monte Carlo valuation of American options*, Math. Finance, 12 (2002), pp. 271–286.
[22] L. C. G. Rogers and E. J. Stapleton, *Fast accurate binomial pricing*, Finance and Stoch., 2 (1997), pp. 3–17.
[23] L. C. G. Rogers and O. Zane, *Valuing moving barrier options*, J. Computational Finance, 1 (1997), pp. 5–11.
[24] E. M. Stein and J. C. Stein, *Stock-price distributions with stochastic volatility: An analytic approach*, Rev. Financial Stud., 4 (1991), pp. 727–752.
[25] R. Stockbridge, *Portfolio optimization in markets having stochastic rates*, Stochastic Theory and Control, 280 (2002), pp. 447–458.
[26] D. Yao, Q. Zhang, and X. Zhou, *A Regime-Switching Model for European Option Pricing*, Working Paper, 2003.
[27] G. Yin and X. Y. Zhou, *Markowitz's mean-variance portfolio selection with regime switching: From discrete-time models to their continuous-time limits*, IEEE Trans. Automat. Control, 49 (2004), pp. 349–360.
[28] Q. Zhang, *Stock trading: An optimal selling rule*, SIAM J. Control Optim., 40 (2001), pp. 64–87.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 6, pp. 2079–2103

SUPERVISORY CONTROL OF DISCRETE EVENT SYSTEMS WITH CTL* TEMPORAL LOGIC SPECIFICATIONS*

SHENGBING JIANG† AND RATNESH KUMAR‡

Abstract. The supervisory control problem of discrete event systems with temporal logic specifications is studied. The full branching time logic of CTL* is used for expressing specifications of discrete event systems. The control problem of CTL* is reduced to the decision problem of CTL*. A small model theorem for the control of CTL* is obtained. It is shown that the control problem of CTL* (resp., CTL) is complete for deterministic double (resp., single) exponential time. A sound and complete supervisor synthesis algorithm for the control of CTL* is provided. Special cases of the control of computation tree logic (CTL) and linear-time temporal logic are also studied.

Key words. discrete event system, supervisory control, temporal logic, computation tree logic, linear-time temporal logic

AMS subject classifications. 93C65, 93B05

DOI. 10.1137/S0363012902409982

1. Introduction.

Discrete event systems (DESs) involve discrete-valued quantities that evolve in response to certain discrete qualitative changes, called *events*. Examples of events include arrival of a customer in a queue, termination of an algorithm in a computer program, loss of a message packet in a communication network, and breakdown of a machine in a manufacturing system. The theory of supervisory control of DESs was introduced by Ramadge and Wonham [28] for designing controllers so that the controlled system satisfies certain desired qualitative constraints, such as a buffer in a manufacturing system should never overflow, or a message sequence in a communication network must be received in the same order as it was transmitted. Many extensions of the basic supervisory control problem such as control with partial observations, decentralized control, modular control, control of nondeterministic systems, and control of infinite behaviors represented by ω-languages, have been studied [16].

In the supervisory control framework for discrete-event systems, an uncontrolled discrete event system, called plant, is modeled as a state machine, the event set of which is finite and is partitioned into the set of controllable and uncontrollable events. The language generated by such a state machine is used to describe the behavior of the plant at the logical level. The control task is formulated as that of the synthesis of a controller, called a supervisor, which exercises control over the plant by dynamically disabling some of the controllable events so that the plant achieves a certain desired behavior, called a specification, which is typically expressed as a formal language.

*Received by the editors June 18, 2002; accepted for publication (in revised form) July 18, 2005; published electronically January 6, 2006. The research was supported in part by the National Science Foundation under grants NSF-ECS-9709796, NSF-ECS-0099851, NSF-ECS-0218207, NSF-ECS-0244732, NSF-EPNES-0323379, and NSF-0424048, a DoD-EPSCoR grant through the Office of Naval Research under grant N000140110621, and a KYDEPSCoR grant. This work was performed while the authors were with the Department of Electrical and Computer Engineering, University of Kentucky.

http://www.siam.org/journals/sicon/44-6/40998.html

†GM, R&D and Planning, Warren, MI 48090-9055 (shengbing.jiang@gm.com).

‡Department of Electrical & Computer Engineering Iowa State University, Ames, IA 50011 (rkumar@iastate.edu).

In this paper, we consider temporal logic [6, 12] as a means to express the control specification.

Temporal logic was studied initially to investigate the manner in which temporal operators are used in natural language arguments [11]. It provided a formal way of qualitatively describing and reasoning about how the truth values of assertions change over time. In [27], Pnueli first argued that temporal logic is appropriate for reasoning about nonterminating concurrent programs such as operating systems and network communication protocols. Now temporal logic is a widely active area of research and has been used in all aspects of concurrent program design, including specification, verification, and mechanical program synthesis.

Temporal logic is an effective means of control specification, and researchers have used it for this purpose. For example, [32, 21, 23, 22, 4] used linear-time temporal logic (LTL); [25, 24] used real-time temporal logic (RTTL) and [2] used metric temporal logic (MTL) (both RTTL and MTL are LTL with real time constraints); [1] used computation tree logic (CTL). Temporal logic was also used in [14, 26, 34, 30, 31] for the study of discrete event systems.

These works on a temporal logic approach for control of discrete event systems are limited in one way or other. For example, the main focus in [32, 21, 22, 25] was verification and analysis (no synthesis was performed). In [23, 24], methods were given for the supervisor synthesis for systems with safety specifications only. In [4] supervisory synthesis for propositional-LTL formulas is considered; no test for the existence of a supervisor is provided, a supervisor is synthesized based only on a "one-step look-ahead," and all controllable events are unobservable. In [2], a sound but not complete (see Remark 7) algorithm was given for the synthesis of supervisors for systems with MTL specifications. In [1], the control problem for systems with CTL specifications was studied. But there are some errors and limitations with the result of [1]. First, the semantics of CTL is defined by using $*$-languages (languages of finite strings [16]) in [1]. This is incorrect, since CTL has a branching-time structure and it is known ([6], and also Example 1) that CTL and $*$-languages are incomparable. Besides, CTL can express liveness properties which cannot be expressed by $*$-languages. Second, only *state-based* supervisors were considered in [1]. (Such a supervisor determines its control based only on the present state, ignoring the information about the state sequence the plant has visited in the past.) Third, the algorithm presented in [1], which works for a restricted class of CTL formulas and has a linear complexity in the number of states in the plant and the length of the CTL formulas, is erroneous (see Remark 6).

The work on "module checking" [20] can be viewed as dual to a supervisory control problem. The goal there is to have an "open system" (a plant in the setting of supervisory control) so that the "closed system" (the controlled system in the setting of supervisory control) satisfies the given CTL* specification for all possible environments (supervisors in the setting of supervisory control). Dually, in the setting of supervisory control, the goal is to have an open system so that the closed system satisfies the given CTL* specification for at least one possible environment. Duality lies in the following equivalence: an open system has the property that all the closed systems (that are induced by the various environments) satisfy a CTL* specification f if and only if it is not the case that there exists an environment so that the closed system satisfies the specification $\neg f$. Note that the former is a module-checking problem whereas the latter is a supervisory control problem.

With the above analogy, our work on supervisory control can be viewed as an *extension* of the work presented in the setting of module checking. In the setting of

module checking, the state set is partitioned into the system states and the environment states, and any subset of the feasible events can occur when the system is in one of its environment states. This, in our setting of supervisory control, translates to having

1. states where either all events are controllable (the environment states of module checking) or all events are uncontrollable (the system states of module checking), and
2. the supervisor (the environment in the setting of module checking) is a *deterministic* system.

Our setting is more general: *all states* can have some events that are controllable and others that are uncontrollable, and the supervisor we design can be a *nondeterministic* system. (See Example 1.)

The setting of "control of reactive systems" [18] has a more ambitious goal: synthesize a controller (which disables events in system states) so that the controlled system satisfies the given CTL* specification for all possible environments (which disables events in environment states). Since it is again possible to disable a set of feasible events in a system state (through a controller), this, in the supervisory control setting, translates to having the following:

1. all events are controllable in all states, and
2. the supervisor is a deterministic system.

As explained above, such restrictions are not present in the setting of supervisory control. It should be noted that in the setting of "control of reactive systems," there are two types of "players," a controller/supervisor and the environment. The supervisory control setting allows only one type of player, namely, a controller/supervisor, whereas the environment is always the "maximal" one (that never disables any event). Thus there are also some differences between the settings.

The work on "robust satisfaction" [19] does consider nondeterministic environments (i.e., supervisors). But the composition mechanism, through which the system and the environment interact, brings about additional restrictions, namely,

1. all events in all states are controllable,
2. exactly one controllable event is enabled in each state, and
3. the environment only "observes" the current state of the system (and not the particular event executed by the system),

The existence of the first two restrictions can be argued as follows: in the setting of "robust satisfaction," the environment, based on its present state, generates a unique output (which is an input for the system) that enables that particular event (and nothing else) in the system. Note that by outputting a certain event, the environment can enable that particular event in the system (equivalently, disable others), thereby making all events controllable in all states of the system. A justification for the third restriction is that the environment updates its state based on only the output generated by the system, which is a function of only the system's state. It should be noted that the setting of "robust satisfaction" allows a type of partial observation since the interacting systems only observe each others' outputs, whereas the supervisory control setting we consider assumes a complete observation of events. Thus there are also some differences between the two settings.

In this paper we study the supervisory control problem for plants possessing *uncontrollable* events with specifications expressed in the full branching time logic of CTL* and allowing supervisors to be *nondeterministic*. The reason for allowing nondeterminism is that the class of nondeterministic supervisors is more powerful than

that of deterministic ones, as is illustrated by Example 1, which makes it possible for a supervisor to exist for a larger class of CTL* specifications. Our approach to supervisor synthesis is based on reduction to satisfiability: We show that a supervisor exists if and only if a certain CTL* specification is satisfiable, and whenever this holds a corresponding satisfying model serves as a supervisor. A corollary of this result is that a deterministic supervisor exists if and only if a deterministic satisfying model exists. Thus the approach developed here can be used to determine the existence of a general nondeterministic as well as a deterministic supervisor, and furthermore following our approach a supervisor (nondeterministic or deterministic) can be obtained when one exists.

Note that randomized nondeterministic control is commonly used in the setting of stochastic systems (see, for example, [15]), whereas the use of nondeterministic supervisors in context of discrete-event systems was first explored in [13]. A formal definition of a nondeterministic control policy, its representation as a nondeterministic state machine, and a means to implement it (also see Remark 2) were first introduced in [17]. The nondeterminism in a supervisor state machine is represented by nondeterministic choices and epsilon-transitions. A nondeterministic choice corresponds to randomly choosing one of the control decisions (from among a set of choices determined off-line) on an observation, whereas an epslion-transition corresponds to randomly changing the control decision (again in accordance with choices determined off-line) without any observation. As explained in [17], a nondeterministic choice can be implemented by a "coin-toss," whereas an epsilon-transition can be implemented using a "random-timer." The results in [17] indicate that when the desired specification is language based, there is no gain to having nondeterministic control (over deterministic control) under complete observation of events. However, the situation is different when there is partial observation—a weaker notion of observability is needed for the existence of the supervisor. Further, this weaker property is algebraically better behaved than observability (such as it is closed under union). The present paper demonstrates that even under complete observation of events there is a gain to having nondeterministic supervisors if the desired specifications are expressed in CTL* (which is more expressive than the language-based specifications).

The paper is organized as follows. First a brief introduction to CTL* is given. Next, the control problem of CTL* is reduced to the decision problem of CTL* and a small model theorem for the control of CTL* is derived. It is further shown that the control problem of CTL* (resp., CTL) is complete for deterministic double (resp., single) exponential time, where a decision problem is said to be complete for a certain computation complexity if both the lower and upper complexity bounds of the problem are the same. A sound and complete supervisor synthesis algorithm for the control of CTL* is provided. Special cases of the control of computation tree logic (CTL) and linear-time temporal logic (LTL) are also studied. For these special cases we are able to provide more efficient algorithms. Finally, an illustrative example is given.

2. Introduction to CTL* and tree automaton. CTL* is also called full branching time logic because of its branching time structure, i.e., at each moment, there may exist alternate courses representing different possible futures. It was proposed in [7] as an unifying framework, subsuming both CTL and LTL, as well as a number of other logic systems. Here we give a brief introduction to CTL*. For a complete introduction to temporal logic, see [6].

Let $M = (Q, AP, R, L)$ be a state transition graph (also called the *Kripke structure* [6]), where Q is the set of states (finite or infinite), AP is a finite set of atomic

proposition symbols, $R \subseteq Q \times Q$ is a total transition relation, i.e., for every $s \in Q$ there is a $s' \in Q$ such that $R(s, s')$, and $L : Q \to 2^{AP}$ is a function that labels each state with a set of atomic propositions that are true at that state. A path in M is defined as an infinite sequence of states, $\pi = (s_0(\pi), s_1(\pi), \ldots)$ such that for every $i \in \{0, 1, \ldots\}$, $(s_i(\pi), s_{i+1}(\pi)) \in R$.

Using the atomic propositions and boolean connectives such as conjunction, disjunction, and negation, we can construct more complex expressions describing properties of states. However, we are also interested in describing the properties of sequences (and more generally of tree structures) of states that the system can visit. Temporal logic is a formalism for describing properties of sequences of states as well as of tree structures of states. Such properties are expressed using *temporal operators* and *path quantifiers* of the temporal logic. These operators and quantifiers can be nested with boolean connectives to generate more complex temporal logic specifications.

The following temporal operators are used for describing the properties along a specific path:

- X ("next time"): requires that a property hold in the next state of the path.
- U ("until"): used to combine two properties. The combined property holds if there is a state on the path where the second property holds, and at every preceding state on the path, the first property holds.
- F ("eventually" or "in the future"): used to assert that a property will hold at some future state on the path.
- G ("always" or "globally"): specifies that a property holds at every state on the path.
- B ("before"): also combines two properties. It requires that if there is a state on the path where the second property holds, then there exists a preceding state on the path where the first property holds.

We have following relations among the above operators, where f denotes a temporal logic specification:

- $Ff \equiv trueUf$,
- $Gf \equiv \neg F \neg f$,
- $fBg \equiv \neg(\neg fUg)$.

Thus one can use X and U to express the other temporal operators.

To describe the branching time structure starting at a particular state, two path quantifiers are used:

- A : for all paths and
- E : for some paths.

These two quantifiers are used in a particular state to specify that all the paths or some of the paths starting at that state have some property. The two quantifiers are related by

- $A \equiv \neg E \neg$.

There are two types of formulas in CTL*: *state formulas* (which are true in a specific state) and *path formulas* (which are true along a specific path). Now we give the definition of CTL* formulas. In the following we assume that p is an atomic proposition, f_1 and f_2 are state formulas, and g_1 and g_2 are path formulas.

Syntax. We inductively define a class of state formulas using rules S1–S3 below and a class of path formulas using rules P1–P3 below:

S1 If $p \in AP$, then p is a state formula.

S2 If f_1 and f_2 are state formulas, then so are $\neg f_1$ and $f_1 \wedge f_2$.

S3 If g_1 is a path formula, then Eg_1 and Ag_1 are state formulas.

P1 Each state formula is also a path formula.

P2 If g_1 and g_2 are path formulas, then so are $\neg g_1$ and $g_1 \wedge g_2$.

P3 If g_1 and g_2 are path formulas, then so are Xg_1 and $g_1 U g_2$.

CTL* formulas are the state formulas generated by the above rules. The length of a formula is the number of boolean, temporal, and path quantifier operators in the formula.

The restricted logic CTL is obtained by restricting the syntax to disallow boolean combinations and nestings of temporal operators. Formally, rules P1–P3 are replaced by

P0 If f_1 and f_2 are state formulas, then Xf_1 and $f_1 U f_2$ are path formulas.

Then CTL formulas are the state formulas generated by rules S1–S3 and P0.

The logic LTL is obtained by removing rules S2–S3, i.e., LTL formulas are state formulas in the form of Ag where g is any path generated by rules S1 and P1–P3. Note instead of defining LTL as path formulas (g) as in [6], we define LTL as state formulas (Ag) as in [3]. This is because for the LTL control problem studied in this paper, we want *all* paths starting from the initial state of the plant to satisfy some required property which can be expressed by a LTL formula of the form Ag.

Note that the only restriction in CTL is that every temporal operator in the formula is immediately preceded by a path quantifier, whereas the only restriction in LTL is that except for the path quantifier A appearing at the beginning of the formula no other path quantifiers exist in the formula. CTL and LTL have different expressive power. For example, the CTL formula $AGEFp$ cannot be expressed by any LTL formula, and the LTL formula $AFGp$ cannot be expressed by any CTL formula, but AGp can be viewed as either a CTL formula or an LTL formula.

Semantics. We define the semantics of CTL* with respect to a state transition graph $M = (Q, AP, R, L)$. For a state formula f, the notation $< M, s >\models f$ (resp., $< M, s >\not\models f$) means that f holds (resp., does not hold) at state s in M. For a path formula g, the notation $< M, \pi >\models g$ (resp., $< M, \pi >\not\models g$) means that g holds (resp., does not hold) along the path π in M. The relation $\models$ is defined inductively as follows:

1. $< M, s >\models p$ if and only if $p \in L(s)$ $\forall p \in AP$.
2. $< M, s >\models \neg f_1$ if and only if $< M, s >\not\models f_1$.
3. $< M, s >\models f_1 \wedge f_2$ if and only if $< M, s >\models f_1$ and $< M, s >\models f_2$.
4. $< M, s >\models Eg_1$ if and only if there exists a path π starting at s such that $< M, \pi >\models g_1$.
5. $< M, s >\models Ag_1$ if and only if for every path π starting at s, we have $< M, \pi >\models g_1$.
6. $< M, \pi >\models f$ if and only if $< M, s_0(\pi) >\models f$, for any state formula f.
7. $< M, \pi >\models \neg g_1$ if and only if $< M, \pi >\not\models g_1$.
8. $< M, \pi >\models g_1 \wedge g_2$ if and only if $< M, \pi >\models g_1$ and $< M, \pi >\models g_2$.
9. $< M, \pi >\models Xg_1$ if and only if $< M, \pi^1 >\models g_1$, where $\pi^1 = (s_1(\pi), s_2(\pi), \ldots)$.
10. $< M, \pi >\models g_1 U g_2$ if and only if there exists a k such that $< M, \pi^k >\models g_2$ and for all $j \in \{0, 1, \ldots, k-1\}$, $< M, \pi^j >\models g_1$, where $\pi^k = (s_k(\pi), s_{k+1}(\pi), \ldots)$.

Remark 1. In the above, the CTL* is interpreted over nonterminating paths. In some cases, we may need to study the systems with terminating behaviors. So the definition of CTL* semantics needs to be extended to finite paths. In this paper, we only consider the systems with nonterminating behaviors and hence use only the above definition.

The following examples show that temporal logic formulas can be used to express

properties such as safety, nonblocking, liveness, and stability.

AGp means that "*for all paths* (A) starting at the present state, *globally* (G) at every state along these paths p is true." It is a safety property.

$AGEFp$ means that "*for all paths* (A) starting from the present state, *globally* (G) for every state along these paths *there exists* (E) a path starting from that state such that *in future* (F) p holds at a state on that path." It is a nonblocking property.

$AG(p_1 \Rightarrow AFp_2)$ means that "*for all paths* (A) starting from the present state, *globally* (G) for every state s along these paths, if p_1 is true at the state s, then p_2 will be true at some *subsequent state along every path* (AF) starting from the state s." It is a liveness property.

$AFGp$ means that "*for all paths* (A) starting from the present state, *eventually* (F) p holds *globally* G". It is a property of stability which requires that the system should eventually reach a set of states where p holds and stay there forever.

DEFINITION 1. *We say that a state formula f is* satisfiable *provided that for some state transition graph M and some state s in M we have $< M, s > \models f$, in which case M is called a* model *for f.*

The decision problem of a temporal logic formula is to test whether the given formula is satisfiable. We have following results for the decision problems of CTL* and CTL.

THEOREM 1 (see [9, 5]). *Given a CTL* formula f, f is satisfiable if and only if it is satisfiable in a finite state transition graph with number of nodes at most double exponential in the length of the formula f.*

THEOREM 2 (see [6]). *The decision problem of CTL* (resp., CTL) is complete for deterministic double (resp., single) exponential time.*

Theorem 1 is called the *small model theorem* for the decision of CTL*. It states that a CTL* formula is satisfiable if and only if it is satisfiable in a *small* finite model, where *small* means that the size of the model is bounded by some function of the length of the given formula. Theorem 2 states that the lower as well as the upper bound of the complexity of the decision problem for CTL* (resp., CTL) is deterministic double (resp., single) exponential in the length of the given formula. (By double (resp., single) exponential we mean $exp(exp(n))$ (resp., $exp(n)$), where $exp(n)$ is a function c^n for some $c > 1$.)

To test the satisfiability of a CTL* formula f, we have the following sound and complete decision procedure [6, 9, 8], the complexity of which is double exponential in the length of the specification CTL* formula.

1. Derive a Rabin tree automaton for the CTL* formula f [9]. The number of states (resp., acceptance condition pairs) of the Rabin tree automaton is double (resp., single) exponential in the length of the formula f.
2. Test the emptiness of the Rabin tree automaton [8]. If the Rabin tree automaton is empty, then the CTL* formula f is not satisfiable; otherwise the formula f is satisfiable, and a model for f can be extracted from the Rabin tree automaton. The complexity of this step is polynomial in the number of states of the Rabin tree automaton and exponential in the number of acceptance condition pairs of the Rabin tree automaton.

The notion of Rabin tree automaton is described below. For simplicity, we consider only the finite automaton on infinite binary trees. The infinite binary tree is the set $T = pr(\{0, 1\}^\omega)$. The elements of T are called nodes, and the empty word ϵ is the root of T. For all $x \in T$, $x \cdot 0$ and $x \cdot 1$ are the left and right successors of x,

respectively. A path π of the tree T is a subset of T such that the root ϵ is in π, and $\forall x \in \pi$, one and only one of $x \cdot 0$ and $x \cdot 1$ is in π. Note that a path of T corresponds a unique word in $\{0,1\}^\omega$. Given an alphabet Σ, a Σ-labeled tree (called Σ-tree) is a function $V : T \to \Sigma$ that maps each node of T to a letter in Σ.

A Rabin tree automaton (on infinite binary Σ-tree) is $\mathcal{A} = (Q, \Sigma, \delta_{\mathcal{A}}, q_0, F)$, where Q is a finite state set, Σ is a finite alphabet set, $\delta_{\mathcal{A}} : Q \times \Sigma \to 2^{Q \times Q}$ is the transition function, $q_0 \in Q$ is the initial state, and $F = \{(G_i, R_i) \mid G_i \cup R_i \subseteq Q, i = 1, \ldots, k\}$ is the Rabin acceptance condition. A run r of $\mathcal{A}$ on an input Σ-tree V is a Q-labeled tree $r : T \to Q$ such that $r(\epsilon) = q_0$ and $\forall\ x \in T$, $(r(x \cdot 0), r(x \cdot 1)) \in \delta_{\mathcal{A}}(r(x), V(x))$. We say that $\mathcal{A}$ accepts an input Σ-tree V if and only if there exists a run r of $\mathcal{A}$ on V such that for each path π of r, there exists a pair (G_i, R_i) in F such that π visits G_i infinitely often and R_i finitely often.

To test the satisfiability of a CTL (a special case of CTL*) formula f, the following more efficient sound and complete decision procedure exists [6], the complexity of which is single exponential in the length of the specification CTL formula:

1. Construct a tableau for the CTL formula f, where a tableau is a state transition structure derived for the given temporal logic formula from which a model of the given formula can be extracted as a subtransition structure whenever that formula is satisfiable. The number of states of the tableau for the CTL formula f is exponential in the length of f.
2. Test the tableau for the existence of a model for f. If there does not exist a model for f in the tableau, then the CTL formula f is not satisfiable; otherwise the formula f is satisfiable, and a model for f can be extracted from the tableau. The complexity of this step is polynomial in the number of states of the tableau.

3. Supervisory control for CTL* specification. In this section, we study the supervisory control problem for systems with CTL* temporal logic specifications. From now on, we assume that the uncontrolled discrete event plant P is modeled by a six tuple: $P = (X, \Sigma, \delta_P, x_0, AP, L_P)$, where X is a finite set of states; Σ is a finite set of event labels that is the disjoint union of Σ_c, the set of controllable events, and Σ_u, the set of uncontrollable events; $\delta_P : X \times \Sigma \to X$ is a partial function defined at each state in X for a subset of Σ; $x_0 \in X$ is the initial state of P; AP is the finite set of atomic proposition symbols with $AP \cap X = \emptyset$; and $L_P : X \to 2^{AP \cup \{\neg p | p \in AP\}}$ is a labeling function such that $\forall x \in X$, $\forall p \in AP$, $p \in L_P(x) \Rightarrow \neg p \notin L_P(x)$. Here for a state x, $p \in L_P(x)$ means that p holds at x, $\neg p \in L_P(x)$ means that p does not hold at x, and if for some atomic proposition p such that neither p nor $\neg p$ is in $L_P(x)$, then it means that p may or may not hold at x. Note from the definition of the transition function δ_P that we are assuming P to be deterministic.

A supervisor S is modeled by a six tuple: $S = (Y, \Sigma, \delta_S, y_0, AP, L_S)$, where Y is a set of states (finite or infinite); Σ and AP are the same sets as given in P; $\delta_S : Y \times \Sigma \to 2^Y$ is a total function defined at each state in Y for each event in Σ; $y_0 \in Y$ is the initial state of S; and $L_S : Y \to 2^{AP \cup \{\neg p | p \in AP\}}$ is a labeling function similar to that in P such that $\forall y \in Y$, $\forall p \in AP$, $p \in L_P(y) \Rightarrow \neg p \notin L_P(y)$. Note from the definition of the transition function δ_S that S is allowed to be nondeterministic. The class of nondeterministic supervisors is more powerful than that of deterministic supervisors, as illustrated by Example 1.

The controlled plant is obtained by the strict synchronous composition of P and S, denoted by $P||S$, which is defined as $P||S = (Z, \Sigma, \delta_{P||S}, z_0, AP, L_{P||S})$, where $Z = X \times Y$ is the state set; Σ and AP are the same sets as given in P; and $\delta_{P||S} : Z \times \Sigma \to 2^Z$

is the state transition function for $P||S$. Let $\sigma \in \Sigma$ and $(x, y) \in X \times Y = Z$; then we define $\delta_{P||S}$ as

$$\delta_{P||S}((x,y),\sigma) = \begin{cases} \{(\delta_P(x,\sigma), z) \mid z \in \delta_S(y,\sigma)\} & \text{if } \delta_P(x,\sigma) \text{ is defined and } \delta_S(y,\sigma) \neq \emptyset; \\ \emptyset & \text{otherwise.} \end{cases}$$

$z_0 = (x_0, y_0) \in Z$ denotes the initial state of $P||S$, and $L_{P||S} : Z \to 2^{AP \cup \{\neg p | p \in AP\}}$ is the labeling function for $P||S$, which is defined as $L_{P||S}(x, y) = L_P(x) \cup L_S(y)$.

We use $M_{P||S} = (Z, R, AP, L)$ to denote the state transition graph of $P||S$, where Z and AP are the same sets as given in $P||S$; $R \subseteq Z \times Z$ is the transition relation with $R = \{(z, z') \mid \exists \sigma \in \Sigma \ s.t. \ z' \in \delta_{P||S}(z, \sigma)\}$; and $L : Z \to 2^{AP}$ is the labeling function which is defined as $\forall z \in Z$, $L(z) = L_{P||S}(z) \cap AP$.

We require that all the supervisors derived should be control-compatible and propositionally consistent with respect to the plant. The control-compatibility of a supervisor requires that when controlling the plant P, the supervisor should never disable an uncontrollable transition in P, where a transition is called an uncontrollable transition if it is labeled by an uncontrollable event. Next, since the propositional labeling of a state $z = (x, y) \in Z$ of $P||S$ is obtained as $L_P(x) \cup L_S(y)$, it is possible that the label of z contains $p \in AP$ as well as its negation (for example, when $p \in L_P(x)$ and $\neg p \in L_S(y)$). We exclude such state machines from being a supervisor by requiring the propositional consistency property defined below.

DEFINITION 2. *A supervisor S is said to be* control-compatible *with respect to a given plant P if for any $s \in \Sigma^*$, $\sigma \in \Sigma_u$, and $z = (x, y) \in \delta_{P||S}(z_0, s)$ such that σ is defined at state x of P, it holds that σ is also defined at state y of S. A supervisor S is said to be* propositionally consistent *with respect to a given plant P if it holds in $P||S$ that for every state $z \in Z$ reachable from z_0, we have $\forall p \in AP$, $p \in L_{P||S}(z) \Rightarrow \neg p \notin L_{P||S}(z)$.*

The supervisory control problem for systems with temporal logic specifications is formulated as follows:

> Let P be a deterministic nonterminating plant with $\Sigma = \Sigma_c \cup \Sigma_u$. For a given CTL* formula f, find a control-compatible and propositionally consistent supervisor S for P such that $P||S$ is nonterminating and $< M_{P||S}, \ z_0 > \models f$, where $M_{P||S}$ is the state transition graph of $P||S$ and z_0 is the initial state of $P||S$.

Before solving the above control problem, we give the definition of the controllability of CTL* formulas.

DEFINITION 3. *Given a nonterminating plant P, a CTL* formula f is said to be controllable with respect to P, also called* P-controllable*, if there exists a control-compatible and propositionally consistent supervisor S such that $P||S$ is nonterminating and $< M_{P||S}, z_0 > \models f$.*

In Definition 3, the supervisor S need not be finite. Through the *small model theorem* derived below, we demonstrate that if a CTL* formula f is controllable, then f can be enforced by a *finite* supervisor. In other words, we don't impose the finiteness of a supervisor a priori in the definition of controllability. Also, the supervisor is allowed to be nondeterministic since in some situations only a nondeterministic supervisor can achieve a given CTL* specification. This is illustrated by the following example.

EXAMPLE 1. *The plant P is shown in Figure* 1(a)*, where $X = \{x_0, x_1, x_2, x_3\}$, $\Sigma = \Sigma_c = \{a, b, c, d, e\}$, $AP = \{p_1, p_2\}$, $L_P(x_0) = L_P(x_1) = AP$, $L_P(x_2) = \{p_1, \neg p_2\}$, and $L_P(x_3) = \{\neg p_1, p_2\}$. (We adopt the following convention for the figures we draw:*

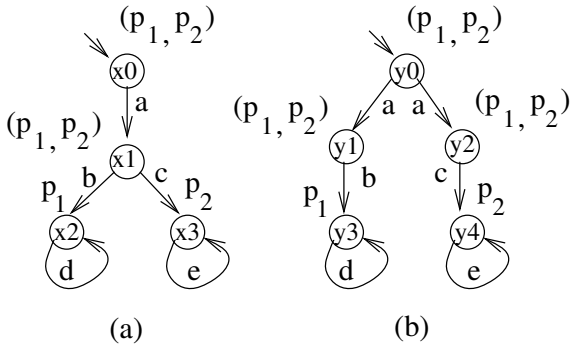


FIG. 1. *Nondeterministic supervisor.*

if an atomic proposition p is not labeled at a state x, then it means that p does not hold at x, i.e., $\neg p \in L_P(x)$.) The specification is described by the CTL *formula $EXAGp_1 \wedge EXAGp_2$, where $EXAGp_i$ has the following meaning: "Exists* (E) *a path (starting from the initial state) such that from the next* (X) *state all paths* (A) *always* (G) *satisfy p_i." Note that the given plant does not satisfy the specification since starting from the only next state x_1, all paths do not always satisfy p_1 and p_2.*

Further, there does not exist a deterministic supervisor that can achieve the specification since AGp_1 and AGp_2 can not be satisfied simultaneously at state x_1. But we can have a nondeterministic supervisor S to achieve the specification, which is shown in Figure 1(b).

Also note that the $*$-language as well as ω-language of the controlled plant is the same as that of the uncontrolled plant, i.e., $L(P||S) = L(P) = a(bd^* + ce^*)$ and $L_\omega(P||S) = L_\omega(P) = a(bd^\omega + ce^\omega)$. This implies that the above CTL *specification can not be expressed by a regular $*$-language or a regular ω-language.*

Remark 2. A formal treatment of nondeterministic control policy, its representation as a state machine, and its implementation are given in [17]. The essential idea is that the control action selection of a nondeterministic supervisor is done on-line nondeterministically from among a set of choices determined off-line. Also, the control action can be changed on-line nondeterministically (before any new observation) in accordance with choices determined off-line. (This feature of nondeterministic control is not being used in the present paper.) The on-line choices, once made, can be used to affect the set of control action choices in future. A nondeterministic control map with above features may be implemented as a control and observation compatible nondeterministic state machine introduced in [17]. (In the context of the present paper, we are assuming a complete observation of events and so only control compatibility is required; observation compatibility is automatically guaranteed.) It is further argued in [17] that to implement a nondeterministic supervisor a mechanism is needed for the on-line nondeterministic selection of the control action (from the set of choices computed off-line), and another mechanism is needed to determine when to nondeterministically change the control action. For the first purpose, a "coin toss" (with as many possible outcomes as the number of control action choices) can be used. For the second purpose, a "random timer" can be used. In the lack of any new observation, the control action is changed if and when the timer goes off.

In the following, we reduce the problem of the control of CTL* to that of the

decision of CTL*, then use the results for the decision of CTL* to solve the control problem of CTL*. We first encode all the controllable sub-trees embedded in the "plant-tree" P by a CTL formula f_P defined as follows.

Add new fresh atomic propositions. Extend AP to $AP' := AP \cup X$. Each state of the plant is viewed as a new atomic proposition. For each $x \in X$, the proposition x holds at state x and at no other state of P.

Encode the initial state of P using formula f_0 defined as

$$f_0 := x_0.$$

This says that in a model for f_0, the atomic proposition x_0 holds at the initial state of the model.

Encode the state set of P using formula $f_1 := f_{11} \wedge f_{12}$ defined as

$$f_{11} := AG\Big(\bigvee_{x\in X} x\Big) \bigwedge_{x\in X} AG\Big(x \Rightarrow \bigwedge_{x'\neq x} \neg x'\Big),$$

$$f_{12} := \bigwedge_{x\in X} AG\Big[x \Rightarrow \bigwedge_{p\in(L_P(x)\cap AP)} p \bigwedge_{\neg p\in(L_P(x)\cap \overline{AP})} \neg p\Big],$$

and $\overline{AP} = \{\neg p \mid p \in AP\}$. In the above, f_{11} states that if M is a model for f_{11}, then every state in M should be labeled with one and only one atomic proposition $x \in X$; f_{12} states that if M is a model for f_{12}, then any atomic proposition which holds (resp., does not hold) at the state x of P should also hold (resp., should not hold) at states in M which are labeled by the proposition x.

Encode the transitions of P using formula f_2 defined as

$$f_2 := \bigwedge_{x\in X} AG\Big(x \Rightarrow AX\Big(\bigvee_{x'\in I_x} x'\Big)\Big),$$

where $I_x = \{x' \mid \exists \sigma \in \Sigma$ such that $x' = \delta_P(x,\sigma)\}$. The formula f_2 states that if M is a model for f_2, s is a state in M labeled with the atomic proposition x, and s' is a successor of s in M labeled with the atomic proposition x', then there must exist a transition from x to x' in P.

Encode the uncontrollable transitions of P using formula f_3 defined as

$$f_3 := \bigwedge_{x\in X} AG\Big(x \Rightarrow \bigwedge_{x'\in I_x^u} EXx'\Big),$$

where $I_x^u = \{x' \mid \exists \sigma \in \Sigma_u$ such that $x' = \delta_P(x,\sigma)\}$. The formula f_3 states that if M is a model for f_3, s is a state in M labeled with the atomic proposition x, and there exists an uncontrollable transition from state x to another state x' in P, then there must exist a successor s' of s in M such that x' is labeled at s'.

Encode all uncontrollable sub-trees of P using the formula f_P defined as

$$f_P := f_0 \wedge f_1 \wedge f_2 \wedge f_3.$$

Remark 3. From the above definition it follows that f_P encodes some information of the plant P. It should be noted that f_P does not contain all the information of P

since from a model M of f_P we cannot reconstruct the plant state machine P. This is because when we encode the transitions (resp., uncontrollable transitions) of P by f_2 (resp., f_3), we require only that the state x' is one step reachable from x, and we ignore all other information such as how many transitions exist between x and x' in P and what are the event labels of these transitions. But the information encoded by f_P is enough for the control of P which is shown in Theorem 3 below.

The following lemma shows that f_P is satisfied by the plant P.

PROPOSITION 1. *Let P be a nonterminating plant and $M_P = (X, R_P, AP', L'_P)$ be the state transition graph of P with $AP' = AP \cup X$, $R_P = \{(x, x') \in X \times X \mid \exists \sigma \in \Sigma,\ x' = \delta_P(x, \sigma)\}$, $L'_P(x) = (L_P(x) \cap AP) \cup \{x\}\ \forall x \in X$. Then it holds that $< M_P, x_0 > \models f_P$, where f_P is as defined above.*

Proof. Since $x_0 \in L'_P(x_0)$, obviously $< M_P, x_0 > \models f_0$. Next, for each state x in M_P, we have

- $[x \in L_P(x)] \bigwedge_{x' \neq x} [x' \notin L_P(x)] \Rightarrow < M_P, x_0 > \models f_{11}$;
- $[\forall p \in (L_P(x) \cap AP), p \in L'_P(x)] \wedge [\forall \neg p \in (L_P(x) \cap \overline{AP}), p \notin L'_P(x)]$ $\Rightarrow < M_P, x_0 > \models f_{12}$;
- $[\forall x' \in \{x' \mid (x, x') \in R_P\}, \exists \sigma \in \Sigma, x' = \delta_P(x, \sigma)] \Rightarrow < M_P, x_0 > \models f_2$;
- $[\forall x' \in \{x' \mid \exists \sigma \in \Sigma_u, x' = \delta_P(x, \sigma)\}, (x, x') \in R_P] \Rightarrow < M_P, x_0 > \models f_3$.

Combining the above implications, we obtain $< M_P, x_0 > \models f_P$. ∎

The following theorem reduces the control problem of CTL* to the decision problem of CTL*.

THEOREM 3. *Given a CTL* formula f and a deterministic nonterminating plant P encoded by the CTL formula f_P, f is P-controllable if and only if the CTL* formula $f \wedge f_P$ is satisfiable.*

Proof. For the necessity, suppose there exists a control-compatible and propositionally consistent supervisor $S = (Y, \Sigma, \delta_S, y_0, AP, L_S)$ such that $< M_{P||S}, z_0 > \models f$. Then we can get a model $M' = (Z, R, AP', L')$ for $f \wedge f_P$ from $M_{P||S} = (Z, R, AP, L)$ as follows: $\forall z = (x, y) \in Z$, $L'(z) = L(z) \cup \{x\}$. Since $< M_{P||S}, z_0 > \models f$, it is obvious that M' is also a model for f, i.e., $< M', z_0 > \models f$. For the formula $f_P = f_0 \wedge f_{11} \wedge f_{12} \wedge f_2 \wedge f_3$, we have the following. Since $z_0 = (x_0, y_0)$, $x_0 \in L'(z_0)$, this implies $< M', z_0 > \models f_0$. Since $M_{P||S}$ can be viewed a subgraph embedded in P, M' is also a subgraph embedded in P. This implies that $< M', z_0 > \models f_{11} \wedge f_2$. From the definition of $L_{P||S}$ and the propositional consistency of S, we know that $< M', z_0 > \models f_{12}$. Further, from the control-compatibility of S, we have $< M', z_0 > \models f_3$. Combining these, we get $< M', z_0 > \models f \wedge f_P$, i.e., $f \wedge f_P$ is satisfiable.

For the sufficiency, let $M = (Q, R, AP', L)$ be a model of $f \wedge f_P$, i.e., $\exists q_0 \in Q$, $< M, q_0 > \models f \wedge f_P$. We can get a supervisor $S = (Y, \Sigma, \delta_S, y_0, AP, L_S)$ from M as follows: $Y \subseteq Q$ is the set of states which are reachable from q_0 in M; $\forall y \in Y,\ \forall \sigma \in \Sigma,$

$$\delta_S(y, \sigma) = \{y' \mid [(y, y') \in R] \wedge [x' = \delta_P(x, \sigma)],\ \text{where } \{x'\} = L(y') \cap X \text{ and } \{x\} = L(y) \cap X\};$$

$y_0 = q_0$; and $\forall y \in Y, L_S(y) = L(y) \cap (AP \cup \{\neg p | p \in AP\})$. Since M is a model of f_P, it ensures that S is control-compatible with respect to P, and further because P is deterministic, S is propositionally consistent with respect to P. Also because P is deterministic, $P||S$ has the same graph as S, and hence it is nonterminating and $< M_{P||S}, z_0 > \models f$. So f is P-controllable. ∎

Now from the small model theorem for the decision of CTL* (Theorem 1), we have the following small model theorem for the control of CTL*.

THEOREM 4. *Given a CTL* formula f and a deterministic nonterminating plant P, f is P-controllable if and only if there exists a* finite *state control-compatible*

and propositionally consistent supervisor S such that $P||S$ is nonterminating and $< M_{P||S}, z_0 > \models f$.

Proof. The sufficiency is obvious. For necessity, from Theorem 3 we know that if f is P-controllable, then $f \wedge f_P$ is satisfiable. Further, from Theorem 1, we have that if $f \wedge f_P$ is satisfiable, then there exists a finite state transition graph $M = (Q, R, AP', L)$ such that $\exists q_0 \in Q,\ < M, q_0 > \models f \wedge f_P$. Using the same method as that in the proof of Theorem 3, we can obtain a finite state control-compatible and propositionally consistent supervisor S from M such that $P||S$ is nonterminating and $< M_{P||S}, z_0 > \models f$. So the theorem holds. □

From Theorem 2, we have the following result for the complexity of control problem for CTL* (resp., CTL).

THEOREM 5. *The control problem for CTL* (resp., CTL) is complete for deterministic double (resp., single) exponential time in the length of the specification formula.*

Proof. From Theorem 3 and the definition of f_P, whose length is polynomial in the number of states of P, we know that the control problem for CTL* (resp., CTL) is polynomial-time reducible to the decision problem for CTL* (resp., CTL). From Theorem 2 we have that the complexity of testing the satisfiability for CTL* (resp., CTL) has an upper bound of deterministic double (resp., single) exponential time in the length of the specification formula. So the control problem for CTL* (resp., CTL) is upper bounded by deterministic double (resp., single) exponential time in the length of the specification formula. This establishes the desired upper bound of the complexity of the control problem.

To establish the desired lower bound of the complexity of the control problem, in view of Theorem 2 it suffices to show that the decision problem can be polynomially reduced to a control problem. For the decision problem of CTL* (resp., CTL), we can view it as a control problem for the plant $P = (X, \Sigma, \delta_P, x_0, AP, L_P)$ with $X = \{x_0\}$; $\Sigma = \Sigma_c = \{\sigma\}$; $x_0 = \delta_P(x_0, \sigma)$; $L_P(x_0) = \emptyset$, where the goal of the control is to find a supervisor that the controlled plant satisfies the given CTL* (resp., CTL) formula. If a supervisor S exists for the above control problem, we can directly use $M_{P||S}$ as the model of the given CTL* (resp., CTL) formula. Since the decision problem for CTL* (resp., CTL) has a lower bound complexity of deterministic double (resp., single) exponential time in the length of the specification formula, we must have that the complexity of the control problem for CTL* (resp., CTL) is lower bounded by deterministic double (resp., single) exponential time in the length of the specification formula. □

From Theorem 3, we know that an algorithm for the supervisor synthesis for CTL* control can be obtained from the decision procedure of CTL*. Let f be a CTL* specification formula and P be a deterministic nonterminating plant; then a supervisor synthesis algorithm is as follows.

ALGORITHM 1. SUPERVISOR SYNTHESIS ALGORITHM FOR CTL* CONTROL.

1. Test the satisfiability of the CTL* formula $f \wedge f_P$. This step is done by using the decision procedure for CTL* as follows:
   (a) Construct a Rabin tree automaton for the CTL* formula f using the method given in [9].
   (b) Construct a tree-automaton for f_P directly from the plant P; this tree automaton has the same state set as P and has no acceptance conditions.
   (c) Construct the Rabin tree automaton for $f \wedge f_P$ from the synchronous composition of the above two tree automata.

(d) Test the emptiness of the set of trees accepted by the Rabin tree automaton for $f \wedge f_P$ [8]. The set of trees accepted by the tree automaton is empty if and only if $f \wedge f_P$ is not satisfiable. If $f \wedge f_P$ is satisfiable, then go to next step; otherwise stop the algorithm and output that "no supervisor exist."

2. If $f \wedge f_P$ is satisfiable, extract a model for the formula $f \wedge f_P$ from its nonempty Rabin tree automaton using the result given in [8].
3. Derive a supervisor from the model for the formula $f \wedge f_P$ by using the method in the proof of Theorem 3.

Remark 4. From Theorem 3, and using an argument similar to the soundness and completeness of the decision procedure for CTL* [9, 8], we can conclude that Algorithm 1 for control synthesis for CTL* is sound and complete. Algorithm 1 has a worst case complexity of double exponential in the length of the CTL* formula f and polynomial in the size of the plant P. This is because the Rabin tree automaton for the specification formula f has a number of states that *is* double exponential in the length of f and has a number of acceptance condition pairs which *is* single exponential in the length of f, and the tree automaton for f_P has the same state set as P and has no acceptance condition, so the final Rabin tree automaton for $f \wedge f_P$ has a number of states which *is* double exponential in the length of f and linear in the number of states of the plant, and it has a number of acceptance condition pairs which *is* single exponential in the length of the specification formula f only.

For an easy synchronous composition of tree automata for f and f_P, it is required that the two tree automata have the same branching degree. To compute the branching degree of a CTL* formula f, we first express it in its *positive normal form* by pushing negations as far inward as possible using De Morgan's law ($\neg(f_1 \vee f_2) \equiv \neg f_1 \wedge \neg f_2$, $\neg(f_1 \wedge f_2) \equiv \neg f_1 \vee \neg f_2$) and the dualities ($\neg AGf_1 \equiv EF\neg f_1$, $\neg A[f_1 U f_2] \equiv E[\neg f_1 B f_2]$, etc.). Then the branching degree of f, denoted by d_f, can be chosen to be the total number of the existential path quantifier "E" in its positive normal form. Similarly, we can get the branching degree of f_P, denoted by d_{f_P}. Then we can choose $d = d_f + d_{f_P}$ as the branching degree of the tree automata models for f and f_P. Next we give an example to illustrate how to compute the branching degree of a CTL* formula and how to derive a tree automaton with a required branching degree for the encoding f_P of P that has the same state set as P.

EXAMPLE 2. *Consider the encoding f_P for the plant P of Example* 1 *and suppose now that $\Sigma_u = \{b\}$. Suppose the specification is given by $f = EXAGp_1$. Then there is one E in the formula f_P because of the uncontrollable transition from x_1 to x_2 in P, and there is one E in f. So the required branching degree of the tree automata for f and f_P can be chosen to be* $1 + 1 = 2$.

A tree automaton for f_P with the required branching degree of 2 *(i.e., the automaton on binary trees) can be obtained as follows: $\mathcal{A} = (X, 2^{AP \cup X}, \delta_{\mathcal{A}}, x_0, \{(X, X)\})$, where X, AP, and x_0 are the same as in P, $\delta_{\mathcal{A}} : X \times 2^{AP \cup X} \to 2^{X^2}$ is given as $\delta_{\mathcal{A}}(x_0, (p_1, p_2, x_0)) = \{(x_1, x_1)\}$, $\delta_{\mathcal{A}}(x_1, (p_1, p_2, x_1)) = \{(x_2, x_2), (x_2, x_3), (x_3, x_2)\}$, $\delta_{\mathcal{A}}(x_2, (p_1, x_2)) = \{(x_2, x_2)\}$, $\delta_{\mathcal{A}}(x_3, (p_2, x_3)) = \{(x_3, x_3)\}$. Note that the uncontrollable transition from x_1 to x_2 in P is captured in $\mathcal{A}$ by requiring that x_2 be included in every state pair in $\delta_{\mathcal{A}}(x_1, (p_1, p_2, x_1))$. It can be verified that any infinite binary tree that is accepted by $\mathcal{A}$ satisfies the formula f_P.*

Remark 5. The supervisory control problem for language-based specifications is typically of two types: (i) the target control problem (where a supervisor is designed so that the controlled language equals the specification language) and (ii) the range

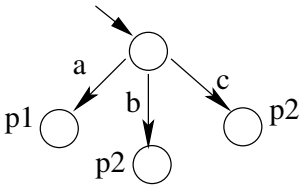


FIG. 2. *A counter example to* [1].

control problem (where a supervisor is designed so that the controlled language is bounded by a lower bound and an upper bound specification languages). Obviously the range control problem is more general since the two bounds can be the same, in which case it is the same as the target control problem. For the range control problem, *any* supervisor is acceptable as long as the controlled language lies in the specified range. If none exists, then one can consider minimal relaxations of the two bounds so that a supervisor will exist.

The situation is even more general for a CTL* specification: a pair of LTL formulae f and g may be chosen to serve as lower and upper bounds for the ω-language of the controlled plant. Then the single LTL formula $\neg f \wedge g$ specifies a range for the controlled ω-language. Of course, more general specifications can be specified in CTL* than just the simple range for ω-language. Similar to the approach taken for the language range control, here we are seeking *any* supervisor that enforces the given CTL* specification. (Algorithm 1 finds one such supervisor.) Now if none exists, then one would like to consider a minimal relaxation of the given CTL* specification for which a supervisor will exist. This topic is not within the scope of the present paper but may be addressed by introducing an order relation over the class of all CTL* formulas defined over a fixed set of atomic propositions using the simulation preorder. We say $f_1 \leq f_2$ if and only if a model M_1 of f_1 is simulated by a model M_2 of f_2. (A simulation relation is a preorder over the set of all models since it is reflexive and transitive but not antisymmetric.) Minimal relaxations of a specification formula can be defined with respect to this order relation.

3.1. Supervisory control for CTL specification. If the specification is given as a CTL formula, we may view it as a CTL* formula and use Algorithm 1 for a supervisor synthesis for CTL control. But this method has a double exponential complexity in the length of the specification formula. From Theorem 3, we know that the control problem for a CTL formula f can be reduced to the decision problem for the formula $f \wedge f_P$. Since f_P by its definition is also a CTL formula, $f \wedge f_P$ is a CTL formula, and so we can get a supervisor synthesis algorithm for the control of the CTL formula f from the decision procedure for the CTL formula $f \wedge f_P$ with a worst-case complexity of single exponential in the length of the CTL specification formula (as opposed to double exponential for the more general case of a CTL* specification). In the appendix, we present a detailed supervisor synthesis algorithm for CTL control.

Remark 6. In [1], the CTL control problem was also studied. But the author restricted the problem by only considering the state-based supervisors and a special class of CTL formulas. Also note that the method in [1] gives wrong results even for some CTL formulas which do belong to the special class of formulas considered in [1]. To see this, consider the example shown in Figure 2, where a, b, c all are controllable

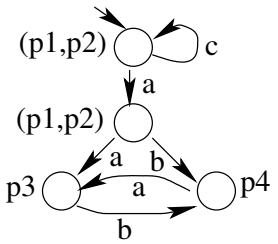


FIG. 3. *An example for the completeness of LTL control.*

events. Then the control action of enabling all a, b, c will let EXp_1 hold at the initial state, and the control action of enabling only b and c will let AXp_2 hold at the initial state. In [1], it was claimed that in order to let $EXp_1 \wedge AXp_2$ hold at the initial state, we may take the conjunction of the control actions for EXp_1 and AXp_2, i.e., enabling b and c would ensure that $EXp_1 \wedge AXp_2$ will hold at the initial state. It is obvious that under this control action, EXp_1 does not hold at the initial state. So the method in [1] gives a wrong result for the above example.

3.2. Supervisory control for LTL specification. Let us next consider the special case of LTL. Recall that LTL is obtained by restricting CTL* in that except for the path quantifier A appearing at the beginning of the formula no other path quantifiers exist in the formula. If the specification f is given as a LTL formula, then we have two different ways to solve the control problem:

1. View the LTL formula as a CTL* formula and directly use Algorithm 1 for the supervisor synthesis of LTL control.
2. First use a tableau construction method such as the one given in [10] to convert the LTL formula into a nondeterministic Buchi automaton; next use the method in [29] to change the nondeterministic Buchi automaton into a deterministic Rabin automaton; next derive a new Rabin automaton from the synchronous composition of the plant automaton and the specification Rabin automaton; and finally use the approach in [33] to solve the control problem on this final Rabin automaton.

These two methods have a same worst-case complexity which is polynomial in the size of the plant and double exponential in the length of the specification LTL formula.

We next propose a supervisor synthesis algorithm for the control of LTL which has a smaller complexity (single exponential in the length of the LTL formula as opposed to double exponential) but it is only sound (and not complete). We first change the LTL formula into a CTL formula by inserting the path quantifier A before every temporal operator in the formula and removing any repeated A; then we apply Algorithm 2 (given in the appendix) for the supervisor synthesis for this CTL formula. From the semantics of CTL and LTL, we know that the supervisor derived does work for the original LTL formula. The worst-case complexity of this method is the same as that for Algorithm 2 which is polynomial in the size of the plant and single exponential in the length of the specification LTL formula.

This method, however, is not complete, i.e., when it answers "no" for the existence of a supervisor, there may still exist a supervisor that can enforce the given LTL specification. Consider, for example, the system shown in Figure 3, for which the specification is given as $A[(p_1 U p_3) \vee (p_2 U p_4)]$. Assuming that the event c is the only controllable event, it is obvious that the specification can be enforced if the supervisor

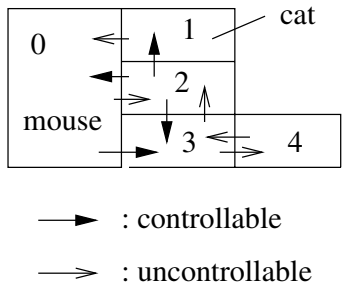


FIG. 4. *Mouse in a maze.*

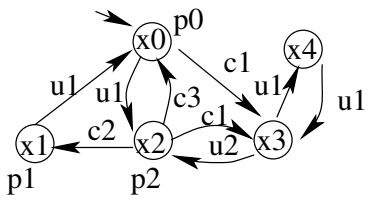


FIG. 5. *Plant model.*

disables c at the initial state. But if we transfer the specification into a CTL formula $A(p_1 U p_3) \vee A(p_2 U p_4)$ using the method described above, then it is easy to verify that no supervisor exists.

Remark 7. The algorithm given in [2] for the control of MTL (LTL together with real-time constraints) is sound but not complete, which was not clarified there. Since an LTL formula is also an MTL formula, we can apply the algorithm given in [2] to the example of Figure 3. The algorithm in [2] will answer "no" for the existence of a supervisor for the above example. But we know that a supervisor does exist, thereby demonstrating the incompleteness of the algorithm given in [2].

4. Illustrative example. In this section, we give a simple example to illustrate our result. This is a traffic control problem of a mouse in a maze. The maze, shown in Figure 4, consists of five rooms connected by various one-way passages, where some of them can be closed through control. There is also a cat which alway stays in room 1. The mouse is initially in room 0, but it can visit other rooms by using one-way passages. Our task is to design a supervisor to control the passages in order to guarantee that

Spec 1 The mouse never visits room 1 where the cat stays (this is a safety property).

Spec 2 The mouse can go to room 0 for play at any time it wants to (this is a nonblocking property).

Spec 3 The mouse shall visit room 2 for food infinitely often (this is a liveness property).

Spec 4 The mouse shall never be locked in a room (this is a nonterminating property).

The above problem can be formulated as a supervisory control problem of a discrete event system with a CTL specification as follows. The system is modeled as a plant $P = (X, \Sigma, \delta_P, x_0, AP, L_P)$, which is shown in Figure 5, where $X = \{x_i, i = 0, 1, 2, 3, 4\}$; $\Sigma = \{c_1, c_2, c_3, u_1, u_2\}$, $\Sigma_c = \{c_1, c_2, c_3\}$; $AP = \{p_0, p_1, p_2\}$; $L_P(x_0) =$

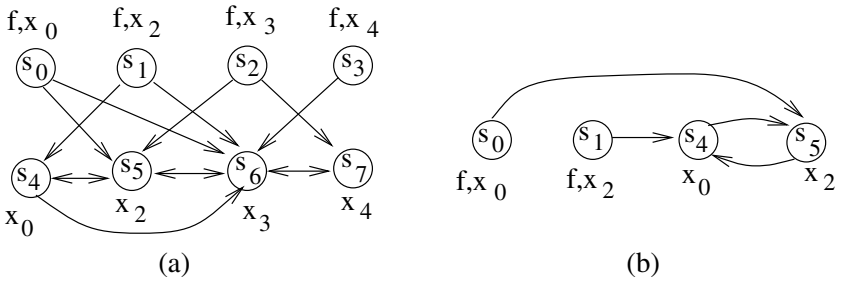


FIG. 6. *Tableau and model for $f \wedge f_P$.*

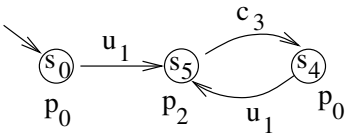


FIG. 7. *Supervisor for the cat-mouse example.*

$\{p_0, \neg p_1, \neg p_2\}$, $L_P(x_1) = \{\neg p_0, p_1, \neg p_2\}$, $L_P(x_2) = \{\neg p_0, \neg p_1, p_2\}$, $L_P(x_3) = L_P(x_4) = \{\neg p_0, \neg p_1, \neg p_2\}$. The specification is given by the CTL formula $f = AG\neg p_1 \wedge AGEFp_0 \wedge AGAFp_2 \wedge AGEXtrue$, where the ith conjunct corresponds to the Spec i.

Now we can use Algorithm 2 for the supervisor synthesis of the above control problem. We first obtain the tableau T for the formula $f \wedge f_P$, where f_P is the CTL formula encoding the plant P (for brevity f_P is omitted here). The tableau $T = (S_T, AP', R_T, L_T)$ is shown in Figure 6(a), where for $0 \leq i \leq 3$, $L_T(s_i) = \{f\} \cup L_T(s_{i+4})$, and for $i > 3$,

$$
\begin{aligned}
L_T(s_4) = \{&p_0, \neg p_1, \neg p_2, AG\neg p_1, AGEFp_0, AGAFp_2, AGEXtrue, AXAG\neg p_1, EFp_0,\\
&AXAGEFp_0, AFp_2, AXAGAFp_2, EXtrue, AXAGEXtrue, AXAFp_2,\\
&x_0, EXx_2\};\\
L_T(s_5) = \{&\neg p_0, \neg p_1, p_2, AG\neg p_1, AGEFp_0, AGAFp_2, AGEXtrue, AXAG\neg p_1, EFp_0,\\
&AXAGEFp_0, AFp_2, AXAGAFp_2, EXtrue, AXAGEXtrue, EXEFp_0, x_2\};\\
L_T(s_6) = \{&\neg p_0, \neg p_1, \neg p_2, AG\neg p_1, AGEFp_0, AGAFp_2, AGEXtrue, AXAG\neg p_1, EFp_0,\\
&AXAGEFp_0, AFp_2, AXAGAFp_2, EXtrue, AXAGEXtrue, EXEFp_0,\\
&AXAFp_2, x_3, EXx_2, EXx_4\};\\
L_T(s_7) = \{&\neg p_0, \neg p_1, \neg p_2, AG\neg p_1, AGEFp_0, AGAFp_2, AGEXtrue, AXAG\neg p_1, EFp_0,\\
&AXAGEFp_0, AFp_2, AXAGAFp_2, EXtrue, AXAGEXtrue, EXEFp_0,\\
&AXAFp_2, x_4, EXx_3\}.
\end{aligned}
$$

Next a model $M = (Q, R, AP', L)$ for $f \wedge f_P$ is derived and this is shown in Figure 6(b), where $Q = \{s_0, s_1, s_4, s_5\} \subset S_T$, $R = R_T|_Q$ and $L = L_T|_Q$, the restriction of R_T and L_T, respectively, to Q.

Finally a supervisor S is obtained from M and is shown in Figure 7. It follows

that the mouse moves between rooms 0 and 2 only, and hence obviously the controlled system $P||S$ satisfies the given specification.

5. Conclusion. We studied the supervisory control problem for systems with temporal logic specifications. The full branching time logic of CTL* is used for expressing the control specifications. The main contributions of the paper are summarized as follows:

1. CTL* temporal logic allows the control constraints on the sequences of states which can be also captured by a regular $*$-language or ω-language, as well as on the more general branching structures of states which cannot be captured by a regular $*$-language or ω-language as shown in Example 1.
2. For the first time a sound and complete supervisory synthesis algorithm for CTL* specifications has been obtained. (Supervisors are allowed to be nondeterministic as this allows for the existence of a supervisor for a larger class of CTL* specifications.)
3. By reducing the control problem to the decision problem, a small model theorem for the CTL* control is derived.
4. The computational complexity of the control algorithms have been derived: the control problem for CTL* (resp., CTL) is complete for deterministic double (resp., single) exponential time in the length of the specification formula. Further, it is polynomial in the number of plant states.
5. Usage of temporal logic specifications does not increase the computational complexity of supervisor synthesis (compared to that of formal language/ automata-based specifications).

The last point above requires further clarification. In some cases, a property may be expressed by either a CTL* formula or by a $*$-language or a ω-language. So for these cases we can compare our method with that based on finite state automaton. If we use a finite state automaton accepting a $*$-language to give the specification, then the supervisor synthesis is polynomial in the product of the number of plant states and the number of the states of the specification automaton. From the known tableau construction methods, we know that the number of states in an automaton model of a temporal logic formula is exponential in the length of the formula (whenever the formula can be represented by an automaton). So if we start with a temporal logic specification (that can be also expressed as an automaton) and convert it to an automaton, and apply the existing supervisory control theory results, then the resulting computational complexity will be polynomial in the number of plant states and single exponential in the length of the temporal logic specification formula. This matches the complexity of our algorithm, and so there is no loss of computational complexity from the approach developed above, yet there is a gain in expressibility since a temporal logic formula is more compact. The use of temporal logic shifts the burden from the user (who gives the specification) to the supervisor designer (who computes the supervisor)—computation of supervisor for a temporal logic specification although more involved, has the same complexity.

A. Supervisor synthesis for CTL specification. We assume that the given CTL formula f is in *positive normal form*. We use $\sim f_1$ to denote the formula in positive normal form equivalent to $\neg f_1$. We begin with a few definitions taken from [6]. The *closure* of f, $\mathrm{cl}(f)$, is the smallest set of formulas containing f and satisfying

- each subformula of f that is a state formula is in $\mathrm{cl}(f)$;
- if EXf_1, EGf_1, $E[f_1Uf_2]$, or $E[f_1Bf_2]$ is in $\mathrm{cl}(f)$, then, respectively, $EXEXf_1$, $EXEGf_1$, $EXE[f_1Uf_2]$, or $EXE[f_1Bf_2]$ is in $\mathrm{cl}(f)$;

- if AFf_1, AGf_1, $A[f_1Uf_2]$, or $A[f_1Bf_2]$ is in $\mathrm{cl}(f)$, then, respectively, $AXAFf_1$, $AXAGf_1$, $AXA[f_1Uf_2]$, or $AXA[f_1Bf_2]$ is in $\mathrm{cl}(f)$.

The *extended closure* of f is defined as $\mathrm{ecl}(f) = \mathrm{cl}(f) \cup \{\sim f_1 | f_1 \in \mathrm{cl}(f)\}$. Note that $|\mathrm{ecl}(f)| = O(|f|)$, where $|f|$ denotes the length of f.

We say that a formula is *elementary* provided that it is a proposition, is the negation of a proposition, or is in the form of AXf_1 or EXf_1. Any other formula is *nonelementary*. Each nonelementary formula may be viewed as either a conjunctive α-formula, $\alpha = \alpha_1 \wedge \alpha_2$, or a disjunctive β-formula, $\beta = \beta_1 \vee \beta_2$. Clearly, $f_1 \wedge f_2$ is an α formula and $f_1 \vee f_2$ is a β formula. A formula such as AGf_1, $A[f_1Uf_2]$, $A[f_1Bf_2]$, etc., may be classified as an α or β formula based on its fix-point characterization; e.g., $AGf_1 = f_1 \wedge AXAGf_1$ is an α formula and $EFf_1 = f_1 \vee EXEFf_1$ is a β formula. The classification for all nonelementary formulas is given as

$\alpha - formula$ $\alpha = \alpha_1 \wedge \alpha_2$,

$$\begin{array}{lll}
\alpha = f_1 \wedge f_2, & \alpha_1 = f_1, & \alpha_2 = f_2, \\
\alpha = A[f_1Bf_2], & \alpha_1 = \sim f_2, & \alpha_2 = f_1 \vee AXA[f_1Bf_2], \\
\alpha = E[f_1Bf_2], & \alpha_1 = \sim f_2, & \alpha_2 = f_1 \vee EXE[f_1Bf_2], \\
\alpha = AGf_1, & \alpha_1 = f_1, & \alpha_2 = AXAGf_1, \\
\alpha = EGf_1, & \alpha_1 = f_1, & \alpha_2 = EXEGf_1;
\end{array}$$

$\beta - formula$ $\beta = \beta_1 \vee \beta_2$,

$$\begin{array}{llll}
\beta = f_1 \vee f_2, & \beta_1 = f_1 & \beta_2 = f_2, \\
\beta = A[f_1Uf_2], & \beta_1 = f_2, & \beta_2 = f_1 \wedge AXA[f_1Uf_2], \\
\beta = E[f_1Uf_2], & \beta_1 = f_2, & \beta_2 = f_1 \wedge EXE[f_1Uf_2], \\
\beta = AFf_1, & \beta_1 = f_1, & \beta_2 = AXAFf_1, \\
\beta = EFf_1, & \beta_1 = f_1, & \beta_2 = EXEFf_1.
\end{array}$$

A state transition graph $M = (Q, R, AP, L)$ is called a *structure* if the relation R is required to be total; otherwise M is called a *prestructure*. An *interior* node of a prestructure is one with at least one successor. A *frontier* node is one with no successors. A prestructure $M_1 = (Q_1, R_1, AP, L_1)$ is said to be *contained* in a structure $M_2 = (Q_2, R_2, AP, L_2)$ whenever $Q_1 \subseteq Q_2$, $R_1 \subseteq R_2$, and $L_1 = L_2|Q_1$, the restriction of L_2 to Q_1; M_1 is said to be *cleanly embedded* in M_2 provided M_1 is contained in M_2, and also every interior node of M_1 has the same set of successors as its corresponding node in M_2.

The following consistency requirements are associated with the labeling function L of a (pre)structure. Since we consider the control of CTL, the definition of L is extended as $L : Q \to 2^{ecl(f)}$, where f is the specification formula. $\forall q \in Q$, we have zero-step consistency rules,

ZS0 $p \in L(q) \Rightarrow \sim p \notin L(q)$;
ZS1 $\alpha \in L(q) \Rightarrow [(\alpha_1 \in L(q)) \wedge (\alpha_2 \in L(q))]$;
ZS2 $\beta \in L(q) \Rightarrow [(\beta_1 \in L(q)) \vee (\beta_2 \in L(q))]$;

one-step consistency rules,

OS0 $AXp \in L(q) \Rightarrow [\forall q' \in Q,\ ((q, q') \notin R) \vee (p \in L(q'))]$;
OS1 $EXp \in L(q) \Rightarrow [\exists q' \in Q,\ ((q, q') \in R) \wedge (p \in L(q))]$.

A *fragment* is a prestructure whose graph is a directed acyclic graph (DAG) such that all its nodes satisfy rules ZS0–ZS2 and OS0 and all its interior nodes satisfy rule OS1.

A formula of the form $A[pUp']$ or $E[pUp']$ is called an *eventuality* formula. Since AFp' and EFp' are special cases of $A[pUp']$ and $E[pUp']$, respectively, they are also eventuality formulas.

An eventuality formula (AFp', $A[pUp']$, EFp', or $E[pUp']$) is said to be *fulfilled* in a structure $M = (Q, AP, R, L)$ if $\forall q \in Q$:

- $AFp' \in L(q)$ (resp., $A[pUp'] \in L(q)$) implies that there is a finite fragment, called DAG$[q, AFp']$ (resp., DAG$[q, A[pUp']]$), rooted at q and *cleanly embedded* in M such that *for all frontier nodes* t of the fragment, $p' \in L(t)$, and for all interior nodes u of the fragment, *true* (resp., p) $\in L(u)$;
- $EFp' \in L(q)$ (resp., $E[pUp'] \in L(q)$) implies that there is a finite fragment, called DAG$[q, EFp']$ (resp., DAG$[q, E[pUp']]$), rooted at q and *cleanly embedded* in M such that for *some frontier node* t of the fragment, $p' \in L(t)$, and there exists one path from q to t in the fragment such that for all interior nodes u along the path, *true* (resp., p) $\in L(u)$.

An eventuality formula (AFp', $A[pUp']$, EFp', or $E[pUp']$) is said to be *pseudo-fulfilled* in a structure $M = (Q, AP, R, L)$ if $\forall q \in Q$,

- $AFp' \in L(q)$ (resp., $A[pUp'] \in L(q)$) implies that there is a finite fragment, called DAG$[q, AFp']$ (resp., DAG$[q, A[pUp']]$), rooted at q and *contained* in M such that *for all frontier nodes* t of the fragment, $p' \in L(t)$, and for all interior nodes u of the fragment, *true* (resp., p) $\in L(u)$;
- $EFp' \in L(q)$ (resp., $E[pUp'] \in L(q)$) implies that there is a finite fragment, called DAG$[q, EFp']$ (resp., DAG$[q, E[pUp']]$), rooted at q and *contained* in M such that for *some frontier node* t of the fragment, $p' \in L(t)$, and there exists one path from q to t in the fragment such that for all interior nodes u along the path, *true* (resp., p) $\in L(u)$.

Now we present a supervisor synthesis algorithm for the control of CTL, which is based on the decision procedure for CTL [6]. The algorithm differs from the decision procedure as follows:

- A modular method is used for the tableau construction. It ensures that the worst-case complexity of the algorithm is polynomial in the size of the plant.
- A supervisor, not a model, is finally synthesized.

Let f be the given CTL specification formula, f_P be the CTL formula encoding the given deterministic nonterminating plant P, and $AP' = AP \cup X$ be the extended atomic proposition set. Since we require the controlled plant to be nonterminating, we can assume that f is in the form of $f = f' \wedge AGEXtrue$. Then the algorithm is given as follows.

ALGORITHM 2. SUPERVISOR SYNTHESIS ALGORITHM FOR CTL SPECIFICATION.

1. Test the satisfiability of the CTL formula $f \wedge f_P$. This step is done by using the decision procedure for CTL as follows:
   (a) Construct a tableau T for the CTL formula $f \wedge f_P$. We use a modular method to obtain the tableau T as follows:
      i. Construct a tableau T_f for the CTL specification formula f. T_f is constructed from a bipartite graph $T_0 = (C \cup D, R_{CD} \cup R_{DC}, AP, L_0)$, where nodes in C are called states, nodes in D are called prestates, and each node is uniquely identified by its label defined by L_0; $R_{CD} \subseteq C \times D$ and $R_{DC} \subseteq D \times C$ are transition relations; $L_0 : C \cup D \rightarrow ecl(f)$ is the labeling function. Initially, C, R_{CD}, and R_{DC} are all empty, and D contains a single prestate d labeled with f. Repeat the following until no more nodes and transitions can be added into T_0: let e be a frontier node of T_0,
         - if $e \in D$, then let $\{L_i \subseteq ecl(f) \mid 1 \leq i \leq k\}$ be the set of all possible labels such that $\forall i \in \{1, 2, \ldots, k\}$, "$[L_i$ is a minimal

superset of $L_0(e)$] $\wedge$ [L_i satisfies rules ZS0-ZS2] $\wedge$ [$\forall p \in AP$, $(p \in L_i) \vee (\neg p \in L_i)$]," and for each L_i create a state c_i with $L_0(c_i) = L_i$, and add c_i into C if $c_i \notin C$, and (e, c_i) into R_{DC};

- if $e \in C$ labeled with the next time formulas

$$\{AXp_1, \ldots, AXp_j, EXp'_1, \ldots, EXp'_k\},$$

then $\forall i \in \{1, \ldots, k\}$, create prestates d_i labeled with $\{p_1, \ldots, p_j, p'_i\}$, and add d_i into D if $d_i \notin D$, and (e, d_i) into R_{CD}.

The tableau T_f is obtained as $T_f = (C_f, R_f, AP, L_f)$, where $C_f = C$, $R_f = R_{CD} \circ R_{DC}$, and $L_f = L_0|_C$, the restriction of L_0 to C.

ii. Derive the tableau T for $f \wedge f_P$ from the synchronous composition of the plant $P = (X, \Sigma, \delta_P, x_0, AP, L_P)$ and the tableau $T_f = (C_f, R_f, AP, L_f)$ as follows: $T = (S_T, R_T, AP', L_T)$, where

- $S_T \subseteq C_f \times X$ is the state set, $S_T = \{(t, x) \in C_f \times X |\ L_f(t)$ and $L_P(x)$ are propositionally consistent$\}$, where "$L_f(t)$ and $L_P(x)$ are propositionally consistent" means that $\forall p \in AP$, $[p \in (L_f(t) \cup L_P(x)) \Rightarrow \neg p \notin (L_f(t) \cup L_P(x))]$;
- $AP' = AP \cup X$;
- $R_T \subseteq S_T \times S_T$ is the transition relation, $R_T = \{((t, x), (t', x')) \in S_T \times S_T |\ (t, t') \in R_f,$ and $\exists \sigma \in \Sigma\ s.t.\ \delta_P(x, \sigma) = x'\}$;
- L_T is the labeling function defined as $\forall (t, x) \in S_T \times S_T$, $L_T((t, x)) = L_f(t) \cup L_P(x) \cup \{x\} \cup \{EXy |\ y \in X, \exists \sigma_u \in \Sigma_u, y = \delta_P(x, \sigma_u)\}$.

(b) Test the tableau T for the existence of a model for $f \wedge f_P$. This is done by first pruning (see below) the tableau T to ensure that the consistency and pseudofulfillment of eventualities are satisfied in T, then checking in the pruned tableau T whether there exists a state s_0 such that $\{f, x_0\} \subseteq L_T(s_0)$. If there exists such a state, then and only then $f \wedge f_P$ is satisfiable. If $f \wedge f_P$ is satisfiable, then go to next step; otherwise stop the algorithm and output that "no supervisor exists."

The pruning of T is achieved by repeatedly applying the following deletion rules until no more nodes can be deleted from T:

- Delete any state which has no successors.
- Delete any state which violates rule OS1.
- Delete any state s such that $\exists r \in L_T(s)$, r is an eventuality formula, and r is not pseudofulfilled at s.

To test the pseudofulfillment of an eventuality formula at each state in T, the following ranking procedure can be used. For an $A[pUq]$ eventuality, initially assign rank 1 to all nodes labeled with q and rank ∞ to all other nodes. Then for each node s and each formula r such that $EXr \in L_T(s)$, define $SUCC_r(s) = \{s' \mid (s, s') \in R, r \in L_T(s')\}$ and compute $\text{rank}(SUCC_r(s)) = \min_{s'}\{\text{rank}(s') \mid s' \in SUCC_r(s)\}$. Now for each node s of rank ∞ such that $p \in L_T(s)$, let $\text{rank}(s) = 1 + \max_r\{\text{rank}(SUCC_r(s)) |\ EXr \in L_T(s)\}$. Since $AGEXtrue$ is contained in f, the formula $EXtrue$ is labeled at every node in T. So the above procedure is well defined. Repeatedly apply the above ranking procedure until stabilization. A node s has a finite rank if and only if $A[pUq]$ is pseudofulfilled at s in T. Testing for the pseudofulfillment of AFq follows from above since it is a special case of

$A[pUq]$. For testing the pseudofulfillment of $E[pUq]$, a similar procedure as above can be applied, with the only modification that $\text{rank}(s) = 1 + min_r\{\text{rank}(SUCC_r(s)) \mid EXr \in L_T(s)\}$. Testing for the pseudofulfillment of EFq is again a special case of $E[pUq]$.

2. Extract a model M for the formula $f \wedge f_P$ from the tableau T. $M = (Q, R, AP', L)$ is extracted from $T = (S_T, R_T, AP', L_T)$ as follows [6]. For each state s in S_T and each eventuality q in $ecl(f)$, we construct a directed acyclic graph rooted at s, DAGG$[s,\ q]$. If the eventuality $q \in L_T(s)$, then DAGG$[s,\ q]$ = DAG$[s,\ q]$; otherwise DAGG$[s, q]$ is taken to be the subgraph consisting of s and a sufficient set of successors to ensure that one-step consistency rules OS0-1 are satisfied. Next we take each DAGG$[s, q]$ and arrange them in a matrix by putting DAGG$[s_j, q_i]$ in the ith row and the jth column of the matrix. The matrix has a dimension of $m \times n$, where m (resp., n) is the number of eventualities (resp., states) in the tableau T. Then we connect all the DAGGs in the matrix together in the following way: for any frontier node s in the ith row, merge it with the corresponding root node s of DAGG$[s, q_{i+1}]$ in the $(i+1)$th row; for any frontier node s in the last row, merge it with the corresponding root node s of DAGG$[s, q_1]$ in the first row. We use $M = (Q, R, AP', L)$ to represent the above finite state transition graph, where Q is the set of states in the graph, R is the transition relation of the graph, and L is the labeling function for each state in the graph which is a natural extension of L_T. M defines a model for $f \wedge f_P$, i.e., $\exists q_0 \in Q$ such that $< M, q_0 > \models f \wedge f_P$.
3. Derive a supervisor S from the model M of $f \wedge f_P$. Since $M = (Q, R, AP', L)$ is a model of $f \wedge f_P$, we know that $\exists q_0 \in Q,\ < M, q_0 > \models f \wedge f_P$. We can get a control-compatible and propositionally consistent supervisor $S = (Y, \Sigma, \delta_S, y_0, L_S)$ from M using the same method as given in the proof of Theorem 3 as follows: $Y \subseteq Q$ is the set of states which are reachable from q_0 in M; $\forall y \in Y,\ \forall \sigma \in \Sigma$,

$$\delta_S(y, \sigma) = \{y' \mid [(y, y') \in R] \wedge [x' = \delta_P(x, \sigma)], \\ \text{where } \{x'\} = L(y') \cap X \text{ and } \{x\} = L(y) \cap X\};$$

$y_0 = q_0$; and $\forall y \in Y, L_S(y) = L(y) \cap (AP \cup \{\neg p | p \in AP\})$.

Remark 8. *From Theorem* 3*, and using an argument similar to the soundness and completeness of the decision procedure for CTL* [6]*, we can conclude that Algorithm* 2 *for control synthesis for CTL is sound and complete. It is easy to check that Algorithm* 2 *has a worst-case complexity of single exponential in the length of the specification CTL formula* f *and polynomial in the number of states of the plant* P*. It matches the lower bound complexity of the CTL control problem given in Theorem* 5.

REFERENCES

[1] M. Antoniotti, *Synthesis and Verification of Discrete Controllers for Robotics and Manufacturing Devices with Temporal Logic and Control-D Systems*, Ph.D. thesis, Department of Computer Science, New York University, New York, 1995.

[2] M. Barbeau, F. Kabaza, and R. St.-Denis, *A method for the synthesis of controllers to handel safety, liveness, and real-time constraints*, IEEE Trans. Automat. Control, 43 (1998), pp. 1543–1559.

[3] E. M. Clarke, O. Grumberg, and D. A. Peled, *Model Checking*, MIT Press, Cambridge, MA, 1999.

[4] A. R. Deshpande and P. Varaiya, *Semantic tableau for control of pltl formulae*, in Proceedings of 35th IEEE conference on Decision and Control, Kobe, Japan, 1996.
[5] E. A. Emerson, *Automata, tableaux, and temporal logic*, in Proceedings of Conference on Logics of Programs, Lecture Notes in Comput. Sci. 193, R. Parikh, ed., Springer-Verlag, Berlin, 1985, pp. 79–88.
[6] E. A. Emerson, *Temporal and modal logic*, in Handbook of Theoretical Computer Science, J. van Leeuwen, ed., Elsevier Science Publishers, New York, 1990.
[7] E. A. Emerson and Y. J. Halpern, *"Sometimes" and "not never" revisited: On branching versus linear time temporal logic*, J. ACM, 33 (1986), pp. 151–178.
[8] E. A. Emerson and C. S. Jutla, *The complexity of tree automata and logics of programs*, in Proceedings of 29th Annual IEEE-CS Symposium on Foundations of Computer Science, pp. 328–337, 1988.
[9] E. A. Emerson and A. P. Sistla, *Deciding full branching time logic*, Inform. Control, 61 (1984), pp. 175–201.
[10] R. Gerth, D. Peled, M. Vardi, and P. Wolper, *Simple on-the-fly automatic verification of linear temporal logic*, in Protocol Specification Testing and Verification, Chapman and Hall, London, 1995, pp. 3–18.
[11] G. E. Hughes and M. J. Creswell, *Introduction to Modal Logic*, Methuen, London, 1977.
[12] M. R. Huth and M. D. Ryan, *Logic in Computer Science: Modeling and Reasoning about Systems*, Cambridge University Press, Cambridge, UK, 2000.
[13] K. Inan, *Nondeterministic supervision under partial observations*, in Lecture Notes in Control and Inform. Sci. 199, G. Cohen and J.-P. Quadrat, eds., Springer-Verlag, New York, 1994, pp. 39–48.
[14] J. F. Knight and K. M. Passino, *Decidability for a temporal logic used in discrete-event system analysis*, Internat. J. Control, 52 (1990), pp. 1489–1506.
[15] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control*, Prentice–Hall, Englewood Cliffs, NJ, 1986.
[16] R. Kumar and V. K. Garg, *Modeling and Control of Logical Discrete Event Systems*, Kluwer Academic Publishers, Boston, 1995.
[17] R. Kumar, S. Jiang, C. Zhou, and W. Qiu, *Polynomial synthesis of supervisor for partially observed discrete event systems by allowing nondeterminism in control*, IEEE Trans. Automat. Control, 50 (2005), pp. 463–475.
[18] O. Kupferman, P. Madhusudan, P. S. Thiagarajan, and M. Y. Vardi, *Open systems and reactive environments: Control and synthesis*, in Proceedings of 11th Conference on Concurrency Theory, Lecture Notes in Comput. Sci. 1877, Springer-Verlag, New York, 2000, pp. 92–107.
[19] O. Kupferman and M. Y. Vardi, *Robust satisfaction*, in Proceedings of 10th Conference on Concurrency Theory, Lecture Notes in Comput. Sci. 1664, Springer-Verlag, New York, 1999, pp. 382–398.
[20] O. Kupferman, M. Y. Vardi, and P. Wolper, *Module checking*, Inform. Comput., 164 (2001), pp. 322–344.
[21] F. Lin, *Analysis and synthesis of discrete event systems using temporal logic*, Control Theory Adv. Tech., 9 (1993), pp. 341–350.
[22] J.-Y. Lin and D. Ionescu, *Verifying a class of nondeterministic discrete event systems in a generalized temporal logic*, IEEE Trans. Systems Man Cybernet., 22 (1992), pp. 1461–1469.
[23] J.-Y. Lin and D. Ionescu, *Reachability synthesis procedure for discrete event systems in a temporal logic*, IEEE Trans. Systems Man Cybernet., 24 (1994), pp. 1397–1406.
[24] J .S. Ostroff, *Synthesis of controllers for real-time discrete event systems*, in Proceedings of 28th IEEE Conference on Decision and Control, Tampa, FL, 1989.
[25] J. S. Ostroff and W. M. Wonham, *A framework for real-time discrete event control*, IEEE Trans. Automat. Control, 35 (1990), pp. 386–397.
[26] K. M. Passino and P. J. Antsaklis, *Branching time temporal logic for discrete event system analysis*, in Proceedings of 1988 Allerton Conference on Communication, Control, and Computing, University of Illinois, Allerton, IL, 1988, pp. 1160–1169.
[27] A. Pnueli, *The temporal logic of programs*, in Proceedings of 18th Annual Symposium on Foundations of Computer Science, Providence, RI, Nov. 1977, pp. 46–57.
[28] P. J. Ramadge and W. M. Wonham, *Supervisory control of a class of discrete event processes*, SIAM J. Control Optim., 25 (1987), pp. 206–230.
[29] S. Safra, *On the complexity of ω-automata*, in Proceedings of 1988 Annual Symposium on the Foundations of Computer Science, White Plains, NY, 1988, pp. 319–327.

[30] K. T. Seow and R. Devanathan, *Temporal framework for assembly sequence representation and analysis*, IEEE Trans. Robotics Automation, 10 (2), pp. 220–229, 1994.

[31] K. T. Seow and R. Devanathan, *A temporal logic approach to discrete event control for the safety cannonical class*, Systems Control Lett., 28 (1996), pp. 205–217.

[32] J. G. Thistle and W. M. Wonham, *Control problems in temporal logic framework*, Internat. J. Control, 44 (1986), pp. 943–976.

[33] J. G. Thistle and W. M. Wonham, *Control of infinite behavior of finite automata*, SIAM J. Control Optim., 32 (1994), pp. 1075–1097.

[34] H. Wong-Toi and D. L. Dill, *Synthesizing processes and schedulers from temporal specifications*, in Proceedings of the 1990 Computer-Aided Verification Workshop, Lecture Notes in Comput. Sci. 531, Springer-Verlag, New York, 1990, pp. 272–281.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 6, pp. 2104–2122

EXISTENCE OF OPTIMAL POLICIES FOR SEMI-MARKOV DECISION PROCESSES USING DUALITY FOR INFINITE LINEAR PROGRAMMING*

DIEGO KLABJAN† AND DANIEL ADELMAN‡

Abstract. Semi-Markov decision processes on Borel spaces with deterministic kernels have many practical applications, particularly in inventory theory. Most of the results from general semi-Markov decision processes do not carry over to a deterministic kernel since such a kernel does not provide "smoothness." We develop infinite dimensional linear programming theory for a general stochastic semi-Markov decision process. We give conditions, general enough to allow deterministic kernels, for solvability and strong duality of the resulting linear programs. By using the developed linear programming theory we give conditions for the existence of a stationary deterministic policy for deterministic kernels, which is optimal among all possible policies.

Key words. semi-Markov decision processes, linear programming, optimal policies

AMS subject classifications. 90C40, 90C45, 90C05

DOI. 10.1137/S0363012903437290

1. Introduction. A semi-Markov decision process (SMDP) on Borel state and action spaces is said to have a Dirac's transition law if the state at the next decision epoch is uniquely determined by a given function evaluated at the current state-action pair. Such models are simple to state but turn out to be even more difficult to study and analyze than their true stochastic counterparts. They have many practical applications, for example, in inventory routing (Adelman (2003)). In a companion paper, Adelman and Klabjan (2005) provide a new SMDP formulation for a widely studied, classical inventory control problem. This problem generalizes the classical economic order quantity problem to a multi-item setting. Most existing SMDP theory does not apply to Dirac's transition laws.

Nearly all approaches to the question of whether there exists an optimal policy require the transition law to be strongly continuous, but a Dirac's transition law is at best only weakly continuous. Strong continuity ensures that "smoothness" is maintained in the optimality equations. For example, existing approaches are based on either the vanishing discount rate methodology (Hernández-Lerma and Lasserre (1990), Vega-Amaya (1993)), or policy iteration (Luque-Vásquez and Hernández-Lerma (1999), Hernández-Lerma and Lasserre (1997)); see also the series of monographs by Hernández-Lerma (1989) and Hernández-Lerma and Lasserre (1996b, 1999). An alternative approach presented by Bhattacharya and Majumdar (1989) is to allow weak continuity of the kernel but to impose equicontinuity of the discounted value functions. Unfortunately, Dirac's transition laws do not provide equicontinuity.

A recent approach in the literature that assumes weak continuity of the transition law is infinite linear programming, developed for the discrete-time case, i.e., Markov

*Received by the editors November 6, 2003; accepted for publication (in revised form) August 4, 2005; published electronically January 6, 2006.
http://www.siam.org/journals/sicon/44-6/43729.html
†Department of Mechanical and Industrial Engineering, University of Illinois at Urbana-Champaign, Urbana, IL (klabjan@uiuc.edu).
‡Graduate School of Business, University of Chicago, Chicago, IL (dan.adelman@ChicagoGSB.edu).

decision processes, by Hernández-Lerma and González-Hernández (1998), Hernández-Lerma and Lasserre (1996a), and Hernández-Lerma and Lasserre (1994). However, existing theory also requires that the expected transition times between decision epochs be lower bounded away from zero. The only exception that we are aware of is the work by Vega-Amaya (2003) in the context of zero-sum semi-Markov games, where the author assumes the transition time to be positive and not necessarily bounded away from zero. This condition is trivially satisfied in the case of discrete time periods. When satisfied in the semi-Markov case it is well known that there exists a transformation of the problem into discrete-time, employed, for instance, by Bhattacharya and Majumdar (1989), Vega-Amaya (1993), and Luque-Vásquez and Hernández-Lerma (1999) although not using linear programming. Unfortunately, for the inventory control applications we have in mind, this condition is violated. It is possible to have multiple decision epochs at the same instant of time.

In this paper, we relax both of the above assumptions. We assume instead that the transition law is weakly continuous, and that the expected transition time *plus* current cost, rather than just the former, is lower bounded away from zero. Therefore, all our results apply to the more restrictive settings in the references above. Instead of seeking transformations to a discrete-time Markov control setting, we work directly with a new infinite linear programming formulation of the SMDP, presented in section 3.1, which extends the formulation of Fox (1966) in finite spaces to Borel spaces and the infinite linear programming formulation of discrete time MDP by Hernández-Lerma and Lasserre (1994). For a general stochastic SMDP, we establish a set of conditions under which this infinite linear program possesses strong duality, i.e., there is no duality gap and primal-dual optimal solutions are attained. Although the infinite linear programming approach to the Borel setting leads to the existence of an optimal policy that is stationary randomized, to date this approach has not been fruitful in showing the existence of an optimal policy that is stationary deterministic. We provide this result when the transition law is Dirac's under a strong recurrence condition. In a companion paper (Adelman and Klabjan (2005)), we show that all the conditions in this paper are verifiable in an inventory application.

We owe a debt of gratitude to Hernández-Lerma and González-Hernández (1998), Hernández-Lerma and Lasserre (1996a), and Hernández-Lerma and Lasserre (1994) for their key insight that infinite linear programming can handle weakly continuous transition laws. This was indeed fortuitous, as what originally prompted our interest in it was one of the authors' use of it in an approximate dynamic programming framework to generate near optimal control policies in inventory routing; see Adelman (2003). In future work, our duality results will prove useful in devising stronger, and possibly even convergent, approximate dynamic programming methodologies.

In section 2 we formulate a general semi-Markov decision process. In section 3 we formulate our primal-dual pair of infinite linear programs and give conditions for strong duality. In section 4, we provide results specialized to the case of a Dirac's transition law.

2. Semi-Markov control model. The semi-Markov control model is defined by $(X, A, \{A(x) : x \in X\}, Q', c')$, where X is the state space and A is the control set. We assume that both X and A are Borel spaces. For each $x \in X$ we are given a nonempty Borel subset $A(x) \subseteq A$, which specifies the set of *admissible controls*, if the state of the system is x. We assume that $K = \{(x, a) : x \in X, a \in A(x)\}$ is a Borel subset of $X \times A$. Let Q' represent the *time-dependant transition law*. If the system is in state $x \in X$ and control action $a \in A(x)$ is taken, then the system's next

state is in B after transition time $t \in T = [0, \infty)$ with probability $Q'(t, B|x, a)$, where $B \subseteq X$ is a Borel set. If the system is in state $x \in X$ and control action $a \in A(x)$ is selected leading to a state x' after a transition time t, then the system incurs a cost $c'(t, x', x, a)$. This cost includes the immediate cost of action a as well as any additional cost occurring during the transition to the next state.

For any Borel set $B \subseteq X$, and for any $(x, a) \in K$ the function $Q'(\cdot, B|x, a)$ is a distribution function, i.e.,

- $Q'(t, B|x, a) = 0$ for every $t \leq 0$,
- $Q'(t, B|x, a)$ is a monotone lower semicontinuous function in t, and
- $\lim_{t\to\infty} Q'(t, X|x, a) = 1$.

We denote by x_n the state of the system at the nth decision time t_n and by $a_n \in A(x_n)$ the corresponding control action. The *transition time* $\delta_{n+1} = t_{n+1} - t_n$ has distribution $F(\cdot|x_n, a_n) = Q'(\cdot, X|x_n, a_n)$. For every Borel set $B \subseteq X$ and for every $(x, a) \in K$ let $Q(B|x, a) = \lim_{t\to\infty} Q'(t, B|x, a)$ denote the probability that the system is in a state from B in the next decision epoch when action a is chosen in state x. We call Q the *transition law*. Observe that Q is a stochastic kernel on X.

We denote by H_n the state of all *admissible histories* until the nth transition. Formally, $H_0 = X$ and $H_n = (K \times T)^n \times X$, where $h_n = (x_0, a_0, \delta_1, \dots, x_{n-1}, a_{n-1}, \delta_n, x_n) \in H_n$ encodes the history of the process.

DEFINITION 1. *A* policy π *is a sequence* $\pi = \{\pi_n\}_{n=0}^{\infty}$ *of stochastic kernels* π_n *on* A *satisfying* $\pi_n(A(x_n)|h_n) = 1$ *for every admissible history* $h_n \in H_n$ *and for every* $n \in \mathbb{N}$. *A policy* π *is a* stationary randomized policy *if there exists a stochastic kernel* ϕ *such that* $\pi_n(\cdot|h_n) = \phi(\cdot|x_n)$ *for each* $h_n \in H_n$ *and for each* $n \in \mathbb{N}$. *A policy* $\pi = \{\pi_n\}_{n=0}^{\infty}$ *is a* stationary deterministic policy *if there exists a measurable function* $f : X \to A$ *such that* $\pi_n(\cdot|h_n)$ *is concentrated at* $f(x_n) \in A(x_n)$ *for each* $n \in \mathbb{N}$. *We denote by* Π *the set of all policies and by* Π_{SD} *the subset of all stationary deterministic policies.*

Every initial distribution ν (which is a probability measure on X) and every policy π determine a unique probability measure P_ν^π and a stochastic process $\{(x_n, a_n, \delta_n), n = 0, 1, \dots\}$ on $\Omega = (X \times A \times T)^\infty$ (theorem of Ionescu Tulcea; see, e.g., Ash (1972, pp. 109) for a proof). We denote by E_ν^π the expectation operator with respect to P_ν^π and for $x \in X$ let E_x^π be equal to E_ν^π, where ν is the Dirac measure concentrated on x. The *mean holding time* in state x under a control $a \in A(x)$ is

$$\tau(x, a) = \int_T t\, F(\mathrm{d}t|x, a) = \int_T t\, Q'(\mathrm{d}t, X|x, a).$$

DEFINITION 2. *Given an initial distribution* ν *and a policy* π, *the* long-run expected average cost *is defined as*

$$J(\pi, \nu) = \limsup_{n\to\infty} \frac{E_\nu^\pi(\sum_{k=0}^{n-1} c'(t_k, x_{k+1}, x_k, a_k)}{E_\nu^\pi(t_n)}.$$

Let $J^* = \inf_\nu \inf_\pi J(\pi, \nu)$. A pair (ν^*, π^*) is a *minimum pair* if $J(\nu^*, \pi^*) = J^*$. The average cost problem is the problem of finding a minimum pair. For $x \in X$ let $J(x) = \inf_{\pi\in\Pi} J(\pi, x)$ be the *optimal average cost function*. A policy π^* is *average cost optimal* if $J(\pi^*, x) = J(x)$ for every $x \in X$. It is easy to see that

$$J(\pi, \nu) = \limsup_{n\to\infty} \frac{E_\nu^\pi(\sum_{k=0}^{n-1} c(x_k, a_k))}{E_\nu^\pi(\sum_{k=0}^{n-1} \tau(x_k, a_k))},$$

where

$$c(x,a) = \int_X \int_T c'(t,x',x,a) Q'(\mathrm{d}t, \mathrm{d}x' | x, a).$$

Note that a Markov control model is a semi-Markov control model with transition times equal to 1 with probability 1.

A special class of semi-Markov processes includes those with the transition law concentrated at a single state.

DEFINITION 3. *The transition law is a* Dirac's transition law *if there exists a measurable function* $s : X \times A \to X$ *such that*

$$Q(B|x,a) = \begin{cases} 1 & s(x,a) \in B, \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

for every Borel measurable set $B \subseteq X$.

DEFINITION 4. *A transition law* Q *is* weakly continuous *if* $h : X \times A \to \mathbb{R}$ *defined by*

$$h(x,a) = \int_X u(y) Q(\mathrm{d}y | x, a)$$

is a continuous bounded function on K *for every continuous bounded function* u *on* X. *Kernel* Q *is* strongly continuous *if* h *is a continuous bounded function on* K *for every measurable bounded function* u *on* X.

If Q is a Dirac's transition law, then $h(x,a) = u(s(x,a))$. In this case Q is weakly continuous if and only if s is a continuous function (consider $u(x) = x$ for all $x \in X$). However, Q is typically not strongly continuous.

For this semi-Markov decision process, the *average cost optimality equation* is

$$u(x) = \inf_{a \in A(x)} \left\{ c(x,a) - g\tau(x,a) + \int_X u(y) Q(\mathrm{d}y | x, a) \right\}.$$

In the case of a Dirac's transition law, this simplifies to

$$u(x) = \inf_{a \in A(x)} \{ c(x,a) - g\tau(x,a) + u(s(x,a)) \}. \tag{2}$$

3. Linear programming and semi-Markov control models with the average cost criterion. In this section we develop a linear programming formulation for the semi-Markov control model. The formulation is based on the prior work on infinite-dimensional linear programming for Markov control models of Hernández-Lerma and Lasserre (1999, pp. 203–249). A thorough coverage of infinite-dimensional linear programs is given by Anderson and Nash (1987).

3.1. Linear programs. Given a Borel space Z and a measurable weight function $f \geq 1$, let $\mathbb{B}_f(Z)$ be the Banach space of measurable functions u with finite *f-norm*

$$\|u\|_f = \sup_Z \frac{|u(s)|}{|f(s)|}.$$

In addition, let $\mathbb{M}_f(Z)$ be the Banach space of signed measures μ on the Borel space on Z with finite f *total variation norm*

$$\|\mu\|_f^{\mathrm{TV}} = \sup_{\|u\|_f \le 1} \left| \int_Z u \, \mathrm{d}\mu \right|.$$

The *total variation norm* of μ is $\|\mu\|_{\mathrm{TV}} = \|\mu\|_1$. It it easy to see that

$$\|\mu\|_{\mathrm{TV}} \le \|\mu\|_f^{\mathrm{TV}}. \tag{3}$$

See, e.g., Hernández-Lerma and Lasserre (1999, pp. 2–3) for a proof that $\mathbb{B}_f(Z)$ and $\mathbb{M}_f(Z)$ are Banach spaces. Let $\mathcal{B}(Z)$ be the Borel σ-algebra on Z and let $\mathbb{C}_b(Z)$ be the set of all continuous, bounded functions on Z.

Let $w : K \longrightarrow \mathbb{R}, w_0(x) : X \longrightarrow \mathbb{R}$ be defined as

$$w(x,a) = \tau(x,a) + c(x,a), \tag{4}$$

$$w_0(x) = \inf_{a \in A(x)} w(x,a). \tag{5}$$

To define linear programs corresponding to the semi-Markov control process, we need the following assumptions.

Assumption A1. *$w(x,a)$ is lower semicontinuous and $\{a \in A(x) : w(x,a) \le r\}$ is compact for every $x \in X$ and $r \in \mathbb{R}$.*

Assumption A2. *τ and c are nonnegative measurable functions.*

Assumption A3. *$w(x,a) \ge 1$ for every $(x,a) \in K$.*

Assumption A4. *There exists a finite constant $k \in \mathbb{R}$ such that*

$$\int_X w_0(y) Q(\mathrm{d}y|x,a) \le k \cdot w(x,a)$$

for every $(x,a) \in K$.

Due to Assumption A1, w_0 is well defined, the infimum can be replaced by the minimum, and, in addition, w_0 is measurable (Rieder (1978)). Assumption A3 can be relaxed to $\tau(x,a) + c(x,a) \ge \epsilon$ for every $(x,a) \in K$ and a given $\epsilon > 0$. By Assumption A3, $\mathbb{B}_{w_0}(X)$ is a well-defined Banach space, and, by Assumption A1, $\mathbb{M}_{w_0}(X)$ is a well defined Banach space. Since every lower semicontinuous function is measurable, it follows from the same two assumptions that $\mathbb{M}_w(K)$ is a well-defined Banach space. Observe also that Assumption A2 implies that $\tau \in \mathbb{B}_w(K)$ and $c \in \mathbb{B}_w(K)$.

Consider the following primal/dual linear programs on the dual pairs defined as $(\mathbb{M}_w(K), \mathbb{B}_w(K)), (\mathbb{R} \times \mathbb{M}_{w_0}(X), \mathbb{R} \times \mathbb{B}_{w_0}(X))$. The primal problem is

$$\inf \int_K c(x,a) \mu(\mathrm{d}(x,a)) \tag{6a}$$

$$\int_K \tau(x,a) \mu(\mathrm{d}(x,a)) = 1 \tag{6b}$$

$$\mu((B \times A) \cap K) - \int_K Q(B|x,a) \mu(\mathrm{d}(x,a)) = 0 \qquad \text{for every } B \in \mathcal{B}(X) \tag{6c}$$

$$\mu \ge 0, \mu \in \mathbb{M}_w(K), \tag{6d}$$

and the dual problem reads

$$\sup \rho \tag{7a}$$

$$\tau(x,a)\rho + u(x) - \int_X u(y) Q(\mathrm{d}y|x,a) \le c(x,a) \qquad \text{for every } (x,a) \in K \tag{7b}$$

$$\rho \in \mathbb{R}, u \in \mathbb{B}_{w_0}(X). \tag{7c}$$

We denote by $\inf(P)$, $\sup(D)$ the optimal value of the primal, dual linear program, respectively.

To see that (6) and (7) are indeed a primal-dual pair consider the following operators. Let $L_0 : \mathbb{M}_w(K) \longrightarrow \mathbb{M}_{w_0}(X)$ be defined as

$$(L_0\mu)(B) = \mu(B \times A) - \int_K Q(B|x,a)\mu(\mathrm{d}(x,a)) \qquad \text{for every } B \in \mathcal{B}(X)$$

and let $L : \mathbb{M}_w(K) \longrightarrow \mathbb{R} \times \mathbb{M}_{w_0}(X)$ be

$$L\mu = \left(\int_K \tau(x,a)\mu(\mathrm{d}(x,a)), L_0\mu \right).$$

The adjoint operator $L^* : \mathbb{R} \times \mathbb{B}_{w_0}(X) \longrightarrow \mathbb{B}_w(K)$ is given by

$$L^*(\rho,u)(x,a) = \tau(x,a)\rho + u(x) - \int_X u(y)Q(\mathrm{d}y|x,a)$$

for every $(\rho,u) \in \mathbb{R} \times \mathbb{B}_{w_0}(X)$ and $(x,a) \in K$. To see that $L^*(\rho,u) \in \mathbb{B}_w(K)$, let $(\rho,u) \in \mathbb{R} \times \mathbb{B}_{w_0}(X)$. Then

$$\left| \frac{\tau(x,a)\rho}{w_(x,a)} \right| \leq |\rho|, \tag{8}$$

$$\left| \frac{u(x)}{w(x,a)} \right| = \left| \frac{u(x)}{w_0(x)} \right| \cdot \frac{w_0(x)}{w(x,a)} \leq \left| \frac{u(x)}{w_0(x)} \right| \leq \|u\|_{w_0}, \tag{9}$$

$$\left| \frac{\int_X u(y)Q(\mathrm{d}y|x,a)}{w(x,a)} \right| \leq \|u\|_{w_0} \frac{\int_X w_0(y)Q(\mathrm{d}y|x,a)}{w(x,a)} \leq \|u\|_{w_0} k, \tag{10}$$

where (8) follows by Assumption A2 and (4), (9) by definition (5), and (10) by Assumption A4. It follows that the linear operator L is continuous with respect to the weak topology—see, e.g., Anderson and Nash (1987, pp. 35–40)—and therefore (7) is a dual linear program to (6). It implies that under A1–A4 we can apply results from Anderson and Nash (1987).

3.2. Results. A linear program is *consistent* if it has a feasible solution and it is *solvable* if there is a feasible solution that attains the optimal objective value. If (6), (7), is solvable, then we can replace inf, sup, in (6a), (7a), by min, max, and we write the corresponding value as $\min(P), \max(D)$, respectively. In this section we discuss the relation between the linear programs and the underlying semi-Markov control model and we give no duality gap and solvability results.

DEFINITION 5. *A function g on Z is a* strictly unbounded function *if there is a nondecreasing sequence of compact sets $Z_n \uparrow Z$ such that* $\lim_{n\to\infty} \inf\{g(x)|x \notin Z_n\} = \infty$.

If Z is compact, then any function is strictly unbounded by considering $Z_n = Z$ for every n. If Z is open but bounded, then a strictly unbounded function must be discontinuous at the boundary of Z.

We need the following additional assumptions.

ASSUMPTION A5. *There is a policy π and an initial distribution ν such that $J(\pi,\nu) < \infty$.*

ASSUMPTION A6. *The transition law is weakly continuous.*

ASSUMPTION A7. *τ is a nonnegative, continuous, bounded function.*

ASSUMPTION A8. *w is strictly unbounded on K.*

Note that Assumptions A1 and A7 imply that c is lower semicontinuous and Assumption A7 yields $\tau \in \mathbb{C}_b(K)$.

Next we give some known results that will be used in subsequent sections. The following theorem is proved in Dynkin and Yushkevich (1979, pp. 88–89).

THEOREM 1. *Let μ be a probability measure on $X \times A$ concentrated on K. Then there exists a stochastic kernel π on A such that*

$$\mu(B \times C) = \int_B \pi(C|x)\hat{\mu}(\mathrm{d}x) \qquad \text{for every } B \in \mathcal{B}(X), C \in \mathcal{B}(A),$$

where $\hat{\mu}(\cdot) = \mu(\cdot \times A)$ is the marginal of μ on X.

DEFINITION 6. *A measure μ on Z is* tight *if for each $\epsilon > 0$ there is a compact set $C \subseteq Z$ such that $\mu(Z \setminus C) < \epsilon$.*

The proofs of the following two theorems are given in Billingsley (1968).

THEOREM 2. *Let Γ be a bounded family of nonnegative measures on Z. Then Γ is tight if and only if there is a strictly unbounded function $g \geq 1$ such that $\sup_{\mu\in\Gamma} \int_Z g \,\mathrm{d}\mu < \infty$. If Γ is a set of probability measures, then the condition $g \geq 1$ can be relaxed to $g \geq 0$.*

THEOREM 3 (due to Prohorov). *Let Γ be a family of probability measures on a Borel space Z. If Γ is tight, then for each sequence $\{\mu_n\}$ in Γ there is a subsequence $\{\mu_m\}$ and a probability measure μ such that*

$$\int_Z u \,\mathrm{d}\mu_m \longrightarrow \int_Z u \,\mathrm{d}\mu \tag{11}$$

for every $u \in \mathbb{C}_b(Z)$.

We say that measures $\{\mu_m\}_m$ converge *weakly* to a measure μ if (11) holds. We will repeatedly use the following corollary.

COROLLARY 1. *Let Γ be a family of nonnegative measures on a Borel space Z. Assume that there exists a constant $K < \infty$ such that $0 < \|\mu\|_{TV} < K$. In addition, let there exist a strictly unbounded function $g \geq 1$ such that $\sup_{\mu\in\Gamma} \int_Z g \,\mathrm{d}\mu < \infty$. Then for each sequence $\{\mu_n\}$ in Γ there is a subsequence $\{\mu_m\}$ and a measure μ such that $\{\mu_m\}$ converges weakly to μ.*

Proof. Let $\{\mu_n\}$ be a sequence in Γ.

If $\liminf_n \|\mu_n\|_{\mathrm{TV}} = 0$, there there exists a subsequence $\{\mu_m\}$ of $\{\mu_n\}$ such that $\lim_m \|\mu_m\|_{\mathrm{TV}} = 0$. But then for any $u \in \mathbb{C}_b(Z)$ and any m we have $|\int_Z u \,\mathrm{d}\mu_m| \leq M\|\mu_m\|_{\mathrm{TV}}$, where $|u(s)| \leq M < \infty$ for any $s \in Z$. Hence $\{\mu_m\}$ converges weakly to the 0 measure.

Let now $\liminf_n \|\mu_n\|_{\mathrm{TV}} > 0$. Without loss of generality we assume that $\|\mu_n\|_{\mathrm{TV}} > m > 0$ for every n. Consider the set $\tilde{\Gamma}$ of probability measures defined as $\{\mu/\|\mu\|_{\mathrm{TV}} : \mu \in \Gamma\}$. We have

$$\sup_{\tilde{\mu}\in\tilde{\Gamma}} \int_Z g \,\mathrm{d}\tilde{\mu} \leq \frac{\sup_{\mu\in\Gamma} \int_Z g \,\mathrm{d}\mu}{m} < \infty$$

by assumption. Therefore by Theorem 2, $\tilde{\Gamma}$ is tight. By Prohorov's theorem we have that there is a weakly convergent subsequence $\{\tilde{\mu}_p\}$ that converges to a measure $\tilde{\mu}$. There is a subsequence $\{\mu_m\}$ of $\{\mu_p\}$ such that $\lim_m \|\mu_m\|_{\mathrm{TV}} = Q$. Clearly

$0 < Q < K$. Now for every $u \in \mathbb{C}_b(Z)$ we have

$$\lim_m \int_Z u \, \mathrm{d}\mu_m = \lim_m \left(\int_Z u \, \mathrm{d}\tilde{\mu}_m \cdot \|\mu_m\|_{\mathrm{TV}} \right) = Q \lim_m \int_Z u \, \mathrm{d}\tilde{\mu}_m = Q \int_Z u \, \mathrm{d}\tilde{\mu}.$$

Therefore $\{\mu_m\}$ converges weakly to $\mu = Q\tilde{\mu}$. □

3.2.1. Consistency and solvability. In this section we give results regarding consistency and solvability of (6) and (7). We first address consistency.

THEOREM 4. *Assume Assumptions* A1–A8 *hold.* (6) *and* (7) *are consistent, and* $\inf(P) = J^*$.

The following lemma is proved by Hernández-Lerma and Lasserre (1999, pp. 225).

LEMMA 1. *Let* $\{\mu_n\}$ *be a sequence of measures on* S *and* μ *a measure on* S *such that* $\{\mu_n\}$ *converges weakly to* μ. *If* $c \geq 0$ *is a lower semicontinuous function on* S, *then*

$$\liminf_n \int_S c \, \mathrm{d}\mu_n \geq \int_S c \, \mathrm{d}\mu.$$

In addition we need the following lemma.

LEMMA 2. *If* $\{\mu_n\}_n$ *converges weakly to* μ, *then for every* $v \in \mathbb{C}_b(X)$ *we have*

$$\lim_{n\to\infty} \int_X v \, \mathrm{d}L_0(\mu_n) = \int_X v \, \mathrm{d}L_0(\mu).$$

Proof. We have

$$\begin{aligned} \int_X v \, \mathrm{d}L_0(\mu) &= \int_K v \cdot \left(1 - \int_X Q(\mathrm{d}y|x,a)\right) \mu(\mathrm{d}(x,a)) \\ &= \lim_{n\to\infty} \int_K v \cdot \left(1 - \int_X Q(\mathrm{d}y|x,a)\right) \mu_n(\mathrm{d}(x,a)) \\ &= \lim_{n\to\infty} \int_X v \, \mathrm{d}L_0(\mu_n) = 0, \end{aligned}$$

where the first equality follows from the definition of the adjoint operator (see section 3.1), and the second equality follows from Assumption A6 and the definition of weak convergence. □

Proof of Theorem 4. (7) is consistent by taking $\rho = 0$, $u = 0$.

Next we address consistency of (6). Consider a policy π and an initial distribution ν such that $J(\pi,\nu) < \infty$. For every integer $n \geq 1$ let us define the probability measure on K as

$$\mu_n(\Omega) = \frac{1}{n} \sum_{i=0}^{n-1} P_\nu^\pi((x_i,a_i) \in \Omega).$$

From Assumption A7 it follows that there exists a constant $M < \infty$ such that $\tau(x,a) \leq M$ for every $(x,a) \in K$. Then

$$\begin{aligned} \int_K w \, \mathrm{d}\mu_n &= \frac{\sum_{k=0}^{n-1} E_\nu^\pi(w(x_k,a_k))}{n} = \frac{\sum_{k=0}^{n-1} E_\nu^\pi(w(x_k,a_k))}{\sum_{k=0}^{n-1} E_\nu^\pi(\tau(x_k,a_k))} \cdot \frac{\sum_{k=0}^{n-1} E_\nu^\pi(\tau(x_k,a_k))}{n} \\ &\leq (J(\pi,\nu)+1) \cdot M < \infty. \end{aligned}$$

This implies that we can use Corollary 1 since by Assumption A8 w is strictly unbounded and $\|\mu_n\|_{\mathrm{TV}} = 1$. Let $\{\mu_m\}_m$ be a subsequence that convergence weakly to μ.

Since every μ_m is a probability measure, so is μ. For a subsequence l of m we have

$$J(\pi,\nu) = \limsup_n \frac{\int_K c\,\mathrm{d}\mu_n}{\int_K \tau\,\mathrm{d}\mu_n} \geq \limsup_m \frac{\int_K c\,\mathrm{d}\mu_m}{\int_K \tau\,\mathrm{d}\mu_m} = \lim_l \frac{\int_K c\,\mathrm{d}\mu_l}{\int_K \tau\,\mathrm{d}\mu_l}. \tag{12}$$

In addition, there exists a subsequence k of l such that

$$\liminf_l \int_K c\,\mathrm{d}\mu_l = \lim_k \int_K c\,\mathrm{d}\mu_k. \tag{13}$$

It follows from (12) that

$$J(\pi,\nu) \geq \lim_k \frac{\int_K c\,\mathrm{d}\mu_k}{\int_K \tau\,\mathrm{d}\mu_k}. \tag{14}$$

By Lemma 1 and (13) we obtain $\lim_k \int_K c\,\mathrm{d}\mu_k \geq \int_K c\,\mathrm{d}\mu$. Since by Assumption A7 $\tau \in \mathbb{C}_b(K)$, we have that $\lim_k \int_K \tau\,\mathrm{d}\mu_k = \int_K \tau\,\mathrm{d}\mu$.

We first show that $\int_K \tau\,\mathrm{d}\mu > 0$. To the contrary, assume that $\int_K \tau\,\mathrm{d}\mu = 0$. Note that $\int_K c\,\mathrm{d}\mu_k = \int_K w\,\mathrm{d}\mu_k \geq \int_K \mathrm{d}\mu_k = 1$. Let us fix an $\epsilon > 0$. There exists an integer k_1 such that for every $k \geq k_1$ we have $\int_K \tau\,\mathrm{d}\mu_k \leq \epsilon$. Then for any $k \geq k_1$ it follows

$$\frac{\int_K c\,\mathrm{d}\mu_k}{\int_K \tau\,\mathrm{d}\mu_k} \geq \frac{1}{\epsilon}.$$

Since ϵ is an arbitrarily small number, it follows that $\lim_k \frac{\int_K c\,\mathrm{d}\mu_k}{\int_K \tau\,\mathrm{d}\mu_k} = \infty$, which contradicts (14) and the assumption that $J(\pi,\nu) < \infty$.

We conclude that $0 < \int_K \tau\,\mathrm{d}\mu < M$. This in turn implies that

$$J(\pi,\nu) \geq \lim_k \frac{\int_K c\,\mathrm{d}\mu_k}{\int_K \tau\,\mathrm{d}\mu_k} = \frac{\lim_k \int_K c\,\mathrm{d}\mu_k}{\lim_k \int_K \tau\,\mathrm{d}\mu_k} \geq \frac{\int_K c\,\mathrm{d}\mu}{\int_K \tau\,\mathrm{d}\mu}.$$

Next we show that μ satisfies (6c). Let $\mathcal{X}$ denote the characteristic or the indicator function of a set. Since for every $B \in \mathcal{B}(X)$ we have

$$P_\nu^\pi(x_i \in B) = E_\nu^\pi(\mathcal{X}_B(x_i)) = E_\nu^\pi(Q(B|x_{i-1}, a_{i-1}))$$

and for every k

$$\int_K Q(B|x,a)\mathrm{d}\mu_k = \frac{1}{k}\sum_{i=0}^{k-1} E_\nu^\pi(Q(B|x_i,a_i)),$$

an easy calculation shows that for every k

$$\mu_k(B \times A) = \int_K Q(B|x,a)\mathrm{d}\mu_k + \frac{P_\nu^\pi(x_{k-1} \in B) - \nu(B)}{k}.$$

Note that the last equality can be rewritten as $L_0(\mu_k) = (P_\nu^\pi(x_{k-1} \in B) - \nu(B))/k$. By considering $v = 1$ in Lemma 2 and the above equality, we obtain that $L_0(\mu) = 0$. Hence μ satisfies (6c).

Consider now the measure

$$\tilde{\mu} = \frac{\mu}{\int_K \tau \, \mathrm{d}\mu}.$$

Clearly $\tilde{\mu}$ satisfies (6b) and by the above argument it satisfies (6c) as well. We also have

$$\int_K w \, \mathrm{d}\tilde{\mu} = \frac{\int_K w \, \mathrm{d}\mu}{\int_K \tau \, \mathrm{d}\mu} = 1 + \frac{\int_K c \, \mathrm{d}\mu}{\int_K \tau \, \mathrm{d}\mu} \leq 1 + J(\pi, \nu) < \infty,$$

showing that $\tilde{\mu} \in \mathbb{M}_w(K)$. Therefore $\tilde{\mu}$ is a feasible solution to (6). Note also that $\int_K c \, \mathrm{d}\tilde{\mu} \leq J(\pi, \nu)$. Since π is an arbitrary policy and ν an arbitrary initial probability distribution, it follows that $\inf(P) \leq J^*$.

It remains to be seen that $J^* \leq \inf(P)$. Since (6) is feasible, there exists a feasible solution μ. If $\int_K c \, \mathrm{d}\mu = \infty$, then there is nothing to prove and therefore we assume that $\int_k c \, \mathrm{d}\mu < \infty$. Then by Assumption A3 and feasibility of μ, $0 < \mu(K) \leq \int_K w \, \mathrm{d}\mu = 1 + \int_K c \, \mathrm{d}\mu < \infty$ and therefore $\mu(K) < \infty$. By Theorem 1, there exists a policy π such that

$$\frac{\mu(B \times C)}{\mu(X \times A)} = \int_B \pi(C|x)\tilde{\mu}(\mathrm{d}x) \qquad \text{for every } B \in \mathcal{B}(X), C \in \mathcal{B}(A). \tag{15}$$

For any randomized stationary policy π, $n \geq 2$, $x \in X$, $B \in \mathcal{B}(X)$, and a measurable function f on K we denote

$$\begin{aligned} f(x, \pi) &= \int_A f(x, a)\pi(\mathrm{d}a|x), \\ Q(B|x, \pi) &= \int_A Q(B|x, a)\pi(\mathrm{d}a|x), \\ Q^n(B|x, \pi) &= P_x^\pi(x_n \in B) = \int_X Q^{n-1}(B|y, \pi)Q(\mathrm{d}y|x, \pi), \\ Q^1(B|x, \pi) &= Q(B|x, \pi). \end{aligned}$$

Then we have

$$\int_K f \, \mathrm{d}\mu/\mu(X \times A) = \int_X f(x, \pi)\tilde{\mu}(\mathrm{d}x), \tag{16}$$

$$E_{\tilde{\mu}}^\pi(f(x_n, a_n)) = \int_X \int_X f(y, \pi)Q^n(\mathrm{d}y|x, \pi)\tilde{\mu}(\mathrm{d}x), \tag{17}$$

$$\tilde{\mu}(B) = \int_X Q^n(B|x, \pi)\tilde{\mu}(\mathrm{d}x), \tag{18}$$

where the first two equalities follow from (15) and aforementioned notation, and the last equality follows by iteratively applying (6c). It follows that

$$J^* \leq J(\pi, \tilde{\mu}) = \frac{\int_K c \, \mathrm{d}\mu/\mu(X \times A)}{\int_K \tau \, \mathrm{d}\mu/\mu(X \times A)} = \int_K c \, \mathrm{d}\mu.$$

Since μ is an arbitrary feasible measure to (6), we conclude that $J^* \leq \inf(P)$. □

Next we discuss solvability.

THEOREM 5. *If Assumptions* A1–A8 *hold, then* (6) *is solvable.*

Proof. Since (6) is consistent by Theorem 4, for every nonnegative integer n there is a feasible measure μ_n to (6) such that

$$\inf(P) \leq \int_K c(x,a)\mu_n(\mathrm{d}(x,a)) \leq \inf(P) + \frac{1}{n} < \infty. \tag{19}$$

Since μ_n is feasible to (6) and from (19) it follows that

$$\|\mu_n\|_w^{\mathrm{TV}} \leq \int_K w \,\mathrm{d}\mu_n = \int_K (\tau + c)\mathrm{d}\mu_n = 1 + \int_K c \,\mathrm{d}\mu_n \leq 2 + \inf(P).$$

If in addition we use (3), we get that $0 < \|\mu_n\|_{\mathrm{TV}} \leq 2 + \inf(P) < \infty$. By Assumption A8 and since $\sup \int_K w \,\mathrm{d}\mu_n$ is bounded, we can use Corollary 1. Let μ_m be a subsequence that converges weakly to a measure μ. We claim that μ is an optimal solution to (6).

From Lemma 1 and (19) it follows that $\int_K c \,\mathrm{d}\mu \leq \inf(P)$. If μ is feasible to (6), then this implies that μ is optimal.

Now we show that μ is feasible to (6). Since by Assumption A7 $\tau \in \mathbb{C}_b(K)$, it follows

$$1 = \int_K \tau \,\mathrm{d}\mu_m \longrightarrow \int_K \tau \,\mathrm{d}\mu$$

and therefore τ satisfies (6b). In turns it implies that

$$\|\mu\|_w^{\mathrm{TV}} \leq \int_K \tau \,\mathrm{d}\mu + \int_K c \,\mathrm{d}\mu \leq 1 + \inf(P)$$

and therefore $\mu \in \mathbb{M}_w(K)$. Since μ_m are feasible, it follows that $L_0(\mu_m) = 0$ for every m and in turn we can apply Lemma 2 with $v = 1$. Therefore μ satisfies (6c). ∎

Next we address solvability of (7). A sequence $\{(\rho_n, u_n)\}_n$ of feasible solutions to (7) is a *maximizing sequence* if $\lim_{n\to\infty} \rho_n = \sup(D)$.

THEOREM 6. *Assume that Assumptions* A1–A4 *hold. If there exists a maximizing sequence* $\{(\rho_n, u_n)\}_n$ *to* (7) *such that* $\|u_n\|_{w_0} \leq r < \infty$ *for a constant* r*, then* (7) *is solvable.*

Proof. Let $\rho = \sup(D)$ and let us define

$$u(x) = \limsup_{n\to\infty} u_n(x).$$

By assumption $\|u\|_{w_0} \leq r$ and therefore $u \in \mathbb{B}_{w_0}(X)$. For every $y \in X$ we have $|u_n(y)| \leq r w_0(y)$ and by Assumption A4 $\int_X w_0(y) Q(\mathrm{d}y|x,a) \leq k w(x,a) < \infty$, which justifies using Fatou's lemma with respect to $Q(\cdot|x,a)$. Since (ρ_n, u_n) satisfies (7b), we have that for every $(x,a) \in K$ and every n

$$\tau(x,a)\rho_n + u_n(x) \leq \int_X u_n(y) Q(\mathrm{d}y|x,a) + c(x,a).$$

After taking lim sup, using $\lim_n \rho_n = \rho$, and applying Fatou's lemma, we obtain

$$\tau(x,a)\rho + u(x) - \int_X u(y) Q(\mathrm{d}y|x,a) \leq c(x,a).$$

Therefore (ρ, u) is a feasible solution to (7) with value $\sup(D)$ and therefore it is an optimal solution. ∎

3.2.2. No duality gap. In this section we prove that under our assumptions there is no duality gap.

THEOREM 7. *If Assumptions* A1–A8 *hold, then* $\sup(D) = \inf(P)$.

Proof. Let

$$H = \left\{ (L\mu, \int_K c \, \mathrm{d}\mu + r) : \mu \in \mathbb{M}_w^+(K), r \geq 0 \right\},$$

where $\mathbb{M}_w^+(K)$ is the set of all nonnegative measures in $\mathbb{M}_w(K)$. By a theorem from Anderson and Nash (1987, pp. 52), if H is closed in the weak topology of $(\mathbb{R} \times \mathbb{M}_{w_0}(X) \times \mathbb{R}, \mathbb{R} \times \mathbb{B}_{w_0}(X) \times \mathbb{R})$, then there is no duality gap.

To this end, let $(D, \geq)$ be a directed set and let $\{\mu_\alpha, r_\alpha\}_{\alpha \in D}$ be a net (see, e.g., Ash (1972) for a definition of directed sets and nets) in $\mathbb{M}_w(K) \times \mathbb{R}_+$ such that

$$\int_K \tau \, \mathrm{d}\mu_\alpha \to r_*, \tag{20}$$

$$\int_X u \, \mathrm{d}L_0(\mu_\alpha) \to \int_X u \, \mathrm{d}\upsilon_* \qquad \text{for every } u \in \mathbb{C}_b(X),$$

$$\int_K c \, \mathrm{d}\mu_\alpha + r_\alpha \to \rho_*. \tag{21}$$

By using Corollary 1 we show that there exists a nonnegative measure $\mu \in \mathbb{M}_w(X)$ and $r \in \mathbb{R}_+$ such that

$$r_* = \int_K \tau \, \mathrm{d}\mu, \tag{22}$$

$$\upsilon_* = L_0(\mu), \tag{23}$$

$$\rho_* = \int_K c \, \mathrm{d}\mu + r_*. \tag{24}$$

Since $r_\alpha \geq 0, \int_K c \, \mathrm{d}\mu_\alpha \geq 0$ and by (21), it follows that $\int_K c \, \mathrm{d}\mu_\alpha$ are bounded for $\alpha \geq \alpha_0$ for an $\alpha_0 \in D$. Therefore by (20) it follows that there exists $\alpha_1 \in D, \alpha_1 > \alpha_0$ such that $\int_K w \, \mathrm{d}\mu_\alpha$ is bounded and positive for $\alpha \geq \alpha_1$. There exists a constant K such that $\|\mu_\alpha\|_w^{\mathrm{TV}} \leq K$ for every $\alpha \geq \alpha_1$. This in turn implies that $\|\mu_\alpha\|_{\mathrm{TV}} \leq \|\mu_\alpha\|_w^{\mathrm{TV}} \leq K$ for every $\alpha \geq \alpha_1$. We conclude that $\{\mu_\alpha\}_{\alpha \geq \alpha_1}$ is bounded. By Assumption A8 and by using Corollary 1 we obtain that there is a subsequence $\{\mu_m\}_m$ that converges weakly to a measure μ.

Since $\tau \in \mathbb{C}_b(K)$ by Assumption A7, it immediately follows that $r_* = \int_K \tau \, \mathrm{d}\mu$. Hence we have (22). By Lemma 1, we have

$$\int_K w \, \mathrm{d}\mu \leq 1 + \liminf_m \int_K c \, \mathrm{d}\mu_m < \infty$$

and therefore $\mu \in \mathbb{M}_w(K)$. Using again Lemma 1 and taking lim inf in (21) we get

$$\rho_* \geq \liminf_m \int_K c \, \mathrm{d}\mu_n + \liminf_m r_m \geq \int_K c \, \mathrm{d}\mu.$$

Thus we can define $r_* = \rho_* - \int_K c \, \mathrm{d}\mu \geq 0$ and we obtain (24). By using Lemma 2 we establish (23) and thus we have shown the theorem. □

3.2.3. Randomized optimal policies and optimality equation on a subset of states. In this section we show, under generous assumptions, that there exists a minimum pair and that that the optimality equation has a solution on a subset of states.

THEOREM 8. *Assume that Assumptions* A1–A8 *hold and that* (7) *is solvable. Let* μ, (ρ, u) *be an optimal solution to* (6), (7), *respectively, and let* $\hat{\mu}$ *be the marginal of* μ *on* X. *Then*

(a) $J^* = \rho$, *and there exists a stationary randomized policy* π^* *and an initial distribution* $\hat{\mu}^*$ *such that* $(\hat{\mu}^*, \pi^*)$ *is a minimum pair, and*

$$J(x, \pi^*) = \rho \tag{25}$$

holds for $\hat{\mu}^*$*-almost all* $x \in X$;

(b) [*complementary slackness*] *and for* μ*-almost all* $(x, a) \in K$ *we have*

$$\tau(x, a)J^* + u(x) = c(x, a) + \int_X u(y)Q(\mathrm{d}y|x, a); \tag{26}$$

(c) *if we denote*

$$S = \{x \in X : \text{there exists } a \in A(x) \text{ such that } (26) \text{ holds for } (x, a)\}, \tag{27}$$

and

$$S^* = S \cap \{x \in S : u(x) < \infty\},$$

and we assume $S^* \neq \emptyset$, *then there exists a stationary policy* $f^* \in \Pi_{SD}$ *such that*

$$\begin{aligned} u(x) &= \min_{a \in A(x)} \left\{ c(x, a) - \tau(x, a)J^* + \int_X u(y)Q(\mathrm{d}y|x, a) \right\} \\ &= c(x, f^*(x)) - \tau(x, f^*(x))J^* + \int_X u(y)Q(\mathrm{d}y|x, f^*(x)) \end{aligned} \tag{28}$$

for every $x \in S^*$.

Proof. We first prove (a). Note that by Theorem 7 we have $\rho = J^*$. Since $0 < \mu(X \times A) \leq \int_{X \times A} w \, \mathrm{d}\mu = 1 + J^* < \infty$, we use Theorem 1 for $\mu/\mu(X \times A)$ to decompose this measure into a policy π^* and initial distribution $\hat{\mu}^*$. It follows from the proof of Theorem 4 that $(\hat{\mu}^*, \pi^*)$ is a minimum pair. The individual ergodic theorem (see, e.g., Yosida (1978)) yields (25).

Next we prove (b). Let q be a measurable function defined by

$$\tau(x, a)J^* + u(x) + q(x, a) = c(x, a) + \int_X u(y)Q(\mathrm{d}y|x, a). \tag{29}$$

Since (ρ, u) is feasible to (7), $q \geq 0$ for every $(x, a) \in K$. After integrating (29) with respect to μ we obtain

$$J^* + \int_K u \, \mathrm{d}\mu + \int_K q \, \mathrm{d}\mu = \int_K c \, \mathrm{d}\mu + \int_K u \, \mathrm{d}\mu, \tag{30}$$

where we have used that μ satisfies (6b) and from (6c) it follows

$$\int_K \left(\int_X u(y)Q(\mathrm{d}y|x, a) \right) \mu(\mathrm{d}(x, a)) = \int_K u \, \mathrm{d}\mu.$$

Since $\mu, (\rho, u)$ are optimal for the primal, dual linear programs, respectively, it follows $J^* = \int_K c \, \mathrm{d}\mu$. This together with

$$\left| \int_K u \, \mathrm{d}\mu \right| \leq k\|u\|_{w_0} \int_K w \, \mathrm{d}\mu = k\|u\|_{w_0}(1 + J^*) < \infty$$

and (30) yields $\int_K q \, \mathrm{d}\mu = 0$. Since q is nonnegative, we get that $q(x, a) = 0$ for μ-almost all (x, a), which completes the proof of the first statement.

It remains to show the last statement. For every $x \in S$ let $\bar{A}(x)$ be the set of all $a \in A(x)$ such that (x, a) satisfies (26). Note that by definition $\bar{A}(x) \neq \emptyset$. After integrating (26) with respect to $\pi^*(\mathrm{d}a|x)$ we obtain

$$u(x) = \int_{\bar{A}(x)} \left[c(x, a) + \int_X u(y) Q(\mathrm{d}y|x, a) - \tau(x, a) J^* \right] \pi^*(\mathrm{d}a|x).$$

Since $u(x) < \infty$ for $x \in S^*$ it follows from the measurable selection theorem of Blackwell and Ryll-Nardzewski (see, e.g., Dynkin and Yushkevich (1979, pp. 255)) that there exists a stationary deterministic policy f^* such that

$$\begin{aligned} &\int_{\bar{A}(x)} \left[c(x, a) + \int_X u(y) Q(\mathrm{d}y|x, a) - \tau(x, a) J^* \right] \pi^*(\mathrm{d}a|x) \\ &\geq c(x, f^*(x)) + \int_X u(y) Q(\mathrm{d}y|x, f^*(x)) - \tau(x, f^*(x)) J^*. \end{aligned}$$

The other inequality follows from feasibility of u to (7). This establishes the second part. □

4. Dirac's transition laws. Next we study Dirac's kernels. Note that in this case Assumption A4 is equivalent to

$$w_0(s(x, a)) \leq k w(x, a)$$

for every $(x, a) \in K$ and Assumption A6 requires s to be continuous. Under a Dirac's transition kernel the corresponding primal linear program is

$$\inf \int_K c(x, a) \mu(\mathrm{d}(x, a)) \tag{31a}$$

$$\int_K \tau(x, a) \mu(\mathrm{d}(x, a)) = 1 \tag{31b}$$

$$\mu((B \times A) \cap K) - \mu(\{(x, a) \in K : s(x, a) \in B\}) = 0 \qquad \text{for every } B \in \mathcal{B}(X) \tag{31c}$$

$$\mu \geq 0, \mu \in \mathbb{M}_w(K), \tag{31d}$$

and the corresponding dual problem reads

$$\sup \rho \tag{32a}$$

$$\tau(x, a)\rho + u(x) - u(s(x, a)) \leq c(x, a) \qquad \text{for every } (x, a) \in K \tag{32b}$$

$$\rho \in \mathbb{R}, u \in \mathbb{B}_{w_0}(X). \tag{32c}$$

By using a stronger version of Theorem 8 and a more stringent assumption we show the existence of a deterministic stationary optimal policies for all the states.

ASSUMPTION A9. *There exist constants $C < \infty$, $\Gamma < \infty$ such that for every measurable subset $S \subseteq X$ there is a measurable function $f : X \setminus S \to A$ with the property that for every $x' \in X \setminus S$ there exists a finite integer N and a set of states $x_0, x_1, \ldots, x_N$ with*

- $x_0 = x'$,
- $a_n = f(x_n) \in A(x_n)$ *for every* $n = 0, \ldots, N-1$,
- $x_{n+1} = s(x_n, a_n)$ *for every* $n = 0, \ldots, N-1$,
- $x_N \in S$,
- $\sum_{n=0}^{N-1} c(x_n, a_n) \leq C$, *and*
- $\sum_{n=0}^{N-1} \tau(x_n, a_n) \leq \Gamma$.

This assumption requires that any two states communicate (select S to be a single state) and the cost and the time of the path between any two states must be uniformly upper bounded.

For Dirac's kernels, we can strengthen Theorem 8 by showing that there exists an optimal policy whose sample path satisfies the average cost optimality equation.

THEOREM 9. *Assume that Assumptions* A1–A8 *hold and that* (32) *is solvable with* (ρ, u) *being an optimal solution. Furthermore, assume that there exists a constant* N *such that* $N > u(x) > -N$ *for every* $x \in X$. *Then there exists a stationary deterministic policy* $f^* \in \Pi_{SD}$ *and a nonempty set* $L \subseteq X$ *such that the average cost optimality equation* (2) *holds for every* $x \in L$ *and*

$$J(x) = J(f^*, x) = J^*$$

for every $x \in L$, *i.e.,* f^* *is an optimal stationary deterministic policy for all* $x \in L$.

The following lemma holds for general kernels.

LEMMA 3. *If Assumption* A3 *holds and* $J(\pi, \nu) < \infty$ *for a policy* π *and initial distribution* ν, *then* $\lim_{n\to\infty} \sum_{i=0}^{n-1} E_\nu^\pi(\tau(x_i, a_i)) = \infty$.

Proof. Suppose that $0 \leq \lim_{n\to\infty} \sum_{i=0}^{n-1} E_\nu^\pi(\tau(x_i, a_i)) < \infty$. Then there is a constant $K \geq 0$ such that

$$\sum_{i=0}^{n-1} E_\nu^\pi(\tau(x_i, a_i)) \leq K$$

for every n. By assumption we have

$$J(\pi, \nu) = \limsup_{n\to\infty} \frac{E_\nu^\pi(\sum_{k=0}^{n-1} c(x_k, a_k))}{E_\nu^\pi(\sum_{k=0}^{n-1} \tau(x_k, a_k))} = \limsup_{n\to\infty} \frac{E_\nu^\pi(\sum_{k=0}^{n-1} w(x_k, a_k))}{E_\nu^\pi(\sum_{k=0}^{n-1} \tau(x_k, a_k))} - 1 < \infty.$$

From Assumption A3 we obtain

$$\infty > 1 + J(\pi, \nu) \geq \limsup_n \frac{n}{K} = \infty,$$

which is a contradiction. □

Proof of Theorem 9. We use the same notation as in the proof of Theorem 8. We first show that there exists a trajectory, whose state-action pairs satisfy the optimality equation. For any $\omega \in \Omega$ let us define $r(\omega) = \sum_{i=1}^{\infty} q(x_i, a_i)$. Since u is dual feasible, we clearly have $r \geq 0$. In addition, let $r_n(\omega) = \sum_{i=1}^{n} q(x_i, a_i)$. We note that $r_1 \leq r_2 \leq r_3 \leq \cdots$ and for any $\omega \in \Omega$ we have $\lim_{n\to\infty} r_n(\omega) = r(\omega)$.

Next we show that for every n we have

$$\int_{\omega\in\Omega} r_n(\omega) P_{\hat{\mu}}^\pi(\mathrm{d}\omega) = 0. \tag{33}$$

We show this by induction. We first note that from (16), (18), and complementary slackness for every n it follows

$$0 = \frac{\int_K q \, \mathrm{d}\mu}{\mu(K)} = \int_X q(x,\pi)\hat{\mu}(\mathrm{d}x) = \int_X \int_X q(y,\pi) Q^n(dy|x,\pi)\hat{\mu}(\mathrm{d}x). \tag{34}$$

For $n = 1$ we have

$$\int_{\omega\in\Omega} r_1(\omega) P^\pi_{\hat{\mu}}(\mathrm{d}\omega) = E^\pi_{\hat{\mu}}(q(x_1,a_1)) = \int_X \int_X q(y,\pi) Q(dy|x,\pi)\hat{\mu}(\mathrm{d}x) = 0,$$

where the second equality follows from (17) and the last one from (34). Assume now that (33) holds for $n-1$. Then

$$\begin{aligned}
\int_{\omega\in\Omega} r_n(\omega) P^\pi_{\hat{\mu}}(\mathrm{d}\omega) &= \int_{\omega\in\Omega} (r_{n-1}(\omega) + q(x_n,a_n)) P^\pi_{\hat{\mu}}(\mathrm{d}\omega) \\
&= \int_{\omega\in\Omega} q(x_n,a_n) P^\pi_{\hat{\mu}}(\mathrm{d}\omega) \qquad (35) \\
&= E^\pi_{\hat{\mu}}(q(x_n,a_n)) = \int_X \int_X q(y,\pi) Q^n(dy|x,\pi)\hat{\mu}(\mathrm{d}x) = 0, \qquad (36)
\end{aligned}$$

where (35) holds by the induction assumption and (36) follows from (17) and (34). Thus we have shown (33) for every n.

By the monotone convergence theorem it follows that

$$\int_{\omega\in\Omega} r(\omega) P^\pi_{\hat{\mu}}(\mathrm{d}\omega) = \lim_{n\to\infty} \int_{\omega\in\Omega} r_n(\omega) P^\pi_{\hat{\mu}}(\mathrm{d}\omega) = 0.$$

Hence there exists ω such that $r(\omega) = 0$, i.e., there is a trajectory that satisfies the optimality equation.

Let L be the set of all $x \in X$ with the property that there exists a trajectory ω with $x_0 = x$ and $r(\omega) = 0$. For every $x \in L$ let $\bar{A}(x) = \{a \in A(x) : q(x,a) = 0, s(x,a) \in L\}$. By definition of L, it follows that $\bar{A}(x) \neq \emptyset$. Now we use the measurable selection theorem of Blackwell and Ryll-Nardzewski as in the proof of Theorem 8. We obtain a stationary deterministic policy f^* satisfying

$$u(x) = c(x, f^*(x)) - \tau(x, f^*(x))J^* + u(s(x, f^*(x))) \tag{37}$$

and such that $q(x, f^*(x)) = 0$ for every $x \in L$. In other words, for every $x \in L$ we have $s(x, f^*(x)) \in L$ and (37) holds.

Let now $x \in L$. Then by iteratively applying (37) for every n it follows that

$$J^* = \frac{\sum_{i=0}^{n-1} c(x_i, f^*(x_i))}{\sum_{i=0}^{n-1} \tau(x_i, f^*(x_i))} + \frac{u(x_n) - u(x)}{\sum_{i=0}^{n-1} \tau(x_i, f^*(x_i))}. \tag{38}$$

If $\tau(x_i, f^*(x_i)) = 0$ for every i, then

$$0 = \sum_{i=0}^{n-1} c(x_i, f^*(x_i)) + u(x_n) - u(x)$$

and in turn by Assumption A3

$$0 = \sum_{i=0}^{n-1} w(x_i, f^*(x_i)) + u(x_n) - u(x) \geq n + u(x_n) - u(x).$$

This can be rewritten as $u(x_n) \leq -n + u(x)$. As n tends to infinity, this yields a contradiction since by assumption u is lower bounded.

We conclude that there exists $\bar{i}$ such that $\tau(x_{\bar{i}}, f^*(x_{\bar{i}})) > 0$. For $n \geq \bar{i}$ we have that $\{\sum_{i=0}^{n-1} \tau(x, f^*(x))\}_n$ is a nondecreasing sequence of positive values and it is therefore bounded away from 0. This in turn implies by taking lim sup in (38) and considering u is bounded that $J(f^*, x) < \infty$. As n goes to infinity, the second term goes to 0 since u is bounded in X and Lemma 3. Therefore $J^* = J(f^*, x) = J(x)$. □

Under the conditions stated in Theorem 9, clearly the conclusions of Theorem 8 hold. Before proving the main result, we need two additional statements.

PROPOSITION 1. *Let $\bar{x} \in X$ be a fixed state. If Assumption* A9 *holds and if u is feasible to* (32)*, then there exists a constant M such that $-M \leq u(x) - u(\bar{x}) \leq M$ for every $x \in X$.*

Proof. Consider $x \in X$ and let (u, ρ) be a feasible solution to (32). Then by Assumption A9 with $x' = x$ and $S = \{\bar{x}\}$ there is a sequence of state-action pairs $(x_i, a_i), a_i \in A(x_i)$ for $i = 0, 1, \ldots, N-1$ such that $x_0 = x, x_N = \bar{x}$. By iteratively using (32b) for $x_i, i = 0, 1, \ldots, N-1$ and then summing up the inequalities we obtain that

$$u(x) \leq \sum_{i=0}^{N-1} c(x_i, a_i) - \rho \sum_{i=0}^{N-1} \tau(x_i, a_i) + u(\bar{x}) \leq C + |\rho| \cdot \Gamma + u(\bar{x}) \leq C + J^* \cdot \Gamma + u(\bar{x}).$$

On the other hand, again by Assumption A9 there exists a sequence of state-action pairs $(x_i, a_i), a_i \in A(x_i)$ for $i = 0, 1, \ldots, N$ with $x_0 = \bar{x}$ and $x_N = x$. Similarly as above we obtain

$$u(\bar{x}) \leq \sum_{i=0}^{N-1} c(x_i, a_i) - \rho \sum_{i=0}^{N-1} \tau(x_i, a_i) + u(x) \leq C + |\rho| \cdot \Gamma + u(x) \leq C + J^* \cdot \Gamma + u(x).$$

This completes the proof by taking $M = C + J^* \cdot \Gamma$. □

We are now ready to prove solvability of (32).

COROLLARY 2. *Under Assumptions* A1–A4 *and Assumption* A9*,* (32) *is solvable.*

Proof. Let $\{\rho_n, u_n\}_n$ be a maximizing sequence. Note that if (ρ, u) is feasible to (32), then for every $r \in \mathbb{R}$ the pair $(\rho, u - r)$ is feasible as well. Therefore $\{\rho_n, \hat{u}_n\}_n$ is a maximizing sequence as well, where $\hat{u}_n = u_n - u_n(\bar{x})$ and $\bar{x} \in X$ is a fixed state. By Proposition 1 $\hat{u}_n$ are bounded since $\hat{u}_n(\bar{x}) = 0$. By Theorem 6, we get that (32) is solvable. □

We summarize the linear programming results in the following proposition.

THEOREM 10. *Assume that Assumptions* A1–A9 *hold. The problems* (31) *and* (32) *are consistent, solvable, and there is no duality gap. There exists a nonempty set $L \subseteq X$, a deterministic stationary policy f^*, and a function $u \in \mathbb{B}_{w_0}(X)$ such that the average cost optimality equation*

$$\begin{aligned} u(x) &= \min_{a \in A(x)} \{c(x, a) - J^* \tau(x, a) + u(s(x, a))\} \\ &= c(x, f^*(x)) - J^* \tau(x, f^*(x)) + u(s(x, f^*(x))) \end{aligned} \tag{39}$$

holds for every $x \in L$ and $s(x, f^*(x)) \in L$ for every $x \in L$. In addition, for every $x \in L$, f^* is the optimal policy and

$$J^* = J(f^*, x) = J(x)$$

for every $x \in L$.

Proof. The first statement has already been proved. The last statement follows from Theorem 9 and Corollary 2. ∎

We are now ready to state the main result in the Dirac's case.

THEOREM 11. *Under Assumptions* A1–A9*, for every* $x_0 = x \in X$ *there exists an optimal deterministic stationary policy* f^**. For every* $x \in X$ *we have* $J(x) = J(f^*, x) = J^*$.

Proof. Let L and f^* be as in Theorem 10 and let f be as in Assumption A9 with respect to this particular L. Consider the deterministic stationary policy $\hat{f}$ defined for any $x \in X$ as

$$\hat{f}(x) = \begin{cases} f(x) & x \in X \setminus L, \\ f^*(x) & x \in L. \end{cases}$$

We claim that the value of this policy is J^* for any $x_0 = x \in X$, which shows the statement.

Let $x_0 = x \in X$ be an initial state. By Assumption A9 policy $\hat{f}$ leads in at most N steps to a state in L and then the policy follows f^*. It is clear that $\sum_{k=0}^{\infty} \tau(x_k, \hat{f}(x_k)) > 0$ and therefore $J(\hat{f}, x) < \infty$. By Lemma 3 it follows that

$$\sum_{k=0}^{\infty} \tau(x_k, \hat{f}(x_k)) = \infty. \tag{40}$$

For any $n \geq N$ we have

$$\begin{aligned} \frac{\sum_{k=0}^{n-1} c(x_k, \hat{f}(x_k))}{\sum_{k=0}^{n-1} \tau(x_k, \hat{f}(x_k))} &= J^* + \frac{\sum_{k=0}^{n-1} \Big(c(x_k, \hat{f}(x_k)) - J^* \tau(x_k, \hat{f}(x_k)) \Big)}{\sum_{k=0}^{n-1} \tau(x_k, \hat{f}(x_k))} \\ &= J^* + \frac{\sum_{k=0}^{N-1} \Big(c(x_k, \hat{f}(x_k)) - J^* \tau(x_k, \hat{f}(x_k)) \Big) + u(\hat{x}) - u(x_n)}{\sum_{k=0}^{n-1} \tau(x_k, \hat{f}(x_k))} \qquad (41) \\ &\leq J^* + \frac{2M}{\sum_{k=0}^{n-1} \tau(x_k, \hat{f}(x_k))}, \end{aligned}$$

where M is as in Proposition 1. (41) follows since for $x_k, k \geq N$ we have $c(x_k, \hat{f}(x_k)) - J^* \tau(x_k, \hat{f}(x_k)) = c(x_k, f^*(x_k)) - J^* \tau(x_k, f^*(x_k)) = u(x_k) - u(x_{k+1})$ by using (39).

Taking the lim sup over n on both sides and considering (40) we obtain $J(\hat{f}, x) \leq J^*$, which completes the proof. ∎

Acknowledgments. The authors thank two anonymous referees for many helpful comments. In particular, the authors acknowledge an anonymous referee for pointing out a subtle error in an early version of the manuscript.

REFERENCES

D. ADELMAN, *Price-directed replenishment of subsets: Methodology and its application to inventory routing*, Manufact. Service Oper. Manage., 5 (2003), pp. 348–371.

D. ADELMAN AND D. KLABJAN, *Duality and existence of optimal policies in generalized joint replenishment*, Math. Oper. Res., 30 (2005), pp. 28–50.

E. J. ANDERSON AND P. NASH, *Linear Programming in Infinite-Dimensional Spaces*, John Wiley, New York, 1987.

R. B. ASH, *Real Analysis and Probability*, Academic Press, New York, 1972.

R. N. Bhattacharya and M. Majumdar, *Controlled semi-Markov models under long-run average rewards*, J. Statist. Planning Inference, 22 (1989), pp. 223–242.
P. Billingsley, *Convergence of Probability Measures*, John Wiley, New York, 1968.
E. B. Dynkin and A. A. Yushkevich, *Controlled Markov Processes*, Springer-Verlag, Berlin, 1979.
B. Fox, *Markov renewal programming by linear fractional programming*, SIAM J. Appl. Math., 14 (1966), pp. 1418–1432.
O. Hernández-Lerma, *Adaptive Markov Control Processes*, Springer-Verlag, Berlin, 1989.
O. Hernández-Lerma and J. González-Hernández, *Infinite linear programming and multichain Markov control processes in uncountable spaces*, SIAM J. Control Optim., 36 (1998), pp. 313–335.
O. Hernández-Lerma and J. B. Lasserre, *Average cost optimal policies for Markov control processes with Borel state space and unbounded costs*, Systems Control Lett., 15 (1990), pp. 349–356.
O. Hernández-Lerma and J. B. Lasserre, *Linear programming and average optimality of Markov control processes on Borel spaces-unbounded costs*, SIAM J. Control Optim., 32 (1994), pp. 480–500.
O. Hernández-Lerma and J. B. Lasserre, *Average optimality in Markov control processes via discounted cost problems and linear programming*, SIAM J. Control Optim., 34 (1996a), pp. 295–310.
O. Hernández-Lerma and J. B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer-Verlag, New York, 1996b.
O. Hernández-Lerma and J. B. Lasserre, *Policy iteration for average cost Markov control processes on Borel spaces*, Acta Appl. Math., 47 (1997), pp. 125–154.
O. Hernández-Lerma and J. B. Lasserre, *Further Topics on Discrete-Time Markov Control Processes*, Springer-Verlag, Berlin, 1999.
F. Luque-Vásquez and O. Hernández-Lerma, *Semi-Markov control models with average costs*, Appl. Math., 26 (1999), pp. 315–331.
U. Rieder, *Measurable selection theorems for optimization problems*, Manuscripta Math., 24 (1978), pp. 115–131.
O. Vega-Amaya, *Average optimality in semi-Markov control models on Borel spaces: Unbounded cost and controls*, Bol. Soc. Mat. Mexicana, 38 (1993), pp. 47–60.
O. Vega-Amaya, *Zero-sum average semi-Markov games: Fixed-point solutions of the Shapley equation*, SIAM J. Control Optim., 42 (2003), pp. 1876–1894.
K. Yosida, *Functional Analysis*, Springer-Verlag, Berlin 1978.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 6, pp. 2123–2188

A REPRESENTATION THEOREM FOR THE ERROR OF RECURSIVE ESTIMATORS*

LÁSZLÓ GERENCSÉR†

Dedicated to Peter Caines in honor of his 60th birthday

Abstract. The ultimate objective of this paper is to develop new techniques that can be used for the analysis of performance degradation due to statistical uncertainty for a wide class of linear stochastic systems. For this we need new technical tools similar to those used in [L. Gerencsér, *Statist. Plann. Inference*, 41 (1994), pp. 303–325]. The immediate technical objective is to extend the previous technical results to the Djereveckii–Fradkov–Ljung scheme with enforced boundedness. Our starting point is a standard approximation of the estimation error used in the asymptotic theory of recursive estimation. Tight control of the difference between the estimation error and its standard approximation, referred to as *residuals*, is a crucial point in our applications. The main technical advance of the present paper is a set of strong approximation theorems for three closely related recursive estimation algorithms in which, for any $q \geq 1$, the L_q-norms of the residual terms are shown to tend to zero with rate $N^{-1/2-\varepsilon}$ with some $\varepsilon > 0$. This is a significant extension of previous results for the recursive prediction error or RPE estimator of ARMA processes given in [L. Gerencsér, *Systems Control Lett.*, 21 (1993), pp. 347–351]. Two useful corollaries will be derived. In the first a standard transform of the estimation-error process for the basic recursive estimation method, Algorithm CR, will be shown to be L-mixing, while in the second the asymptotic covariance matrix of the estimator for the same method will be given. Applications to multivariable adaptive prediction and the minimum-variance self-tuning regulator for ARMAX systems will be described.

Key words. adaptive prediction, stochastic complexity, recursive estimation, L-mixing processes, asymptotic covariance, stochastic adaptive control

AMS subject classifications. 93E12, 93E35

DOI. 10.1137/S0363012991217421

1. Introduction.

The ultimate objective of this paper is to develop new techniques for the analysis of performance degradation due to statistical uncertainty for a wide class of linear stochastic systems. Performance degradation due to statistical uncertainty, called *regret*, following [46], can be computed at a single time moment, yielding instantaneous regret, or it can be summed over time, yielding cumulative regret. The objective of the paper is to develop new techniques that can be used for analyzing the pathwise (almost sure) asymptotics of the cumulative regret for a class of adaptive prediction and stochastic adaptive control problems.

A number of examples on the interaction of identification and control are available in the *identification for control* literature, see [29, 40, 41]. While those papers contain fundamentally new ideas, the analysis they present contains heuristic elements. In particular, they assume the independence of actually weakly dependent quantities in order to simplify the computation of the instantaneous regret. The present paper lays the foundations for a rigorous discussion of these heuristic arguments. Special examples of these new technical tools have been developed in the context of adaptive prediction of ARMA processes in [24].

*Received by the editors August 12, 1991; accepted for publication (in revised form) April 5, 2005; published electronically January 26, 2006. This research was supported by the Natural Sciences and Engineering Research Council of Canada under grant 01329 and by the National Research Foundation of Hungary under grant T 047193.

http://www.siam.org/journals/sicon/44-6/21742.html

†MTA SZTAKI (Computer and Automation Research Institute of the Hungarian Academy of Sciences), Kende 13-17, H-1111 Budapest, Hungary (gerencser@sztaki.hu).

The immediate *technical objective* is a detailed analysis of the Djereveckii–Fradkov–Ljung (DFL) scheme with enforced boundedness, given as *Algorithm DFL*, (3.53)–(3.54). This is a practically useful recursive estimation method introduced in [11, 12, 51] with a wide range of applications; see [3, 53]. The algorithm in its original form is given under (3.50) and (3.51) which is a potentially divergent procedure. To ensure convergence the original method is modified by enforced boundedness, a device that has been widely used in practice and rigorously analyzed in [19]. The study of the DFL scheme is reduced to the study of two related stochastic approximation methods, Algorithm DR (discrete-time recursion) and Algorithm CR (continuous-time recursion), described in section 3. The conditions under which these methods are analyzed are very close to what we had in [19]. However, a critical condition imposed on the initialization of the process has been significantly simplified. Our conditions will be compared with other conditions used in the literature, in particular with those in [3], with emphasis on the so-called "boundedness condition."

Asymptotic properties of recursive estimation processes are established in classical theory by using a series of approximations (see, e.g., [54]). Thus we get a standard approximation of the error term, see, e.g., [65] for a lucid exposition, for which limit results are easily established. Tight control of the difference between the estimation error and its standard approximation, that will be referred to as *residuals*, is crucial in the analysis of performance degradation due to statistical uncertainty; see [24].

The *main technical advance* of the present paper is a set of strong approximation theorems for three closely related recursive estimation algorithms, given as Theorems 4.1–4.3, in which, for any $q \geq 1$, the L_q-norms of the residual terms are shown to tend to zero with rate $N^{-1/2-\varepsilon}$ with some $\varepsilon > 0$. This is a significant extension of a previous result given in [19], where only the rate of convergence for the L_q-norms of the estimation error has been established and the explicit approximation of the estimation error and the residual term is not discussed at all. It extends also the result of [22] on the residual of the recursive prediction error estimator for ARMA processes. The proof is quite demanding: in addition to some basic inequalities developed in [17] the proof relies on [19] and uses a nontrivial moment inequality for weighted multiple integrals of L-mixing processes given in [21]. Preliminary versions of the results of section 4 have been formulated in [20].

In comparison the material of sections 5 and 6 are relatively straightforward corollaries demanding numerous small steps, though. In Theorem 5.1 a standard transform of the estimation-error process for the basic recursive estimation method, Algorithm CR, will be shown to be L-mixing, while in Theorem 6.2 the asymptotic covariance matrix of the estimator for the same method will be given.

The *significance* of the results of the present paper is demonstrated by describing two applications in section 7. In the first example the pathwise cumulative regret is quantified for an online adaptive predictor of multivariable linear stochastic systems. In the second example a similar measure of performance degradation for the minimum-variance self-tuning regulator is computed. Both applications follow the arguments of [24], but heavily rely on the results of the present paper. A further application for indirect adaptive control of multivariable linear stochastic systems is given in [27]. We think that the results are tailored to the needs of the users and they will pave the way to many further applications.

To motivate the studies carried out in this paper we will first give an illuminative application of less known technical results on off-line prediction error identification methods for ARMA processes. The application, given as Theorem 2.1, provides the answer to a basic problem of the theory of *stochastic complexity*, developed by Rissa-

nen; see [58]: the performance degradation of adaptive predictors. The extension of this results to adaptive predictors using online estimation requires the extension of the relevant technical tools. First, a strong approximation result for recursive prediction error identification methods for ARMA processes will be given as Theorem 2.4, this is also the starting point for the investigations of the present paper. Two important corollaries are Theorems 2.5 and 2.6. The relevance of these results in analyzing performance degradation in the context of online adaptive prediction of ARMA processes will be described, culminating in Theorem 2.7. This theorem will be considered as a benchmark in future applications.

2. Adaptive prediction. Basic notions and conditions. An adaptive predictor for ARMA processes is obtained if we use estimated system-parameters in the prediction equation at time n as if it was the true value. Then we may ask, how much do we lose in prediction accuracy due to the inexact knowledge of the parameters. First, we consider adaptive predictors using off-line estimation and indicate the nature of technical results that are needed for the analysis. Then, using the strong approximation result (2.23) we arrive at analogous technical results for recursive estimation, which in turn can be applied to derive interesting properties of real-time adaptive predictors.

The set of real numbers will be denoted by $\mathbb{R}$, the p-dimensional Euclidean space will be denoted by $\mathbb{R}^p$. The Euclidean norm of $x \in \mathbb{R}^p$ will be denoted by $|x|$. We shall often use subscripts to indicate partial derivatives.

Let (y_n), $0 \le n < \infty$, be a wide-sense stationary ARMA (p, q) process satisfying the difference equation

$$\sum_{i=0}^{p} b_i^* y_{n-i} = \sum_{j=0}^{q} c_j^* e_{n-j},$$

or in shorthand notation $B^*y = C^*e$, where B^* and C^* are polynomials of the backward-shift operator of degree p and q, respectively. Define the polynomial $B^*(z^{-1}) = \sum_{i=0}^{p} b_i^*(z^{-i})$ and similarly $C^*(z^{-1})$. To estimate the system-parameters b_i^*, c_j^* from observed data (y_n) using the prediction error method the following technical assumption is assumed.

CONDITION 2.1. *The polynomials B^*, C^* are stable and relative prime, $b_0^* = c_0^* = 1$ and $b_p^* \neq 0$, $c_q^* \neq 0$.*

The condition $b_p^* \neq 0$, $c_q^* \neq 0$ has been assumed to allow the extension of our results to cases when the degree of one of the polynomials B^* or C^*, but not both, is overestimated. The relevant work that we use is [1]. To characterize the noise process we shall need the following definition that has been introduced in [17].

DEFINITION 2.1. *We say that a discrete-time $\mathbb{R}^p$-valued stochastic process (u_n) is* M-bounded *if, for all $1 \le q < \infty$,*

$$M_q(u) := \sup_{n \ge 0} \mathrm{E}^{1/q} |u_n|^q < \infty. \tag{2.1}$$

In this case we also write $u_n = O_M(1)$. For a stochastic process $(z_n), n \ge 0$, and a positive sequence (c_n) we write $z_n = O_M(c_n)$ if $u_n = z_n/c_n = O_M(1)$.

A basic tool that we will use is the theory of L-mixing processes, elaborated in [17] and used to solve some hard problems in system identification; see [18, 19, 22, 38, 43]. This concept is a generalization of what is called "exponentially stable processes" in the system identification literature, see Definition 3.1 in Section 8.3 of [6] or [53]. We

give the definition here for discrete-time processes. Let a probability space $(\Omega, \mathcal{F}, P)$ be given together with a pair of families of σ-algebras $(\mathcal{F}_n, \mathcal{F}_n^+), n = 0, 1, \dots$, such that (i) $\mathcal{F}_n \subset \mathcal{F}$ is monotone increasing, (ii) $\mathcal{F}_n^+ \subset \mathcal{F}$ is monotone decreasing, and (iii) $\mathcal{F}_n$ and $\mathcal{F}_n^+$ are independent for all n. For $n < 0$ we set $\mathcal{F}_n^+ = \mathcal{F}_0^+$.

DEFINITION 2.2. *A stochastic process $u = (u_n)$, $n = 0, 1, \dots$, is L-mixing with respect to $(\mathcal{F}_n, \mathcal{F}_n^+)$ if it is $\mathcal{F}_n$-adapted, M-bounded, and for all $q \geq 1$, with $\tau = 0, 1, \dots$ and*

$$\gamma_q(\tau, u) = \gamma_q(\tau) = \sup_{n \geq \tau} \mathrm{E}^{1/q} |u_n - \mathrm{E}\ (u_n | \mathcal{F}_{n-\tau}^+)|^q,$$

we have

$$\Gamma_q = \Gamma_q(u) = \sum_{\tau=0}^{\infty} \gamma_q(\tau) < \infty. \tag{2.2}$$

The process u is L^+-mixing if, in addition, for all $q \geq 1$ there exist $C_q, c_q > 0$ such that for all nonnegative integers τ,

$$\gamma_q(\tau, u) \leq C_q (1 + \tau)^{-1-c_q}.$$

Discussion of L-mixing. The prime example for L-mixing processes is a sequence of i.i.d. random variables with finite moments of all orders. The response of an exponentially stable linear filter, with an L-mixing process as its input, is L-mixing. Products of L-mixing processes are also L-mixing. These properties make sure that the *verification* of L-mixing is typically easy in problems of system identification. The same invariance properties hold for the class of L^+-mixing processes. For "exponentially stable processes" we would require that $\gamma_q(\tau)$ converges to 0 geometrically fast, at least for some values of q, typically for $q = 4$. We shall need conditions for higher order moments to derive sharp bounds for the error terms in certain uniform laws of large numbers.

CONDITION 2.2. *The system-noise process (e_n), $0 \leq n < \infty$, defined over an underlying probability space $(\Omega, \mathcal{F}, P)$, is an M-bounded process. Moreover there is an increasing sequence of σ-fields $(\mathcal{F}_n)$, $0 \leq n < \infty$, $(\mathcal{F}_n) \subset \mathcal{F}$, such that (e_n) is a martingale difference process with constant conditional variance:*

$$\mathrm{E}(e_n | \mathcal{F}_{n-1}) = 0, \qquad \mathrm{E}(e_n^2 | \mathcal{F}_{n-1}) = \sigma^2 = \mathrm{const.}$$

almost surely. Finally, we assume that (e_n) is L-mixing with respect to a pair of families of σ-algebras $(\mathcal{F}_n, \mathcal{F}_n^+)$.

It follows that (e_n) is a wide-sense stationary orthogonal process. Conditions 2.1 and 2.2 together will be called the *standard conditions* for ARMA processes.

Discussion of the moment condition. The difference between our conditions and the conditions given in standard works such as [6] or [35] is that, there only the condition

$$M_4(e) := \sup_{n \geq 0} \mathrm{E}^{1/4} |e_n|^4 < \infty$$

is required. See the comment to Definition 3.1 in Section 8.3 of [6], or condition (4.1.20) of [35]. Thus our condition is much stronger, but our conclusions given in Theorem 2.2 will be also significantly stronger than the results of [6, 35] given in

terms of classical concepts such as strong consistency, central limit theorem, or the law of iterated logarithm, which all follow from our result and the corresponding result for martingales. M-boundedness could be relaxed by requiring the uniform boundedness of moments of *sufficiently high order*, however the order of the moments would depend on the order of the ARMA process, i.e., on p and q. This is due to the fact that in our proof we rely on Kolmogorov's continuity theorem for random fields to get sharp bounds for the error terms in uniform laws of large numbers, which requires the existence of finite moments up to an order strictly greater than $(p+q)$ in the present application; see Theorem 8.3.

Set $\theta^* = (b_1^*, \dots, b_p^*, c_1^*, \dots, c_q^*)^T$. Let $D_C \subset \mathbb{R}^q$ denote the set of vectors $(c_1, \dots, c_q)$ such that the corresponding polynomial C^* is stable, let $D_B = \mathbb{R}^p$ and let

$$D_\theta = D_B \times D_C \subset \mathbb{R}^{p+q}.$$

Let $D_{\theta 0} \subset D_\theta$ be a compact domain such that $\theta^* \in \operatorname{int} D_{\theta 0}$. Then the *prediction error method* for estimating the parameter θ^* is defined as follows (cf., e.g., [6, 35]): first take an arbitrary $\theta \in D_{\theta 0}$ and define an estimated prediction error process $\bar{\varepsilon} = (\bar{\varepsilon}_n(\theta))$ by the inverse equation

$$C\bar{\varepsilon} = By \tag{2.3}$$

using zero initial conditions. Define the cost function

$$V_N(\theta) = \frac{1}{2} \sum_{n=1}^{N} \bar{\varepsilon}_n^2(\theta).$$

Minimizing $V_N(\theta)$ over $D_{\theta 0}$ yields an estimate $\widehat{\theta}_N$.

A precise definition of $\widehat{\theta}_N$ taking into account the possibility of the existence of several local minima can be given as follows: let $\Omega' \subset \Omega$ be a measurable set such that the equation

$$\frac{\partial}{\partial \theta} V_N(\theta) = 0$$

has a unique solution in the interior of $D_{\theta 0}$ denoted by $\operatorname{int} D_{\theta 0}$ on the event $\Omega' \subset \Omega$. Then this solution will be accepted as $\widehat{\theta}_N$ on Ω', while $\widehat{\theta}_N$ is defined as an arbitrary $D_{\theta 0}$-valued random variable on $\Omega \setminus \Omega'$. It can be shown that we can take Ω' so that $P(\Omega') > 1 - C_q N^{-q}$ for any $q > 0$, see [18, Lemma 2.1], the proof of which is partially based on [1].

Remark. The uniqueness result of [1] remains valid if we redefine D_θ so that the degree of one of the polynomials B^* or C^*, but not both, is overestimated. This is why $b_p^* \neq 0$, $c_q^* \neq 0$ has been assumed in Condition 2.1.

The quantity to be studied in the context of adaptive prediction is the prediction error $\bar{\varepsilon}_n(\widehat{\theta}_{n-1})$. We ask how much do we lose in prediction accuracy due to the statistical uncertainty present in $\widehat{\theta}_{n-1}$. A basic result says that, assuming that the standard conditions, Conditions 2.1 and 2.2, are satisfied then the excess in mean prediction error, also called the *regret*, see [46], satisfies

$$\mathrm{E}(\bar{\varepsilon}_n^2(\widehat{\theta}_{n-1}) - e_n^2) = \frac{\sigma^2(e)}{n}(p+q)(1+o(1)). \tag{2.4}$$

This result is given in [24] and, under different conditions in [64]. It extends the result of [8] for AR processes. A similar result for the cumulative regret for Gaussian linear regression was proved in [56] (see also Theorem 5.3 in [58]).

Summation over n in (2.4) gives that the left-hand side is asymptotically equivalent to $\sigma^2(e)(p+q)\log N$. It has been shown in Theorem 1.1. of [24] that we can remove the expectation operator and we get the following *pathwise* result for the cumulative regret.

THEOREM 2.1. *Assume that the standard conditions, Conditions* 2.1 *and* 2.2*, are satisfied. Then*

$$\lim_{N\to\infty}\frac{1}{\log N}\lim_{N\to\infty}\sum_{n=1}^{N}(\bar{\varepsilon}_n^2(\widehat{\theta}_{n-1})-e_n^2)=\sigma^2(e)(p+q)\quad\text{a.s.} \tag{2.5}$$

This result, under different conditions, was given for AR processes in [36, 37]. The much more difficult ARMA case was solved in [24, 64], using different conditions and different methods. Note that classical limit theorems are not suitable to derive (2.5). Both results, (2.4) and (2.5), play prominent role in the theory of stochastic complexity. The quantity

$$C_{1,N}=\sum_{n=1}^{N}\bar{\varepsilon}_n^2(\widehat{\theta}_{n-1}) \tag{2.6}$$

is called a predictive stochastic complexity in [58].

Technical tools: Strong approximations. Now we come to some technical details that are essential in the proof of the above results. A key point is a characterization of the estimation error process which is more accurate than previously known results. Define the asymptotic cost function by

$$W(\theta)=\lim_{n\to\infty}\frac{1}{2}\mathrm{E}\,\bar{\varepsilon}_n^2(\theta).$$

In the Gaussian case this is the asymptotic log-likelihood function modulo constants. It is easy to see that $W_\theta(\theta^*)=0$, where $_\theta$ denotes differentiation with respect to θ. Also it is well known that

$$R^*=W_{\theta\theta}(\theta^*)=\lim_{n\to\infty}\mathrm{E}\,\bar{\varepsilon}_{\theta n}(\theta^*)\bar{\varepsilon}_{\theta n}(\theta^*)^T$$

is nonsingular and in fact positive definite. Then we have the following representation of the estimation error (cf. [18]).

THEOREM 2.2. *Assume that the standard conditions for ARMA processes, Conditions* 2.1 *and* 2.2*, are satisfied, then we have*

$$\widehat{\theta}_N-\theta^*=-(R^*)^{-1}\frac{1}{N}\sum_{n=1}^{N}\bar{\varepsilon}_{\theta n}(\theta^*)e_n+O_M(N^{-1}). \tag{2.7}$$

The *main contribution* of Theorem 2.2 is that the residual term has been shown to be of the order of magnitude $O_M(N^{-1})$. This is an improvement over the classical results of [50, 59] in the ARMA case. The significance of this improvement is easily demonstrated: the residual term is sufficiently small so that limit theorems such as the law of iterated logarithms (LIL) and invariance principles for the estimator process

can be immediately derived using martingale limit theory (cf. [33]). But the real motivation behind Theorem 2.2 had been the need to verify Rissanen's tail condition for Gaussian ARMA processes, introduced in the seminal paper [57], which in turn can be used to derive a lower bound for the cumulative loss in performance of any adaptive predictor for Gaussian ARMA processes (cf. [26]).

Discussion of mixing conditions. There are a number of other notions of mixing. The best known notion is ϕ-mixing, an excellent and concise introduction to which is given in Chapter 7.2 of [15]. The measure of mixing for two σ-algebras $\mathcal{G}$ and $\mathcal{H}$ is defined for any $1 \leq p \leq \infty$ as follows:

$$\phi_p(\mathcal{G}|\mathcal{H}) = \sup_{A \in \mathcal{G}} ||P(A|\mathcal{H}) - P(A)||_p, \tag{2.8}$$

where $||\xi||_p$ denotes the L_p-norm of the random variable ξ. It can be shown that for $p = 1$ we have $\frac{1}{2}\phi_1(\mathcal{G}|\mathcal{H}) = \alpha(\mathcal{G}, \mathcal{H})$, where

$$\alpha(\mathcal{G}, \mathcal{H}) = \sup_{A \in \mathcal{G}, B \in \mathcal{H}} |P(AB) - P(A)P(B)| \tag{2.9}$$

is the familiar measure of strong mixing. Similarly, we have for $p = \infty$

$$\phi_\infty(\mathcal{G}|\mathcal{H}) = \sup_{A \in \mathcal{G}, B \in \mathcal{H}, P(B) > 0} |P(A|B) - P(A)| \tag{2.10}$$

which is the familiar measure of uniform mixing. A stochastic process (x_n) is then ϕ_p-mixing if with

$$\phi_p(n) = \phi_p(\mathcal{F}_n|\mathcal{F}_n^+),$$

where now $\mathcal{F}_n = \sigma\{x_i, i \leq n\}$, $\mathcal{F}_n^+ = \sigma\{x_i, i \geq n\}$, we have $\lim_{n\to\infty} \phi_p(n)$.

In contrast to L-mixing, the *verification* of even the weakest form of ϕ-mixing, which is called for historical reasons strong mixing or α-mixing, is nontrivial even for Gaussian processes (see Chapter 17 of [42]). On the other hand, measurable static functions of ϕ-mixing processes are ϕ-mixing, while this may not be the case, e.g., for discontinuous functions of L-mixing processes. (For a positive statement see [23, Theorem II.7].)

From the point of view of *usefulness* both notions are equally useful for *off-line estimation*. Namely, the key technical device in analyzing off-line estimators is a kind of improved Hölder inequality, see Lemma 8.1 of section 8, or Chapter 7.2 of [15], or Appendix III of [33]. In fact, it can be shown that the theorem remains valid even if the assumption that (e_n) is L-mixing is completely removed, since the remaining conditions imply the validity of certain improved Hölder inequalities. The situation is quite different for recursive estimation methods, where L-mixing is heavily exploited. Further discussion on this will be given in section 3.

The first step in the proof of Theorem 2.1 is to consider a second-order Taylor series expansion of the terms on the left-hand side. The estimation error process is handled using a standard transformation in the stochastic approximation literature. Define a piecewise constant continuous-time extension of $\widehat{\theta}_n$, and, denoting the time variable by t, introduce a new process by first normalizing $(\widehat{\theta}_t - \theta^*)$ to $t^{1/2}(\widehat{\theta}_t - \theta^*)$ and then using an exponential change of time-scale $t = e^s$. Thus we get a new process

$$\psi_s = e^{s/2}(\widehat{\theta}_{e^s} - \theta^*). \tag{2.11}$$

A key observation is that the *transformed process* (ψ_s) is L-mixing with respect to $(\mathcal{F}_{e^s}, \mathcal{F}^+_{e^s})$. For the definition of L-mixing in continuous time see the next section. The proof of this fact is based on Theorem 2.2 and the following simple result given as Theorem 3.3 in [24].

LEMMA 2.1. *Let $(u_t), t \geq 0$, be a zero-mean L-mixing process with respect to some pair of families of σ-algebras $(\mathcal{F}_t, \mathcal{F}^+_t)$. Let*

$$x_T = T^{-1/2} \int_1^T u_t dt. \tag{2.12}$$

Then the process $(y_s) = (x_{e^s})$ is L-mixing with respect to the pair of families of σ-algebras $(\mathcal{F}_{e^s}, \mathcal{F}^+_{e^s})$.

With this observation it can be shown that the process $\bar{\varepsilon}_n(\theta^*)$ and its gradient are asymptotically independent of $(\widehat{\theta}_N - \theta^*)$, which is exploited in proving (2.4). On the other hand, it can be shown that the result given in (2.5) is essentially a law of large numbers in the new time-scale.

Now we come to the extensions of the above results for the case of online or *recursive estimation* of θ^*. The most widely used recursive estimation methods for ARMA processes is the recursive prediction error (RPE) method, which in the case of Gaussian processes reduces to the recursive maximum-likelihood (RML) method; see [6, 53]. This procedure serves as a *benchmark* for the general theory to be developed in section 3, in particular it is a prime example for the Djereveckii–Fradkov–Ljung scheme or DFL scheme. For both theoretical and practical reasons we consider RPE estimator processes $\widehat{\widehat{\theta}}_n$ using a resetting mechanism to enforce the boundedness of the estimator. The convergence analysis for such a procedure has been given in Theorem 4.2 of [19].

We will now give the details of the RPE method for ARMA processes and a set of technical conditions that we use to guarantee convergence. The conditions are simpler than those given in section 4 of [19]. We will shortly indicate how the present conditions imply the conditions given for the DFL scheme in the next section. Most of the discussion of these conditions will be deferred to the next section.

The definition of the RPE method *without resetting* is

$$\widehat{\widehat{\theta}}_{n+1} = \widehat{\widehat{\theta}}_n - \frac{1}{n+1} (\widehat{\widehat{R}}_n)^{-1} \frac{\partial}{\partial \theta} \varepsilon_{n+1} \cdot \varepsilon_{n+1}, \tag{2.13}$$

$$\widehat{\widehat{R}}_{n+1} = \widehat{\widehat{R}}_n + \frac{1}{n+1} \left(\left(\frac{\partial}{\partial \theta} \varepsilon_{n+1} \right) \left(\frac{\partial}{\partial \theta} \varepsilon_{n+1} \right)^T - \widehat{\widehat{R}}_n \right) \tag{2.14}$$

with some initial conditions $(\widehat{\widehat{\theta}}_0, \widehat{\widehat{R}}_0)$, where ε_n and $\frac{\partial}{\partial \theta} \varepsilon_n$ denote online estimates of $\bar{\varepsilon}_n(\theta^*)$ and $\frac{\partial}{\partial \theta} \bar{\varepsilon}_n(\theta)|_{\theta=\theta^*}$. These are obtained by using the most recent estimations of B^* and C^* in the linear filters defining the current values of $\bar{\varepsilon}_n(\theta^*)$ and $\frac{\partial}{\partial \theta} \bar{\varepsilon}_n(\theta)|_{\theta=\theta^*}$.

Thus, e.g., ε_{n+1} is defined by the time-varying filter

$$(\widehat{\widehat{C}}_n \varepsilon)_{n+1} = (\widehat{\widehat{B}}_n y)_{n+1}. \tag{2.15}$$

For further details see [53].

The RPE method without resetting is a special case of a general recursive estimation scheme, called the DFL scheme, to be described in details in the next section; see (3.50)–(3.51). Note that together with θ^* we also estimate the matrix R^*.

It is well known from simulation examples that the RPE method may diverge, unless some precaution is taken. This difficulty is often dealt with a controversial "boundedness condition" first formulated in [51]. This will be discussed in detail in the context of the DFL method. A convergent *truncated* RPE method has been given in section 4 of [19], which we now describe. Let

$$D_R = \mathbb{R}^+(p \times p) \quad \text{and} \quad D = D_\theta \times D_R,$$

where $\mathbb{R}^+(p \times p)$ denotes the set of symmetric, positive definite $p \times p$ matrices. Let $D_{\theta 0} \subset D_\theta$ be a compact set containing θ^* in its interior, and similarly let $D_{R0} \subset D_R$ be a compact set containing R^* in its interior and let $D_0 = D_{\theta 0} \times D_{R0}$.

Resetting. If at any time n the next estimator $(\hat{\hat{\theta}}_{n+1}, \hat{\hat{R}}_{n+1})$ would leave D_0, then we redefine its value by resetting it to the initial value, i.e.,

$$\text{for} \quad (\hat{\hat{\theta}}_{n+1}, \hat{\hat{R}}_{n+1}) \notin \operatorname{int} D_0 \quad \text{reset as} \quad (\hat{\hat{\theta}}_{n+1}, \hat{\hat{R}}_{n+1}) := (\hat{\hat{\theta}}_0, \hat{\hat{R}}_0). \tag{2.16}$$

To avoid being trapped to the boundary of the truncation domain the initial value $(\hat{\hat{\theta}}_0, \hat{\hat{R}}_0)$ must be aligned to $D_{\theta 0} \times D_{R0}$, as described in Condition 3.4. This condition is given in terms of the so-called associated ordinary differential equation (ODE). Define for $\theta \in D_\theta$

$$R^*(\theta) = \lim_{n \to \infty} \mathrm{E} \bar{\varepsilon}_{\theta n}(\theta) \bar{\varepsilon}_{\theta n}(\theta)^T. \tag{2.17}$$

Then obviously $R^* = R^*(\theta^*)$. With this notation the associated ODE, with the time variable v, is defined as

$$\begin{aligned} \dot{\theta}_v &= -R_v^{-1} \frac{\partial}{\partial \theta} W(\theta_v), \\ \dot{R}_v &= R^*(\theta_v) - R_v. \end{aligned} \tag{2.18}$$

The right-hand side is defined in $D_\theta \times D_R$. This is the usual way of defining the associated ODE; see [3, 53]. However in [19] as well as later in this paper we will define the associated ODE by using a change of time-scale $t = e^v$.

The condition ensuring that resetting works for the general recursive estimation methods given in section 3, including the DFL scheme is Condition 3.4. Following the arguments of section 4 of [19] it is easy to see that the first part of Condition 3.4, requiring a certain kind of asymptotic stability of the associated ODE, follows for the RPE method. Namely, it follows directly from [1] that (2.18) has a unique stationary point in $D_\theta \times D_R$, which is (θ^*, R^*). It is also easy to see that this equilibrium point is asymptotically stable, since the eigenvalues of the Jacobian matrix of the right-hand side of the ODE at (θ^*, R^*) are all -1. Now it is easy to show that the associated ODE is globally asymptotically stable in $D_\theta \times D_R$. For the proof we need the observation that $W(\theta_v)$ is nonincreasing as long as R_v is positive definite, and R_v is bounded and positive definite as long as θ_v belongs to a fixed compact set.

Let $x_n = (\hat{\hat{\theta}}_n, \hat{\hat{R}}_n)$ denote the estimator at time n, let $z = (\theta, R)$ denote a running parameter, and let $z(v, u, \xi)$ denote the solution of (2.18) with initial value ξ at time u. Then it follows that for every $\xi \in D_0$, $v \geq u \geq 0$ the solution $z(v, u, \xi) \in D$ is defined for $1 \leq s \leq t < \infty$, it converges to $x^*(\theta^*, R^*)$ for $t \to \infty$ and we have, with some C_0 and $\alpha = 1 - c$ with arbitrary small $c > 0$,

$$\left\| \frac{\partial}{\partial \xi} z(v, u, \xi) \right\| \leq C_0 e^{-\alpha(v-u)}. \tag{2.19}$$

It follows, using a change of time-scale $t = e^v$, that the first part of Condition 3.4 is satisfied. Here $\|\cdot\|$ denotes the operator norm of a matrix.

To ensure the validity of the second part of Condition 3.4 we have to assume that some a priori knowledge of the system-parameters θ^* and the Hessian R^*, say $\xi_0 = (\widehat{\widehat{\theta}}_0, \widehat{\widehat{R}}_0)$ are available. They can be obtained, e.g., from an off-line estimation.

CONDITION 2.3. *Let $D_0 = D_{\theta 0} \times D_{R0} \subset D_\theta \times D_R$ be a compact truncation domain such that $x^* = (\theta^*, R^*) \in \text{int } D_0$.* (i) *It is assumed that there exists a compact convex domain $D_0' \subset D$ such that for all $v \geq u \geq 0$,*

$$z(v,u,\xi) \in D_0' \quad \text{for } \xi \in D_0 \quad \text{and } z(v,u,\xi) \in D \quad \text{for } \xi \in D_0'. \tag{2.20}$$

(ii) *It is assumed that we have an initial estimate $\xi_0 = (\widehat{\widehat{\theta}}_0, \widehat{\widehat{R}}_0)$ such that for any $v \geq u \geq 0$ we have $z(v,u,\xi_0) \in \text{ int } D_0$.*

Remark. Since our objective is to restate Theorem 4.2 of [19], part (iii) of Condition 3.4 of the present paper need not be required at this time, since it was not required in [19] either; it is special addition for the present paper.

To ensure the stability of the time-varying filter (2.15) given as $(\widehat{\widehat{C}}_n\varepsilon)_{n+1} = (\widehat{\widehat{B}}_n y)_{n+1}$ we need a second condition imposed on the truncation domain (cf. Condition 3.7 given for the DFL method). Let us consider a fixed state-space realization of the inverse system (2.3) and let the state-transition matrix be denoted by $\widetilde{C}$. In [19] this is given as the so-called companion matrix corresponding to the polynomial C (see Condition 4.5 of [19]). Let $D_{B0} \subset D_B$ and $D_{C0} \subset D_C$ be compact domains and let

$$D_{\theta 0} = D_{B0} \times D_{C0}.$$

Now Condition 3.7 would read as follows.

CONDITION 2.4. *Let $D_{\tilde{C}0}$ denote the set of matrices $\tilde{C}$, when C is taken from D_{C0}. Then $D_{\tilde{C}0}$ is jointly stable in the sense that there exists a single $q \times q$ symmetric positive definite matrix U and $0 < \lambda < 1$ such that for all $\tilde{C} \in D_{\tilde{C}0}$,*

$$\widetilde{C}^T U \widetilde{C} \;\leq\; \lambda U.$$

It follows that there exists some $c > 0$ such that for any sequence $(\tilde{C}_n)$ with $\tilde{C}_n \in D_{\tilde{C}0}$ we have

$$||\tilde{C}_n \cdots \tilde{C}_0|| \leq c\lambda^{n/2}. \tag{2.21}$$

A discussion of the joint stability condition. Condition 2.4 above is required only to ensure that (2.21) holds. In the system identification literature it had been occasionally implicitly assumed that the individual stability of each $\tilde{C} \in D_{\tilde{C}0}$ implies (2.21); see, e.g., [34]. This is easily seen to be wrong. One way to ensure Condition 2.4 is to choose the truncation domain D_0 small, but this is obviously not practical. A better way is to use a suitable realization of the inverse system (2.3). To indicate the potential of alternative realizations let us consider a Gilbert–Kalman realization of the inverse system (see [44]). Assume that the roots of the polynomial $C = C(z^{-1})$ are all real and simple and let them be denoted by λ_i. Then we will have

$$\tilde{C} = \text{diag}(\lambda_i),$$

and obviously any compact set of matrices $D_{\tilde{C}0}$ is jointly stable. Potentially useful alternative realizations are given in [55]. A second way of ensuring the validity of

(2.21) is given in [3]. This will be discussed in connection with the DFL scheme in the next section.

Finally we will need two additional conditions for the noise process. First, the M-boundedness of (e_n) is further strengthened by assuming the existence and boundedness of certain exponential moments.

CONDITION 2.5. *We assume that $|e_n|^2$ is in class M^*; i.e., for some $\varepsilon > 0$ we have*

$$\sup_n \mathrm{E}\ \exp \varepsilon |e_n|^2 < \infty.$$

This condition is certainly satisfied if (e_n) is a stationary Gaussian process. The role of this condition will be discussed in section 3 in the context of the DFL scheme.

Secondly, we need to be more specific on the mixing rate of (e_n). The condition to follow is motivated by Lemma 3.1 in [24], which states for continuous-time L-mixing processes (cf. Definition 3.2) that, if (u_t) is an L-mixing process, then $\gamma_q(\tau, u) \leq 4\Gamma_q(u)/\tau$ for all $q \geq 1$ and $\tau \geq 0$. The validity of a slightly stronger inequality is required by the following condition in discrete time (cf. Condition 3.9 given for the DFL scheme).

CONDITION 2.6. *We assume that (e_n) is L^+-mixing with respect to a pair of families of σ-algebras $(\mathcal{F}_n, \mathcal{F}_n^+)$.*

The role of this condition is in the analysis of the difference between the "frozen parameter" process $\bar{\varepsilon}_n(\theta)$ evaluated at $\theta = \widehat{\theta}_n$ and its online estimate ε_n, see [19, Lemma 5.6] restated as Lemma 3.2 of the present paper. From the purely technical point of view, L^+-mixing is used in [19, Theorem 6.1]. In view of the general theorem for the DFL scheme, given as Theorem 3.3, we get the following result (see also Theorem 4.2 of [19]):

THEOREM 2.3. *Let (y_n) be an ARMA process satisfying the standard conditions, Conditions* 2.1 *and* 2.2. *Consider the RPE estimator $(\widehat{\widehat{\theta}}_N, \widehat{\widehat{R}}_N)$ defined by* (2.13), (2.14), *modified by a resetting mechanism given under* (2.16). *Let the truncation domain be of the form*

$$D_0 = D_{\theta 0} \times D_{R0} \quad \text{with} \quad D_{\theta 0} = D_{B0} \times D_{C0}.$$

Assume that D_0 satisfies Condition 2.3 *and D_{C0} satisfies Condition* 2.4. *Finally let the innovation process satisfy the additional conditions, Conditions* 2.5 *and* 2.6. *Then for the recursive estimators $(\widehat{\widehat{\theta}}_N, \widehat{\widehat{R}}_N)$ we have*

$$\widehat{\widehat{\theta}}_N - \theta^* = O_M(N^{-1/2}) \quad \textit{and} \quad \widehat{\widehat{R}}_N - R^* = O_M(N^{-1/2}). \tag{2.22}$$

One of the special features of this result is that the *moments* of the estimation error are bounded from above. While the above theorem is certainly of interest, it is obviously much weaker than the characterization of the off-line estimator given in Theorem 2.2. But Theorem 2.3 is a key technical tool in deriving a strong approximation theorem relating the RPE estimator to the off-line prediction error estimator. This result is given in [22], stating that under the conditions of Theorem 4.2 of [19] (and thus under the conditions of Theorem 2.3) we have

$$\widehat{\widehat{\theta}}_N - \widehat{\theta}_N = O_M\left(\frac{\log N}{N}\right). \tag{2.23}$$

Combining (2.23) with Theorem 2.2 we get the following result.

THEOREM 2.4. *Under the conditions of Theorem* 2.3 *we have*

$$\widehat{\widehat{\theta}}_N - \theta^* = -(R^*)^{-1}\frac{1}{N}\sum_{n=1}^{N}\overline{\varepsilon}_{\theta n}(\theta^*)e_n + O_M\left(\frac{\log N}{N}\right). \tag{2.24}$$

This strong approximation result provides a very precise characterization of $\widehat{\widehat{\theta}}_N$. The control of moments of the residual term is an essential feature of the result that is very much exploited in deriving Theorems 2.7 and 2.8. A direct corollary of the above theorem is the following.

THEOREM 2.5. *Under the conditions of Theorem* 2.3 *we have*

$$\mathrm{E}N(\widehat{\widehat{\theta}}_N - \theta^*)(\widehat{\widehat{\theta}}_N - \theta^*)^T = \sigma^2(R^*)^{-1} + O(N^{-1/2}\log N). \tag{2.25}$$

Finally, taking into account Theorem 2.2, (2.23), and Lemma 2.1 we get the following result.

THEOREM 2.6. *Under the conditions of Theorem* 2.3 *the transformed process*

$$\psi_s = e^{s/2}(\widehat{\widehat{\theta}}_{e^s} - \theta^*) \tag{2.26}$$

is L-mixing *with respect to* $(\mathcal{F}_{e^s}, \mathcal{F}^+_{e^s})$.

The above three results, Theorems 2.4, 2.5, and 2.6, are the key tools in extending Theorem 2.1 to adaptive predictors using recursive estimators rather than off-line estimators (cf. [24]). Thus we get the following *key result*.

THEOREM 2.7. *Under the conditions of Theorem* 2.3 *we have*

$$\lim_{N\to\infty}\frac{1}{\log N}\lim_{N\to\infty}\sum_{n=1}^{N}(\overline{\varepsilon}_n^2(\widehat{\widehat{\theta}}_{n-1}) - e_n^2) = \sigma^2(e)(p+q) \quad \text{a.s.} \tag{2.27}$$

In addition, the above proposition remains valid, if we replace $\overline{\varepsilon}_n(\widehat{\widehat{\theta}}_{n-1})$ by its online computed approximation ε_n; see (2.15).

THEOREM 2.8. *Under the conditions of Theorem* 2.3 *we have*

$$\lim_{N\to\infty}\frac{1}{\log N}\lim_{N\to\infty}\sum_{n=1}^{N}(\varepsilon_n^2 - e_n^2) = \sigma^2(e)(p+q) \quad \text{a.s.} \tag{2.28}$$

The main *contribution* of the present paper is the extension of the technical results given as Theorems 2.4, 2.5, and 2.6 to general recursive estimation schemes that include the DFL scheme with enforced boundedness, given as (3.53)–(3.54). The extension of Theorem 2.4 uses the results of [17, 19] but requires an additional technical tool given in [21]. This extension will be given in section 3. The extensions of Theorems 2.5 and 2.6 are obtained using straightforward, though numerous, approximations in sections 5 and 6. The present paper actively uses the results of [17, 19, 21]. To facilitate reading, these relevant results are summarized in section 8.

3. General recursive estimation schemes. The prime objective of this section is to formulate a general recursive estimation method, the DFL scheme with enforced boundedness, together with conditions that ensure its convergence. It is given as Algorithm DFL under (3.53)–(3.54), developed in [11, 12, 51], see also the books [3, 13, 53].

But first we present two closely related recursive algorithms: Algorithm CR (continuous-time recursion), (3.16), and Algorithm DR (discrete-time recursion), (3.34),

which can be interpreted as "frozen parameter" approximations to the DFL scheme. The main results of the paper will be formulated and proved for the continuous-time method, Algorithm CR. The connection between the continuous-time and the discrete-time algorithms is straightforward. In contrast, the connection between Algorithm DR and the DFL scheme is not straightforward at all, but it has been worked out in [19, sections 6 and 7]. Details will be given while discussing the DFL method.

Our first tentative general method is a continuous-time recursive estimation process without resetting, given by a random differential equation of the form

$$\dot{x}_t = \frac{1}{t}(H(t, x_t, \omega) + \delta H(t, \omega)), \qquad x_1 = \xi_1, \tag{3.1}$$

defined over the underlying probability space $(\Omega, \mathcal{F}, P)$. Here x_t indicates an estimator sequence and $H = (H(t, x, \omega))$ is a random field defined in $[1, \infty) \times D$, where D is a bounded open domain in $\mathbb{R}^p \times \Omega$ and $\delta H(t, \omega))$ is a perturbation term to be described later. The advantage of continuous time is that some calculations can be carried out more easily than in discrete time.

The technical conditions that we impose on $H(t, x, \omega)$ will be tuned to fit the DFL scheme, given by (3.53) and (3.51) below. A continuous-time example for a random field $H(t, x, \omega)$ that is motivated by the DFL scheme is the following:

$$H(t, x, \omega) = \varepsilon(t, x, \omega)\eta(t, x, \omega), \tag{3.2}$$

where $\varepsilon(t, x, \omega)$ and $\eta(t, x, \omega)$ are stationary, jointly Gaussian processes, defined by finite-dimensional stable linear filters applied to a standard Wiener-process (w_s):

$$\varepsilon(t, x, \omega) = \int_{-\infty}^{t} h_\varepsilon(t - s, x) dw_s, \qquad \eta(t, x, \omega) = \int_{-\infty}^{t} h_\eta(t - s, x) dw_s, \tag{3.3}$$

such that in an appropriate state-space representation the state-space matrices corresponding to the impulse responses $h_\varepsilon(\tau, x)$ and $h_\eta(\tau, x)$ are sufficiently smooth functions of the parameter x. In the recursive maximum likelihood identification method for discrete-time Gaussian ARMA processes $\varepsilon(n, x, \omega)$ would be the estimated input noise, with x being the system-parameter and $\eta(n, x, \omega)$ would be its negative gradient with respect to x, assuming stationary initialization for both processes. To specify the conditions to be imposed we need some preliminary technical details. The notion of M-bounded processes will now be extended to parameter-dependent, continuous-time processes.

DEFINITION 3.1. *Let $D_0 \subset \mathbb{R}^p$ be a compact set and let $(u_t(x))$ be an $\mathbb{R}^k$-valued measurable stochastic process defined on $\Omega \times \mathbb{R}^+ \times D_0$, where $\mathbb{R}^+ = \{t : t \geq 0\}$. We say that $(u_t(x))$ is M-bounded (in D_0) if for all q with $1 \leq q < \infty$ we have*

$$M_q(u) = \sup_{\substack{t \geq 0 \\ x \in D_0}} \mathrm{E}^{1/q} |u_t(x)|^q < \infty. \tag{3.4}$$

If $(u_t(x))$ is M-bounded, then we write $= O_M(1)$. We shall use the same terminology if x or t degenerates into a single point. If c_t is a sequence of positive numbers, then we write $u_t(x) = O_M(c_t)$ if $u_t(x)/c_t = O_M(1)$.

The notion of L-mixing will now be extended to parameter-dependent, continuous-time processes. Let a probability space $(\Omega, \mathcal{F}, P)$ be given together with a pair of families of σ-algebras $(\mathcal{F}_t, \mathcal{F}_t^+)$ such that (i) $\mathcal{F}_t \subset \mathcal{F}$ is monotone increasing, (ii) $\mathcal{F}_t^+ \subset \mathcal{F}$

is monotone decreasing and $\mathcal{F}_t^+$ is right continuous in t, i.e., $\mathcal{F}_s^+ = \sigma\{\bigcup_{0<\varepsilon} \mathcal{F}_{s+\varepsilon}^+\}$, and (iii) $\mathcal{F}_t$ and $\mathcal{F}_t^+$ are independent for all t. For $s<0$ we set $\mathcal{F}_s^+ = \mathcal{F}_0^+$.

DEFINITION 3.2. *Let $D_0 \subset \mathbb{R}^p$ be a compact set and let $(u_t(x))$ be an $\mathbb{R}^k$-valued measurable stochastic process defined on $\Omega \times \mathbb{R}^+ \times D_0$. We say that $u = (u_t(x))$ is L-mixing with respect to $(\mathcal{F}_t, \mathcal{F}_t^+)$, uniformly in x for $x \in D_0$, if it is $\mathcal{F}_t$-progressively measurable, M-bounded (in D_0) and if for all $q \geq 1$ with*

$$\gamma_q(\tau, u) = \gamma_q(\tau) = \sup_{\substack{t \geq \tau \\ x \in D_0}} \mathrm{E}^{1/q} |u_t(x) - \mathrm{E}\,(u_t(x) | \mathcal{F}_{t-\tau}^+)|^q, \qquad \tau \geq 0,$$

we have

$$\Gamma_q = \Gamma_q(u) = \int_0^\infty \gamma_q(\tau) d\tau < \infty. \tag{3.5}$$

We say that $(u_t(x))$, $t \geq 0$, $x \in D_0$, is L^+-mixing with respect to $(\mathcal{F}_t, \mathcal{F}_t^+)$, uniformly in x for $x \in D_0$, if in addition for all $q \geq 1$ there exist $C_q, c_q > 0$ such that for all $\tau \geq 0$

$$\gamma_q(\tau, u) \leq C_q (1+\tau)^{-1-c_q}. \tag{3.6}$$

The definition extends to parameter-free processes (u_t) and to discrete-time processes $(u_n(x))$. In the latter case we set

$$\Gamma_q = \Gamma_q(u) = \sum_{\tau=0}^\infty \gamma_q(\tau) < \infty. \tag{3.7}$$

CONDITION 3.1. *The process $H = (H(t,x,\omega))$ is assumed to be defined in $\Omega \times \mathbb{R}^+ \times D$, where $D \subset \mathbb{R}^p$ is an open domain, it is three times continuously differentiable with respect to x for $x \in D$ almost surely and for any compact set $D_0 \subset D$ H and its derivatives up to order 3 are M-bounded in D_0. Furthermore $(H(t,x,\omega))$ and its first derivative $H_x = (H_x(t,x,\omega))$ are L^+-mixing with respect to $(\mathcal{F}_t, \mathcal{F}_t^+)$, uniformly in $x \in D_0$.*

In [19] we used the finite difference field of $H = (H(t,x,\omega))$ to capture the smoothness of $H(t,x,\omega)$. In general, we considered the process

$$\Delta u/\Delta x\ (t, x, x+h, \omega) = |u_t(x+h) - u_t(x)|/|h|$$

defined for $t \geq 0$, $x \neq x+h \in D$. We say that $u = (u_t(x))$ is *M-Lipschitz continuous* with respect to x in D_0, if the process $\Delta u/\Delta x$ defined above is M-bounded; i.e., if for all $1 \leq q < \infty$ we have

$$M_q(\Delta u/\Delta x) = \sup_{\substack{t \geq 0 \\ x \neq x+h \in D_0}} \mathrm{E}^{1/q} |u_t(x+h) - u_t(x)|^q/|h| < \infty.$$

Condition 1.1. of [19] is then as follows.

CONDITION H. *The processes $(H(t,x,\omega))$ and $(\Delta H/\Delta x(t,x,x+h,\omega))$ are assumed to be separable and L^+-mixing with respect to $(\mathcal{F}_t, \mathcal{F}_t^+)$, uniformly in $x, x+h \in D$.*

It is easy to see that Condition H is implied by Condition 3.1.

A discussion of L-mixing. We give further details for comparing L-mixing and ϕ-mixing as described in Chapter 7.2 of [15]. In L-mixing we consider projections on the relative future defined by $\mathcal{F}_{t-\tau}^+$ and the resulting approximation error is

$$\mathrm{E}^{1/q} |u_t - \mathrm{E}\,(u_t | \mathcal{F}_{t-\tau}^+)|^q \leq \gamma_q(\tau, u) \tag{3.8}$$

for $\tau \geq 0$. In ϕ-mixing we consider projections on the past and the corresponding error from the mean, defined as

$$||P(A|\mathcal{H}) - P(A)||_p \tag{3.9}$$

for $A \in \mathcal{G}$. Assuming that there is a random variable Φ such that

$$||P(A|\mathcal{H}) - P(A)|| \leq \Phi \tag{3.10}$$

for all $A \in \mathcal{G}$, we have the following proposition (see Proposition 2.6, (2.23), of Chapter 7.2 of [15]). Let Z be a $\mathcal{G}$-measurable random variable such that $||Z||_s$ is finite and let $r, s > 1$ be such that $r^{-1} + s^{-1} = 1$. Then

$$||\mathrm{E}(Z|\mathcal{H}) - \mathrm{E}(Z)||_p \leq 2||\Phi||_p^{1/r}||\mathrm{E}(|Z|^s|\mathcal{H})||_p^{1/s}. \tag{3.11}$$

Now if (u_t), $t \geq 0$, is L-mixing with respect to $(\mathcal{F}_t, \mathcal{F}_t^+)$, then taking the conditional expectation of $u_t - \mathrm{E}\ (u_t|\mathcal{F}_{t-\tau}^+)$ with respect to $\mathcal{F}_{t-\tau}$ (cf. (3.8)), we get by Jensen's inequality and the assumed independence of $\mathcal{F}_{t-\tau}$ and $\mathcal{F}_{t-\tau}^+$

$$\mathrm{E}^{1/q}|\mathrm{E}\ (u_t|\mathcal{F}_{t-\tau}) - \mathrm{E}\ (u_t)|^q \leq \gamma_q(\tau, u). \tag{3.12}$$

In this respect the two notions of mixing lead to similar conclusions.

We will need to strengthen the condition on the M-boundedness of H_x as follows for reasons that will be discussed later, following Condition 3.8.

CONDITION 3.2. *$H(t, x, \omega)$ is piecewise continuous in t almost surely, and for any compact set $D_0 \subset D$ there exists a random variable $L_t = L_t(\omega) \geq 0$ such that for all $x \in D_0$*

$$|H_x(t, x, \omega)| \leq L_t(\omega)$$

and here L_t is in class M^*, i.e., for some $\varepsilon > 0$ we have

$$\sup_t \mathrm{E} \exp(\varepsilon L_t) < \infty. \tag{3.13}$$

In [19] we had a weaker condition (see Condition 1.2 of [19]).

CONDITION L. *$H(t, x, \omega)$ is piecewise continuous in t, and for any compact set $D_0 \subset D$ is Lipschitz-continuous in x for $x \in D_0$ almost surely with a (t, ω)-dependent Lipschitz constant $L_t = L_t(\omega) \geq 0$; i.e., for $x, x' \in D_0$ we have*

$$|H(t, x, \omega) - H(t, x', \omega)| \leq L_t(\omega)|x - x'|,$$

where L_t is in class M^*.

Assuming that $(\delta H(t, \omega))$ is piecewise continuous in t almost surely, a solution (x_t) of (3.1) exists almost surely in some finite or infinite interval. A central role in the analysis of (x_t) is played by the mean-field of $H(t, x, \omega)$. To simplify the presentation it is assumed that the mean-field is essentially independent of t, but a small perturbation is allowed: we have $\mathrm{E}H(t, x, \omega) = G(x) + \delta G(t, x)$, where $\delta G(t, x)$ is small in a sense to be specified below.

CONDITION 3.3. *We have for any compact set $D_0 \subset D$ and $t \geq 0$, $x \in D_0$*

$$\mathrm{E}H(t, x, \omega) = G(x) + \delta G(t, x),$$

where $\delta G(t,x) = O(t^{-1/2-\varepsilon})$ *uniformly in* $x \in D_0$, *with some* $\varepsilon > 0$. $G(y)$ *has continuous and bounded partial derivatives up to third order. Finally, we assume that*

$$G(x) = 0 \tag{3.14}$$

has a unique solution x^* *in* D.

Remark. In [19] the slightly weaker condition $\delta G(t,x) = O(t^{-1/2})$ has been used (see Condition 1.3 of [19]). Also only differentiability up to order 2 was required.

Let us now consider the associated ODE

$$\dot{y}_t = \frac{1}{t} G(y_t), \qquad y_s = \xi,\ s \geq 1. \tag{3.15}$$

Under the condition above, (3.15) has a unique solution in some finite or infinite interval, which we denote by $y(t,s,\xi)$. It is well known that $y(t,s,\xi)$ is a twice continuously differentiable function of ξ. The celebrated ODE principle states that the solution trajectories of the random differential equation (3.1), under additional conditions, follow the solution trajectories of the associated ODE (3.15).

Interpreting (3.1) as a continuous-time stochastic approximation method for solving the nonlinear algebraic equation $G(x) = 0$, an obvious difference compared to classical theory (see [54]) is that G is not defined on the whole space. Thus we are led to the study of recursive estimation methods *constrained* to a fixed domain D. In fact for theoretical reasons it is better to assume that the estimator process is constrained to a *compact* domain $D_0 \subset D$. One way to enforce boundedness of the estimation process is to restart it whenever it would leave D_0. Such a *truncated* version of (3.1) is described by Algorithm CR below, following [19]. A short discussion on the resetting mechanism to follow will be given in the context of the DFL scheme.

Algorithm CR. Consider a continuous-time recursion given by a random differential equation

$$\dot{x}_t = \frac{1}{t}(H(t,x_t,\omega) + \delta H(t,\omega)), \qquad x_1 = \xi_1 \tag{3.16}$$

combined with the following *resetting* mechanism. Let $D_0 \subset D$ denote a compact truncation domain such that $x^* \in \text{ int } D_0$. Let us initialize (3.16) at some time $\sigma \geq 1$ and let $x_\sigma = \xi_1 \in \text{int} D_0$. Let

$$\tau(\sigma) = \min\{t : t > \sigma, x_t \in \partial D_0\}, \tag{3.17}$$

where ∂D_0 denotes the boundary of D_0. Then we reset x to $x_1 = \xi_1$, which is formally stated by requiring that the right-hand side limit of x_t at $t = \tau = \tau(\sigma)$ will be ξ_1:

$$x_{\tau+} = \xi_1. \tag{3.18}$$

Thus we get a piecewise continuous trajectory (x_t) defined in some finite or infinite interval.

Remark. An alternative resetting mechanism, used in the analysis of discrete-time processes, is obtained by putting

$$x_t = \xi_1 \quad \text{for} \quad n < t \leq n+1 \ \text{ if } \ x_\tau \in \partial D_0 \ \text{ for } \ n < \tau \leq n+1. \tag{3.19}$$

To ensure that the estimator sequence is not bounced back and forth by resetting we need to impose some condition on the shape and relative position of the truncation

domain, x^* and ξ_1, which is captured via the flow induced by the ODE. For this, first we need to define the star-like closure of the set D_0, relative to x^*, as follows:

$$D_0^* = \{y : y = x^* + \lambda(x - x^*),\ 0 \le \lambda \le 1, x \in D_0\}.$$

The condition below is a simplified and corrected version of Condition 1.5. of [19]. The simplification is that the condition on the position of the initial value $x_1 = \xi_1$ has been relaxed, while the correction is that an additional compact convex set D_0' containing the truncation domain has been introduced that has been implicitly used in the final step of the proof of Theorem 1.1. of [19]; see (2.10) of [19].

CONDITION 3.4. *Let $D_0 \subset D$ be a compact truncation domain such that $x^* \in \text{int}D_0$. We assume the following.* (i) *There exists a compact convex set $D_0' \subset D$ such that*

$$y(t,s,\xi) \in D_0' \quad \text{for } \xi \in D_0 \quad \text{and } y(t,s,\xi) \in D \quad \text{for } \xi \in D_0' \tag{3.20}$$

for all $t \ge s \ge 1$. In addition $\lim_{t\to\infty} y(t,s,\xi) = x^*$ for $\xi \in D$ and

$$\left\| \frac{\partial}{\partial \xi} y(t,s,\xi) \right\| \le C_0 (s/t)^\alpha \tag{3.21}$$

with some $C_0 \ge 1, \alpha > 0$ for all $\xi \in D_0'$ and $t \ge s \ge 1$. (ii) *We have an initial estimate $x_1 = \xi_1$ such that for all $t \ge s \ge 1$ we have $y(t,s,\xi_1) \in$ int D_0.* (iii) *Finally, for the star-like closure of the set D_0 we have $D_0^* \subset D$.*

In [19] we had a the following stability condition (Condition 1.5 of [19] with minor corrections added).

CONDITION D. (i) *For every $\xi \in D_0$, $t \ge s \ge 1$, $y(t,s,\xi) \in D$ is defined for $1 \le s \le t < \infty$ and converges to x^* for $t \to \infty$ and we have with some $C_0, \alpha > 0$*

$$\left\| \frac{\partial}{\partial \xi} y(t,s,\xi) \right\| \le C_0 (s/t)^\alpha. \tag{3.22}$$

(ii) *We assume that the initial condition ξ_1 is in* int D_{00}, *where $D_{00} \subset$* int D_0 *is a compact domain which is invariant for* (3.15) *such that for any $t > s \ge 1$,*

$$y(t,s,D_{00}) = \{y(t,s,\xi) : \xi \in D_{00}\} \subset \text{ int } D_{00}.$$

Remark. Since our objective is to restate Theorem 4.2 of [19], part (iii) of Condition 3.4 of the present paper need not be verified for this purpose, since it was not required in [19]; it is special addition for the present paper.

Remark. The condition on the existence of D_0' can be removed if D itself is convex. Indeed, the ODE given by (3.15) becomes autonomous after a change of time-scale $t = e^v$ (see below), thus the remaining condition in part (i) of Condition 3.4 implies that the set

$$D_0'' = \{y : y = y(t,s,\xi), \xi \in D_0,\ t \ge s \ge 1\}$$

is invariant for the ODE. It is easy to see that it is also compact, so we can take for D_0' the convex envelope of D_0''. We will show below that part (ii) of Condition D follows from part (ii) of Condition 3.4. Finally, part (iii) of Condition 3.4 is a minor additional technical condition needed for the present paper.

We shall use subscripts to indicate partial derivatives below. Using a change of time-scale $t = e^v$, $s = e^u$, the inequality (3.21) is equivalent to the condition that, for the solutions of the differential equation

$$\frac{d}{dv} z_v = G(z_v), \qquad z_u = \xi,\ u \geq 0, \tag{3.23}$$

denoted by $z(v, u, \xi)$ we have the stability condition

$$\|z_\xi(v, u, \xi)\| \leq C_0 e^{-\alpha(u-v)}.$$

It can be shown that if for $\xi = x^*$ we can verify $\|z_\xi(v, u, x^*)\| \leq C_0' e^{-\alpha(u-v)}$ with some C_0', then the above stability condition follows from the remaining components of Condition 3.4. Equivalently, it can be shown that if $\|y_\xi(t, s, x^*)\| \leq C_0'(s/t)^\alpha$ with some C_0', then (3.21) follows from the remaining components of Condition 3.4.

Setting

$$A^* = \frac{\partial G(x)}{\partial x}\Big|_{x=x^*}, \tag{3.24}$$

we have $y_\xi(t, s, x^*) = e^{A^*(\log t - \log s)}$. The exponent α can be related to the eigenvalues of the Jacobian matrix A^* as follows. Let

$$\alpha^* = \min_i\{-\Re\lambda_i(A^*)\}, \quad i = 1, \ldots, p, \tag{3.25}$$

where $\lambda_i(A^*)$ denote the eigenvalues of A^* and $\Re$ denotes real part. Then, denoting the spectral norm by $||.||_{sp}$ we have $||e^{A^*(\log t - \log s)}||_{sp} = e^{-\alpha^*(\log t - \log s)} = (s/t)^{\alpha^*}$. Since for any square matrix B we have $\lim_n ||B^n||^{1/n} = ||B||_{sp}$, we conclude that by taking

$$\alpha = \alpha^*_-, \tag{3.26}$$

where α^*_- denotes any number that is smaller than α^*, we have

$$||e^{A^*(\log t - \log s)}|| \leq C_0 e^{-\alpha(\log t - \log s)} = C_0(s/t)^\alpha \tag{3.27}$$

with some $C_0 > 0$. If the Jordan form of A^* is diagonal, then we can take $\alpha = \alpha^*$.

LEMMA 3.1. *Condition* 3.4(ii) *implies Condition* D(ii).

Proof. Both Condition 3.4(ii) and Condition D(ii) can be trivially rewritten in terms of the solutions of the ODE (3.23), denoted by $z(v, u, \xi)$. Let $D_0^+ \subset D$ be a small open neighborhood of D_0 such that

$$I(D_0^+) = \bigcup_{\xi \in D_0^+, t \geq 0} z(t, 0, \xi) \subset D.$$

Then it is easily seen that $I(D_0^+)$ is an open invariant set containing D_0. It is a well-known result of Krasovskii (see Theorem 5.3 of his 1963 book) that Conditions 3.3 and 3.4 imply the existence of a C^2 Lyapunov function V with domain of definition $I(D_0^+)$ such that $V(z) > 0$ for $z \neq x^*$, $V(x^*) = 0$, and

$$\frac{d}{dv} V(z(v, u, \eta)) < 0 \quad \text{for} \quad \eta \in D_0,\ z(v, u, \eta) = z \neq x^*,$$

and $V(z)$ tends to $+\infty$ when z tends to the boundary of $I(D_0^+)$. Since the Jacobian $A^* = (\partial/\partial z)\ G(z)|_{z=x^*}$ is strictly stable, V can be chosen so that its Hessian $(\partial^2/\partial z^2)\ V(z)|_{z=x^*}$ is positive definite.

Let us consider the level sets

$$U_V(c) = \{z : V(z) \le c\} \quad \text{and} \quad S_V(c) = \{z : V(z) = c\}.$$

For sufficiently small c the set $S_V(c)$ is C^1 isomorphic to a sphere. Let us now reverse time and consider the differential equation

$$\frac{d}{du}\overline{z}_u = -G(\overline{z}_u), \qquad \overline{z}_0 = \zeta \in S_V(c),$$

the solution of which is denoted by $\overline{z}(u, 0, \zeta)$. Let the solution $\overline{z}(u, 0, \zeta) \in D_0$ exist for $u \le u^*(\zeta) \le \infty$. Obviously we have, with $\eta \in D_0,\ \zeta = z(v, 0, \eta) \in S_V(c)$,

$$\overline{z}(v, 0, \zeta) = \eta.$$

For any $\zeta \in S_V(c)$ we choose a finite backward travel time denoted by $w = w(\zeta)$ such that $0 < w(\zeta) \le u^*(\zeta)$, and such that w is a C^1 function. Consider the set

$$D_{00} = \{z : z = \overline{z}(u, 0, \zeta) \ \text{ with some } \ \zeta \in S_V(c),\ 0 \le u \le w(\zeta)\} \ \cup \ U_V(c).$$

It is easy to see that D_{00} is compact and invariant for the ODE (3.23). Furthermore it is easy to see that for any $\zeta \in S_V(c)$ and $0 \le u < w(\zeta)$ (with strict inequality) the point $\overline{z}(u, 0, \zeta)$ is in the interior of D_{00}.

Now let $c \le V(\xi_1)$ and let ζ_1 be the point where the trajectory $z(v, 0, \xi_1)$ hits $S_V(c)$, say for $v = v(\xi_1)$. Choosing the backward travel time w so that $w(\zeta_1) > v(\xi_1)$, the above defined set D_{00} will satisfy the second part of Condition D, and the lemma follows. □

Finally, consider the perturbation term $\delta H(t, \omega)$. Following [19] and motivated by the application for the DFL scheme, we will use the following condition.

CONDITION 3.5. *$(\delta H(t, \omega))$ is a measurable M-bounded process, which is piecewise continuous in t almost surely, moreover there exists an $\varepsilon > 0$ such that for any fixed $q > 1$ and for any $s \ge 1$,*

$$\sup_{s \le \sigma \le qs} \int_{\sigma}^{\tau(\sigma) \wedge q\sigma} \frac{1}{r} |\delta H(r, \omega)| dr = O_M(s^{-1/2-\varepsilon}). \tag{3.28}$$

It is no loss of generality to assume that $\varepsilon < 1/2$. We assume that the ε's showing up here and in Condition 3.3 *are identical.*

Remark. In [19] the slightly weaker condition

$$\sup_{s \le \sigma \le qs} \int_{\sigma}^{\tau(\sigma) \wedge q\sigma} \frac{1}{r} |\delta H(r, \omega)| dr = O_M(s^{-1/2}) \tag{3.29}$$

was required (see Condition 1.6 of [19]). This is sufficient to establish a rate of convergence result for the moments.

The above condition seems to be hard to verify, since it involves $\tau(\sigma)$, which itself is defined in terms of the process (x_t). In fact, the condition seems to be artificially tuned so that the proof can be carried out. An alternative, seemingly more useful, condition implying Condition 3.5 would be

$$\sup_{s \le \sigma \le qs} \int_{\sigma}^{q\sigma} \frac{1}{r} |\delta H(r, \omega)| dr = O_M(s^{-1/2-\varepsilon}), \tag{3.30}$$

which is independent of the stopping time $\tau(\sigma)$. The latter is certainly satisfied if $\delta H(r,\omega) = O_M(r^{-1/2-\varepsilon})$. The prominent role of Condition 3.5 will become clear in the context of the DFL scheme, see (3.56) and Lemma 3.2. The following is given in Theorem 1.1 of [19].

THEOREM 3.1. *Consider the continuous-time recursive estimation process defined by* (3.16) *with the resetting mechanism* (3.17) *and* (3.18). *Assume that Conditions* 3.1–3.5 *are satisfied, moreover Condition* 3.4 *is satisfied with* $\alpha > 1/2$. *Then the solution* (x_t) *is defined for all* $t \in [1,\infty)$ *with probability* 1 *and* $x_t = O_M(t^{-1/2})$. *Moreover, the following stronger result also holds: for any fixed* $1 < q < \infty$ *we have*

$$x_t^* = \sup_{t \le s \le qt} |x_s| = O_M(t^{-1/2}).$$

As has been noted at the end of section 2 in [19], using the alternative resetting method (3.19) does not affect the validity of Theorem 1.1 of [19].

Definition of $\overline{\alpha}$. In subsequent analysis a crucial role will be played by the *gap* between α, introduced in Condition 3.4, and $1/2$, therefore we introduce a separate notation: we write

$$\overline{\alpha} = \alpha - 1/2. \tag{3.31}$$

An example: a recursive estimation method is called a stochastic Newton method if the Jacobian matrix of the right-hand side of the associated ODE at $x = x^*$ is $-I$, where I is an identity matrix. Then we can take $\alpha^* = \alpha = 1$ and $\overline{\alpha} = 1/2$.

Let us now consider *discrete-time* processes of the form

$$x_{n+1} = x_n + \frac{1}{n+1}(H(n+1,x_n,\omega) + \delta H(n+1,\omega)), \quad x_0 = \xi_0 \in \text{ int } D_0. \tag{3.32}$$

Boundedness of the estimator sequence will be enforced by a *resetting* mechanism. Let $D_0 \subset D$ be a compact domain. If x_{n+1} leaves D_0, then we redefine x_{n+1} to be x_0. To formalize this: at any time n let x_{n+1-} denote the value of x computed at time $n+1$ by (3.32) and let

$$B_{n+1} = \{\omega : x_{n+1-} \notin \text{int} D_0\}. \tag{3.33}$$

Algorithm DR. A discrete-time recursive estimation process with resetting is defined as follows: we define x_{n+1} by

$$x_n + (1 - \chi_{B_{n+1}})\frac{1}{n+1}(H(n+1,x_n,\omega) + \delta H(n+1,\omega)) + \chi_{B_{n+1}}(x_0 - x_n). \tag{3.34}$$

Remark. Note that the correction term on the right-hand side was $H(n,x_n,\omega)$ in [19]. The present notation fits the applications better: the estimator based on observations up to time n is updated by a new observation received at time $n+1$.

A standard way of analyzing this algorithm is to use continuous-time imbedding and this route has been followed in [19]. A more recent approach, in which the error that arises via this imbedding procedure is eliminated, is a discrete-time ODE method, developed in [25]. Here we follow the approach of [19], with a minor modifications. Let $(H^c(t,x,\omega))$ be the piecewise constant continuous-time extensions of $(H(n,x,\omega))$:

$$H^c(t,x,\omega) = H(n,x,\omega) \quad \text{for} \quad 1 \le n \le t < n+1. \tag{3.35}$$

Define $\delta H^c(t,x,\omega)$ in a similar manner. Let the exit time $\tau(\sigma)$ for any nonnegative integer σ be defined as

$$\tau(\sigma) = \min\{n : n \text{ integer},\ n > \sigma, x_{n-} \notin \operatorname{int} D_0\}. \tag{3.36}$$

CONDITION 3.6. *$(\delta H(n,\omega))$ is a measurable M-bounded process, moreover there exists an $\varepsilon > 0$ such that for any fixed $q > 1$ and for any integers $s \geq \sigma \geq 1$, with $[x]$ denoting integer part, we have*

$$\sup_{s \leq \sigma \leq [qs]} \sum_{r=\sigma}^{\tau(\sigma)\wedge[q\sigma]} \frac{1}{r} |\delta H(r,\omega)| dr = O_M(s^{-1/2-\varepsilon}). \tag{3.37}$$

It is easy to see that Condition 3.5 follows with the modified resetting mechanism (3.19). The following result is an easy corollary of Theorem 3.1 and has been established in [19] as Theorem 1.2.

THEOREM 3.2. *Consider the discrete-time recursive estimation process with resetting defined by* (3.34). *Let $(H^c(t,x,\omega))$ be the piecewise constant continuous-time extensions of $(H(n,x,\omega))$ defined under* (3.35). *Assume that $H^c(t,x,\omega)$ satisfies Conditions* 3.1–3.4 *and the latter condition is satisfied with $\alpha > 1/2$. Let $\delta H^c(n,\omega)$ satisfy Condition* 3.6, *with $\tau(\sigma)$ defined as in* (3.36). *Then we have $x_n = O_M(n^{-1/2})$.*

Let us now consider a general recursive estimation scheme developed in [11, 12, 51], see also [3, 13, 53], which will be called the DFL scheme. Its basic building block is a parameter-dependent vector-valued process $(\overline{\phi}_n(x))$, with $x \in D \subset \mathbb{R}^p$, where D is an open domain, defined by the state-space equation

$$\overline{\phi}_{n+1}(x) = A(x)\overline{\phi}_n(x) + B(x)e_n, \tag{3.38}$$

with some nonrandom initial condition $\overline{\phi}_1(x)$, the value of which is often assumed to be zero. The dimensionality of $\overline{\phi}_n(x)$ will be denoted by r. In the analysis of [19], as in all other works on the analysis of the DFL scheme we have to ensure that for any choice of $x = x_n \in D$ the time-varying system

$$\phi_{n+1} = A(x_n)\phi_n + B(x_n)e_n, \quad \phi_0 = 0, \tag{3.39}$$

is bounded input-bounded output (BIBO) stable. This is ensured by the following condition.

CONDITION 3.7. *The functions $A(x), B(x)$ are three times continuously differentiable in D. Moreover, the family of matrices $A(x)$, $x \in D_0$, with D_0 being the preselected truncation domain, is jointly stable in the sense that there exist a single symmetric positive definite $r \times r$ matrix V and $0 < \lambda < 1$ such that for all $x \in D_0$,*

$$A^T(x)VA(x) \leq \lambda V.$$

Discussion of the joint stability condition. In the case of recursive estimation of linear stochastic systems the joint stability condition can be satisfied by an appropriate realization of system (3.38). Namely, in these cases (3.38) has the structure

$$\overline{\phi}_{1,n+1} = A_1\overline{\phi}_{1,n} + B_1 e_n, \tag{3.40}$$

$$\overline{\phi}_{2,n+1}(x) = A_2(x)\overline{\phi}_{2,n}(x) + B_2(x)\overline{\phi}_{1,n+1}, \tag{3.41}$$

where $\overline{\phi}_{1,n}$ is independent of x and is observable. Thus it is sufficient to ensure the joint stability of (3.41), which has an observable input. For any fixed x and

nonsingular $T = T(x)$ we have the system equivalence

$$(A_2(x), B_2(x), I) = (T(x)A_2(x)T^{-1}(x), T(x)B_2(x), T(x)^{-1}), \tag{3.42}$$

and the latter realization can also be used to compute $\overline{\phi}_{2,n+1}(x)$. Assume that $A(x)$ is stable for all $x \in D$. Choosing $T(x)$ so that $T(x)A_2(x)T^{-1}(x)$ is a contraction for all $x \in D$ and assuming that $T(x)$ is continuous in x, it is easy to see that Condition 3.7 is satisfied for the transformed system with any compact $D_0 \subset D$. In addition, assuming that $(A_2(x), B_2(x), I)$ uniquely determines x, the same holds for the equivalent system $(T(x)A_2(x)T^{-1}(x), T(x)B_2(x), T(x)^{-1})$.

Assuming joint stability of $(A(x))$, it follows that there exists some $c > 0$ such that for any sequence $(A(x_n))$ with $x_n \in D_0$ we have

$$||A(x_n) \cdots A(x_0)|| \le c\lambda^{n/2}. \tag{3.43}$$

In fact, this is the key property that we need in the analysis. An alternative method for ensuring the validity of (3.43) used in [3] is to require that the sequence $(A(x_n))$, or equivalently the sequence (x_n), is *slowly varying*. This method will be discussed later.

The input noise (e_n) is assumed to satisfy two conditions (see Conditions 2.5 and 2.6 of [19].)

CONDITION 3.8. *We assume that (e_n) is a wide-sense stationary process and that $|e_n|^2$ is in class M^*; i.e., for some $\varepsilon > 0$ we have*

$$\sup_n \mathrm{E}\ \exp \varepsilon |e_n|^2 < \infty.$$

Condition 3.8 is standard in the Chinese school for recursive estimation (see, e.g., [7]) and is certainly satisfied for wide-sense stationary Gaussian sequences. The weaker condition that (e_n) is M-bounded is assumed also in the special case of (3.53), given as Example 1, p. 215 of [3], (see Condition (A'5) on p. 290 of [3]). The existence of finite moments of all orders for certain state variables is required also in the general model of recursive estimation of [3], see Condition (A'5) on p. 290 of [3].

Discussion of Condition 3.8. Assume $\delta H(r, \omega) = 0$ identically and that no resetting takes place in the interval $[1, t]$. Then we have

$$x_t - y_t = \int_1^t \frac{1}{r}(H(r, x_r, \omega) - G(y_r))\ dr. \tag{3.44}$$

Now we can bound the right-hand side from above in two ways as

$$\begin{aligned} &\left| \int_1^t \frac{1}{r}(H(r, x_r, \omega) - G(x_r))\ dr \right| + \int_1^t \frac{1}{r} L|x_r - y_r|\ dr, \\ &\left| \int_1^t \frac{1}{r}(H(r, y_r, \omega) - G(y_r))\ dr \right| + \int_1^t \frac{1}{r} L_r|x_r - y_r|\ dr. \end{aligned} \tag{3.45}$$

In both cases we can apply the Bellman–Gronwall lemma. In the first case we need only the Lipschitz continuity of G, while H may be even discontinuous (which is the case, e.g., for the signed least mean squares (LMS) methods), but the first term is hard to analyze, unless H is a Markov process for any fixed x (see Chapter 1 of Part II of [3]). In the second case we need the Lipschitz continuity of H and Condition 3.8 has to be imposed on L_r to ensure that the application of the Bellman–Gronwall lemma gives

meaningful result. On the other hand, the analysis of the first term is significantly simpler, since it is essentially the integral of a zero-mean L-mixing process.

CONDITION 3.9. *We assume that* (e_n) *is* L^+*-mixing with respect to a pair of families of* σ*-algebras* $(\mathcal{F}_n, \mathcal{F}_n^+)$.

The role of this condition will be discussed in connection with Lemma 3.2, see also [19, Lemma 5.6 and Theorem 6.1]. Now we are ready to define a random field $H(n, x, \omega)$ in terms of $\overline{\phi}_n(x)$ as follows:

$$H(n, x, \omega) = Q(\overline{\phi}_n(x)), \tag{3.46}$$

where for the sake of simplicity Q is a quadratic function from $\mathbb{R}^r$ to $\mathbb{R}^p$. An alternative, more general definition would be

$$H(n, x, \omega) = F(Q(\overline{\phi}_n(x)), x), \tag{3.47}$$

where Q is quadratic and F is linear in Q and three times continuously differentiable in its second variable x. Also define the mean-field

$$G(x) = \lim_{n\to\infty} \mathrm{E}\, Q(\overline{\phi}_n(x)). \tag{3.48}$$

It is easy to see that $G(x)$ is well defined, since $\overline{\phi}_n(x)$ is asymptotically wide-sense stationary: in fact $\overline{\phi}_n(x) = \overline{\phi}_{*n}(x) + O_M(\beta^n)$, where $\overline{\phi}_{*n}(x)$ is wide-sense stationary and $0 < \beta < 1$, and thus

$$G(x) = \mathrm{E}\, Q(\overline{\phi}_n(x)) + O(\beta^n). \tag{3.49}$$

The estimation problem in the context of the DFL scheme is then to solve the nonlinear algebraic equation

$$G(x) = 0$$

based on observations of $Q(\overline{\phi}_n(x))$. It is assumed that a unique solution x^* exists in D and in fact $x^* \in D_0$. In identification problems the estimation of x^* can be carried out in an off-line fashion, but this is not the case in stochastic adaptive control. Thus we focus on recursive estimation of x^*.

It is not difficult to see (cf. [19]) that under Conditions 3.7, 3.8, and 3.9 the piecewise constant continuous-time extension of the random field $H(n, x, \omega)$ defined by (3.46) satisfies Conditions 3.1, 3.2, and 3.3 with G defined under (3.48). In fact, in the latter condition $\delta G(t, x)$ decays exponentially fast to zero.

We use an iterative procedure, in which the estimate of x^* at time n will be denoted by x_n. To update this estimate we should use the correction term $Q(\overline{\phi}_n(x_n))$, but this frozen parameter value cannot be easily computed. In fact in stochastic adaptive control problems it cannot be computed at all. Hence we generate an online approximation of $Q(\overline{\phi}_n(x_n))$ and thus we arrive at the following first version of the DFL method: define recursively

$$\phi_{n+1} = A(x_n)\phi_n + B(x_n)e_n, \tag{3.50}$$

$$x_{n+1} = x_n + \frac{1}{n} Q(\phi_{n+1}) \tag{3.51}$$

with initial conditions $x_0 = \xi_0 \in \mathrm{int} D_0$ and ϕ_0 a constant, nonrandom initial state. It is assumed that $Q(\phi_{n+1})$ is *computable* by coupling a physical system with our computer.

Discussion on the DFL scheme. The applicability of this general estimation scheme in the theory of recursive identification of linear stochastic systems has been discussed in more details [53], albeit its analysis has not been complete. Further examples of application are given in [3]. Here also a rigorous and detailed analysis of a nonlinear modification of the DFL scheme is given, using a Markovian dynamics in generating the state sequence (ϕ_n). This setup extends the range of applicability of the method, but the verification of the existence of the solution of a Poisson equation, (see Condition (A.4) of Chapter 1.1, Part II in [3]) seems to be hard. A special case of the DFL scheme is stochastic linear regression, in which $\overline{\phi}_n$ does not depend on x at all and $H(n, x, \omega)$ is of the form

$$H(n, x, \omega) = Q(l(\overline{\phi}_n, x)) \tag{3.52}$$

with l being linear in both $\overline{\phi}$ and x, has been analyzed in [7, 14, 47].

It is well known from simulations that the DFL scheme may diverge, unless some precaution is taken. The above procedure will therefore be modified so that the estimates x_n will be enforced to stay in a compact domain $D_0 \subset D$, such that $x^* \in \mathrm{int}D_0$. This will be achieved by a *resetting* mechanism: if x_{n+1} leaves D_0 we redefine it to be $x_0 = \xi_0$. To formalize this procedure let x_{n+1-} denote the value of x computed at time $n+1$ by (3.51). Then if $x_{n+1-} \notin \mathrm{int}D_0$, then we reset it to its initial value ξ_0. To formalize the procedure let

$$B_{n+1} = \{\omega : x_{n+1-} \notin \mathrm{int}D_0\}.$$

Then we define the following algorithm.

Algorithm DFL. The DFL scheme with resetting:

$$\phi_{n+1} = A(x_n)\phi_n + B(x_n)e_n, \tag{3.53}$$

$$x_{n+1} = x_n + (1 - \chi_{B_{n+1}})\frac{1}{n+1}Q(\phi_{n+1}) + \chi_{B_{n+1}}(x_0 - x_n). \tag{3.54}$$

An additional stopping time is used in [3] to ensure the validity of (3.43) by ensuring that the sequence $(A(x_n))$, or equivalently the sequence (x_n), is *slowly varying.* Following [3, (3.1.2) on p. 291] for any positive integer σ define the stopping time

$$\nu(\sigma) = \min\{n : n \text{ integer}, \ n > \sigma, |x_{n+1} - x_n| > \delta\}, \tag{3.55}$$

where σ is some fixed positive number. It is well known that if δ is sufficiently small, then (3.43) holds. However, the a priori determination of a right value of δ seems to be hard.

Discussion of the "boundedness condition." The eventual divergence of the DFL scheme is often dealt with the controversial "boundedness condition" first formulated in [51] requiring that the estimator process visits a compact domain of attraction of the ODE infinitely often. A lucid exposition of the underlying principle is given in [52, Lemma 1.12], which is considered there as the key tool for the ODE method. Almost sure convergence using the above "boundedness condition" has also been established for a nonlinear, Markovian extension of the DFL method in [3, Part II, Chapter 1.9, Theorem 15]. Unfortunately, the "boundedness condition" is much too restrictive: it is a condition on the process itself that we analyze and it is not clear at all if it is satisfied even for basic methods such as RPE for ARMA processes.

One way to enforce the boundedness of the estimation process is to consider a compact truncation domain containing the true parameter in its interior and to

"project" the estimator back to this domain if it would leave it; see [51, 45]. It is easy to see that this procedure may fail even for deterministic algorithms, namely the ODE, which approximates the evolution of the discrete-time algorithm, may force us to move out of the truncation domain.

A rigorous treatment of the boundedness problem has been given in [3], where the estimator process is stopped if it leaves a prescribed compact domain containing the true parameter in its interior. Denoting by $\Omega' \subset \Omega$ the event that the estimator process is never stopped, the almost sure convergence of the estimator process has been established on Ω'; see [3, Part II, Chapter 1.6, Proposition 11]. But convergence with probability strictly smaller than 1 is not satisfactory from the practical point of view. The above *truncated* version of the DFL methods has been given and analyzed in [19].

The definition of the truncation domain requires some a priori knowledge of the system-parameters no matter what truncation procedure we use. This may seem to be a restrictive assumption, but even deterministic iterative methods for optimization may fail without good initialization.

In practice we start with an initial value and a truncation domain which may or may not satisfy our conditions. If it does not and the solution trajectory of the associated ODE starting at $x_0 = \xi_0$ does hit the boundary of D_0, then a heuristic argument, following [19], shows that the estimator process will be likely to hit the neighborhood of the same point of the boundary of the truncation domain. This phenomenon can be detected during the computations and a larger truncation domain can be chosen. Such an adaptive choice of the truncation domain has not yet been studied. A special case when the boundedness problem does not arise is the use of a stochastic regression approach, such as extended least squares (ELS); see [53].

To connect the DFL scheme with Algorithm DR define

$$\delta H(n,\omega) = Q(\phi_n) - Q(\overline{\phi}_n(x_n)). \tag{3.56}$$

Then (3.54) can be written in the form of (3.34). A critical point in the analysis of the DFL scheme is that the perturbation term $\delta H(n,\omega)$ is *not given a priori*, rather it is defined via the recursive procedure itself. In fact, the analysis of $\delta H(n,\omega)$ is a substantial component of the convergence analysis of the DFL-method, which has been worked out in [19, sections 5 and 6], leading to the following result (cf. Lemma 5.6 of [19]).

LEMMA 3.2. *Consider the DFL scheme defined by* (3.53)–(3.54). *Assume that Conditions* 3.7, 3.8, *and* 3.9 *are satisfied. In addition assume that Condition* 3.4 *is satisfied with* $\alpha > 1/2$. *Then* $(\delta H(n,\omega))$ *defined by* (3.56) *is an* M*-bounded process, moreover there exists an* ε *with* $0 < \varepsilon < 1/2$ *such that for any fixed* $q > 1$ *and for any integer* $s \geq 1$ *and integers* σ,

$$\sup_{s \leq \sigma \leq [qs]} \sum_{\sigma}^{\tau(\sigma) \wedge [q\sigma]} \frac{1}{r} |\delta H(r,\omega)| dr = O_M(s^{-1/2-\varepsilon}). \tag{3.57}$$

In short, $(\delta H(n,\omega))$ satisfies Condition 3.6. Postulating the validity of Condition 3.4 we conclude that all conditions of Theorem 3.2 are satisfied and thus we get the following result.

THEOREM 3.3. *Consider the DFL scheme defined by* (3.53)–(3.54). *Assume that Conditions* 3.7, 3.8, *and* 3.9 *are satisfied. In addition assume that Condition* 3.4 *is satisfied with* $\alpha > 1/2$. *Then we have* $x_n = O_M(n^{-1/2})$.

Discussion of the result. A special feature of the above result is that the *moments* of the estimation error are bounded from above. The only alternative result

on the moments of the estimation error in the context of the DFL scheme seems to be Proposition 24 of [3, Part II, Chapter 1.10] where the L_2 moments of the error of the stopped process is shown to be of the order $1/n$.

Almost sure convergence of the DFL scheme has been stated in [51] using the controversial "boundedness condition," requiring that the estimator process visits a compact domain of attraction of the ODE infinitely often. See also [52, Lemma 1.12] for a related result. Almost sure convergence using the above "boundedness condition" has also been established for a nonlinear, Markovian extension of the DFL method in [3, Part II, Chapter 1.8, Theorem 15]. The almost sure convergence of the estimator process has been established on a set $\Omega' \subset \Omega$ of probability strictly less than 1; see [3, Part II, Chapter 1.6, Proposition 11 and Chapter 3.4, Theorem 17].

An alternative set of results are obtained for stochastic regression models developed in [47]. Results on the rate of almost sure convergence are given in [7, 14]. See also Theorem 2 of [4] or Theorem 1 of [10]. The main shortcoming of stochastic regression, such as ELS, see [53], compared to the DFL scheme is that its range of applicability is limited. For example, in estimating an ARMA process by ELS we must impose the condition that the polynomial $C - 1/2$ is positive real.

Further discussion on mixing conditions. L-mixing and ϕ-mixing can both be used to derive two main results of [19], (Theorems 1.1 and 1.2), restated here as Theorems 3.1 and 3.2. In both results the key technical device is an improved Hölder inequality, see Lemma 8.1 of section 8, or Chapter 7.2 of [15], or Appendix III of [33]. An improved Hölder inequality of [15] is restated as Lemma 8.2. The situation is quite different for the DFL scheme, where L^+-mixing has been heavily exploited for deriving Theorem 3.3, in particular in proving Lemma 3.2 (see sections 5 and 6 of [19], in particular Theorem 6.1 in [19]).

4. Strong approximation of the estimation error. The main result of the present paper is a significant extension of Theorem 2.4 for the three, closely related recursive estimation schemes presented in the previous section. These extensions will be stated and proved in this section. The analysis will be carried out in detail for Algorithm CR, given by (3.16) and the resetting mechanism (3.17) and (3.18); see Theorem 4.1. The proof is nontrivial and relies on the results of [17, 19, 21]. The corresponding results for Algorithms DR and DFL will then follow by relatively simple arguments. The extension of the two other main results for the RPE method, given in section 2 as Theorems 2.5 and 2.6, will be given in the next two sections. Note that the conditions for the next theorem are identical with the conditions of Theorem 3.1.

THEOREM 4.1. *Consider the continuous-time recursive estimation scheme, Algorithm CR, given by* (3.16) *with the resetting mechanism* (3.17) *and* (3.18). *Assume that Conditions* 3.1–3.5 *are satisfied and Condition* 3.4 *is satisfied with* $\alpha > 1/2$. *Then the solution of* (3.16), (x_t), *is defined for all* $t \in [1, \infty)$ *with probability* 1 *and we have with*

$$\varepsilon_x = \min(\overline{\alpha}, \varepsilon)_-,$$

where c_- *is any number smaller than* c, $\overline{\alpha}$ *is given by* (3.31), *and* ε *is given in Condition* 3.5,

$$x_t - x^* = \int_1^t \frac{\partial}{\partial \xi} y(t, s, x^*) \frac{1}{s} H(s, x^*, \omega) ds + O_M(t^{-1/2-\varepsilon_x}). \tag{4.1}$$

Discussion of the result. The bound $O_M(t^{-1/2-\varepsilon_x})$ cannot be improved in general. Indeed, let $\delta H(t, \omega) = 0$, then $\varepsilon_x = \overline{\alpha}_- = \alpha_- - 1/2$, where $\alpha = \alpha^*_-$ (see (3.31), (3.26),

and (3.25)). Thus

$$-1/2 - \varepsilon_x = -\alpha^*_-.$$

Consider now a linear process with additive, state-independent, bounded noise, i.e., let $H(t,x,\omega) = A^*x + u_t$, where (u_t) is a zero-mean L-mixing bounded process. Then Algorithm CR reads

$$\dot{x}_t = \frac{1}{t}(A^*x_t + u_t), \qquad x_1 = \xi_1. \tag{4.2}$$

Assuming that A^* is stable, the boundedness of (u_t) implies the boundedness of (x_t), hence taking a sufficiently large truncation domain no resetting will take place ever. Obviously we have $x^* = 0$ and we can write the exact equality

$$x_t = \left(\frac{\partial}{\partial\xi}y(t,1,x^*)\right)\cdot\xi_1 + \int_1^t \frac{\partial}{\partial\xi}y(t,s,x^*)\frac{1}{s}H(s,x^*,\omega)ds \tag{4.3}$$

with

$$\frac{\partial}{\partial\xi}y(t,s,x^*) = e^{A^*(\log t - \log s)}.$$

Thus the residual term, the first term on the right-hand side of (4.3), is $e^{A^*\log t}\xi_1$, since now $s = 1$. Thus we have

$$||e^{A^*\log t}||_{sp} = t^{-\alpha^*}, \tag{4.4}$$

and since for any square matrix B we have $||B|| \geq ||B||_{sp}$, we conclude that

$$||e^{A^*\log t}|| \geq t^{-\alpha^*}. \tag{4.5}$$

Thus there exists a ξ_1 such that

$$|e^{A^*\log t}\xi_1| \geq t^{-\alpha^*}|\xi_1|, \tag{4.6}$$

implying that the result of the theorem is sharp.

To interpret this result note that the matrix $(\frac{\partial}{\partial\xi})y(t,s,x^*)$ is the sensitivity matrix, which indicates the relative effect of a perturbation of the initial condition at time s on the solution of (3.15) at time t. Thus the dominant term on the right-hand side represents the cumulative effect of the ideal correction terms $\frac{1}{s}H(s,x^*,\omega)$ at time t. A similar representation of the error $x_t - x^*$ for classical Robbins–Monroe processes, had been implicitly used already in [54]. The above dominant term has been explicitly presented for a class of stopped stochastic approximation processes in Lemma 3.1 of [65].

The novelty of the present result is that it is stated for a general recursive estimation scheme, that can handle the widely used DFL scheme, a crucial boundedness assumption enforced by a resetting mechanism and a tight upper bound for the residual term has been obtained. A relatively straightforward corollary of Theorem 4.1 is the following discrete-time result, in which the conditions are identical with the conditions of Theorem 3.2.

THEOREM 4.2. *Consider the discrete-time recursive estimation process, Algorithm DR, with resetting defined by* (3.34). *Let* $(H^c(t,x,\omega))$ *be the piecewise constant continuous-time extension of* $(H(n,x,\omega))$ *defined under* (3.35). *Assume that*

$(H^c(t,x,\omega))$ satisfies Conditions 3.1–3.4 *and Condition* 3.4 *is satisfied with* $\alpha > 1/2$. *Let* $\delta H(n,\omega)$ *satisfy Condition* 3.6, *with* $\tau(\sigma)$ *defined in* (3.36). *Then we have, with* $\varepsilon_x = \min(\overline{\alpha},\varepsilon)_-$, *where* $\overline{\alpha}$ *is given by* (3.31) *and* ε *is given by Condition* 3.6,

$$x_N - x^* = \sum_{n=1}^{N} \frac{\partial y}{\partial \xi}(N,n,x^*)\frac{1}{n}H(n,x^*,\omega) + O_M(N^{-1/2-\varepsilon_x}).$$

Specializing the last result to the DFL scheme we get a result that is very useful for applications (see section 7).

THEOREM 4.3. *Consider the DFL scheme defined by* (3.53)–(3.54). *Assume that the state-space equation* (3.38) *satisfies Condition* 3.7, *the noise process* (e_n) *satisfies Conditions* 3.8 *and* 3.9, *and the associated ODE satisfies Condition* 3.4 *with* $\alpha > 1/2$. *Let* $\varepsilon_x = \min(\overline{\alpha},\varepsilon)_-$, *where* $\overline{\alpha}$ *is defined under* (3.31) *and* ε *is given by Lemma* 3.2. *Then we have*

$$x_N - x^* = \sum_{n=1}^{N} \frac{\partial y}{\partial \xi}(N,n,x^*)\frac{1}{n}Q(\overline{\phi}_n(x^*)) + O_M(N^{-1/2-\varepsilon_x}).$$

Remark. The proof of Lemma 5.6 in [19], based on Theorem 6.1 of the same paper, implies that in Condition 3.5 we have $\varepsilon < 1/2$. Thus in the present case it is not our choice to have $\varepsilon < 1/2$. It follows that the upper bound for the residual term cannot be as small as $O_M(N^{-1})$, in contrast to what we had for the off-line prediction error method for ARMA processes; see Theorem 2.2.

The above results take a particularly attractive form for partially *stochastic Newton methods*. A recursive estimation method is called a partially stochastic Newton method if the Jacobian matrix of the right-hand side of the associated ODE at $x = x^*$ is of the form

$$\begin{pmatrix} -I & 0 \\ X & Y \end{pmatrix},$$

where I is an identity matrix. An example, the standard recursive prediction error estimation of ARMA processes, in which both the system-parameter θ^* and the Hessian of asymptotic cost function R^* are estimated and the estimates of the system-parameters are updated using Newton-like steps, is a partially stochastic Newton method with respect to the system-parameters.

The above decomposition of the Jacobian is in one-to-one correspondence with the splitting of the parameter vector x as $x = (x^1, x^2)$. With this notation it is easy to see that

$$\frac{\partial}{\partial \xi^1} y(t,s,\xi)_{|\xi=x^*} = \left(\frac{s}{t} I, 0\right)$$

for $s \le t$ and the statement of Theorem 4.3 simplifies to the following.

THEOREM 4.4. *Assume that the conditions of Theorem* 4.3 *are satisfied and that we can split the parameter vector* x *as* $x = (x^1, x^2)$ *so that the estimation method is a partially stochastic Newton method with respect to* x^1. *Let* (Q^1, Q^2) *be the corresponding splitting of* Q. *Then we have with the same* ε_x *as in Theorem* 4.3

$$x_N^1 - x^{1*} = \frac{1}{N}\sum_{n=1}^{N} Q^1(\overline{\phi}_n(x^*)) + O_M(N^{-1/2-\varepsilon_x}).$$

Theorem 4.4 is an extension of Theorem 2.4 to general partially stochastic Newton methods, but with a weaker error term, since $\varepsilon_x < 1/2$.

The result given as (2.23) can also be extended. Let the off-line estimator $\hat{x}_N$ of x^* be defined as the solution of

$$U_N(x) = \sum_{n=1}^{N} Q^1(\overline{\phi}_n(x)) = 0$$

with respect to x. The handling of multiple solutions is precisely described in [18]. Then it is easy to see that Theorem 2.2 can be extended and noting that the Jacobian matrix of the right-hand side of the associated ODE at $x = x^*$ is of the form given above, we get for the first component of $\hat{x}$

$$\hat{x}_N^1 - x^{1*} = \frac{1}{N} \sum_{n=1}^{N} Q^1(\overline{\phi}_n(x^*)) + O_M(N^{-1}).$$

Combining this with Theorem 4.4 and writing $\widehat{\widehat{x}}_N = x_N$ we get

$$\widehat{\widehat{x}}_N^1 - \hat{x}_N^1 = O_M(N^{-1/2-\varepsilon_x}) \tag{4.7}$$

which is an extension of (2.23), albeit with a weaker error term.

Discussion of the result. To *compare* these results with the results of [3, 45] we note that the limit results of [3] are of classical nature: weak convergence and CLT (central limit theorem), which are not strong enough for calculating performance degradation that we called pathwise cumulative regret. The same remark applies to the weak-convergence results of [45].

In the case of stochastic regression methods, developed in [47] and extended in [7, 14], tight bounds for the almost sure rate of convergence of the estimator process are given. But even these results are not applicable in general to get exact asymptotic results for the pathwise cumulative regret, except in very special cases, such as the minimum-variance self-tuning regulator for ARX systems; see [49]. For ARMAX systems these techniques yield only qualitative results; see [48].

Further discussion on mixing conditions. The proof of Theorem 4.1 relies on a moment inequality for weighted multiple integrals of L-mixing processes given in [21]. It is likely that this result can be extended to ϕ-mixing processes, since it is based on the repeated use of an improved Hölder inequality, which does have its variant for ϕ-mixing processes; see Lemmas 8.1 and 8.2 of section 8 and Chapter 7.2 of [15] for further results. Thus it is likely that L-mixing and ϕ-mixing can both be used to derive the results of the present section for Algorithms CR and DR, given as Theorems 4.1 and 4.2.

The situation is quite different for the DFL scheme, where L^+-mixing has already been heavily exploited for getting the rate of convergence of higher order moments; see Theorem 3.3. Furthermore, L^+-mixing is very much used in the context of all three algorithms (Algorithm CR, Algorithm DR, and the DFL scheme) in deriving the results of sections 4 and 5. Moreover, the formulation of the main result of section 5 is given in terms of the concept of L-mixing. It is not clear if a similar result holds in the context of ϕ-mixing. Even the following simple related problem seems to be open: under what conditions is the response of an exponentially stable linear filter, with a ϕ_p-mixing process as its input, ϕ_p-mixing?

Proof of Theorem 4.1. Assume $x^* = 0$. Also we can assume that $\delta G = 0$, namely the term $\delta G(t, x_t)$ can be merged with $\delta H(t, \omega)$. Indeed, the condition that $\delta G(t, x) = O(t^{-1/2-\varepsilon})$ uniformly in x for $x \in D_0$, see Condition 3.3, implies that Condition 3.5 remains valid when $\delta H(t, \omega)$ is replaced by $\delta H(t, \omega) + \delta G(t, x_t)$.

Let us consider the process (x_t) on the interval $[s, qs)$ with $s \geq 1$, $q > 1$, and let $\overline{y}_t$ denote the solution of the ODE (3.15) starting from x_s at time s. Let C_s denote the event that x_t hits ∂D_0 in $[s, qs)$. Then we can write $x_t - \overline{y}_t$ as

$$(1 - \chi_{C_s}) \int_s^t \frac{\partial}{\partial \xi} y(t, r, x_r) \cdot \frac{1}{r} \left(\overline{H}(r, x_r, \omega) + \delta H(r, \omega) \right) \, dr + \chi_{C_s}(x_t - \overline{y}_t) \tag{4.8}$$

with $\overline{H}(r, x, \omega) = H(x, r, \omega) - G(r, x)$ by using Lemma 8.6. Let us now take into account the fact that $y_\xi(t, r, x)$ and $\overline{H}(r, x, \omega)$ are continuously differentiable with respect to x. Hence we can write

$$\frac{\partial}{\partial \xi} y(t, r, x_r) = \frac{\partial}{\partial \xi} y(t, r, 0) + \int_0^1 \frac{\partial^2}{\partial \xi^2} y(t, r, \lambda x_r) d\lambda \cdot x_r$$

and

$$\overline{H}(r, x_r, \omega) = \overline{H}(r, 0, \omega) + \int_0^1 \frac{\partial}{\partial x} \overline{H}(r, \lambda x_r, \omega) d\lambda \cdot x_r.$$

Substituting into (4.8) we get that the first integral on the right-hand side of (4.8) can be written as the sum of the following five terms:

$$\begin{aligned}
I_1 &= \int_s^t \frac{\partial}{\partial \xi} y(t, r, 0) \cdot \frac{1}{r} \overline{H}(r, 0, \omega) dr, \\
I_2 &= \int_s^t \frac{\partial}{\partial \xi} y(t, r, 0) \cdot \frac{1}{r} \int_0^1 \frac{\partial}{\partial x} \overline{H}(r, \lambda x_r, \omega) d\lambda \cdot x_r dr, \\
I_3 &= \int_s^t \int_0^1 \frac{\partial^2}{\partial \xi^2} y(t, r, \lambda x_r) d\lambda \cdot x_r \cdot \frac{1}{r} \overline{H}(r, 0, \omega) dr, \\
I_4 &= \int_s^t \int_0^1 \frac{\partial^2}{\partial \xi^2} y(t, r, \lambda x_r) d\lambda \cdot x_r \cdot \frac{1}{r} \int_0^1 \frac{\partial}{\partial x} \overline{H}(r, \lambda' x_r, \omega) d\lambda' \cdot x_r dr, \\
I_5 &= \int_s^t \frac{\partial}{\partial \xi} y(t, r, x_r) \cdot \frac{1}{r} \delta H(r, \omega) dr.
\end{aligned}$$

We will later also write $I_1 = I_{1,t} = I_{1,t,s}$ when we want to emphasize the dependence of I_1 on t and s. Then we can write

$$x_t - \overline{y}_t = (1 - \chi_{C_s})(I_1 + I_2 + I_3 + I_4 + I_5) + \chi_{C_s}(x_t - \overline{y}_t). \tag{4.9}$$

We will approximate I_2 and I_3 so that we replace λx_r and $\lambda' x_r$ by 0 and define

$$\begin{aligned}
I_2^* &= \int_s^t \frac{\partial}{\partial \xi} y(t, r, 0) \cdot \frac{1}{r} \int_0^1 \frac{\partial}{\partial x} \overline{H}(r, 0, \omega) d\lambda \cdot x_r dr, \\
I_3^* &= \int_s^t \int_0^1 \frac{\partial^2}{\partial \xi^2} y(t, r, 0) d\lambda \cdot x_r \cdot \frac{1}{r} \overline{H}(t, 0, \omega) dr.
\end{aligned}$$

For the sake of notational homogenity we will also write $I_1 = I_1^*$.

LEMMA 4.1. *We have for fixed q and any $s \le t \le qs$,*

$$x_t - \overline{y}_t = I_1^* + I_2^* + I_3^* + O_M(s^{-1/2-\varepsilon}). \tag{4.10}$$

Remark. It will be clear from the proof of the lemma that in the case $\delta H(t,\omega) = 0$ the last term becomes $O_M(s^{-1})$. Indeed the error term $O_M(s^{-1/2-\varepsilon})$ shows up only in the last step of the proof, in the estimation of the effect of I_5. Thus a key factor in the accuracy of the ODE approximation is the perturbation term $\delta H(t,\omega)$.

Proof. Estimation of I_2. We claim that for $s \le t \le qs$ we have

$$I_2 = I_2^* + O_M(s^{-1}). \tag{4.11}$$

Indeed, fix λ and integrate first with respect to r. We expand $\frac{\partial}{\partial x}\overline{H}^i(r, \lambda x_r, \omega)$ (for $i = 1, \dots, p$) into a Taylor series about 0 once more to obtain

$$\frac{\partial}{\partial x}\overline{H}^i(r, \lambda x_r, \omega) = \frac{\partial}{\partial x}\overline{H}^i(r, 0, \omega) + \left(\int_0^1 \frac{\partial^2}{\partial x^2}\overline{H}^i(r, \lambda'\lambda x_r, \omega)d\lambda'\right) \cdot x_r.$$

The expression under the integral term here can be shown to be $O_M(1)$ by the same argument that we used above, since $\overline{H}$ is assumed to have continuous third derivatives almost surely which are also M-bounded. Thus we get $\frac{\partial}{\partial x}\overline{H}(r, \lambda x_r, \omega) = \frac{\partial}{\partial x}\overline{H}(r, 0, \omega) + O_M(r^{-1/2})$. Integration with respect to λ from 0 to 1 and multiplication by $r^{-1}x_r = O_M(r^{-3/2})$ yield an error term $O_M(r^{-2})$. Finally, since $\|\frac{\partial}{\partial \xi}y(t, r, 0)\| \le C_0(r/t)^\alpha$ we get

$$I_2 = \int_s^t \frac{\partial}{\partial \xi}y(t, r, 0) \cdot \frac{1}{r}\frac{\partial}{\partial x}\overline{H}(r, 0, \omega) \cdot x_r dr + O_M(s^{-1}) \tag{4.12}$$

as stated. Note that the dominant term can be estimated by using the moment inequality given as Theorem 8.1. Thus we also get $I_2 = O_M(s^{-1/2})$.

Estimation of I_3. We claim that for $s \le t \le qs$ we have

$$I_3 = I_3^* + O_M(s^{-1}). \tag{4.13}$$

Indeed, in the inner integrand of I_3 we can write

$$\frac{\partial^2}{\partial \xi^2}y(t, r, \lambda x_r) = \frac{\partial^2}{\partial \xi^2}y(t, r, 0) + \left(\int_0^1 \frac{\partial^3}{\partial \xi^3}y(t, r, \lambda'\lambda x_r)d\lambda'\right) \cdot x_r, \tag{4.14}$$

where the last term is to be interpreted as the product of a 4-tensor with a 1-tensor yielding a 3-tensor, thus interpreting $\cdot$ as a tensor product. Substituting (4.14) into the expression of I_3 we get for fixed λ, λ' the product of the following two terms:

$$\frac{\partial^2}{\partial \xi^2}y(t, r, 0) \cdot x_r \cdot \frac{1}{r}\overline{H}(r, 0, \omega) = O_M(r^{-3/2}),$$
$$\frac{\partial^3}{\partial \xi^3}y(t, r, \lambda'\lambda x_r) \cdot x_r \cdot x_r \cdot \frac{1}{r}\overline{H}(r, 0, \omega) = O_M(r^{-2}),$$

where we used the fact that the partial derivatives of $y(t, r, \xi)$ with respect to ξ are bounded by a deterministic constant; see Lemma 8.8. Integrating from s to t the contribution of the integral of the second term is $O_M(s^{-1})$, thus we get

$$I_3 = \int_s^t \left(\frac{\partial^2}{\partial \xi^2}y(t, r, 0) \cdot x_r\right) \cdot \frac{1}{r}\overline{H}(r, 0, \omega)dr + O_M(s^{-1}) \tag{4.15}$$

as stated. Note that the expected upper bound $I_3 = O_M(s^{-1/2})$ cannot be readily derived from the above approximation: we cannot use the moment inequality given as Theorem 8.1 since the weights x_r are random!

Estimation of I_4. We claim that for $s \le t \le qs$ we have

$$I_4 = O_M(t^{-1}). \tag{4.16}$$

Indeed, by Theorem 3.1 we have $x_r = O_M(r^{-1/2})$, hence for fixed λ, λ' the contribution of the term $r^{-1}x_r \cdot x_r$, interpreted as an appropriate tensor product, is $O_M(r^{-2})$. On the other hand, $\|y_{\xi\xi}(t,r,x)\| \le C_0, (r/t)^\alpha \le C_1'$ with some $C_0', C_1' > 0$, uniformly in x for $x \in D_0$ (cf. Lemma 8.8). Third,

$$\left\| \frac{\partial}{\partial x} H(r, \lambda' x_r, \omega) \right\| \le \sup_{x \in D_0^*} \left\| \frac{\partial}{\partial x} H(r, x, \omega) \right\| \triangleq H_x^*(x, r, \omega), \tag{4.17}$$

where D_0^* denotes the star-like closure of D_0. Since by assumption $D_0^* \subset \mathrm{int} D$ and the partial derivative $H_{xx}(r,x,\omega)$ exists and is continuous almost surely and is M-bounded, we get by the maximal inequality, given as Theorem 8.3, that the right-hand side of (4.17) is $O_M(1)$. Hence we finally get, using the triangle inequality, that

$$I_4 = O_M\left(\int_s^t C_0 (r/t)^\alpha r^{-2} dr \right) = O_M(s^{-1})$$

as stated.

Estimation of the effect of $I_{5,t}$. We claim that for $s \le t \le qs$

$$(1 - \chi_{C_s})\ I_{5,t} = O_M(s^{-1/2-\varepsilon}). \tag{4.18}$$

Indeed, we have

$$(1 - \chi_{C_s})\ |I_{5,t}| \le (1 - \chi_{C_s}) \int_s^{t \wedge \tau(s)} \left\| \frac{\partial}{\partial \xi} y(t, r, 0) \right\| \ \frac{1}{r} |\delta H(r, \omega)|\ dr,$$

since for $t > \tau(s)$ we have $1 - \chi_{C_s} = 0$. Noting that $\|y_\xi(t,r,0)\| \le C_0$ and taking into account Condition 3.5 we get the claim.

Write now $x_t - \overline{y}_t$ as

$$I_1 + I_2 + I_3 + I_4 + (1 - \chi_{C_s}) I_5 - \chi_{C_s}(I_1 + I_2 + I_3 + I_4) + \chi_{C_s}(x_t - \overline{y}_t), \tag{4.19}$$

and estimate the contribution of the last two terms. Note that

$$I_1 + I_2 + I_3 + I_4 = \int_s^t \frac{\partial}{\partial \xi} y(t, r, x_r) \cdot \frac{1}{r} \overline{H}(r, x_r, \omega)\ dr = O_M(1). \tag{4.20}$$

Indeed, $||y_\xi(t,r,x_r)|| \le C_0$ and $\overline{H}(r, x_r, \omega)$ is M-bounded; see the argument leading to (4.17). Similarly $|x_t - \overline{y}_t| = O_M(1)$, actually we have $|x_t - \overline{y}_t| = O(1)$. As for χ_{C_s} we have the following lemma that has been given as Lemma 2.3 in [19].

LEMMA 4.2. *Consider the continuous-time recursive estimation scheme given by* (3.16) *with the resetting mechanism* (3.17) *and* (3.18). *Assume that Conditions* 3.1–3.5 *are satisfied. Let* C_s *denote the event that* x_t *hits* ∂D_0 *in the interval* $[s, qs)$. *Then for any* $m \ge 1$ *we have* $P(C_s) = O(s^{-m})$.

Thus the contribution of the last two terms in (4.19) is $O_M(s^{-m})$ for any $m \geq 1$ and with this Lemma 4.1 has been proved. □

Our next step is to show that the dominant term is I_1^*; i.e., the terms I_2^* and I_3^* are negligible. This is stated in the next lemma, which is the *key lemma* for the proof of Theorem 4.1. Its proof requires a new tool, specially designed for the present application: moment inequalities for double integrals of L-mixing processes.

LEMMA 4.3. *We have for* $s \leq t \leq qs$

$$x_t - \overline{y}_t = I_1^* + O_M(s^{-1/2-\varepsilon}).$$

Proof. In order to obtain sharper estimates of I_2^* and I_3^* let us write

$$I_2^* + I_3^* = \int_s^t g_r x_r \, dr, \tag{4.21}$$

where the matrix-valued process (g_r) is defined by

$$g_r = \frac{\partial}{\partial \xi} y(t, r, 0) \cdot \frac{1}{r} \frac{\partial}{\partial x} \overline{H}(r, 0, \omega) + \frac{\partial^2}{\partial \xi^2} y(t, r, 0) \cdot \frac{1}{r} \overline{H}(r, 0, \omega). \tag{4.22}$$

Thus we can write Lemma 4.1 as

$$x_t = \overline{y}_t + I_1^* + \int_s^t g_r x_r dr + O_M(s^{-1/2-\varepsilon}). \tag{4.23}$$

If we had $x_r = x$ a small constant, then we could write the integral on the right-hand side of (4.23) as $\int_s^t g_r dr \cdot x$, which then could be estimated by the moment inequality given as Theorem 8.1, since both $\overline{H}_x(r, 0, \omega)$ and $\overline{H}(r, 0, \omega)$ are zero-mean L-mixing processes. If x is small, then the contribution of this term will be negligible. To show that the second term in (4.23) is indeed negligible we iterate (4.23), i.e., substitute x_r by the expression that is given by (4.23). Writing $I_1^* = I_{1,t}^*$ we get for x_t the expression

$$\overline{y}_t + I_{1,t}^* + \int_s^t g_r \left(\overline{y}_r + I_{1,r}^* + \int_s^r g_p x_p dp + O_M(s^{-1/2-\varepsilon}) \right) dr + O_M(s^{-1/2-\varepsilon}). \tag{4.24}$$

Let us set

$$J_1 = \int_s^t g_r \overline{y}_r dr, \qquad J_2 = \int_s^t g_r I_{1,r}^* dr, \qquad J_3 = \int_s^t g_r \int_s^r g_p x_p \, dp \, dr.$$

The last term of the double integral in (4.24) yields $\int_s^t g_r O_M(s^{-1/2-\varepsilon}) dr = O_M(s^{-1/2-\varepsilon})$ since $\int_s^t g_r dr = O_M(1)$, therefore the effect of this term can be merged into the final residual term of (4.24). Thus we get

$$x_t - \overline{y}_t = I_{1,t}^* + J_1 + J_2 + J_3 + O_M(s^{-1/2-\varepsilon}). \tag{4.25}$$

We show that $J_1 + J_2 + J_3 = O_M(s^{-1})$.

Estimation of J_1. To estimate J_1 write it as

$$J_1 = \int_s^t g_r y(r, s, x_s) dr = L_1(x_s) \quad \text{with} \quad L_1(x) = \int_s^t g_r y(r, s, x) dr.$$

Note that $L_1(0) = 0$. To estimate $L_1(x_s)$ consider a Taylor series expansion of $L_1(x)$ around 0:

$$L_1(x_s) = \int_0^1 L_{1x}(\lambda x_s)d\lambda \cdot x_s, \tag{4.26}$$

where $L_{1x} = (\partial/\partial x)L_1(x)$. It is easy to see that in computing L_{1x} differentiation and integration can be interchanged, thus we can write

$$L_{1x}(x) = \int_s^t \left(\frac{\partial}{\partial \xi} y(t,r,0) \cdot \frac{1}{r} \frac{\partial}{\partial x} \overline{H}(r,0,\omega) + \frac{\partial^2}{\partial \xi^2} y(t,r,0) \cdot \frac{1}{r} \overline{H}(r,0,\omega) \right) \cdot \frac{\partial}{\partial x} y(r,s,x)\ dr. \tag{4.27}$$

Since $y_\xi(t,r,0)$, $y_{\xi\xi}(t,r,0)$, and $y_x(r,s,x)$ are deterministic and bounded and $\overline{H}_x(r,0,\omega)$ and $\overline{H}(r,0,\omega)$ are zero-mean L-mixing processes we get by the moment inequality, given as Theorem 8.1, that for each fixed $x \in D_0$

$$L_{1x}(x) = O_M\left(\int_s^t \frac{1}{r^2} dr \right)^{1/2} = O_M(s^{-1/2}).$$

Using similar arguments and taking into account that G is three-times continuously differentiable, we obtain that $L_{1xx}(x) = (\partial^2/\partial x^2)L_1(x) = O_M(s^{-1/2})$. Using now the maximal inequality, given as Theorem 8.3, we get

$$\|L_{1x}(\lambda x_s)\| \le \sup_{x \in D_0^*} \|L_{1x}(x)\| = O_M(s^{-1/2}).$$

Taking into account that $x_s = O_M(s^{-1/2})$ we finally get

$$J_1 = L_1(x_s) = O_M(s^{-1}). \tag{4.28}$$

Estimation of J_2. To estimate J_2 let us use the definition of $I_{1,t}^*$ and write

$$J_2 = \int_s^t g_r \int_s^r f_{1,v} \overline{H}(v,0,\omega) dv,$$

where the modulating function $f_{1,v}$ is

$$f_{1,v} = \frac{\partial}{\partial \xi} y(r,v,0) \cdot \frac{1}{v}.$$

Write g_r as

$$g_r = f_{2,1,r} \frac{\partial}{\partial x} \overline{H}(r,0,\omega) + f_{2,2,r} \overline{H}(r,0,\omega),$$

where the modulating functions are

$$f_{2,1,r} = \frac{\partial}{\partial \xi} y(t,r,0) \cdot \frac{1}{r} \quad \text{and} \quad f_{2,2,r} = \frac{1}{r} \frac{\partial^2}{\partial \xi^2} y(t,r,0) \cdot \frac{1}{r}.$$

Noting that $y_\xi(t,r,0)$ and $y_{\xi\xi}(t,r,0)$ are bounded and applying the moment inequality for double integrals of L-mixing processes, given as Theorem 8.2 in section 8, we get

$$J_2 = O_M(s^{-1}). \tag{4.29}$$

Estimation of J_3. For J_3 we get, after interchanging the order of integration,

$$J_3 = \int_s^t g_p x_p \left(\int_p^t g_r dr \right) \, dp.$$

For the inner integral we have

$$\int_p^t g_r dr = O_M(p^{-1/2})$$

by the moment inequality given as Theorem 8.1. Since $g_p = O_M(p^{-1})$ we have $g_p x_p = O_M(p^{-3/2})$ and thus the integrand of the outer integral is of the order of magnitude $O_M(p^{-2})$. It follows that

$$J_3 = O_M(s^{-1}). \tag{4.30}$$

Thus we conclude that indeed $J_1 + J_2 + J_3 = O_M(s^{-1})$, and substituting this into (4.25) the proof of Lemma 4.3 is complete. □

Pasting together. Let us now take a subdivision of the half-line $[1, \infty)$ by the points q^i with $q > 1$ and let us consider an interval $q^n \le t < q^{n+1}$. Let us define for $i \ge 1$

$$\delta_i = I^*_{1,q^i,q^{i-1}} = \int_{q^{i-1}}^{q^i} \frac{\partial}{\partial \xi} y(q^i, r, 0) \frac{1}{r} \overline{H}(r, 0, \omega) dr.$$

Note that $\delta_i = O_M(q^{-i/2})$.

LEMMA 4.4. *We have for* $q^n \le t < q^{n+1}$

$$x_t - y_t = \sum_{i=1}^n \frac{\partial}{\partial \xi} y(t, q^i, 0) \delta_i + I^*_{1,t,q^n} + O_M(q^{-n(1/2+\varepsilon_x)}). \tag{4.31}$$

Proof. Using Lemma 8.7 of section 8 with $s_i = q^i$, $i = 0, 1, \ldots, n$, $s_{n+1} = t$, we get the following expression for $x_t - y_t$:

$$\sum_{i=1}^n \int_0^1 \frac{\partial}{\partial \xi} y(t, q^i, w(\lambda)) \, d\lambda \cdot \left(x_{q^i} - y(q^i, q^{i-1}, x_{q^{i-1}}) \right) + \left(x_t - y(t, q^n, x_{q^n}) \right), \tag{4.32}$$

where $w(i, \lambda) = (1 - \lambda) y(q^i, q^{i-1}, x_{q^{i-1}}) + \lambda x_{q^i}$. Taking into account Lemma 4.3 write the ith local tracking error $x_{q^i} - y(q^i, q^{i-1}, x_{q^{i-1}})$ in the form $I^*_{1,q^i,q^{i-1}} + O_M(q^{-(i-1)(1/2+\varepsilon)}) = \delta_i + O_M(q^{-(i-1)(1/2+\varepsilon)})$ to get

$$\begin{aligned} x_t - y_t = \sum_{i=1}^n \int_0^1 \frac{\partial}{\partial \xi} y(t, q^i, w(\lambda)) d\lambda \cdot \left(\delta_i + O_M(q^{-(i-1)(1/2+\varepsilon)}) \right) \\ + I^*_{1,t,q^n} + O_M(q^{-n(1/2+\varepsilon)}). \end{aligned} \tag{4.33}$$

To estimate the cumulative effect of the error terms $O_M(q^{-(i-1)(1/2+\varepsilon)})$ note that we have

$$\| y_\xi(t, q^i, w(i, \lambda)) \| \le C_0 (q^i/t)^\alpha \le C_0 (q^i/q^n)^\alpha = C_0 q^{-\alpha(n-i)},$$

thus we get an upper bound

$$\sum_{i=1}^{n}\int_0^1 C_0 q^{-\alpha(n-i)}\cdot O_M(q^{-(i-1)(1/2+\varepsilon)})+O_M(q^{-n(1/2+\varepsilon)}). \tag{4.34}$$

This expression can be estimated from above by using the remark after Lemma 8.5 of section 8, given as (8.6), applied for the sequences $(q^{-\alpha i})$ and $(q^{-(1/2+\varepsilon)i})$, the convolution of which is bounded from above by $C\ \max(q^{-\alpha n}, q^{-(1/2+\varepsilon)n})$ assuming that $\alpha\neq 1/2+\varepsilon$. Since $\max(-\alpha,-(1/2+\varepsilon))=-\min(\alpha,1/2+\varepsilon)=-(1/2+\min(\overline{\alpha},\varepsilon))$, we get

$$x_t-y_t=\sum_{i=1}^{n}\int_0^1\frac{\partial}{\partial\xi}y(t,q^i,w(i,\lambda))d\lambda\cdot\delta_i+I^*_{1,t,q^n}+O_M(q^{-n(1/2+\varepsilon_x)}). \tag{4.35}$$

To further simplify the right-hand side of (4.34) we replace $w(i,\lambda)$ by 0. Note that by Lemma 8.8 of section 8

$$\left\|\frac{\partial}{\partial\xi}y(t,q^i,w(i,\lambda))-\frac{\partial}{\partial\xi}y(t,q^i,0)\right\|\le C_0'(q^i/t)^\alpha|w(i,\lambda)|,$$

and hence cumulative error of this approximation is majorized by $C_0'\sum_{i=1}^n(q^i/t)^\alpha\cdot|w(i,\lambda)|\cdot\delta_i$. Note that $w(i,\lambda)=O_M(q^{-i/2})$, uniformly in λ since $x_{q^i}=O_M(q^{-i/2})$ by Theorem 3.1 and $|\overline{y}_{q^i}|=|y(q^i,q^{i-1},x_{q^{i-1}})|\le C_0|x_{q^{i-1}}|$, therefore $w(i,\lambda)\cdot\delta_i=O_M(q^{-i})$ uniformly in λ. Thus the the cumulative error of the last approximation is bounded from above by

$$C_0'\sum_{i=1}^{n}(q^i/t)^\alpha\cdot O_M(q^{-i})\le\sum_{i=1}^{n}q^{-\alpha(n-i)}\cdot O_M(q^{-i})=O_M(q^{-n\alpha'})$$

with $\alpha'=\min(\alpha,1)_-$, by the remark after Lemma 8.5, given as (8.6). Since $1/2+\varepsilon<1$, we have $\alpha'\ge\min(\alpha,1/2+\varepsilon)_-=\varepsilon_x$ and with this the proof of the lemma is complete. □

Now the ith term on right-hand side of (4.31) can be written as

$$\begin{aligned}\frac{\partial}{\partial\xi}y(t,q^i,0)\int_{q^{i-1}}^{q^i}\frac{\partial}{\partial\xi}y(q^i,r,0)\frac{1}{r}\overline{H}(r,0,\omega)dr\\ =\int_{q^{i-1}}^{q^i}\frac{\partial}{\partial\xi}y(t,r,0)\frac{1}{r}\overline{H}(r,0,\omega)dr,\end{aligned} \tag{4.36}$$

thus the cumulative contribution of the dominant terms in (4.31) is exactly what is the dominant term in Theorem 4.1. Since $y_t=O(t^{-\alpha})=O(t^{-(1/2+\overline{\alpha})})$ the term can be merged into the residual term $O_M(q^{-n(1/2+\varepsilon_x)})$ and thus the proof of Theorem 4.1 has been completed. □

Proof of Theorem 4.2. Let $(H^c(t,x,\omega))$ be the piecewise constant extension of $(H(n,x,\omega))$ defined under (3.35) and define a piecewise linear extension of (x_n) for $1\le n\le t\le n+1$ by

$$x^l_t=(t-n)x_n+(n+1-t)x_{n-1}\quad\text{if}\quad x_{n-}\in\mathrm{int}D_0,\quad x^l_1=x_0=\xi_0. \tag{4.37}$$

On the other hand, if $x_{n-} \notin \text{int}D_0$, then we reset x^l to its initial value ξ_0 at time $t = n$ and put a hold on the recursion until $t = n + 1$; i.e., we set

$$x_t^l = \xi_0 \quad \text{for} \quad 1 \le n < t \le n+1 \quad \text{if} \quad x_{n-} \notin \text{int}D_0. \tag{4.38}$$

Now it is easy to see that in intervals $n \le t \le n+1$ where no resetting takes place (x_t^l) satisfies a differential equation of the form

$$\dot{x}_t^l = \frac{1}{t}(H^c(t, x, \omega) + \delta H(t, \omega)), \tag{4.39}$$

where $\delta H(t, \omega) = \delta H^c(n, \omega) + O_M(t^{-1})$, cf. [19, (2.13)]. The conditions of Theorem 4.1 can be easily verified for the above procedure, except that we use the alternative resetting mechanism given by (3.17) and (3.19). Thus we get by Theorem 4.1

$$x_t^l - x^* = \int_1^t \frac{\partial}{\partial \xi} y(t, s, x^*) \frac{1}{s} H^c(s, x^*, \omega) ds + O_M(t^{-1/2-\varepsilon_x}). \tag{4.40}$$

Let $t = N$ be an integer and let $n \le s < n + 1$, with n being integer. We have

$$\left\| \frac{\partial}{\partial \xi} y(t, s, x^*) - \frac{\partial}{\partial \xi} y(t, n, x^*) \right\| \le C_0' \left(\frac{s}{t}\right)^{\alpha} \frac{1}{t}. \tag{4.41}$$

Indeed, by Lemma 8.8

$$||y_{r\xi}(t, r, x^*)|| \le C_0'(r/t)^{\alpha} \cdot \left\| \frac{1}{t} G_{\xi}(x^*) \right\|.$$

Integrating $y_{r\xi}(t, r, \xi)$ between n and s we get (4.41). Now replacing $y_{\xi}(t, s, x^*)$ by $y_{\xi}(t, n, x^*)$ in (4.40), noting that $\frac{1}{s} - \frac{1}{n} = O(\frac{1}{s^2})$ and taking into account that $\overline{H}^c(t, x^*, \omega)$ is M-bounded we get that the cumulative error is of the order of magnitude

$$O_M\left(\int_1^t \left(\frac{s}{t}\right)^{\alpha} \cdot \frac{1}{s^2}\, ds \right) = O_M(t^{-1}), \tag{4.42}$$

which can be merged into the residual term $O_M(t^{-1/2-\varepsilon_x})$ and thus the proof of Theorem 4.2 is complete. □

Proof of Theorem 4.3. Defining H and δH as in (3.46) and (3.56), the conditions of Theorem 4.1 have been verified in section 5 of [19]. In particular, the critical Condition 3.5 is verified in Lemma 5.6 in [19] (restated as Lemma 3.2 in the present paper), thus the claim follows. □

5. The transformed error process is *L*-mixing. In this section we derive a useful corollary of Theorem 4.1, stating that an appropriate transformation of the error process $x_t - x^*$ is L-mixing. Define the transformed process

$$\tilde{x}_r = e^{r/2}(x_{e^r} - x^*). \tag{5.1}$$

The weak limit of the shifted process $(\tilde{x}_{r+\rho})$, when $\rho \to \infty$, is established in [5] and Theorem 13, Chapter 4.5, Part II of [3], under conditions, which are different from the conditions of the present paper. It is proven that $(\tilde{x}_{r+\rho})$ converges weakly to the solution of the linear stochastic differential equation

$$d\tilde{z}_r = (A^* + I/2)\tilde{z}_r + d\tilde{w}_r, \tag{5.2}$$

with zero initial condition; in short,

$$(\tilde{x}_{r+\rho}) \to (\tilde{z}_r) \tag{5.3}$$

in a weak sense, where (cf. (3.24))

$$A^* = \frac{\partial G(x)}{\partial x}\Big|_{x=x^*},$$

assuming that $(A^* + I/2)$ is stable. Here $d\tilde{w}_r$ is the stochastic differential of a Wiener-process, with some covariance matrix $P^* dt$. The weak limit $(\tilde{z}_r)$ is an L-mixing process with respect to the pair of σ-algebras $(\tilde{\mathcal{F}}_r, \tilde{\mathcal{F}}_r^+)$ generated by the past and future increments of the Wiener-process $(\tilde{w}_r)$, respectively. Hence it is indicated, but not implied by (5.3), that the transformed process $(\tilde{x}_r)$ itself is also L-mixing with respect to some pair of σ-algebras $(\tilde{\mathcal{F}}_r, \tilde{\mathcal{F}}_r^+)$. We prove that this is indeed the case.

The emphasis is on the nonasymptotic nature of our result. An analogous result for off-line prediction error estimators of ARMA parameters has been proved in [24]. It extends to RPE estimators due to the strong approximation result given in [22]. It has also been shown in [24] that this result is instrumental in deriving a pathwise characterization of performance degradation of an online adaptive predictor. Like in section 4, we assume that $\delta G(t, y) = 0$, which implies $\mathrm{E} H(s, x^*, \omega) = 0$ exactly for all s.

THEOREM 5.1. *Consider the continuous-time recursive estimation scheme given by* (3.16) *with the resetting mechanism* (3.17) *and* (3.18). *Assume that the conditions of Theorem* 4.1 *are satisfied. Then the transformed process* $(\tilde{x}_r)$ *is* L*-mixing with respect to* $(\mathcal{F}_{e^r}, \mathcal{F}_{e^r}^+)$.

Proof. Approximation, dynamic representation, and discretization of the process $(\tilde{x}_r)$. By Theorem 4.1 the dominant term in the error process is

$$\int_1^t \frac{\partial}{\partial \xi} y(t, s, x^*) \frac{1}{s} H(s, x^*, \omega) ds,$$

and the error term is $O_M(t^{-1/2+\varepsilon_x})$. We transform the dominant term and the residual term $O_M(t^{-1/2+\varepsilon_x})$ in the same way as the error process itself (cf. (5.1)): we multiply by $t^{1/2}$ and introduce the new variables $t = e^r$ and $s = e^p$.

Now, since $\frac{\partial}{\partial \xi} y(t, s, x^*)$ is the solution of the variational equation

$$\frac{\partial}{\partial t} \frac{\partial}{\partial \xi} y(t, s, x^*) = \frac{1}{t} A^* \frac{\partial}{\partial \xi} y(t, s, x^*)$$

with initial condition $\frac{\partial}{\partial \xi} y(s, s, x^*) = I$, we get, using an exponential change of time-scale followed by an inverse change of time-scale, that

$$\frac{\partial}{\partial \xi} y(t, s, x^*) = e^{A^* \log(t/s)}.$$

Thus the dominant term in the error process gets transformed into

$$\tilde{x}_{1,r} = e^{r/2} \int_0^r e^{A^*(r-p)} H(e^p, x^*, \omega) dp. \tag{5.4}$$

Now Theorem 4.1 implies the following.

CLAIM. *We have*

$$\tilde{x}_r - \tilde{x}_{1,r} = O_M(e^{-\varepsilon_x r}). \tag{5.5}$$

A dynamic representation of $\tilde{x}_{1,r}$ is obtained by differentiating (5.4) with respect to r. Then we get that $\tilde{x}_{1,r}$ satisfies the differential equation

$$\frac{d}{dr}\tilde{x}_{1,r} = (A^* + I/2)\tilde{x}_{1,r} + e^{r/2}H(e^r, x^*, \omega), \qquad r \geq 0. \tag{5.6}$$

The dynamics satisfied by $(\tilde{x}_{1,r})$ is similar to the dynamics satisfied by (z_r), given by (5.2), but the process $e^{r/2}H(e^r, x^*, \omega)$ is not a good approximation to the increments of a Wiener-process; it is not even M-bounded. This difficulty can be avoided using discretization and averaging. Take a small, fixed positive number h and consider the discrete-time sampled process $\tilde{x}_{1,nh}$. It satisfies the discrete-time dynamics

$$\tilde{x}_{1,(n+1)h} = e^{(A^*+I/2)h}\tilde{x}_{1,nh} + \int_{nh}^{(n+1)h} e^{(A^*+I/2)((n+1)h-p)}e^{p/2}H(e^p, x^*, \omega)dp.$$

Note that the input process is obtained as a weighted average of the input process of (5.6) over the interval $[nh, (n+1)h]$. Denote the second term on the right-hand side, which is the input process for the discretized system, by $(\tilde{u}_{1,n})$; i.e., set

$$\tilde{u}_{1,n+1} = \int_{nh}^{(n+1)h} e^{(A^*+I/2)((n+1)h-p)}e^{p/2}H(e^p, x^*, \omega)dp. \tag{5.7}$$

The discrete-time dynamics.

$$\tilde{x}_{1,(n+1)h} = e^{(A^*+I/2)h}\tilde{x}_{1,nh} + \tilde{u}_{1,n+1}, \qquad n \geq 0, \tag{5.8}$$

with zero initial condition. In what follows we develop a series of approximations of the process $\tilde{u}_{1,n+1}$.

An averaging effect for the process $(\tilde{u}_{1,n})$. Going back to the original time-scale in (5.7) we can write $\tilde{u}_{1,n+1}$ as

$$\tilde{u}_{1,n+1} = \int_{e^{nh}}^{e^{(n+1)h}} \left(\frac{s}{e^{(n+1)h}}\right)^{(-A^*-I/2)} \frac{1}{s^{1/2}} H(s, x^*, \omega)ds. \tag{5.9}$$

CLAIM U1. *For the order of magnitude of* $(\tilde{u}_{1,n+1})$ *we have*

$$\tilde{u}_{1,n+1} = O_M(h^{1/2}). \tag{5.10}$$

Indeed, using the moment inequality given as Theorem 8.1 we get that for any $q \geq 1$

$$\mathrm{E}^{1/q}|\tilde{u}_{1,n+1}|^q \leq C_q\left(\int_{e^{(nh}}^{e^{(n+1)h}} \left\|\left(\frac{s}{e^{(n+1)h}}\right)^{(-A^*-I/2)} \frac{1}{s^{1/2}}\right\|^2 ds\right)^{1/2} \cdot M_q^{1/2}(H(x^*))\Gamma_q^{1/2}(H(x^*)).$$

Note that for $e^{nh} \leq s \leq e^{(n+1)h}$, $0 < h \leq h_0$, with some $0 < h_0$ fixed,

$$\left\|\left(\frac{s}{e^{(n+1)h}}\right)^{(-A^*-I/2)}\right\| = ||e^{(p-(n+1)h)(-A^*-I/2)}|| \leq C, \tag{5.11}$$

where C is independent of n and h, since the set of matrices $e^{(p-(n+1)h)(-A^*-I/2)}$ with p varying between nh and $(n+1)h$ is compact. Thus we get

$$\mathrm{E}^{1/q}|\tilde{u}_{1,n+1}|^q \le C_q\left(\int_{e^{nh}}^{e^{(n+1)h}} C^2\frac{1}{s}ds\right)^{1/2}\cdot M_q^{1/2}(H(x^*))\Gamma_q^{1/2}(H(x^*)),$$

and the right-hand side is $O(h^{1/2})$ indeed, as stated.

Truncated averaging: The process $(\tilde{u}_{2,n})$ and choosing δ_n and ε_δ. To eliminate the dependence in the process $(\tilde{u}_{1,n})$ we follow standard procedures, as described, e.g., in [42]. First we remove a small portion of the integral by decreasing the upper limit of the integration to $e^{(n+1)h}-\delta_{n+1}$ with some positive δ_{n+1}. Since the original range of the integration has length $e^{(n+1)h}-e^{nh}=O(he^{nh})$ a reasonable choice for δ_{n+1} is

$$\delta_{n+1} = he^{\varepsilon_\delta nh} \tag{5.12}$$

with $0<\varepsilon_\delta<1$. Thus we define

$$\tilde{u}_{2,n+1} = \int_{e^{nh}}^{e^{(n+1)h}-\delta_{n+1}}\left(\frac{s}{e^{(n+1)h}}\right)^{(-A^*-I/2)}\frac{1}{s^{1/2}}H(s,x^*,\omega)ds. \tag{5.13}$$

Claim U2. *We have for $h>0$*

$$\tilde{u}_{1,n+1}-\tilde{u}_{2,n+1} = O_M(h^{1/2}e^{-(1-\varepsilon_\delta)nh/2}) \quad\text{and}\quad \tilde{u}_{1,n+1}-\tilde{u}_{2,n+1}=O_M(h^{1/2}). \tag{5.14}$$

For the proof first note that

$$e^{(n+1)h}-\delta_{n+1}\ge e^{nh}. \tag{5.15}$$

Indeed, this is equivalent to $\delta_{n+1}\le e^{(n+1)h}-e^{nh}=e^{nh}(e^h-1)$ and since $\delta_{n+1}<he^{nh}$ and $h<(e^h-1)$, the validity of (5.15) follows.

The error of the approximation is

$$\tilde{u}_{1,n+1}-\tilde{u}_{2,n+1} = \int_{e^{(n+1)h}-\delta_{n+1}}^{e^{(n+1)h}}\left(\frac{s}{e^{(n+1)h}}\right)^{(-A^*-I/2)}\frac{1}{s^{1/2}}H(s,x^*,\omega)ds,$$

which can be estimated by the moment inequality given as Theorem 8.1. Thus we get that $\mathrm{E}^{1/q}|\tilde{u}_{1,n+1}-\tilde{u}_{2,n+1}|^q$ is bounded from above by

$$C_q\left(\int_{e^{(n+1)h}-\delta_{n+1}}^{e^{(n+1)h}}\left\|\left(\frac{s}{e^{(n+1)h}}\right)^{(-A^*-I/2)}\frac{1}{s^{1/2}}\right\|^2 ds\right)^{1/2}\cdot M_q^{1/2}(H(x^*))\Gamma_q^{1/2}(H(x^*)).$$

Taking into account the kernel estimate given above as (5.11) we get for $\mathrm{E}^{1/q}|\tilde{u}_{1,n+1}-\tilde{u}_{2,n+1}|^q$ the upper bound

$$C_q\left(\int_{e^{(n+1)h}-\delta_{n+1}}^{e^{(n+1)h}}\frac{C^2}{s}ds\right)^{1/2}\cdot M_q^{1/2}(H(x^*))\Gamma_q^{1/2}(H(x^*)). \tag{5.16}$$

For the integral term we have

$$\left(\int_{e^{(n+1)h}-\delta_{n+1}}^{e^{(n+1)h}}\frac{1}{s}ds\right)^{1/2} = (\log e^{(n+1)h}-\log(e^{(n+1)h}-\delta_{n+1}))^{1/2},$$

which is majorized by

$$\left(\delta_{n+1}/(e^{(n+1)h}-\delta_{n+1})\right)^{1/2}.$$

Since $e^{(n+1)h}-\delta_{n+1}\geq e^{nh}$, we can continue the above inequality to get

$$\left(\int_{e^{(n+1)h}-\delta_{n+1}}^{e^{(n+1)h}}\frac{1}{s}ds\right)^{1/2}\leq\left(\delta_{n+1}/e^{nh}\right)^{1/2}.$$

Taking into account the definition of δ_{n+1} we get

$$\left(\delta_{n+1}/e^{nh}\right)^{1/2}=\left(he^{\varepsilon_\delta nh}/e^{nh}\right)^{1/2}=h^{1/2}e^{(\varepsilon_\delta-1)nh/2}.$$

Combining the latter inequalities with (5.16) we get the first part of Claim U2, given as (5.14), while the second part is a trivial consequence.

The independent sequence $(\tilde{u}_{3,n})$. This is a key step in our arguments. We complete the construction of an approximating process of $(\tilde{u}_{1,n})$ by projecting $\tilde{u}_{2,n+1}$ on the relative future $\mathcal{F}^+_{e^{nh}-\delta_n}$. In fact, assuming that the conditional expectation operator and integration can be interchanged, we define

$$\tilde{u}_{3,n+1}=\int_{e^{nh}}^{e^{(n+1)h}-\delta_{n+1}}\left(\frac{s}{e^{(n+1)h}}\right)^{(-A^*-I/2)}\frac{1}{s^{1/2}}\mathrm{E}(H(s,x^*,\omega)|\mathcal{F}^+_{e^{nh}-\delta_n})ds. \tag{5.17}$$

It is obvious that $(\tilde{u}_{3,n})$ constitutes an independent sequence of random variables adapted to $\mathcal{F}_{e^{nh}}$.

Remark. We will now approximate the process $\tilde{u}_{2,n+1}$ and get two dual error bounds. The first error bound ensures that the error is exponentially decaying, but there is multiplicative factor h^{-c} with $c>0$, while the second bound ensures that the approximating process itself is of the order $O_M(h^{1/2})$.

Claim U3. *We have with* $c>0$ *that shows up in Condition* 3.1 *(see the definition of* L^+*-mixing), the following two estimates:*

$$\tilde{u}_{2,n+1}-\tilde{u}_{3,n+1}=O_M(h^{-c}e^{-(1/2+c\varepsilon_\delta)nh}) \text{ and } \tilde{u}_{2,n+1}-\tilde{u}_{3,n+1}=O_M(h^{1/2}). \tag{5.18}$$

First we show that $(\tilde{u}_{3,n+1})$ is an M-bounded sequence. Indeed, write $\tilde{u}_{3,n+1}$ as

$$\tilde{u}_{3,n+1}=\mathrm{E}\left(\int_{e^{nh}}^{e^{(n+1)h}-\delta_{n+1}}\left(\frac{s}{e^{(n+1)h}}\right)^{(-A^*-I/2)}\frac{1}{s^{1/2}}H(s,x^*,\omega)ds \;\middle|\; \mathcal{F}^+_{e^{nh}-\delta_n}\right)$$

and estimate the L_q-norm of the right-hand side using Jensen's inequality. Taking into account (5.10), modified so that upper limit of the integration is reduced to the nonrandom upper limit $e^{(n+1)h}-\delta_{n+1}$, we get the claimed M-boundedness of $(\tilde{u}_{3,n+1})$ and in fact we get

$$\tilde{u}_{3,n+1}=O_M(h^{1/2}), \tag{5.19}$$

and thus the second part of the Claim U3 is proved.

To bound the approximation error more accurately define for $e^{nh}\leq s\leq e^{(n+1)h}-\delta_{n+1}$

$$v_s=\left|\left(\frac{s}{e^{(n+1)h}}\right)^{(-A^*-I/2)}\frac{1}{s^{1/2}}\left(H(s,x^*,\omega)-\mathrm{E}(H(s,x^*,\omega)|\mathcal{F}^+_{e^{nh}-\delta_n})\right)\right|.$$

Then obviously

$$|\tilde{u}_{2,n+1} - \tilde{u}_{3,n+1}| \le \int_{e^{nh}}^{e^{(n+1)h}-\delta_{n+1}} v_s ds.$$

and by the triangle inequality for the L_q-norm for $q \ge 1$,

$$\mathrm{E}^{1/q}|\tilde{u}_{2,n+1} - \tilde{u}_{3,n+1}|^q \le \int_{e^{nh}}^{e^{(n+1)h}-\delta_{n+1}} \mathrm{E}^{1/q} v_s^q ds. \tag{5.20}$$

To estimate v_s, note that $||(s/e^{(n+1)h})^{(-A^*-I/2)}|| \le C$ with some C for all s, n, and h with $0 < h \le h_0$ and $s^{-1/2} \le e^{-nh/2}$. On the other hand, we have for any $q \ge 1$

$$\mathrm{E}^{1/q}|(H(s,x^*,\omega) - \mathrm{E}(H(s,x^*,\omega)|\mathcal{F}^+_{e^{nh}-\delta_n}))|^q \le \gamma_q(s-(e^{nh}-\delta_n), H(x^*)),$$

and thus

$$\mathrm{E}^{1/q} v_s^q \le Ce^{-nh/2}\gamma_q(s-(e^{nh}-\delta_n), H(x^*)).$$

It follows that $\mathrm{E}^{1/q}|\tilde{u}_{3,n+1} - \tilde{u}_{2,n+1}|^q$ can be bounded from above by

$$Ce^{-nh/2}\int_{e^{nh}}^{e^{(n+1)h}-\delta_{n+1}} \gamma_q(s-(e^{nh}-\delta_n), H(x^*))ds. \tag{5.21}$$

By Condition 3.1 $\gamma_q(s-(e^{nh}-\delta_n), H(x^*)) \le C(1+\delta_n)^{-1-c}$. Furthermore, note that the range of $\tau(s) = s-(e^{nh}-\delta_n)$ is included in the semi-infinite interval $[\delta_n, \infty)$; thus

$$\begin{aligned}\int_{e^{nh}}^{e^{(n+1)h}-\delta_{n+1}} \gamma_q(s-(e^{nh}-\delta_n), H(x^*))ds &\le \int_{\delta_n}^{\infty} \gamma_q(\tau, H(x^*))d\tau \\ &\le C'(1+\delta_n)^{-c} < C'\delta_n^{-c}.\end{aligned}$$

Combining this with (5.21) and taking into account the definitions of the lag δ_n given by (5.12), we get for any $1 \le q < \infty$

$$\mathrm{E}^{1/q}|\tilde{u}_{2,n+1} - \tilde{u}_{3,n+1}|^q \le C_q e^{-nh/2} h^{-c} e^{-c\varepsilon_\delta nh}$$

with some C_q, which is independent of n and h and this is equivalent to the first part of Claim U3, given as (5.18).

The final approximating process $(\tilde{x}_{3,nh})$. We are going to define a final approximation to $\tilde{x}_{nh}$ that plays a key role in subsequent analysis. This is obtained from the discrete-time dynamics (5.8) so that $\tilde{u}_{1,n+1}$ is replaced by $\tilde{u}_{3,n+1}$. Thus we define the process $(\tilde{x}_{3,(n+1)h})$ by

$$\tilde{x}_{3,(n+1)h} = e^{(A^*+I/2)h}\tilde{x}_{3,nh} + \tilde{u}_{3,n+1}, \qquad n \ge 0, \tag{5.22}$$

with zero initial condition. Let

$$\tilde{\mathcal{F}}_r = \mathcal{F}_{e^r} \quad \text{and} \quad \tilde{\mathcal{F}}^+_r = \mathcal{F}^+_{e^r}.$$

We claim that the approximating process $(\tilde{x}_{3,nh})$ is L-mixing with respect to $(\tilde{\mathcal{F}}_{nh}, \tilde{\mathcal{F}}^+_{nh})$.

Indeed, since the real parts of the eigenvalues of A^* are less than or equal to α^* and $\alpha < \alpha^*$, the spectral norm of $e^{(A^*+I/2)h}$ is less than $e^{-\overline{\alpha}h}$ and hence there exists a $C > 0$ such that for any positive integer m

$$||e^{(A^*+I/2)mh}|| \le Ce^{-\overline{\alpha}mh}. \tag{5.23}$$

The input process $(\tilde{u}_{3,n})$ is an M-bounded, independent, $\tilde{\mathcal{F}}_{nh}$-adapted sequence, hence it is L-mixing with respect to $(\tilde{\mathcal{F}}_{nh}, \tilde{\mathcal{F}}^+_{nh})$. Thus the output process $(\tilde{x}_{3,nh})$ is L-mixing with respect to $(\tilde{\mathcal{F}}_{nh}, \tilde{\mathcal{F}}^+_{nh})$, by Lemma 8.4, as stated.

To get an accurate bound for the estimation error $\tilde{x}_{nh} - \tilde{x}_{3,nh}$ let us introduce the notations

$$\varepsilon_{x2} = \min(\overline{\alpha}, (1 - \varepsilon_\delta)/2), \tag{5.24}$$

$$\varepsilon_{x3} = \min(\overline{\alpha}, 1/2 + c\varepsilon_\delta). \tag{5.25}$$

Obviously $\varepsilon_{x2}, \varepsilon_{x3} > 0$. To formulate the next result note that if (ξ_t) and (η_t) are stochastic processes such that $\xi_t = O_M(c_t)$ and $\eta_t = O_M(d_t)$, where $c_t, d_t > 0$, then, trivially,

$$\xi_t + \eta_t = O_M(c_t + d_t). \tag{5.26}$$

LEMMA 5.1. *The final approximation error $\tilde{x}_{(n+1)h} - \tilde{x}_{3,(n+1)h}$ is given by*

$$\tilde{x}_{(n+1)h} - \tilde{x}_{3,(n+1)h} = O_M(e^{-\varepsilon_x nh} + h^{1/2}e^{-\varepsilon_{x2}nh} + h^{-c}e^{-\varepsilon_{x3}nh}) = O_M(1). \tag{5.27}$$

Proof. The proof is almost trivial. It is easy to see, using the moment inequality given as Theorem 8.1, that both $(\tilde{x}_{nh})$ and $(\tilde{x}_{3,nh})$ are M-bounded, which implies the second part of the claim. To prove the first part, first note that the first term on the right-hand side comes from (5.5). Next note that the error process $(\tilde{x}_{1,(n+1)h} - \tilde{x}_{3,(n+1)h})$ satisfies

$$(\tilde{x}_{1,(n+1)h} - \tilde{x}_{3,(n+1)h}) = e^{(A^*+I/2)h}(\tilde{x}_{1,nh} - \tilde{x}_{3,nh}) + (\tilde{u}_{1,n+1} - \tilde{u}_{3,n+1}), \qquad n \ge 0,$$

with zero initial conditions. For the input process $(\tilde{u}_{1,n+1} - \tilde{u}_{3,n+1})$ the combination of the upper bounds given in Claims U2 and U3, or equivalently in (5.14) and (5.18), is used. Applying Lemma 8.5 we get the second and third terms on the right-hand side of (5.27), which is thus proved. □

To complete the proof of Theorem 5.1 we first note that defining

$$r_n = \tilde{x}_{3,nh} - \tilde{x}_{nh},$$

this residual process is L-mixing with respect to $(\tilde{\mathcal{F}}_{nh}, \tilde{\mathcal{F}}^+_{nh})$. Indeed, (r_n) is M-bounded and $\tilde{\mathcal{F}}_{nh}$-measurable. On the other hand, writing (5.27) in the form $r_n = O_M(e^{-\varepsilon'_x nh})$ we get for any integer $\tau \ge 0$

$$\gamma_q(\tau, r) = \sup_{n \ge \tau} \mathrm{E}^{1/q}|r_n - \mathrm{E}\,[r_n|\mathcal{F}^+_{n-\tau}]\,|^q \le 2 \sup_{n \ge \tau} \mathrm{E}^{1/q}|r_n|^q \le 2C_q e^{-\varepsilon'_x \tau h} \tag{5.28}$$

with some finite C_q. The right-hand side is obviously summable over τ and thus we get the claim.

Since the class of L-mixing processes is closed under addition, it follows that $\tilde{x}_{nh}$ is also L-mixing with respect to $(\tilde{\mathcal{F}}_{nh}, \tilde{\mathcal{F}}^+_{nh})$. The second remark we need is that the

processes $(\tilde{x}_{nh+d})$ and $(\tilde{x}_{3,nh+d})$, with $0 \le d < h$ fixed, can be analyzed similarly and it is easy to see that all the relevant estimates are valid uniformly in d. Thus we conclude that the processes $(\tilde{x}_{nh+d})$ are L-mixing with respect to $(\tilde{\mathcal{F}}_{nh+d}, \tilde{\mathcal{F}}^+_{nh+d})$, uniformly in d for $0 \le d < h$. Applying Corollary 3.5 of [24], restated as Lemma 8.3 in section 8, implies that the continuous-time process $(\tilde{x}_r)$ itself is L-mixing with respect to $(\tilde{\mathcal{F}}_r, \tilde{\mathcal{F}}^+_r) = (\mathcal{F}_{e^r}, \mathcal{F}^+_{e^r})$ and the proof is complete. □

6. The asymptotic covariance matrix. The asymptotic covariance matrix for Algorithm DFL, (3.53)–(3.54), has been rigorously derived in Theorem 13, Chapter 4.5, Part II of [3] in a *series* model, where the initial time tends to infinity, and thus the probability of exiting the truncation domain tends to 0. The asymptotic covariance matrix of Robbins–Monroe-type recursive estimators has been known for long time; cf., e.g., [54]. Here the correction term $H(n, x, \omega)$ is assumed to form an independent sequence; see Condition A.3 in Chapter 2.3 of [54]. The asymptotic covariance matrix for the RPE estimator of ARMA processes has been first given in [60] using the eventually false a priori assumption that the nontruncated estimator sequence converges almost surely. It is likely that the analysis of the cited paper carries over to truncated estimators.

The purpose of this section is to derive the asymptotic covariance matrix for the general continuous-time recursive estimator process, Algorithm CR, given in (3.16) equipped with a resetting mechanism defined under (3.17) and (3.18). The study of the discrete-time procedure, Algorithm DR, given in (3.34) and Algorithm DFL, given under (3.53)–(3.54), with resetting mechanisms defined in section 3, can be reduced to the study of Algorithm CR, as pointed out in sections 3 and 4. The main advance of this section relative to the cited result of [3] is that the asymptotic covariance matrix for the DFL scheme with enforced boundedness is obtained for a *single* process.

We also get a rate of convergence for the covariance-matrix sequence, which is useful in applications such as the analysis of performance degradation to statistical parametric uncertainty. For the present section we need the following additional condition.

CONDITION 6.1. *We assume that* $(H(s, x^*, \omega))$ *is asymptotically wide-sense stationary in the following sense: there exists a zero-mean, wide-sense stationary process* $(H_0(s, x^*, \omega))$ *such that*

$$\eta_s = H(s, x^*, \omega) - H_0(s, x^*, \omega) = O_M(s^{-1-\varepsilon_H}) \tag{6.1}$$

with some $\varepsilon_H > 0$.

This condition is easily verified in system identification. In fact, if we consider the general estimation scheme of section 3 defined by (3.53)–(3.54), then it is easy to see that we have $\eta_s = O_M(e^{-\beta s})$ with some $\beta > 0$. Now we have the following modification of Theorem 4.1.

THEOREM 6.1. *Consider the continuous-time recursive estimation scheme given by* (3.16) *with the resetting mechanism* (3.17) *and* (3.18). *Assume that the conditions of Theorem* 4.1 *are satisfied and in addition Condition* 6.1 *is also satisfied. Recall that* $\varepsilon_x = \min(\overline{\alpha}, \varepsilon)_-$, *where* $\overline{\alpha}$ *is defined under* (3.31) *and* ε *is given in Condition* 3.5. *Then we have*

$$x_t - x^* = \int_1^t \frac{\partial}{\partial \xi} y(t, s, x^*) \frac{1}{s} H_0(s, x^*, \omega) ds + O_M(t^{-1/2-\varepsilon_x}), \tag{6.2}$$

and the wide-sense stationary process $(H_0(s, x^*, \omega))$ *is* L^+*-mixing.*

Proof. Consider the expression for the error $x_t - x^*$ that has been given in Theorem 4.1, or in (4.1). The difference between (4.1) and (6.2) is in the dominant terms and this difference can be majorized by

$$\int_1^t \left| \frac{\partial}{\partial \xi} y(t, s, x^*) \frac{1}{s} \eta_s \right| ds \le \int_1^t C_0 (s/t)^\alpha \left| \frac{1}{s} \eta_s \right| ds,$$

due to Condition 3.4. Taking the L_q-norm of both sides with some $q \ge 1$ and applying the triangle inequality for L_q-norms we get an upper bound of the form

$$C_q \int_1^t C_0 (s/t)^\alpha \frac{1}{s} s^{-1-\varepsilon_H} ds,$$

which is majorized by $C_q' t^{-1-\varepsilon_H}$. Thus the difference between the dominant terms is certainly $O_M(t^{-1/2-\varepsilon_x})$ and thus (6.2) follows.

To prove that $(H_0(s, x^*, \omega))$ is L^+-mixing, note that repeating the argument leading to (5.28) gives that for any integer $\tau \ge 0$

$$\gamma_q(\tau, \eta) \le 2 C_q \tau^{-1-\varepsilon_H},$$

and hence (η_s) is L^+-mixing. Since the class of L^+-mixing processes is closed under addition, it follows that $(H_0(s, x^*, \omega))$ is also L^+-mixing and the proof is complete. □

Remark. There is no loss of generality to assume that

$$\gamma_q(\tau, H_0) \le C_q (1 + \tau)^{-1-c_q}$$

for all $\tau \ge 0$ with the same C_q, c_q as in Condition 3.1 requiring that H and $\Delta H / \Delta x$ be L^+-mixing.

To formulate the basic result of this section we need some notations. Denoting the autocovariance matrix of $H_0(s, x^*, \omega)$ by $\rho(\tau)$, i.e., setting

$$\rho(\tau) = \mathrm{E}\,[H_0(s + \tau, x^*, \omega) H_0^T(s, x^*, \omega)] = \mathrm{E}\,[H_0(\tau, x^*, \omega) H_0^T(0, x^*, \omega)],$$

we define a basic quantity:

$$P^* = \int_{-\infty}^{\infty} \rho(\tau) d\tau. \tag{6.3}$$

Since the process $(H_0(s, x^*, \omega))$ is L-mixing, the above integral converges. Indeed, since $H_0 = (H_0(s, x^*, \omega))$ is a wide-sense stationary zero-mean L-mixing process, using Lemma 8.1 with $p = q = 2$, we get

$$\rho(\tau) \le C \gamma_2(|\tau|, H_0) \tag{6.4}$$

with some $C > 0$, thus integrability follows.

It is easy to see, cf. Lemma 6.4 below, that the matrix P^* is the asymptotic covariance matrix of the arithmetic mean

$$\frac{1}{2T} \int_{-T}^{T} H_0(s, x^*, \omega) ds,$$

i.e., we have

$$P^* = \lim_{T \to \infty} 2T\ \mathrm{E}\ \left[\left(\frac{1}{2T} \int_{-T}^{T} H_0(s, x^*, \omega) ds \right) \left(\frac{1}{2T} \int_{-T}^{T} H_0(s, x^*, \omega) ds \right)^T \right]. \tag{6.5}$$

We will also need the notation introduced in (3.24):

$$A^* = \left.\frac{\partial G(x)}{\partial x}\right|_{x=x^*}.$$

The value of the asymptotic covariance matrix can be easily guessed. Namely, the assumed validity of (5.3) implies that, $t^{1/2}(x_t - x^*)$ is asymptotically normally distributed with zero mean and covariance matrix S^*, which satisfies the Lyapunov equation (6.6) below. This result on the asymptotic covariance matrix of the estimator has strong roots in the classical theory of stochastic approximation; see [54]. The closest to our result is Theorem 13, Chapter 4.5, Part II of [3].

THEOREM 6.2. *Consider the continuous-time recursive estimation scheme given by* (3.16) *with the resetting mechanism* (3.17) *and* (3.18). *Assume that the conditions of Theorem* 4.1 *are satisfied and in addition* $(H(s, x^*, \omega))$ *satisfies Condition* 6.1. *Then the asymptotic covariance matrix of the error process* $(x_t - x^*)$, *defined by*

$$S^* = \lim_{t\to\infty} t\mathrm{E}[(x_t - x^*)(x_t - x^*)^T],$$

exists and it satisfies the Lyapunov equation

$$(A^* + I/2)S^* + S^*(A^* + I/2)^T + P^* = 0, \tag{6.6}$$

where A^* *is defined, (see also* (3.24)*) and* P^* *is defined by* (6.3). *More exactly we have with some* $\varepsilon_{xx} > 0$

$$\mathrm{E}[(x_t - x^*)(x_t - x^*)^T] = \frac{1}{t}S^* + O(t^{-1-\varepsilon_{xx}}).$$

Remark. In the case of a stochastic Newton method, i.e. when $A^* = -I$, we get

$$S^* = P^*.$$

In the context of Algorithm DFL, (3.53)–(3.54), this can be directly seen from Theorem 4.4.

Take the example of the recursive least squares (LSQ) estimation of an AR(p) process given

$$y_n = (\theta^*)^T\phi_n + e_n,$$

where θ^* is the p-dimensional AR-parameter, $\phi_n = (-y_{n-1}, \dots, -y_{n-p})^T$, and e_n is the noise term with variance $\sigma^2(e)$. AR processes are special in the sense that the off-line LSQ estimator can be computed *exactly* in a recursive fashion, thus the off-line and online estimators, if properly initialized, coincide and their asymptotic covariance is the same. A nontrivial corollary of Theorem 6.2 is that this is still the case if both estimators are forced to stay inside a compact domain using truncation for the off-line estimator and resetting for the online estimator.

Let

$$R^* = \mathrm{E}\phi_n\phi_n^T$$

assuming stationarity of ϕ_n. Then, under well-known conditions the asymptotic covariance matrix of the LSQ estimator is known to be

$$S^* = \sigma^2(e)(R^*)^{-1}.$$

For the recursive least squares (RLSQ) estimator we have the updating term, with $x = \theta$,

$$H_n(s, \theta, \omega) = R^{-1}\phi_n(y_n - \phi_n^T\theta)$$

from which we get

$$G(\theta) = \theta^* - \theta,$$

thus the RLSQ method is a stochastic Newton method. Since

$$H_n(s, \theta^*, \omega) = R^{-1}\phi_n e_n,$$

we get

$$P^* = \sigma^2(e)(R^*)^{-1}$$

which indeed agrees with S^*.

Remark. For the discrete-time method, Algorithm DR, we have

$$x_t^l - x^* = \int_1^t \frac{\partial}{\partial \xi} y(t, s, x^*)\frac{1}{s}H^c(s, x^*, \omega)ds + O_M(t^{-1/2-\varepsilon_x});$$

see (4.40) and the analysis given below is applicable. Note however that now we get the familiar expression, see [54],

(6.7)
$$P^* = \int_{-\infty}^{\infty} \mathrm{E}\,[H_0^c(\tau, x^*, \omega)H_0^{cT}(0, x^*, \omega)]d\tau = \sum_{-\infty}^{\infty} \mathrm{E}\,[H_0(m, x^*, \omega)H_0^T(0, x^*, \omega)].$$

Proof of Theorem 6.2.

Reduction to the process $(\tilde{x}_{3,nh})$. The claim of the theorem can be reformulated in terms of the transformed process, with $t = e^r$, as follows: we have with some $\varepsilon_{xx} > 0$

$$\mathrm{E}[\tilde{x}_r\tilde{x}_r^T] = S^* + O(e^{-\varepsilon_{xx}r}). \tag{6.8}$$

Now by Lemma 5.1 we have with $r = nh$ the following expression for $x_t - x^* = e^{-r/2}\tilde{x}_r$:

$$\frac{1}{e^{nh/2}}\tilde{x}_{3,nh} + \frac{1}{e^{nh/2}}O_M(e^{-\varepsilon_x nh} + h^{1/2}e^{-\varepsilon_{x2}nh} + h^{-c}e^{-\varepsilon_{x3}nh}). \tag{6.9}$$

Multiplying both sides by $e^{nh/2}$, squaring them, and taking into account the second part of Lemma 5.1, we get the following key lemma.

Lemma 6.1. *We have with* $r = nh$,

$$\mathrm{E}[\tilde{x}_r\tilde{x}_r^T] = \mathrm{E}[\tilde{x}_{3,nh}\tilde{x}_{3,nh}^T] + O(e^{-\varepsilon_x nh} + h^{1/2}e^{-\varepsilon_{x2}nh} + h^{-c}e^{-\varepsilon_{x3}nh}). \tag{6.10}$$

This error estimate seems to be fragile, due to Terms 3 and 4 on the right-hand side, in view of the fact that the left-hand side is $O(h)$, but this weakness will be eliminated at the very end of the proof of Theorem 6.2 by appropriate choice of h.

Thus the study of the covariance matrix of x_t is reduced to the study of the covariance matrix of $\tilde{x}_{3,nh}$, which will be denoted by $R_{3,n}^{\tilde{x}}$:

$$R_{3,n}^{\tilde{x}} = \mathrm{E}\,[\tilde{x}_{3,nh}\tilde{x}_{3,nh}^T].$$

Now change n to $n+1$ and note that $\tilde{x}_{3,(n+1)h}$ is defined via the discrete-time dynamical system (5.22), in which the input process $(\tilde{u}_{3,n+1})$ consists of a sequence of independent random variables. The covariance matrix of $\tilde{u}_{3,m}$ will be denoted by

$$R^{\tilde{u}}_{3,m} = \mathrm{E}[\tilde{u}_{3,m}\tilde{u}^T_{3,m}].$$

In what follows we shall develop a sequence of approximations of $\tilde{u}_{3,m}$ to get a nice approximation for $R^{\tilde{u}}_{3,m}$.

The approximating process $(\tilde{u}_{4,m})$. Let us recall, see (5.9), that in the original time-scale we have

$$\tilde{u}_{1,m+1} = \int_{e^{mh}}^{e^{(m+1)h}} \left(\frac{s}{e^{(m+1)h}}\right)^{(-A^*-I/2)} \frac{1}{s^{1/2}} H(s, x^*, \omega) ds.$$

Approximate $\tilde{u}_{1,m+1}$ by replacing the kernel within the integrand by 1; i.e., set

$$\tilde{u}_{4,m+1} = \int_{e^{mh}}^{e^{(m+1)h}} \frac{1}{s^{1/2}} H(s, x^*, \omega) ds.$$

Claim U4. *We have*

$$\tilde{u}_{1,m+1} - \tilde{u}_{4,m+1} = O_M(h^{3/2}). \tag{6.11}$$

To prove the claim note that the approximation error can be written as

$$\tilde{u}_{1,m+1} - \tilde{u}_{4,m+1} = \int_{e^{mh}}^{e^{(m+1)h}} \left(\left(\frac{s}{e^{(m+1)h}}\right)^{(-A^*-I/2)} - I\right) \frac{1}{s^{1/2}} H(s, x^*, \omega) ds.$$

Using the moment inequality given as Theorem 8.1 we get for any $q \geq 2$ that $\mathrm{E}^{1/q}|\tilde{u}_{1,m+1} - \tilde{u}_{4,m+1}|^q$ is bounded from above by

$$C_q \left(\int_{e^{mh}}^{e^{(m+1)h}} \left\| \left(\left(\frac{s}{e^{(m+1)h}}\right)^{(-A^*-I/2)} - I\right) \frac{1}{s^{1/2}} \right\|^2 ds\right)^{1/2} M_q^{1/2}(H(x^*))\Gamma_q^{1/2}(H(x^*)).$$

Now, $\|(s/e^{(m+1)h})^{(-A^*-I/2)} - I\| = \|e^{(-A^*-I/2)(\log s - (m+1)h)} - I\| \leq ch$, with some c, which depends only on A^*, for $0 < h \leq h_0$, since $-h \leq (\log s - (m+1)h) \leq 0$. (Apply a Taylor series expansion of the matrix exponential to get the desired inequality.) Thus we get

$$\mathrm{E}^{1/q}|\tilde{u}_{1,m+1} - \tilde{u}_{4,m+1}|^q \leq C_q \left(\int_{e^{mh}}^{e^{(m+1)h}} (ch)^2 \frac{1}{s} ds\right)^{1/2} M_q^{1/2}(H(x^*))\Gamma_q^{1/2}(H(x^*)),$$

and from here

$$\mathrm{E}^{1/q}|\tilde{u}_{1,m+1} - \tilde{u}_{4,m+1}|^q \leq Ch^{3/2},$$

where C is independent of h and thus Claim U4 follows.

The approximating process $(\tilde{u}_{5,m+1})$. This approximation is obtained by replacing $\frac{1}{s^{1/2}}$ within the integral by a constant:

$$\tilde{u}_{5,m+1} = \frac{1}{e^{mh/2}} \int_{e^{mh}}^{e^{(m+1)h}} H(s, x^*, \omega) ds.$$

CLAIM U5. *We have*

$$\tilde{u}_{4,m+1} - \tilde{u}_{5,m+1} = O_M(h^{3/2}). \tag{6.12}$$

Indeed, we have

$$\tilde{u}_{5,m+1} - \tilde{u}_{4,m+1} = \int_{e^{mh}}^{e^{(m+1)h}} \left(\frac{1}{e^{mh/2}} - \frac{1}{s^{1/2}} \right) H(s, x^*, \omega) ds,$$

and we can apply the moment inequality given as Theorem 8.1. For this purpose we estimate the integrand

$$0 \le \left(\frac{1}{e^{mh/2}} - \frac{1}{s^{1/2}} \right) \le \left(\frac{1}{e^{mh/2}} - \frac{1}{e^{(m+1)h/2}} \right) \le \frac{1}{(e^{mh/2})^2} (e^{(m+1)h/2} - e^{mh/2})$$

$$\le \frac{1}{e^{mh/2}} (e^{h/2} - 1) \le \frac{1}{e^{mh/2}} h$$

for small h. Thus we get the following upper bound for $\mathrm{E}^{1/q} |\tilde{u}_{5,m+1} - \tilde{u}_{4,m+1}|^q$ with $q \ge 2$:

$$C_q \left(\int_{e^{mh}}^{e^{(m+1)h}} \left(\frac{1}{e^{mh/2}} - \frac{1}{s^{1/2}} \right)^2 ds \right)^{1/2} M_q^{1/2}(H(x^*)) \Gamma_q^{1/2}(H(x^*))$$

$$\le C_q \left(\int_{e^{mh}}^{e^{(m+1)h}} \frac{h^2}{e^{mh}} ds \right)^{1/2} M_q^{1/2}(H(x^*)) \Gamma_q^{1/2}(H(x^*)) = O(h^{3/2}),$$

and the claim follows.

The approximating process $(\tilde{u}_{6,m+1})$. This approximation is obtained by replacing $H(s, x^*, \omega)$ by $H_0(s, x^*, \omega)$ in the definition of $\tilde{u}_{6,m+1}$, i.e., we define

$$\tilde{u}_{6,m+1} = \frac{1}{e^{mh/2}} \int_{e^{mh}}^{e^{(m+1)h}} H_0(s, x^*, \omega) ds. \tag{6.13}$$

CLAIM U6. *We have*

$$\tilde{u}_{5,m+1} - \tilde{u}_{6,m+1} = O_M(h e^{-mh(1/2+\varepsilon_H)}) \quad \textit{and} \quad \tilde{u}_{5,m+1} - \tilde{u}_{6,m+1} = O_M(h^{1/2}). \tag{6.14}$$

Indeed, we have, using (6.1),

$$\tilde{u}_{5,m+1} - \tilde{u}_{6,m+1} = \frac{1}{e^{mh/2}} \int_{e^{mh}}^{e^{(m+1)h}} \eta_s ds$$

$$= \frac{1}{e^{mh/2}} (e^{(m+1)h} - e^{mh}) O_M(e^{-mh(1+\varepsilon_H)}) = O_M(h e^{-mh(1/2+\varepsilon_H)})$$

and the first part of the claim follows. The second part is a direct consequence of Theorem 8.1.

Summarizing the equations expressing the approximation errors between the successive values $\tilde{u}_3, \tilde{u}_2, \tilde{u}_1, \tilde{u}_4, \tilde{u}_5, \tilde{u}_6$ given by (5.14), (5.18), (6.11), (6.12), (6.14) we get the following lemma.

LEMMA 6.2. *Let c be as in Condition* 3.1*, requiring that H be L^+-mixing. Then we have*

$$\tilde{u}_{3,m} = \tilde{u}_{6,m} + \delta\tilde{u}, \tag{6.15}$$

where

$$\delta\tilde{u} = O_M(h^{-c}e^{-(1/2+c\varepsilon_\delta)mh} + h^{1/2}e^{-(1-\varepsilon_\delta)mh/2} + h^{3/2} + he^{-mh(1/2+\varepsilon_H)}),$$

and all error terms are also $O_M(h^{1/2})$.

Squaring this equation we get, for $R^{\tilde{u}}_{3,m+1} = \mathrm{E}\,[\tilde{u}_{3,m}\tilde{u}^T_{3,m}]$,

$$R^{\tilde{u}}_{3,m+1} = \mathrm{E}\,[\tilde{u}_{6,m}\tilde{u}^T_{6,m}] + \delta R, \tag{6.16}$$

where

$$\delta R = O(h^{1/2-c}e^{-(1/2+c\varepsilon_\delta)mh} + he^{-(1-\varepsilon_\delta)mh/2} + h^2 + h^{3/2}e^{-mh(1/2+\varepsilon_H)}).$$

The covariance matrix of $\tilde{u}_{6,m+1}$. Next we show that the covariance matrix of the approximation $\tilde{u}_{6,m+1}$ can be expressed in terms of the matrix P^*. This is no surprise in view of the assumed validity of (6.5), but to capture the rate of convergence extra work is needed.

LEMMA 6.3. *Let c be as in Condition* 3.1, *requiring that H be L^+-mixing. Then we have*

$$\mathrm{E}[\tilde{u}_{6,m+1}\tilde{u}^T_{6,m+1}] = hP^* + O(h^{1-c}e^{-cmh}). \tag{6.17}$$

Proof. Consider normalized arithmetic means of the form

$$s_{A,B} = \frac{1}{(B-A)^{1/2}}\int_A^B H_0(s, x^*, \omega)ds$$

with $A < B$. It is obvious that

$$\mathrm{E}\,[s_{A,B}\ s^T_{A,B}] = \frac{1}{B-A}\int_A^B\int_A^B \rho(s-s')\ ds\ ds',$$

where $\rho(\tau)$ is the autocovariance function of the process $H_0 = (H_0(s, x^*, \omega))$.

Note that if $H_0 = (H_0(s, x^*, \omega))$ is a wide-sense stationary zero-mean L^+-mixing process then we have, using (6.4) and the inequality $\gamma_2(|\tau|, H_0) \le C(1+|\tau|)^{-c}$,

$$\rho(\tau) \le C(1+|\tau|)^{-c} \tag{6.18}$$

with some $C, c > 0$. Applying Lemma 6.4 below with

$$A = e^{mh} \quad \text{and} \quad B = e^{(m+1)h},$$

we have $(B-A) = e^{mh}(h + O(h^2))$ for small h. Thus we get, using the inequality $(1+B-A)^{-c} < C'(B-A)^{-c}$,

$$\frac{1}{B-A}\int_A^B\int_A^B \rho(s-s')\ ds\ ds' = P^* + O(e^{-cmh}h^{-c}),$$

and from here we get for the covariance matrix of $\tilde{u}_{6,m+1}$,

$$\mathrm{E}[\tilde{u}_{6,m+1}\tilde{u}^T_{6,m+1}] = \frac{1}{e^{mh}}\int_{e^{mh}}^{e^{(m+1)h}}\int_{e^{mh}}^{e^{(m+1)h}} \rho(s-s')dsds' = \frac{B-A}{e^{mh}}(P^* + O(e^{-cmh}h^{-c})),$$

and Lemma 6.3 follows. ∎

LEMMA 6.4. *Let $(\rho(\tau))$, $-\infty < \tau < \infty$, be a matrix-valued measurable function process, satisfying $||\rho(\tau)|| \le C(1+|\tau|)^{-c}$ with some $C, c > 0$. Then we have for any $A < B$*

$$\frac{1}{B-A}\int_A^B \int_A^B \rho(s-s')ds\ ds' = P^* + O((1+B-A)^{-c}).$$

Proof. Introduce the new variables $\tau = s - s', \mu = s + s'$. This change of coordinates has a Jacobian with determinant 2, i.e., $d\tau\ d\mu = 2ds\ ds'$. The new variable τ takes its values between $-(B-A)$ and $B-A$ and for each fixed τ the possible values of μ are in the interval $(2A+|\tau|, 2B-|\tau|)$. Thus

$$\begin{aligned} P_{A,B} &= \frac{1}{B-A}\int_{-(B-A)}^{B-A}\int_{2A+|\tau|}^{2B-|\tau|} \rho(\tau)\frac{1}{2}d\tau\ d\mu \\ &= \frac{1}{B-A}\int_{-(B-A)}^{B-A}(2B-2A-2|\tau|)\rho(\tau)\frac{1}{2}d\tau\ d\mu = \int_{-(B-A)}^{B-A}\left(1-\frac{|\tau|}{B-A}\right)\rho(\tau)d\tau. \end{aligned}$$

From here it follows immediately that $||P_{A,B}||$ can be written as

$$\left\|\frac{1}{B-A}\int_A^B\int_A^B \rho(s-s')ds\ ds'\right\| \le \int_{-(B-A)}^{B-A}||\rho(\tau)||d\tau \le \int_{-\infty}^{\infty}||\rho(\tau)||d\tau. \tag{6.19}$$

This inequality will be used subsequently. Now, write

$$\int_{-(B-A)}^{B-A}\left(1-\frac{|\tau|}{B-A}\right)\rho(\tau)d\tau = \int_{-(B-A)}^{B-A}\rho(\tau)d\tau - \int_{-(B-A)}^{B-A}\frac{|\tau|}{B-A}\rho(\tau)d\tau.$$

Then

$$\begin{aligned} \int_{-(B-A)}^{B-A}\rho(\tau)d\tau - P^* &= \int_{-(B-A)}^{B-A}\rho(\tau)d\tau - \int_{-\infty}^{\infty}\rho(\tau)d\tau \\ &= -\int_{-\infty}^{-(B-A)}\rho(\tau)d\tau - \int_{B-A}^{\infty}\rho(\tau)d\tau. \end{aligned}$$

Taking into account that $||\rho(\tau)|| \le C(1+|\tau|)^{-1-c}$, we get that

$$\left\|-\int_{-\infty}^{-(B-A)}\rho(\tau)d\tau - \int_{B-A}^{\infty}\rho(\tau)d\tau\right\| \le \frac{2C}{c}(1+B-A)^{-c}. \tag{6.20}$$

On the other hand,

$$\int_{-(B-A)}^{B-A}\frac{|\tau|}{B-A}||\rho(\tau)||d\tau \le \int_{-(B-A)}^{B-A}\frac{|\tau|}{B-A}C(1+|\tau|)^{-1-c}d\tau.$$

Write $|\tau|C(1+|\tau|)^{-1-c} \le (1+|\tau|)C(1+|\tau|)^{-1-c} = C(1+|\tau|)^{-c}$ and use the symmetry of the last integrand above to get the upper bound

$$\begin{aligned} 2\int_0^{B-A}\frac{1}{B-A}C(1+\tau)^{-c}d\tau &\le \frac{2}{B-A}\frac{C}{(-c+1)}(1+\tau)^{-c+1}|_0^{B-A} \\ &= \frac{2}{B-A}\frac{C}{(-c+1)}((1+B-A)^{-c+1}-1)) \le C'(1+B-A)^{-c} \end{aligned}$$

and combining this with (6.20) the proposition of the lemma follows. □

The final approximation of $R^{\tilde{u}}_{3,m+1}$. Combining (6.16) and (6.17) we get

$$\begin{aligned} R^{\tilde{u}}_{3,m+1} &= \mathrm{E}[\tilde{u}_{3,m}\tilde{u}^T_{3,m}] \\ &= hP^* + O(h^{1-c}e^{-cmh} + h^{1/2-c}e^{-(1/2+c\varepsilon_\delta)mh} \\ &\quad + he^{-(1-\varepsilon_\delta)mh/2} + h^2 + h^{3/2}e^{-mh(1/2+\varepsilon_H)}). \end{aligned} \tag{6.21}$$

To simplify notations write the residual terms in the form $h^{\beta_i}e^{-\varepsilon_i mh}$, $i = 1, \dots, 5$, with

$$\begin{aligned} &\beta_1 = 1 - c, && \varepsilon_1 = c, \\ &\beta_2 = 1/2 - c, && \varepsilon_2 = 1/2 + c\varepsilon_\delta, \\ &\beta_3 = 1, && \varepsilon_3 = (1 - \varepsilon_\delta)/2, \\ &\beta_4 = 2, && \varepsilon_4 = 0, \\ &\beta_5 = 3/2, && \varepsilon_5 = 1/2 + \varepsilon_H. \end{aligned} \tag{6.22}$$

Obviously we have $\varepsilon_i > 0$ for $i \neq 4$. For $i = 4$ we have $\varepsilon_4 = 0$, but then $\beta_4 = 2$. With this notations we can formulate the following lemma.

LEMMA 6.5. *We have with* $h^{\beta_i}e^{-\varepsilon_i mh}$, $i = 1, \dots, 5$, *defined under* (6.22)

$$R^{\tilde{u}}_{3,m+1} = \mathrm{E}\,[\tilde{u}_{3,m}\tilde{u}^T_{3,m}] = hP^* + \sum_{i=1}^{5} O(h^{\beta_i}e^{-\varepsilon_i mh}). \tag{6.23}$$

The discrete-time Lyapunov equation. Consider the discrete-time dynamics followed by $(\tilde{x}_{3,nh})$, given by (5.22). Since the input process is a sequence of independent random variables it follows that the covariance matrix of $\tilde{x}_{3,nh}$, denoted by $R^{\tilde{x}}_{3,n}$ $R^{\tilde{x}}_{3,n+1}$, satisfies the Lyapunov equation

$$R^{\tilde{x}}_{3,n+1} = e^{(A^*+I/2)h}R^{\tilde{x}}_{3,n}e^{(A^*+I/2)^T h} + R^{\tilde{u}}_{3,n+1},$$

with zero initial condition. Substituting $R^{\tilde{u}}_{3,n+1}$ from (6.23) and setting $n = m$, we get

$$R^{\tilde{x}}_{3,m+1} = e^{(A^*+I/2)h}R^{\tilde{x}}_{3,m}e^{(A^*+I/2)^T h} + hP^* + \sum_{i=1}^{5} O(h^{\beta_i}e^{-\varepsilon_i mh}). \tag{6.24}$$

Solving this iteratively in the range $0 \le m \le n$ we get

$$\begin{aligned} R^{\tilde{x}}_{3,n+1} &= \sum_{m=0}^{n} e^{(A^*+I/2)(n-m)h}\ hP^*\ e^{(A^*+I/2)^T(n-m)h} \\ &\quad + \sum_{m=0}^{n} e^{(A^*+I/2)(n-m)h}\left(\sum_{i=1}^{5} h^{\beta_i}e^{-\varepsilon_i mh}\right)e^{(A^*+I/2)^T(n-m)h}. \end{aligned} \tag{6.25}$$

The contributions of the terms $h^{\beta_i}e^{-\varepsilon_i mh}$, $i = 1, \dots, 5$, are estimated as follows:

$$\begin{aligned} \Delta_i &= \left\| \sum_{m=0}^{n} e^{(A^*+I/2)(n-m)h}Ch^{\beta_i}e^{-\varepsilon_i mh}e^{(A^*+I/2)^T(n-m)h} \right\| \\ &\le C' \sum_{m=0}^{n} e^{-2\overline{\alpha}(n-m)h} \cdot h^{\beta_i}e^{-\varepsilon_i mh}. \end{aligned}$$

Applying Lemma 8.5 we get, assuming that $2\overline{\alpha} \neq \varepsilon_i$, with

$$\overline{\varepsilon}_i = \min(2\overline{\alpha}, \varepsilon_i) \tag{6.26}$$

the upper bound

$$\Delta_i \leq C' h^{\beta_i} e^{-\overline{\varepsilon}_i nh} / |e^{2\overline{\alpha}h} - e^{\varepsilon_i h}| = O(h^{\beta_i - 1} e^{-\overline{\varepsilon}_i nh}), \tag{6.27}$$

and thus

$$\begin{aligned} R^{\tilde{x}}_{3,n+1} &= \sum_{m=0}^{n} e^{(A^* + I/2)(n-m)h}\ hP^*\ e^{(A^* + I/2)^T (n-m)h} \\ &\quad + \sum_{m=0}^{n} O(h^{\beta_i - 1} e^{-\overline{\varepsilon}_i nh}). \end{aligned} \tag{6.28}$$

Obviously we have $\overline{\varepsilon}_i > 0$ for $i \neq 4$. For $i = 4$ we have $\overline{\varepsilon}_4 = 0$, but then $\beta_4 = 2$.

Next we consider the first dominant term on the right-hand side of (6.28) and define its approximation by setting $m' = n - m$ and extending the summation to ∞:

$$R^{\tilde{x}d}_{3,n+1} = \sum_{m=0}^{n} e^{(A^* + I/2)(n-m)h} hP^* e^{(A^* + I/2)^T (n-m)h}, \tag{6.29}$$

$$R^{\tilde{x}}_{3*} = \sum_{m'=0}^{\infty} e^{(A^* + I/2)m'h} hP^* e^{(A^* + I/2)^T m'h}. \tag{6.30}$$

Claim. *We have*

$$R^{\tilde{x}d}_{3,n+1} - R^{\tilde{x}}_{3*} = O(e^{-2\overline{\alpha}nh}). \tag{6.31}$$

Indeed, writing $m' = n - m$ and taking out the left factor $e^{(A^* + I/2)(n+1)h}$ and the right factor $e^{(A^* + I/2)^T (n+1)h}$ we have

$$\begin{aligned} R^{\tilde{x}d}_{3,n+1} - R^{\tilde{x}}_{3*} &= \sum_{m'=n+1}^{\infty} e^{(A^* + I/2)m'h} hP^* e^{(A^* + I/2)^T m'h} \\ &= e^{(A^* + I/2)(n+1)h} \left(\sum_{m=0}^{\infty} e^{(A^* + I/2)mh} hP^* e^{(A^* + I/2)^T mh} \right) e^{(A^* + I/2)^T (n+1)h}, \end{aligned}$$

the operator norm of which is obviously majorized by $C' e^{-2\overline{\alpha}nh}$, as claimed.

Lemma 6.6. *We have*

$$R^{\tilde{x}}_{3*} - S^* = O(h). \tag{6.32}$$

Proof. The covariance matrix $R^{\tilde{x}}_{3*}$ is the solution of the algebraic Lyapunov equation

$$R^{\tilde{x}}_{3*} = e^{(A^* + I/2)h} R^{\tilde{x}}_{3*} e^{(A^* + I/2)^T h} + hP^*.$$

Taking into account the equality $e^{(A^* + I/2)h} = I + (A^* + I/2)h + O(h^2)$, this can be written as

$$R^{\tilde{x}}_{3*} = (I + (A^* + I/2)h + O(h^2)) R^{\tilde{x}}_{3*} (I + (A^* + I/2)^T h + O(h^2)) + hP^*,$$

which is simplified to

$$0 = (A^* + I/2)R_{3*}^{\tilde{x}} + R_{3*}^{\tilde{x}}(A^* + I/2)^T + P^* + O(h),$$

and the stability of $(A^* + I/2)$ implies the claim. □

Combining (6.28), (6.31), and (6.32) we get, assuming that $2\overline{\alpha} \neq \varepsilon_i$,

$$R_{3,n+1}^{\tilde{x}} = S^* + O(e^{-2\overline{\alpha}nh} + h) + \sum_{i=1}^{5} O(h^{\beta_i - 1} e^{-\overline{\varepsilon}_i nh}). \tag{6.33}$$

The final approximation of $R_{3,n+1}^{\tilde{x}}$. For a given r we choose h and n in the following way: let $\varepsilon_h > 0$ and let h satisfy

$$e^{-\varepsilon_h r} \leq h \leq 2e^{-\varepsilon_h r}, \tag{6.34}$$

and in addition let r be an integer multiple of h, say, $r = nh$. Then from (6.33) we get

$$R_{3,n+1}^{\tilde{x}} = S^* + O(e^{-2\overline{\alpha}nh} + e^{-\varepsilon_h nh}) + \sum_{i=1}^{5} O(e^{-(\beta_i - 1)\varepsilon_h nh} e^{-\overline{\varepsilon}_i nh}). \tag{6.35}$$

Combining this with (6.10) and substituting $h = e^{-\varepsilon_h r} = e^{-\varepsilon_h nh}$ we get

$$\begin{aligned} \mathrm{E}[\tilde{x}_{nh}\tilde{x}_{nh}^T] &= S^* + O(e^{-2\overline{\alpha}nh} + e^{-\varepsilon_h nh}) + \sum_{i=1}^{5} O(e^{-(\beta_i - 1)\varepsilon_h nh} e^{-\overline{\varepsilon}_i nh}) \\ &\quad + O(e^{-\varepsilon_x nh} + e^{-\varepsilon_h nh/2} e^{-\varepsilon_{x2} nh} + e^{c\varepsilon_h nh} e^{-\varepsilon_{x3} nh}). \end{aligned} \tag{6.36}$$

The generic form of the error terms is $O(e^{-\gamma nh})$, where the values of γ are the following:

$$\begin{array}{lll} 2\overline{\alpha}, & \varepsilon_h, & (\beta_i - 1)\varepsilon_h + \overline{\varepsilon}_i,\ i = 1, \ldots, 5, \\ \varepsilon_x, & \varepsilon_h/2 + \varepsilon_{x2}, & c\varepsilon_h - \varepsilon_{x3}. \end{array}$$

Obviously for sufficiently small ε_h all these constants are positive and thus (6.8) and the claim of Theorem 6.2 follows. □

7. Two applications. The usefulness of the results of the present paper is demonstrated by describing two applications. In the first example the pathwise cumulative regret is quantified for an online adaptive predictor of multivariable linear stochastic systems; see (7.8). It is a previously unpublished result, presented at MTNS '96. In the second example a similar measure of performance degradation for the minimum-variance self-tuning regulator is considered. This problem, that had been formulated as far back as 1971 in [2] in a slightly different context from ours, has been solved only in 1994; see [28]. The result of [28] is restated in (7.19). A further application for indirect adaptive control of multivariable linear stochastic systems is given in [27]. All these applications rely on the results of the present paper, in particular Theorems 4.3, 5.1, and 6.2.

Multivariable adaptive prediction. Let (y_n), $0 \leq n < \infty$, be a vector-valued, wide-sense stationary stochastic process defined by a finite-dimensional linear stochastic system:

$$y = H(\theta^*)e. \tag{7.1}$$

Here $H(\theta) = I + C(\theta)(q^{-1}I - A(\theta))^{-1}B(\theta)$ is a square, causal, rational transfer function of the backward shift operator q^{-1}.

CONDITION 7.1. *$H(\theta)$ is defined for $\theta \in D$, where $D \subset \mathbb{R}^p$ is an open set and in its state-space realization the matrices $(A(\theta), B(\theta), C(\theta))$ are twice continuously differentiable functions of θ. Moreover, $H(\theta)$ is stable and inverse stable.*

CONDITION 7.2. *The system-noise process (e_n), $0 \leq n < \infty$, is an M-bounded, vector-valued wide-sense stationary orthogonal process. In addition there is an increasing sequence of σ-fields $(\mathcal{F}_n)$, $0 \leq n < \infty$, such that (e_n) is a martingale difference process with constant conditional covariance:*

$$\mathrm{E}\,[e_n|\mathcal{F}_{n-1}] = 0, \qquad \mathrm{E}(e_n e_n^T|\mathcal{F}_{n-1}) = \Lambda^*$$

almost surely, with $\Lambda^* > 0$.

These conditions will be called the standard conditions for multivariable linear stochastic systems. In the multivariable version of the prediction error method we have to estimate θ^* and Λ^* jointly to improve efficiency. Let $\theta \in D$ and let Λ be a symmetric positive definite matrix and then define the second order stationary process $\bar{\varepsilon}(\theta)$ by

$$\bar{\varepsilon}(\theta) = H^{-1}(\theta)y.$$

Then define the cost function

$$V_N(\theta, \Lambda) = \frac{1}{2}\sum_{n=1}^{N} \bar{\varepsilon}_n^T(\theta)\Lambda^{-1}\bar{\varepsilon}_n(\theta) + \frac{N}{2}\log\ \det\Lambda. \tag{7.2}$$

If (e_n) is an i.i.d. sequence of Gaussian random vectors with distribution $N(0, \Lambda^*)$, then $V_N(\theta, \Lambda)$ is the negative conditional log-likelihood function, except for an additive constant. This cost function will be minimized in (θ_N, Λ_N) and the minimizing value, the off-line estimator of (θ^*, Λ^*) will be denoted by $(\widehat{\theta}_N, \hat{\Lambda}_N)$. A more precise definition of $(\widehat{\theta}_N, \hat{\Lambda}_N)$, taking into account the possibility of the existence of several local minima, can be given following [18].

Define the asymptotic cost function by

$$W(\theta, \Lambda) = \lim_{n\to\infty} \frac{1}{2}\mathrm{E}\,[\bar{\varepsilon}_n^T(\theta)\Lambda^{-1}\bar{\varepsilon}_n(\theta)] + \frac{1}{2}\log\ \ \det\Lambda. \tag{7.3}$$

It is easy to see that for any symmetric, positive definite Λ,

$$W_\theta(\theta^*, \Lambda) = 0. \tag{7.4}$$

The Hessian of W with respect to θ at (θ^*, Λ^*) is

$$R^* = W_{\theta\theta}(\theta^*, \Lambda^*) = \lim_{n\to\infty} \mathrm{E}\,[\bar{\varepsilon}_{\theta n}^T(\theta^*)(\Lambda^*)^{-1}\bar{\varepsilon}_{\theta n}(\theta^*)]. \tag{7.5}$$

The above cost function can be treated with the extension of the DFL scheme indicated by an alternative definition of the random filed $H(n, x, \omega)$ in (3.47).

CONDITION 7.3. *Equation* (7.4) *has a unique solution $\theta = \theta^*$ for any symmetric, positive definite Λ and the Hessian matrix $W_{\theta\theta}(\theta^*, \Lambda^*)$ is positive definite.*

The *performance index* of interest is the squared absolute value of the prediction error. Let $\Sigma_{\theta\theta}$ be the asymptotic covariance matrix of the off-line prediction error estimator $\widehat{\theta}_n$. Then it is well known that $\Sigma_{\theta\theta} = (R^*)^{-1}$. Let

$$T^* = 2\frac{\partial^2}{\partial\theta^2} \lim_{n\to\infty} \mathrm{E}\,[\bar{\varepsilon}_n^T(\theta)\bar{\varepsilon}_n(\theta)]\ \big|_{\theta=\theta^*} \tag{7.6}$$

be the second order sensitivity matrix of the performance index. Then we have the following result.

THEOREM 7.1. *Let us consider a multivariable system satisfying Conditions* 7.1, 7.2, *and* 7.3. *In addition assume that* (e_n) *is L-mixing. Then we have almost surely*

$$\lim_{N\to\infty}\sum_{n=1}^{N}(|\overline{\varepsilon}_n(\widehat{\theta}_{n-1})|^2-|e_n|^2)/\log N=\frac{1}{2}\mathrm{Tr}T^*\Sigma_{\theta\theta}. \tag{7.7}$$

The expression $\frac{1}{2}\mathrm{Tr}T^*$ will be called the *normalized cost of adaptation*. An important difference between ARMA and multivariable systems is that, unless $\Lambda^*\neq cI$, with c being a scalar, the trace formula given on the right-hand side of (7.7) cannot be further simplified. However, it can be shown that $\frac{1}{2}\mathrm{Tr}T^*\Sigma_{\theta\theta}$ is invariant with respect to diffeomorphic transformation of the parameter space, while restriction of the parameter space, i.e., writing $\theta=g(\eta)$ with $\dim\eta<\dim\theta$, with g being a smooth function, reduces the normalized cost of adaptation. Note that, the normalized cost of adaptation is not determined solely by structural parameters, it may depend also on the actual multivariable system, unlike in the ARMA case.

To extend this result for adaptive predictors defined in terms of recursive estimators we rely on Theorem 4.3 and we get the following result.

CLAIM. *Let* $\widehat{\widehat{\theta}}_n$ *be a recursive estimator of* θ^* *with asymptotic covariance matrix* $\overline{\Sigma}_{\theta\theta}$. *Then under appropriate technical conditions, obtained by specializing the conditions of Theorem* 4.3, *we have*

$$\lim_{N\to\infty}\sum_{n=1}^{N}(|\overline{\varepsilon}_n(\widehat{\widehat{\theta}}_{n-1})|^2-|e_n|^2)/\log N=\frac{1}{2}\mathrm{Tr}T^*\overline{\Sigma}_{\theta\theta} \tag{7.8}$$

almost surely. In analogy with the ARMA case, if we use a stochastic Newton method, then we have

$$\overline{\Sigma}_{\theta\theta}=\Sigma_{\theta\theta}.$$

The minimum-variance self-tuning regulator. Consider now a stochastic control system in ARMAX(n,m,p) representation defined by the relation

$$A^*(q^{-1})y=q^{-1}B^*(q^{-1})u+C^*(q^{-1})e, \tag{7.9}$$

where $A^*(q^{-1})$, $B^*(q^{-1})$, and $C^*(q^{-1})$ are polynomials of the backward shift operator q^{-1} of degree n,m,p, respectively. Their coefficients are denoted by a_i^*,b_i^*,c_i^*, respectively, with $a_0^*=1$, $a_n^*\neq 0$, $b_0^*\neq 0$, $b_m^*\neq 0$, $c_0^*=1$, $c_p^*\neq 0$. Here u is the input process, e is the noise process, and y is the output process. The notation u is a shorthand for $(u(t))$, $0\le t\le\infty$. Assume that the polynomials B^* and C^* are stable and that $\deg C^*\le\deg A^*$. By extending the vector $(c_1^*,\dots,c_p^*)$ with zeros, if necessary, we can actually assume that $\deg A^*=\deg C^*$. The stochastic process e is a zero-mean wide-sense stationary orthogonal process; i.e., for all $t,s\ge 0$ we have $\mathrm{E}e(t)=0$ and $\mathrm{E}\,[e(s)e(t)]=\sigma^2(e)\delta_{st}$, where δ_{st} is the Kronecker symbol.

The minimum-variance control for the ARMAX system given under (7.9) is given by (cf. [2])

$$q^{-1}B^*u=(A^*-C^*)y. \tag{7.10}$$

Using this control law we get, under the assumption that the initial values are all zero, $y(t) = e(t)$. Equation (7.10) can be written in the form

$$u(t-1) = -(\eta^*)^T \phi(t), \tag{7.11}$$

where

$$\eta^* = \frac{1}{b_0^*}(a_1^* - c_1^*, \ldots, a_n^* - c_n^*, b_1^*, \ldots, b_m^*)^T \tag{7.12}$$

and

$$\phi(t) = (-y(t-1), \ldots, -y(t-n), u(t-2), \ldots, u(t-m-1)). \tag{7.13}$$

If the values of the parameters of the stochastic control system are unknown, then a stochastic adaptive control procedure will be needed. Within stochastic adaptive control a special procedure is the self-tuning regulation, that has been proposed in [2] for minimum-variance control. For a new perspective of this procedure see [63]. This is a stochastic approximation procedure defined as follows: let $\hat{\eta}(0)$ be an initial estimate of η^* and let $\hat{\eta}(t-1)$ be an estimate computed at time $t-1$. Then define the control action by

$$u(t-1) = -\hat{\eta}(t-1)\, \phi(t). \tag{7.14}$$

This is followed by observing $y(t)$ which is generated by (7.9). Finally we generate the next estimates $\hat{\eta}(t)$ by

$$\hat{\eta}(t) = \hat{\eta}(t-1) + R^{-1}\frac{1}{t}\, \phi(t) y(t), \tag{7.15}$$

where R is a symmetric positive definite matrix. A basic question in the context of stochastic adaptive control is the characterization of the performance degradation

$$y^2(t) - e^2(t) \tag{7.16}$$

and to establish its pathwise properties. This problem was first formulated in [2]. It has been open for a long time, until a solution was presented in [28], using the results of the present paper.

The performance of the minimum-variance self-tuning regulator had been studied in [48, 49]. In [48] the right order of magnitude for the so-called cumulative regret was found for general ARMAX systems. In [49] the right constant in a tight upper bound for cumulative regret had been obtained for ARX systems. For a survey see [46]. Note, however, that in these papers the so-called indirect adaptive control procedures had been considered, where identifiability is ensured by the injection of rare shocks with diminishing frequency into the system. Similar results were obtained in [31, 32].

Let $D \subset \mathbb{R}^{n+m}$ be a set of candidate controller parameters to be specified below. For any $\eta \in D$ and for $t \geq 0$ we consider the control law

$$u(t-1) = -\eta^T \phi(t),$$

where $\phi(t)$ is defined above in (7.13). Thus we get a closed-loop system in which both u and y depend on η. To stress this dependence we write $u(t) = \bar{u}(t, \eta)$ and

$y(t) = \bar{y}(t, \eta)$. Let D denote the open set of η's in $\mathbb{R}^{m+n}$ such that the closed-loop system is stable. Define the nonlinear vector-valued function

$$G(\eta) \triangleq \lim_{t\to\infty} \mathrm{E}\,[\,\bar{\phi}(t,\eta)\bar{y}(t,\eta)]. \tag{7.17}$$

It is easy to see that we have $G(\eta^*) = 0$.

Let S^* denote the asymptotic covariance matrix of $\hat{\eta}(t)$; i.e., let

$$S^* = \lim_{t\to\infty} t \cdot \mathrm{E}\,[(\hat{\eta}(t) - \eta^*)(\hat{\eta}(t) - \eta^*)^T],$$

assuming that the limit exists. Define the *second order sensitivity matrix*

$$T^* = \lim_{t\to\infty} \mathrm{E}\left[\frac{\partial^2}{\partial\eta^2}_{|\eta=\eta^*} \bar{y}^2(t,\eta)\right]. \tag{7.18}$$

Claim (see [28]). *Consider the minimum-variance self-tuning regulator for an ARMAX(n, m, p) system given by* (7.15). *Then, under appropriate technical conditions, obtained by specializing the conditions of Theorem* 4.3, *we have the following pathwise characterization of the cumulative performance degradation:*

$$\lim_{N\to\infty} \sum_{t=1}^{N} (y^2(t) - e^2(t))/\log N = \frac{1}{2}\mathrm{Tr}\; T^*S^* \tag{7.19}$$

almost surely. Moreover, for any symmetric positive definite R we have

$$\frac{1}{2}\mathrm{Tr}\; T^*S^* \geq \sigma^2(e)(m+n). \tag{7.20}$$

The inequality (7.20) *is an equality if and only if $R = -G_\eta(\eta^*)$ and $C^* = 1$.*

The proof of (7.19) follows [24]. We note in passing that it has been a common belief that $G_\eta(\eta^*)$ is not computable. However, using a technique of Hjalmarsson (cf. [39]) it can be shown that for certain interesting physical systems $G_\eta(\eta^*)$ is in fact computable. The proof of (7.19) follows [24].

Conclusion. Performance degradation due to statistical uncertainty, also called regret, is of great interest in adaptive prediction and control of stochastic systems. To quantify the pathwise cumulative regret we need technical tools similar to those developed in [24] in the context of adaptive prediction of ARMA processes. These new tools have been developed in this paper. The usefulness of the results in stochastic adaptive control has been demonstrated for the minimum-variance self-tuning regulator for ARMAX systems in section 7; see also [28]. A further application for indirect adaptive control of multivariable linear stochastic systems is given in [27].

The results can be also applied in the context of identification for control; see [29, 40, 41]. For any fixed feedback strategy the covariance matrix of the estimation error and consequently the cumulative regret over any finite horizon will depend on the feedback strategy. The cumulative regret over finite horizon distorts the performance of the controller and this distortion can be precisely characterized using the results of the present paper. Thus a controller with optimal overall performance over a fixed finite horizon can be developed, at least in theory, i.e., pretending that we know the systems dynamics.

A further potential area of application is adaptive experimental design, see [30], in which the objective function to be minimized is the trace of the covariance matrix

of the estimation error, which can be computed experimentally for any fixed input pattern.

Another more classical possible application is the derivation of limit results such as LIL and invariance principles along the lines of [38].

The scope of applications can be enlarged by extending the technical results themselves. The extension of the results of the present paper to Kiefer–Wolfwitz-type stochastic approximation procedures, such as the simultaneous perturbation stochastic approximation, or SPSA, method due to Spall [61, 62] seems to be possible.

8. Auxiliary results.

LEMMA 8.1. *Let (x_t), $t \geq 0$, be a zero-mean L-mixing process with respect to $(\mathcal{F}_t, \mathcal{F}_t^+)$ and let y be an $\mathcal{F}_s$-measurable random variable for some $0 \leq s \leq t$, such that its moments, which appear in the inequality below, are finite. Then for every $1 \leq p, q \leq \infty$ such that $1/p + 1/q = 1$ we have*

$$|\mathrm{E} x_t y| \leq 2\gamma_p(t-s, x)\mathrm{E}^{1/q}|y|^q.$$

Analogous inequalities for strong mixing stationary sequences are given in [9] and for uniformly mixing stationary sequences in [42]. A concise survey of these inequalities is given in Chapter 7.2 of [15] and Appendix III of [33]. Here we restate an improved Hölder inequality under the weakest condition on mixing, namely strong-mixing or α-mixing (cf. Corollary 2.5 of Chapter 7.2 of [15]).

LEMMA 8.2. *Let $p, q, r > 1$ be such that $p^{-1} + q^{-1} + r^{-1} = 1$. Let Y and Z be $\mathcal{H}$-measurable and $\mathcal{G}$-measurable random variables such that $||Y||_q$ and $||Z||_r$ are finite, respectively. Then*

$$|\mathrm{E}[YZ] - \mathrm{E}[Y]\mathrm{E}[Z]| \leq C\alpha(\mathcal{H}, \mathcal{G})^{1/p}||Y||_q||Z||_r. \tag{8.1}$$

The improved Hölder inequality of Lemma 8.1 plays a key role in deriving the following moment inequality (cf. Theorem 1.1 in [17]).

THEOREM 8.1. *Let (u_t), $t \geq 0$, be a zero-mean L-mixing process. Let (f_t) be a function in $L_2[0, T]$. Then we have for all $m \geq 2$ with $C_m = 2(m-1)^{1/2}$*

$$\mathrm{E}^{1/m}\left|\int_0^T f_s u_s ds\right|^m \leq C_m \left(\int_0^T f_t^2 dt\right)^{1/2} M_m^{1/2}(u) \cdot \Gamma_m^{1/2}(u).$$

Extension of the statement to vector-valued processes is an elementary exercise, but obviously the constant C_m will be different. Extension to random (f_t) is not possible in general, but an extension is possible for multiple integrals with deterministic kernel (cf. [21]). Here we need only the following special result.

THEOREM 8.2. *Let (u_t) and (v_t) be zero-mean L-mixing processes. Then we have*

$$I_{T_0} = \int_{T_0}^T \frac{1}{t} u_t \int_{T_0}^t \frac{1}{s} v_s ds dt = O_M(T_0^{-1}).$$

The following simple lemma is stated as Corollary 3.5 in [24].

LEMMA 8.3. *Let $(\mathcal{F}_t, \mathcal{F}_t^+)$ be a pair of families of σ-algebras as in section 3 and let (x_t), $t \geq 0$, be an $\mathcal{F}_t$-adapted, measurable stochastic process. Then (x_t) is L-mixing with respect to $(\mathcal{F}_t, \mathcal{F}_t^+)$ if and only if the processes (x_{n+d}) are L-mixing with respect to $(\mathcal{F}_{n+d}, \mathcal{F}_{n+d}^+)$, uniformly in d for $0 \leq d < 1$.*

Let us consider a stochastic process $(u_n(\theta))$ with $\theta \in D \subset \mathbb{R}^p$, where D is an open set, which is measurable, separable, M-bounded, and M-Lipschitz continuous in θ for $\theta \in D$. By Kolmogorov's theorem the realizations of $(x_n(\theta))$ are continuous in θ with probability 1, hence we can define for almost all ω

$$u_n^* = \max_{\theta \in D_0} |u_n(\theta)|,$$

where $D_0 \subset D$ is a compact domain. The following result is given as Theorem 3.4 in [17].

THEOREM 8.3. *Assume that $(u_n(\theta))$ is a stochastic process which is measurable, separable, M-bounded, and M-Lipschitz continuous in θ for $\theta \in D$. Let u_n^* be the random variable defined above. Then we have for all positive integers q and $s > p$*

$$M_q(u^*) \le C(M_{qs}(u) + M_{qs}(\Delta u/\Delta\theta)),$$

where C depends only on p, q, s, and D_0, D.

A continuous-time version of the following lemma was given in [17] as Lemma 2.4.

LEMMA 8.4. *Let (u_n), $n \ge 0$, be a zero-mean L-mixing $\mathbb{R}^p$-valued process and define another $\mathbb{R}^p$-valued process (x_n) by*

$$x_{n+1} = Ax_n + u_n, \qquad x_0 = 0,$$

where the spectral norm of A is smaller than 1, say, we have $||A^n|| \le C\alpha^n$ with some $C > 0$ and $0 < \alpha < 1$. Then the output process (x_n) is L-mixing.

The first part of the following result was stated in Lemma 7.4 of [19]. The second part of the quoted lemma was not correctly stated and is therefore restated and proved here.

LEMMA 8.5. *Let (u_n), $n \ge 0$, be an M-bounded process and define a process (x_n) by*

$$x_{n+1} = \lambda x_n + \rho^n u_n, \qquad x_0 = 0, \tag{8.2}$$

where $0 < \lambda < \rho$. Then for any $m \ge 1$ we have

$$\mathrm{E}^{1/m}|x_n|^m \le \frac{\rho^n}{\rho - \lambda} M_m(u).$$

On the other hand, if $0 < \rho < \lambda$, then we have

$$\mathrm{E}^{1/m}|x_n|^m \le \frac{\lambda^n}{\lambda - \rho} M_m(u).$$

Proof. Let $0 < \lambda < \rho$ and set $z_n = \rho^{-n} x_n$. Then we have, after multiplying (8.2) by $\rho^{-(n+1)}$,

$$z_{n+1} = \lambda\rho^{-1} z_n + \rho^{-1} u_n,$$

which can be solved explicitly for z_n to get

$$z_n = \sum_{i=0}^{n-1} (\lambda\rho^{-1})^i \rho^{-1} u_{n-1-i}.$$

Using the triangle inequality for the $L_m(\Omega, \mathcal{F}, P)$-norm and the condition $0 < \lambda < \rho$ we get

$$M_m(z) \le (1 - \lambda\rho^{-1})^{-1}\rho^{-1}M_m(u)$$

from which the first proposition follows.

A useful reformulation of the above argument is the following: writing

$$x_n = \rho^n z_n = \sum_{i=0}^{n-1} \lambda^i \rho^{n-1-i} u_{n-1-i} \tag{8.3}$$

we have

$$\mathrm{E}^{1/m}|x_n|^m \le \sum_{i=0}^{n-1} \lambda^i \rho^{n-1-i} \mathrm{E}^{1/m}|u_{n-1-i}|^m \le \sum_{i=0}^{n-1} \lambda^i \rho^{n-1-i} M_q(u). \tag{8.4}$$

Thus it is sufficient to establish that for $0 < \lambda < \rho$

$$\sum_{i=0}^{n-1} \lambda^i \rho^{n-1-i} \le \frac{\rho^n}{\rho - \lambda} \tag{8.5}$$

and this is obtained from the above argument with $u_n = 1$ for all n. The advantage of this reformulation is that the left-hand side is the convolution of the sequences (λ^n) and (ρ^n) and thus it is symmetric in λ and ρ.

In the case when $0 < \rho < \lambda$ we use (8.4) to estimate $\mathrm{E}^{1/m}|x_n|^m$, but the role of λ and ρ is interchanged; thus we get

$$\mathrm{E}^{1/m}|x_n|^m \le \frac{\lambda^n}{\lambda - \rho} M_m(u). \qquad \square$$

Remark. A simple corollary is that

$$\sum_{i=0}^{n-1} \lambda^i \rho^{n-1-i} \le \frac{\max\ (\lambda^n, \rho^n)}{|\rho - \lambda|}. \tag{8.6}$$

The lemma below has been used for ODE analysis of stochastic approximation processes in [16]. The conditions are similar to Condition 3.4(i). Consider the ODE

$$\dot{y}_t = F(t, y_t), \qquad y_s = \xi,\ s \ge 1. \tag{8.7}$$

The solution of the above ODE will be denoted by $y(t, s, \xi)$ in the time interval where it exists and is unique.

CONDITION 8.1. *$F = (F(t, y))$ is defined for $t \ge 1$, $y \in D$, where $D \subset \mathbb{R}^p$ is an open set and F is continuously differentiable in (t, y). It is assumed that there exists a compact domain $D_0' \subset D$ such that $y(t, s, \xi) \in D$ for all $\xi \in D_0'$ and $1 \le s \le t < \infty$.*

LEMMA 8.6. *Assume that Condition 8.1 is satisfied. Let (x_t), $1 \le t < \infty$, be a continuous, piecewise continuously differentiable curve such that $x_t \in D_0'$ for $t \ge 1$ and $x_1 = y_1 = \xi \in D_0'$. Then for $t \ge 1$*

$$x_t - y_t = \int_1^t \frac{\partial}{\partial \xi} y(t, r, x_r)\,(\dot{x}_r - F(r, x_r))\ dr. \tag{8.8}$$

Proof. Write $z_r = y(t, r, x_r)$. Obviously the left-hand side of (8.8) can be written as $z_t - z_1$ and we have

$$z_t - z_1 = \int_1^t \dot{z}_r dr = \int_1^t (y_r(t, r, x_r) + y_\xi(t, r, x_r)\dot{x}_r)\ dr. \tag{8.9}$$

Taking into account the equality $y_r(t, r, x_r) = -y_\xi(t, r, x_r) \cdot F(t, x_r)$ we get the lemma. □

A discretized version of the above lemma has been used implicitly in the final step of the proof of Theorem 1.1 of [19], see (2.10) of [19]. We now formulate this lemma with explicit conditions. It has been used in the proof of Lemma 4.4.

CONDITION 8.2. *Let $D_0' \subset D$ be a compact domain as in Condition* 8.1. *Assume that D_0' is convex and that there exists a compact set $D_0 \subset D_0'$ such that for all $x \in D_0$ and $t \geq s \geq 1$ we have $y(t, s, x) \in D_0'$.*

Let $1 = s_0 \leq s_2 \leq \cdots \leq s_n \leq s_{n+1} = t$ and let $(x_{s_i}) \in D_0$, $i = 0, 1, \ldots, n$, be a sequence such that $x_1 = y_1 = \xi \in D_0$. These points are considered as approximations to $y_{s_i} = y(s_i, 1, \xi)$. We will estimate the tracking error $x_t - y_t$ in terms of *local tracking* errors

$$(x_{s_i} - y(s_i, s_{i-1}, x_{s_{i-1}})).$$

LEMMA 8.7. *Let $F = (F(t, y))$ satisfy Conditions* 8.1 *and* 8.2 *and let $(x_{s_i}) \in D_0$, $i = 0, 1, \ldots, n$, be a sequence such that $x_1 = y_1 = \xi$. Then*

$$x_t - y_t = (x_t - y(t, s_n, x_{s_n})) + \sum_{i=1}^n \int_0^1 \frac{\partial}{\partial \xi} y(t, s_i, w(i, \lambda))\ d\lambda \cdot (x_{s_i} - y(s_i, s_{i-1}, x_{s_{i-1}})),$$

where $w(i, \lambda) = (1 - \lambda)y(s_i, s_{i-1}, x_{s_{i-1}}) + \lambda x_{s_i}$.

Proof. Consider the sequence $z_i = y(t, s_i, x_{s_i})$, $i = 0, 1, \ldots, n$. Then $z_0 = y_t$ and we can write

$$\begin{aligned} x_t - y_t &= (x_t - z_n) + \sum_{i=1}^n (z_i - z_{i-1}) \\ &= (x_t - y(t, s_n, x_{s_n})) + \sum_{i=1}^n (y(t, s_i, x_{s_i}) - y(t, s_{i-1}, x_{s_{i-1}})). \end{aligned} \tag{8.10}$$

Now for $1 \leq s \leq s' \leq t$ we have $y(t, s, x) = y(t, s', y(s', s, x))$. Setting $s = s_{i-1}$, $s' = s_i$, $x = x_{s_{i-1}}$, the ith term of the right-hand side of (8.10) thus becomes

$$\begin{aligned} &y(t, s_i, x_{s_i}) - y(t, s_i, y(s_i, s_{i-1}, x_{s_{i-1}})) \\ &\qquad = \int_0^1 \frac{\partial}{\partial \xi} y(t, s_i, w(i, \lambda))\ d\lambda \cdot (x_{s_i} - y(s_i, s_{i-1}, x_{s_{i-1}})) \end{aligned}$$

with $w(i, \lambda) = (1 - \lambda)y(s_i, s_{i-1}, x_{s_{i-1}}) + \lambda x_{s_i}$ for $0 \leq \lambda \leq 1$. Note that $w(i, \lambda) \in D_0'$ for $i = 1, \ldots, n$ since D_0' is convex and thus $y(t, s_i, w(i, \lambda))$ is well defined, and the lemma follows. □

Let $G = (G(y))$ be defined in an open set $D \subset \mathbb{R}^p$ and consider the ODE

$$\dot{y}_t = \frac{1}{t} G(y_t), \qquad y_s = \xi,\ s \geq 1. \tag{8.11}$$

We will have conditions that ensure that the above ODE has a unique solution in some finite or infinite interval, which we denote by $y(t,s,\xi)$. We assume the validity of the following condition, which is weaker than Conditions 3.3 and 3.4.

CONDITION 8.3. *G has continuous partial derivatives up to second order for $y \in D$. There exist compact sets $D_0 \subset D_0' \subset D$ such that for all $\xi \in D_0$, $t \geq s \geq 1$, we have $y(t,s,\xi) \in D_0'$ and*

$$\|y_\xi(t,s,\xi)\| \leq C_0(s/t)^\alpha \tag{8.12}$$

with some $C_0 \geq 1$, $\alpha > 0$. Let $||\partial^i G(y)/\partial y^i|| \leq L$ for $y \in D_0'$ and $i = 0,1,2$.

We prove that the stability expressed by the condition above is in a sense inherited by the second order derivatives of $y(t,s,\xi)$.

LEMMA 8.8. *Let G satisfy Condition 8.3. Then for all $\xi \in D_0$, $t \geq s \geq 1$,*

$$\|y_{\xi\xi}(t,s,\xi)\| \leq L\alpha^{-1}C_0^3 \cdot (s/t)^\alpha,$$
$$\|y_{s\xi}(t,s,\xi)\| \leq (L\alpha^{-1}+1)LC_0^3 \cdot \frac{1}{s}(s/t)^\alpha.$$

Remark. From the proof below it follows that if G is three times continuously differentiable, then with some constant C_0' we have $\|y_{\xi\xi\xi}(t,s,\xi)\| \leq C_0'(s/t)^\alpha$.

Proof. Use a change of time-scale $t = e^v$, $s = e^u$, and consider the differential equation

$$\frac{d}{dv}z_v = G(z_v), \qquad z_u = \xi,\ u \geq 0,$$

with its solution being denoted by $z(v,u,\xi)$, $v \geq u \geq 0$. Then (8.12) implies

$$\|z_\xi(v,u,\xi)\| \leq C_0 e^{-\alpha(u-v)}, \tag{8.13}$$

and the propositions of the lemma are equivalent to, after the substitution $u = \log t$ and $v = \log s$,

$$\|z_{\xi\xi}(v,u,\xi)\| \leq L\alpha^{-1}C_0^3 \cdot e^{-\alpha(v-u)},$$
$$\|z_{u\xi}(v,u,\xi)\| \leq (L\alpha^{-1}+1)LC_0^3 \cdot e^{-\alpha(v-u)}.$$

Now we have

$$\frac{\partial}{\partial v}z_\xi(v,u,\xi) = G_y(z(v,u,\xi)) \cdot z_\xi(v,u,\xi), \qquad z_\xi(u,u,\xi) = I. \tag{8.14}$$

It is easy to see that $z_{\xi\xi}(v,u,\xi)$ exists and is continuous in (v,u,ξ). From (8.14) we get

$$\frac{\partial}{\partial v}z_{\xi\xi}(v,u,\xi) = G_{yy}(z(v,u,\xi)) \cdot z_\xi(v,u,\xi)z_\xi(v,u,\xi) + +G_y(z(v,u,\xi)) \cdot z_{\xi\xi}(v,u,\xi)$$

with $z_{\xi\xi}(u,u,\xi) = 0$. Since the operator norm of the first term is majorized by $LC_0^2e^{-2\alpha(u-v)}$ and since the time-varying linear differential equation with transition matrix $G_y(z(v,u,\xi))$ is exponentially stable due to (8.13), we get the first claim of the lemma from the identity

$$\int_0^t e^{-\alpha(v-r)}e^{-2\alpha r}dr = e^{-\alpha v}\int_0^v e^{-\alpha r}dr < \alpha^{-1}e^{-\alpha v}.$$

To estimate the mixed derivatives, take into account $z_u(v,u,\xi) = -z_\xi(v,u,\xi)\cdot G(\xi)$ to get

$$z_{u\xi}(v,u,\xi) = -z_{\xi\xi}(v,u,\xi)\cdot G(\xi) - z_\xi(v,u,\xi)\cdot G_\xi(\xi),$$

from which the second claim follows using (8.13) and the proven first part of the lemma. □

Acknowledgments. The author expresses his thanks to Peter Caines for arranging a long-term visit to McGill University and for numerous fruitful discussions on stochastic systems. The help of Zsuzsanna Vágó, Zalán Mátyás, and Zsanett Orlovits in rereading the manuscript of the paper is also gratefully acknowledged.

REFERENCES

[1] K. J. Åström and T. Söderström, *Uniqueness of the maximum likelihood estimates of the parameters of an ARMA model*, IEEE Trans. Automat. Control, 19 (1974), pp. 769–773.
[2] K. J. Åström and B. Wittenmark, *Problems of identification and control*, J. Math. Anal. Appl., 34 (1971), pp. 90–113.
[3] A. Benveniste, M. Métivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, Berlin, 1990.
[4] B. Bercu, *Weighted estimation and tracking for ARMAX models*, SIAM J. Control Optim., 33 (1995), pp. 89–106.
[5] A. N. Borodin, *A stochastic approximation procedure in the case of weakly dependent observations*, Theory Probab. Appl., 24 (1979), pp. 34–52.
[6] P. E. Caines, *Linear Stochastic Systems*, Wiley, New York, 1988.
[7] H. F. Chen and L. Guo, *Identification and Stochastic Adaptive Control*, Birkhäuser Boston, Boston, MA, 1991.
[8] L. D. Davisson, *Prediction error of stationary Gaussian time series of unknown covariance*, IEEE Trans. Inform. Theory, 19 (1965), pp. 783–795.
[9] Yu. A. Davydov, *Convergence of distributions generated by stationary stochastic processes*, Theory Probab. Appl., 13 (1968), pp. 691–696.
[10] B. Delyon, *General results on the convergence of stochastic algorithms*, IEEE Trans. Automat. Control, 41 (1996), pp. 1245–1255.
[11] D. P. Djereveckii and A. L. Fradkov, *An application of the theory of Markov processes to the analysis of the dynamics of adaptation algorithms*, Autom. Remote Control, 2 (1974), pp. 39–48.
[12] D. P. Djereveckii and A. L. Fradkov, *Two models for analyzing the dynamics of adaptation algorithms*, Autom. Remote Control, 1 (1974), pp. 67–75.
[13] D. P. Djereveckii and A. L. Fradkov, *Applied Theory of Discrete Adaptive Control Systems*, Nauka, Moscow, 1981 (in Russian).
[14] T. E. Duncan and B. Pasik-Duncan, *Some methods for the adaptive control of continuous time linear stochastic systems*, in Topics in Stochastic Systems: Modelling, Estimation and Adaptive Control, L. Gerencsér and P. E. Caines, eds., Springer-Verlag, Berlin, Heidelberg, 1991, pp. 242–267.
[15] S. N. Ethier and T. G. Kurtz, *Markov Processes. Characterization and Convergence*, Wiley, New York, 1986.
[16] S. Geman, *Some averaging and stability results for random differential equations*, SIAM J. Appl. Math., 36 (1979), pp. 86–105.
[17] L. Gerencsér, *On a class of mixing processes*, Stochastics, 26 (1989), pp. 165–191.
[18] L. Gerencsér, *On the martingale approximation of the estimation error of ARMA parameters*, Systems Control Lett., 15 (1990), pp. 417–423.
[19] L. Gerencsér, *Rate of convergence of recursive estimators*, SIAM J. Control Optim., 30 (1992), pp. 1200–1227.
[20] L. Gerencsér, *A representation theorem for the error of recursive estimators*, in Proceedings of the 31st IEEE Conference on Decision and Control, Tucson, AZ, 1992, pp. 2251–2256.
[21] L. Gerencsér, *Multiple integrals with respect to L-mixing processes*, Statist. Probab. Lett., 17 (1993), pp. 73–83.
[22] L. Gerencsér, *Strong approximation of the recursive prediction error estimator of the parameters of an ARMA process*, Systems Control Lett., 21 (1993), pp. 347–351.

[23] L. Gerencsér, *Fixed gain off-line estimators of ARMA parameters*, J. Math. Systems Estim. Control, 4 (1994), pp. 249–252. Retrieval code for full electronic manuscript, 66945.
[24] L. Gerencsér, *On Rissanen's predictive stochastic complexity for stationary ARMA processes*, J. Statist. Plann. Inference, 41 (1994), pp. 303–325.
[25] L. Gerencsér, *Stability of random iterative mappings*, in Modeling Uncertainty. An Examination of Its Theory, Methods, and Applications, M. Dror, P. Lécuyer, and F. Szidarovszky, eds., Kluwer, Dordrecht, The Netherlands, 2002, pp. 359–371.
[26] L. Gerencsér and J. Rissanen, *A prediction bound for Gaussian ARMA processes*, in Proceedings of the 25th IEEE Conference on Decision and Control, Vol. 3, Athens, Greece, 1986, pp. 1487–1490.
[27] L. Gerencsér and Zs. Vágó, *Adaptive control of multivariable linear stochastic systems. A strong approximation approach*, in Proceedings of the European Control Conference, Karlsuhe, Germany, 1999, p. F587.
[28] L. Gerencsér, J. H. van Schuppen, J. Rissanen, and Zs. Vágó, *Stochastic complexity, selftuning and optimality*, in Proceedings of the 33rd IEEE Conference on Decision and Control, Orlando, FL, 1994, pp. 652–654.
[29] M. Gevers, *Towards a joint design of identification and control?*, in Essays on Control: Perspectives in the Theory and Its Applications, H. L. Trentelman and J. C. Willems, eds., Birkhäuser Boston, Boston, MA, 1993, pp. 111–151.
[30] M. Gevers, X. Bombois, B. Codrons, F. De Bruyne, and G. Scorletti, *The role of experimental conditions in model validation for control*, in Robustness in Identification and Control, Lecture Notes in Control and Inform. Sci. 245, A. Garulli, A. Tesi, and A. Vicino, eds., Springer-Verlag, London, 1999, pp. 72–86.
[31] L. Guo, *The logarithm law of self-tuning regulators*, in Proceedings of the 12th IFAC World Congress, Vol. I, Sydney, Australia, 1993, pp. 227–232.
[32] L. Guo, *Further results on least squares based adaptive minimum variance control*, SIAM J. Control Optim., 32 (1994), pp. 187–212.
[33] P. Hall and C. C. Heyde, *Martingale Limit Theory and Its Applications*, Academic Press, New York, 1980.
[34] E. J. Hannan, *The convergence of some time-series recursions*, Ann. Statist., 4 (1976), pp. 1258–1270.
[35] E. J. Hannan and M. Deistler, *The Statistical Theory of Linear Systems*, Wiley, New York, 1988.
[36] E. J. Hannan, A. J. McDougall, and D. S. Poskitt, *Recursive estimation of autoregressions*, J. Roy. Statist. Soc. Ser. B, 51 (1989), pp. 217–233.
[37] E. M. Hemerley and M. A. H. Davis, *Strong consistency of the PLS criterion for order determination of autoregressive processes*, Ann. Statist., 17 (1989), pp. 941–946.
[38] A. Heunis, *Rates of convergence for an adaptive filtering algorithm driven by stationary dependent data*, SIAM J. Control Optim., 32 (1994), pp. 116–139.
[39] H. Hjalmarsson, *Efficient tuning of linear multivariable controllers using iterative feedback tuning*, Internat. J. Adapt. Control Signal Process., 13 (1990), pp. 553–572.
[40] H. Hjalmarsson, M. Gevers, and F. De Bruyne, *For model-based control design, closed-loop identification gives better performance*, Automatica J. IFAC, 32 (1996), pp. 1659–1673.
[41] H. Hjalmarsson and K. Lindqvist, *Identification for control: L_2 and L_∞ methods*, in Proceedings of the 40th IEEE Conference on Decision and Control, Orlando, FL, 2001, pp. 2701–2706.
[42] I. A. Ibragimov and Yu. A. Linnik, *Independent and Stationary Sequences of Random Variables*, Wolters and Nordhoff, Groningen, The Netherlands, 1971.
[43] J. A. Joslin and A. J. Heunis, *Law of the iterated logarithm for a constant-gain linear stochastic gradient algorithm*, SIAM J. Control Optim., 39 (2000), pp. 533–570.
[44] T. Kailath, *Linear Systems*, Prentice–Hall, Englewood Cliffs, NJ, 1980.
[45] H. J. Kushner and G. Yin, *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, New York, 1997.
[46] T. Z. Lai, *Information bounds, certainty equivalence and learning in asymptotically efficient adaptive control of time-invariant stochastic systems*, in Topics in Stochastic Systems: Modelling, Estimation and Adaptive Control, L. Gerencsér and P. E. Caines, eds., Springer-Verlag, Berlin, Heidelberg, 1991, pp. 268–299.
[47] T. Z. Lai and C. Z. Wei, *Least squares estimates in stochastic regression models with application to identification and control of dynamic systems*, Ann. Statist., 10 (1982), pp. 154–165.
[48] T. Z. Lai and C. Z. Wei, *Extended least squares and their applications to adaptive control of dynamic systems*, IEEE Trans. Automat. Control, 31 (1986), pp. 898–906.
[49] T. Z. Lai and C. Z. Wei, *Asymptotically efficient self-tuning regulators*, SIAM J. Control Optim., 25 (1987), pp. 466–481.

[50] L. Ljung, *On consistency and identifiability*, Math. Programming Stud., 5 (1976), pp. 169–190.
[51] L. Ljung, *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, 22 (1977), pp. 551–575.
[52] L. Ljung, G. Pflug, and H. Walk, *Stochastic Approximation and Optimization of Random Systems*, DMV Seminar 17, Birkhäuser-Verlag, Basel, 1992.
[53] L. Ljung and T. Söderström, *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, MA, 1983.
[54] M. B. Nevel'son and R. Z. Has'minskii, *Stochastic Approximation and Recursive Estimation*, AMS, Providence, RI, 1976.
[55] R. Ober, *Balanced realizations: Canonical form, parametrization, model reduction*, Internat. J. Control, 46 (1987), pp. 643–670.
[56] J. Rissanen, *A predictive least squares principle*, IMA J. Math. Control Inform., 3 (1986), pp. 211–222.
[57] J. Rissanen, *Stochastic complexity and predictive modelling*, Ann. Statist., 14 (1986), pp. 1080–1100.
[58] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*, World Scientific Publisher, Teaneck, NJ, 1989.
[59] J. Rissanen and P. E. Caines, *The strong consistency of maximum likelihood estimators for ARMA processes*, Ann. Statist., 7 (1979), pp. 297–315.
[60] V. Solo, *The second order properties of a time series recursion*, Ann. Statist., 9 (1981), pp. 307–317.
[61] J. C. Spall, *Multivariate stochastic approximation using a simultaneous perturbation gradient approximation*, IEEE Trans. Automat. Control, 37 (1992), pp. 332–341.
[62] J. C. Spall, *Adaptive stochastic approximation by the simultaneous perturbation method*, in Proceedings of the 37th IEEE Conference on Decision and Control, Tampa, FL, 1998, pp. 3872–3879.
[63] J. H. van Schuppen, *Tuning of Gaussian Stochastic Control Systems*, Report BS-R9223, CWI, Amsterdam, 1992.
[64] S. M. Veres, *Relations between information criteria for model-structure selection Part 3. Strong consistency of the predictive least squares criterion*, Internat. J. Control, 52 (1990), pp. 737–751.
[65] G. Yin, *A stopping rule for the Robbins-Monro method*, J. Optim. Theory Appl., 67 (1990), pp. 151–173.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 6, pp. 2189–2209

STATE FEEDBACK IMPULSE ELIMINATION FOR SINGULAR SYSTEMS OVER A HERMITE DOMAIN*

DANIEL COBB†

Abstract. We reduce the problem of impulse elimination via state feedback in singular differential equations to algebra. Our results are developed for systems over an arbitrary Hermite domain. We show that the established theories for the time-invariant and the real analytic time-varying settings can be unified in this way. Besides the constant and real analytic functions, several other function rings are considered. Our algebraic theory is applied to these cases, providing solutions to the impulse elimination problem for classes of systems not previously studied. In particular, our work allows the restriction of the feedback matrix to certain function rings.

Key words. singular systems, impulse elimination, algebraic systems, state feedback

AMS subject classifications. 93B25, 93B52, 93B55

DOI. 10.1137/040618515

1. Introduction.

We are interested in the problem of designing a state feedback law $u = K(t)\,x$ for a time-varying singular differential equation

$$E(t)\,\dot{x} = A(t)\,x + B(t)\,u \tag{1}$$

such that the closed-loop system

$$E(t)\,\dot{x} = (A(t) + B(t)\,K(t))\,x \tag{2}$$

exhibits no impulsive transients. The matrices E, A, and B are assumed to have entries in an appropriate set of functions on $\mathbb{R}$ (possibly constant) with $E(t), A(t) \in \mathbb{R}^{n\times n}$, $B(t) \in \mathbb{R}^{n\times m}$, and $K(t) \in \mathbb{R}^{m\times n}$. This problem has been treated in a variety of contexts over the past 25 years [10], [16], [12], [13], [4], [17], [18]. For example, we originally posed and solved the problem for the time-invariant (i.e., constant matrix) case in [10].

For time-invariant systems, the fact that solutions of (2) can exhibit impulsive behavior was originally established in [14] and [15, Ch. 22]. One method of analysis is based on the Weierstrass decomposition [8, Thm. 3, p. 28]: Given E, A with $\det(sE - A) \not\equiv 0$, there exist nonsingular $P, Q \in \mathbb{R}^{n\times n}$ such that

$$PEQ = \begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix}, \quad PAQ = \begin{bmatrix} X & 0 \\ 0 & I \end{bmatrix},$$

where N is nilpotent. If $N \neq 0$, the solution of (1) contains an impulsive term of the form

$$z = -\sum \delta^{(k-1)} N^k z_o. \tag{3}$$

(See [19] for details.) More generally, when $E(t)$ and $A(t)$ are analytic functions, it is shown in [3] that an expression similar to (3) holds under mild assumptions.

*Received by the editors November 8, 2004; accepted for publication (in revised form) August 5, 2005; published electronically January 26, 2006.
http://www.siam.org/journals/sicon/44-6/61851.html
†Department of Electrical and Computer Engineering, University of Wisconsin, 1415 Engineering Drive, Madison, WI 53706-1691 (cobb@engr.wisc.edu).

Since impulses must be interpreted as unbounded, conventional notions of closed-loop stability dictate that K be chosen to make (2) impulse free. For the time-invariant case, we established a necessary and sufficient condition ([10, Thm. 6]) under which such a matrix K exists. This condition can be written

$$\operatorname{Im} E + A \operatorname{Ker} E + \operatorname{Im} B = \mathbb{R}^n.$$

Since then, two alternative proofs of this result have appeared. (See [12, Thm. 2.5.1] and [13, Thm. 3-2.1].)

The work of Campbell and Petzold [3] extended the theory of singular systems (1) to the time-varying setting, where E, A, and B are matrices over the real analytic functions on $\mathbb{R}$. More recently, the corresponding impulse elimination problem was solved by Wang in [4, Thm. 4.1]. In this case, necessary and sufficient conditions for impulse elimination are

$$\operatorname{Im} E(t) + A(t) \operatorname{Ker} E(t) + \operatorname{Im} B(t) = \mathbb{R}^n \quad \forall t,$$
$$\operatorname{rank} E(t) = \text{constant}.$$

Our contention is that the impulse elimination problem is primarily a problem in algebra. Indeed, after careful examination (and some modification), the arguments in [4] can be reduced to algebraic manipulations over a certain class of rings. Pursuing this idea not only leads to a unification of the time-invariant and analytic time-varying theories, but also yields a more general framework in which the impulse elimination problem for other classes of time-varying systems can be solved with little extra effort.

An important consequence of our approach is that it allows the entries of K to be restricted to certain function rings (although E, A, and B must share the same restriction). Hence, we are able to solve a wide variety of constrained feedback problems which have not been considered in the literature.

Our algebraic theory is the subject of sections 2 and 3. In section 4, we apply our results to various types of time-varying singular systems.

2. Algebraic preliminaries. Let R be a commutative ring (with identity). If $x_1, \ldots, x_k \in R$, a *Bezout identity* is an equation of the form $\sum a_i x_i = 1$ ($a_i \in R$). For a matrix $M \in R^{p \times q}$, let

$$\operatorname{rank} M = \max \left\{ k \;\middle|\; M \text{ has a nonzero } k\text{th-order minor} \right\} \tag{4}$$

and

$$\rho M = \max \left\{ k \;\middle|\; \text{the } k\text{th-order minors of } M \text{ satisfy a Bezout identity} \right\}. \tag{5}$$

Obviously, $\operatorname{rank} M \geq \rho M$ for any M. It can be shown that $\operatorname{rank} M$ and ρM are invariant under left and right unimodular transformations. (See [1, p. 25].) If $R = \mathbb{R}$, then $\operatorname{rank} M = \rho M$. We denote this common value by $\operatorname{rank}_{\mathbb{R}} M$.

Consider the set G of all triples (P, Q, D), where $P, Q, D \in R^{n \times n}$ and P, Q are unimodular. Define the binary operation

$$(P_1, Q_1, D_1) * (P_2, Q_2, D_2) = (P_2 P_1, Q_1 Q_2, D_1 Q_2 + Q_1 D_2).$$

It is routine to verify that G has the structure of a group. Now consider pairs (E, A), where $E, A \in R^{n \times n}$. We may define a right group action on the set of all (E, A) according to

$$(E, A) \cdot (P, Q, D) = (PEQ, P(AQ + ED)). \tag{6}$$

The *orbit* of particular (E, A) is the set of all pairs $(\widetilde{E}, \widetilde{A})$ such that $(\widetilde{E}, \widetilde{A}) = (E, A) \cdot (P, Q, D)$ for some P, Q, D. It is easy to verify that the set of all orbits forms a partition of $R^{n\times n} \times R^{n\times n}$.

Following the terminology of Campbell and Petzold [3], we say (E, A) is in *standard canonical form* if

$$E = \begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix}, \quad A = \begin{bmatrix} X & 0 \\ 0 & I \end{bmatrix}, \tag{7}$$

where N is strictly upper triangular with E, A identically partitioned. Similar to their notion of "analytic solvability" for systems (1), we say $(E, A) \in R^{n\times n} \times R^{n\times n}$ is *algebraically solvable* if its orbit under (6) contains a member in standard canonical form. (The degenerate cases (I, X) and (N, I) are also allowed.) We say that (E, A) has *unit index* if the orbit of (E, A) contains a member in standard canonical form with $N = 0$. It is clear from the definitions that algebraic solvability is invariant under the group action (6).

The question arises whether a unit index orbit can contain a member in standard canonical form with $N \neq 0$. Fortunately, the next result answers this question in the negative.

THEOREM 2.1. *Suppose* (E, A) *has unit index and* $(E, A)\cdot(P, Q, D)$ *is in standard canonical form* (7). *Then* $N = 0$.

Proof. Suppose (E, A) belongs to an orbit with two members in standard canonical form, one with $N = 0$ and the other with $N \neq 0$. Then there exist D and unimodular P and Q such that

$$P \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix} Q^{-1},$$

$$P \left(\begin{bmatrix} X & 0 \\ 0 & I \end{bmatrix} + \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} DQ^{-1} \right) = \begin{bmatrix} Y & 0 \\ 0 & I \end{bmatrix} Q^{-1}$$

for some X and Y and some strictly upper triangular $N \neq 0$. Let

$$\begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} = DQ^{-1}, \quad P_1 = P \begin{bmatrix} I & D_{12} \\ 0 & I \end{bmatrix},$$

and $X_1 = X + D_{11}$. Then

$$P_1 \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} = P \begin{bmatrix} I & D_{12} \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} = P \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix} Q^{-1}, \tag{8}$$

$$\begin{aligned} P_1 \begin{bmatrix} X_1 & 0 \\ 0 & I \end{bmatrix} &= P \begin{bmatrix} I & D_{12} \\ 0 & I \end{bmatrix} \begin{bmatrix} X_1 & 0 \\ 0 & I \end{bmatrix} \\ &= P \begin{bmatrix} X + D_{11} & D_{12} \\ 0 & I \end{bmatrix} \\ &= P \left(\begin{bmatrix} X & 0 \\ 0 & I \end{bmatrix} + \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} DQ^{-1} \right) \\ &= \begin{bmatrix} Y & 0 \\ 0 & I \end{bmatrix} Q^{-1}. \end{aligned} \tag{9}$$

Let

$$\begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} = P_1, \quad \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} = Q^{-1}.$$

From (8), $P_{21} = NQ_{21}$ and $NQ_{22} = 0$. From (9), $P_{22} = Q_{22}$. Since N is strictly triangular and nonzero, there exist an integer $q > 1$ and x such that $N^q = 0$ and $N^{q-1}x \neq 0$. Since P_1 is unimodular, there exist y and z such that

$$\begin{bmatrix} 0 \\ x \end{bmatrix} = P_1 \begin{bmatrix} y \\ z \end{bmatrix},$$

$$x = P_{21}y + P_{22}z = NQ_{21}y + Q_{22}z.$$

Multiplying by N^{q-1} yields

$$N^{q-1}x = N^q Q_{21}y + N^{q-1}Q_{22}z = 0,$$

which is a contradiction. □

In practice, algebraic solvability may be difficult to establish, so we introduce a more direct condition that will suit our purposes just as well. We say that (E, A) is *presolvable* if *any one* of the following conditions holds:

(PS1) $\operatorname{Im} E + A \operatorname{Ker} E = R^n$,

(PS2) $\operatorname{Im} E \cap A \operatorname{Ker} E \neq 0$,

(PS3) $\operatorname{Ker} E \cap \operatorname{Ker} A \neq 0$.

Algebraic solvability and standard canonical form are related to existence and uniqueness of solutions of (1), as discussed in [3]. However, presolvability is a purely algebraic condition, having no simple connection to the dynamics of (1). Nevertheless, we can prove the following.

THEOREM 2.2.

(1) *Algebraic solvability implies presolvability.*

(2) *Presolvability is invariant under the group action* (6).

Proof. (1) There exist P, Q, and D that put (E, A) in standard canonical form. Suppose $N = 0$. Then

$$\begin{aligned} P(\operatorname{Im} E + A \operatorname{Ker} E) &= P\left(\operatorname{Im} E + A \operatorname{Ker} E + EDQ^{-1} \operatorname{Ker} E\right) \\ &= \operatorname{Im} PEQ + PAQ \operatorname{Ker} PEQ + PED \operatorname{Ker} PEQ \\ &\supset \operatorname{Im} PEQ + P(AQ + ED) \operatorname{Ker} PEQ \\ &= \operatorname{Im} \begin{bmatrix} I \\ 0 \end{bmatrix} + \operatorname{Im} \begin{bmatrix} 0 \\ I \end{bmatrix} \\ &= R^n, \end{aligned}$$

so (PS1) holds.

If $N \neq 0$, there exists an integer $q > 1$ such that $N^q = 0$ and $N^{q-1} \neq 0$. Choose any $x \in R^n$ such that $N^{q-1}x \neq 0$, set $y = N^{q-2}x$, and $z = Ny$. Then $z \neq 0$. Let

$$v = Q \begin{bmatrix} 0 \\ y \end{bmatrix} - D \begin{bmatrix} 0 \\ z \end{bmatrix}, \quad w = Q \begin{bmatrix} 0 \\ z \end{bmatrix}.$$

Then $w \neq 0$ and

$$P(Ev - Aw) = PEQ \begin{bmatrix} 0 \\ y \end{bmatrix} - P(AQ + ED) \begin{bmatrix} 0 \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ Ny - z \end{bmatrix} = 0,$$

so $Ev = Aw$. Also,

$$P(Ew) = PEQ \begin{bmatrix} 0 \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ Nz \end{bmatrix} = 0,$$

so $w \in \operatorname{Ker} E$ and $Aw \in \operatorname{Im} E \cap A \operatorname{Ker} E$. If $Aw \neq 0$, (PS2) holds; if $Aw = 0$, (PS3) holds.

(2) To prove invariance of presolvability, first suppose (PS1) holds for (E, A). Then

$$\begin{aligned}\operatorname{Im} PEQ + P(AQ + ED) \operatorname{Ker} PEQ &= P\left(\operatorname{Im} E + \left(A + EDQ^{-1}\right) \operatorname{Ker} E\right) \\ &= P\left(\operatorname{Im} E + \left(A + EDQ^{-1}\right) \operatorname{Ker} E + EDQ^{-1} \operatorname{Ker} E\right) \\ &\supset P\left(\operatorname{Im} E + \left(\left(A + EDQ^{-1}\right) - EDQ^{-1}\right) \operatorname{Ker} E\right) \\ &= R^n,\end{aligned}$$

so (PS1) also holds for $(E, A) \cdot (P, Q, D)$ and $(E, A) \cdot (P, Q, D)$ is presolvable.

Now assume that (PS2) holds for (E, A), but not for $(E, A) \cdot (P, Q, D)$. Then there exist x, y such that $Ex = 0$ and $Ey = Ax \neq 0$. Hence, $x \neq 0$,

$$P(AQ + ED) Q^{-1} x = PE\left(y + DQ^{-1}x\right) \in \operatorname{Im} PEQ \cap P(AQ + ED) \operatorname{Ker} PEQ = 0,$$

$$0 \neq Q^{-1}x \in \operatorname{Ker} PEQ \cap \operatorname{Ker} P(AQ + ED). \tag{10}$$

This establishes (PS3), and therefore, presolvability relative to $(E, A) \cdot (P, Q, D)$.

Finally, suppose that (PS3) holds for (E, A), but $(E, A) \cdot (P, Q, D)$ fails to satisfy (PS2). Then there exists $x \neq 0$ such that $Ex = Ax = 0$ and

$$P(AQ + ED) Q^{-1} x = PEDQ^{-1}x \in \operatorname{Im} PEQ \cap P(AQ + ED) \operatorname{Ker} PEQ = 0.$$

Hence, (10) again holds, verifying (PS3) and presolvability of $(E, A) \cdot (P, Q, D)$. □

If (E, A) has unit index, it turns out that the matrix D plays no essential role in establishing standard canonical form. This is made precise in the next theorem.

THEOREM 2.3. *If (E, A) has unit index, then there exists a unimodular $Q \in R^{n\times n}$ such that,* for every *$D \in R^{n\times n}$, there exists a unimodular $P \in R^{n\times n}$ which yields standard canonical form* (7) *with $N = 0$.*

Proof. Suppose (P_1, Q_1, D_1) achieves standard canonical form for some X_1 and with $N = 0$. Let $Q = Q_1$ and let D be given. Setting

$$\begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} = Q_1^{-1}(D - D_1), \quad P_2 = \begin{bmatrix} I & -D_{12} \\ 0 & I \end{bmatrix},$$

$X = X_1 + D_{11}$, and $P = P_2 P_1$ yields

$$PEQ = P_2(P_1 E Q_1) = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix},$$

$$\begin{aligned} P(AQ + ED) &= P_2\left(P_1(AQ_1 + ED_1) + (P_1 E Q_1) Q_1^{-1}(D - D_1)\right) \\ &= \begin{bmatrix} I & -D_{12} \\ 0 & I \end{bmatrix} \left(\begin{bmatrix} X_1 & 0 \\ 0 & I \end{bmatrix} + \begin{bmatrix} D_{11} & D_{12} \\ 0 & 0 \end{bmatrix} \right) \\ &= \begin{bmatrix} X & 0 \\ 0 & I \end{bmatrix}. \end{aligned}$$ □

For an arbitrary commutative ring R, we can establish necessary conditions under which (E, A) has unit index. First we need a lemma.

LEMMA 2.4. *Let $M \in R^{n\times n}$. If there exist unimodular $P, Q \in R^{n\times n}$ such that*

$$PMQ = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \tag{11}$$

then $\operatorname{rank} M = \rho M$.

Proof. Suppose the identity matrix in (11) is $n_1 \times n_1$. Then it is clear by definitions (4) and (5) that

$$\operatorname{rank} PMQ = n_1 = \rho PMQ.$$

The result follows from invariance of rank and ρ under unimodular transformations. □

THEOREM 2.5. *If (E, A) has unit index, then*

(1) $\operatorname{rank} E = \rho E$,

(2) $\operatorname{Im} E + A \operatorname{Ker} E = R^n$,

(3) (E, A) *is presolvable.*

Proof. (1) This follows from standard canonical form and Lemma 2.4.

(2) Invoking standard canonical form,

$$\begin{aligned} P\left(\operatorname{Im} E + A \operatorname{Ker} E\right) &= P\left(\operatorname{Im} E + A \operatorname{Ker} E + EDQ^{-1} \operatorname{Ker} E\right) \\ &= \operatorname{Im} PEQ + PAQ \operatorname{Ker} PEQ + PED \operatorname{Ker} PEQ \\ &\supset \operatorname{Im} PEQ + P\left(AQ + ED\right) \operatorname{Ker} PEQ \\ &= \operatorname{Im} \begin{bmatrix} I \\ 0 \end{bmatrix} + \begin{bmatrix} X & 0 \\ 0 & I \end{bmatrix} \operatorname{Im} \begin{bmatrix} 0 \\ I \end{bmatrix} \\ &= \operatorname{Im} \begin{bmatrix} I \\ 0 \end{bmatrix} + \operatorname{Im} \begin{bmatrix} 0 \\ I \end{bmatrix} \\ &= R^n. \end{aligned}$$

(3) This is obvious from part (2). □

Let $B \in R^{n\times m}$. The group action (6) may be extended to triples (E, A, B) according to

$$(E, A, B) \cdot (P, Q, D) = \left(PEQ, P\left(AQ + ED\right), PB\right). \tag{12}$$

In [16] we introduced the concept of "impulse controllability," which is fundamental to the study of state feedback in singular systems. We can adapt this idea to the algebraic setting by taking its feedback characterization as the definition. We say that $K \in R^{m\times n}$ is *impulse eliminating* if $(E, A + BK)$ has unit index. The triple (E, A, B) is *impulse controllable* if there exists an impulse eliminating K.

THEOREM 2.6. *Impulse controllability is invariant under* (12).

Proof. Suppose (E, A, B) is impulse controllable, and let K be impulse eliminating. Choose any P, Q, D, and let $K_1 = KQ$. Then

$$\left(PEQ, P\left(AQ + ED\right) + \left(PB\right) K_1\right) = \left(PEQ, P\left(\left(A + BK\right) Q + ED\right)\right),$$

which lies in the same orbit as $(E, A + BK)$ and, hence, has unit index. Thus $\left(PEQ, P\left(AQ + ED\right), PB\right)$ is impulse controllable. □

THEOREM 2.7. *If (E, A, B) is impulse controllable, then*

(1) $\operatorname{rank} E = \rho E$,

(2) $\operatorname{Im} E + A \operatorname{Ker} E + \operatorname{Im} B = R^n$.

Proof. Suppose $(E, A+BK)$ has unit index. From Theorem 2.5, part (1), $\operatorname{rank} E = \rho E$. By Theorem 2.5, part (2),

$$\begin{aligned}\operatorname{Im} E + A \operatorname{Ker} E + \operatorname{Im} B &\supset \operatorname{Im} E + A \operatorname{Ker} E + BK \operatorname{Ker} E \\ &\supset \operatorname{Im} E + (A+BK) \operatorname{Ker} E = R^n. \qquad \square\end{aligned}$$

We conclude this section by proving a pair of lemmas which will be useful in what follows, and which hold for any commutative ring.

LEMMA 2.8. *Let $M \in R^{p\times q}$. The following statements are equivalent:*

(1) $\operatorname{Im} M = R^p$,

(2) *M has a right inverse,*

(3) $\rho M = p$.

Proof. (1)⇒(2) Let $e_1, \dots, e_p$ be the canonical unit vectors in R^p. Since $\operatorname{Im} M = R^p$, there exist $x_1, \dots, x_p \in R^q$ such that $Mx_i = e_i$. Let $L = \begin{bmatrix} x_1 & \cdots & x_p \end{bmatrix}$. Then $MLe_i = Mx_i = e_i$, so $ML = I$.

(2)⇒(3) Suppose $ML = I$. From the Binet–Cauchy formula,

$$\sum_{1\le j_1<\cdots<j_p\le q} M \begin{pmatrix} 1 & \cdots & p \\ j_1 & \cdots & j_p \end{pmatrix} L \begin{pmatrix} j_1 & \cdots & j_p \\ 1 & \cdots & p \end{pmatrix} = \det I = 1,$$

so $\rho M = p$.

(3)⇒(1) There exist $x_{j_1\cdots j_p} \in R$ such that

$$\sum_{1\le j_1<\cdots<j_p\le q} x_{j_1\cdots j_p} M \begin{pmatrix} 1 & \cdots & p \\ j_1 & \cdots & j_p \end{pmatrix} = 1. \tag{13}$$

Traversing the ith row and expanding by minors yields

$$M \begin{pmatrix} 1 & \cdots & p \\ j_1 & \cdots & j_p \end{pmatrix} = \sum_{l=1}^{p} (-1)^{i+j_l} m_{ij_l} M \begin{pmatrix} 1 & \cdots & i-1 & i+1 & \cdots & p \\ j_1 & \cdots & j_{l-1} & j_{l+1} & \cdots & j_p \end{pmatrix}, \tag{14}$$

where $M = [m_{ij}]$. Combining (13) and (14), we obtain $y_{ij} \in R$ such that $\sum_j y_{ij} m_{ij} = 1$. Let $k \ne i$ and replace the ith row of M with the kth row. This yields the calculation

$$\begin{aligned}\sum_{l=1}^{p} (-1)^{i+j_l} m_{kj_l} M \begin{pmatrix} 1 & \cdots & i-1 & i+1 & \cdots & p \\ j_1 & \cdots & j_{l-1} & j_{l+1} & \cdots & j_p \end{pmatrix} \\ = M \begin{pmatrix} 1 & \cdots & i-1 & k & i+1 & \cdots & p \\ & j_1 & & \cdots & & j_p & \end{pmatrix} = 0.\end{aligned}$$

Hence, $\sum_j y_{ij} m_{kj} = 0$. Let

$$y_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{iq} \end{bmatrix}.$$

Then My_i is equal to the ith unit vector e_i. Let $x \in R^p$ and

$$z = \begin{bmatrix} y_1 & \cdots & y_p \end{bmatrix} x.$$

Then

$$Mz = \begin{bmatrix} My_1 & \cdots & My_p \end{bmatrix} x = \begin{bmatrix} e_1 & \cdots & e_p \end{bmatrix} x = x.$$

Since x is arbitrary, $\operatorname{Im} M = R^p$. □

LEMMA 2.9. *Let*

$$E = \begin{bmatrix} E_{11} & 0 \\ 0 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix},$$

where $E_{11} \in R^{n_1 \times n_1}$, $A_{ij} \in R^{n_i \times n_j}$, *and* $B_i \in R^{n_i \times m}$.

(1) (E, A) *has unit index iff* E_{11} *and* A_{22} *are unimodular.*

(2) $\rho E = n_1$ *and* $\operatorname{Im} E + A \operatorname{Ker} E + \operatorname{Im} B = R^n$ *iff* E_{11} *is unimodular and* $\rho \begin{bmatrix} A_{22} & B_2 \end{bmatrix} = n_2$.

Proof. (1) (Necessary) From Theorem 2.3, there exist unimodular P and Q so that (PEQ, PAQ) is in standard canonical form with $N = 0$. Let

$$\begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} = P, \quad \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} = Q^{-1}.$$

Then

$$PE = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} Q^{-1}$$

implies $Q_{12} = 0$, so Q_{11} and Q_{22} are unimodular. Also, $P_{11}E_{11} = Q_{11}$ and $P_{21}E_{11} = 0$, so E_{11} is unimodular and $P_{21} = 0$. It follows from

$$PA = \begin{bmatrix} X & 0 \\ 0 & I \end{bmatrix} Q^{-1}$$

that $P_{22}A_{22} = Q_{22}$, so A_{22} is unimodular.

(Sufficient) Let

$$P = \begin{bmatrix} E_{11}^{-1} & -E_{11}^{-1}A_{12}A_{22}^{-1} \\ 0 & A_{22}^{-1} \end{bmatrix}, \quad Q = \begin{bmatrix} I & 0 \\ -A_{22}^{-1}A_{21} & I \end{bmatrix},$$

and $D = 0$. Then

$$PEQ = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \quad P(AQ + ED) = \begin{bmatrix} E_{11}^{-1}\left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right) & 0 \\ 0 & I \end{bmatrix}.$$

(2) (Necessary) Unimodularity of E_{11} follows from the definition of ρ. Thus

$$\operatorname{Ker} E = \operatorname{Im} \begin{bmatrix} 0 \\ I \end{bmatrix}, \quad \operatorname{Im} \begin{bmatrix} E_{11} & A_{12} & B_1 \\ 0 & A_{22} & B_2 \end{bmatrix} = \operatorname{Im} E + A \operatorname{Ker} E + \operatorname{Im} B = R^n.$$

For any $w \in R^{n_2}$, there exist x, y, z such that

$$\begin{bmatrix} 0 \\ w \end{bmatrix} = \begin{bmatrix} E_{11} & A_{12} & B_1 \\ 0 & A_{22} & B_2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix},$$

so

$$w = \begin{bmatrix} A_{22} & B_2 \end{bmatrix} \begin{bmatrix} y \\ z \end{bmatrix}$$

and $\operatorname{Im} \begin{bmatrix} A_{22} & B_2 \end{bmatrix} = R^{n_2}$. The result follows from Lemma 2.8.

(Sufficient) The definition of ρ gives $\rho E = n_1$. Let $v \in R^{n_1}$ and $w \in R^{n_2}$. Then there exist y and z such that $A_{22}y + B_2 z = w$. Set $x = E_{11}^{-1}(v - A_{12}y - B_1 z)$. Then

$$\begin{bmatrix} E_{11} & A_{12} & B_1 \\ 0 & A_{22} & B_2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} v \\ w \end{bmatrix},$$

so

$$\operatorname{Im} E + A \operatorname{Ker} E + \operatorname{Im} B = \operatorname{Im} \begin{bmatrix} E_{11} & A_{12} & B_1 \\ 0 & A_{22} & B_2 \end{bmatrix} = R^n. \qquad \square$$

3. Pencils over a Hermite domain. We say R is a *Hermite domain* if it is an integral domain and, for every $a, b \in R$, there exist $u, v, x, y \in R$ such that $ux + vy = 1$ and $ax + by = 0$ [2, p. 469]. It should be noted that the definition of a Hermite domain varies in the literature. For example, [6, p. 345] gives a definition which is different from, but is implied by, the one given in [2]. In particular, every Bezout domain is Hermite [2, Thm. 3.2], and, therefore, every principal ideal domain, field, etc. is also a Hermite domain. For the remainder of this section, our standing assumption is that R is a Hermite domain (as in [2]).

One advantage of working in a Hermite domain is that matrices over R can be triangularized: For any $M \in R^{p\times q}$ $(p \neq q)$, there exists a lower triangular $L \in R^{\min\{p,q\}\times\min\{p,q\}}$ and a unimodular $Q \in R^{q\times q}$ such that

$$MQ = \begin{cases} \begin{bmatrix} L & 0 \end{bmatrix}, & p < q, \\ \begin{bmatrix} L \\ 0 \end{bmatrix}, & p > q. \end{cases}$$

A similar result, in which $\operatorname{Ker} M$ plays a special role, was established for real analytic functions in [5]. The arguments used in [5] are essentially algebraic and can be adapted to any Hermite domain. Since these ideas are central to our results, we develop the underlying algebraic arguments in detail, culminating in Theorem 3.3 and its corollary.

LEMMA 3.1. *Let $M \in R^{2\times 2}$ with at least one first-row entry nonzero. There exists a unimodular $Q \in R^{2\times 2}$ such that MQ is lower triangular with its $1,1$ entry nonzero.*

Proof. Let $\begin{bmatrix} a & b \end{bmatrix}$ be the first row of M and choose $u, v, x, y \in R$ such that $ux + vy = 1$ and $ax + by = 0$. Let

$$Q = \begin{bmatrix} v & x \\ -u & y \end{bmatrix}.$$

Then MQ is lower triangular and $\det Q = 1$, so Q is unimodular. The first row of MQ is $\begin{bmatrix} a & b \end{bmatrix} Q \neq 0$, but the $1,2$ entry of MQ is zero, so its $1,1$ entry must be nonzero. $\square$

LEMMA 3.2. *Let $M \in R^{p\times q}$ with at least one first-row entry nonzero. There exists a unimodular $Q \in R^{q\times q}$ such that MQ has the form*

$$MQ = \begin{bmatrix} a & 0 \\ b & C \end{bmatrix} \tag{15}$$

with $a \neq 0$.

Proof. Q will be constructed as a series of column permutations and transformations of the form

$$\begin{bmatrix} v & & x & \\ & I & & \\ -u & & y & \\ & & & I \end{bmatrix},$$

where u, v, x, and y are as in the proof of Lemma 3.1. The product of such transformations is unimodular.

Begin operating on M by permuting its columns so that either the $1,1$ or $1,2$ entry is nonzero. Applying Lemma 3.1 to the upper left 2×2 submatrix yields a matrix of the form

$$\begin{bmatrix} d & 0 & e \\ f & g & h \\ j & k & L \end{bmatrix},$$

where $d, f, g \in R$ and $d \neq 0$. The $1,3$ entry may be brought to zero by applying Lemma 3.1 to the 2×2 submatrix formed from the first two rows and the first and third columns. Proceeding inductively across the first row, we achieve the form (15) with $a \neq 0$. □

THEOREM 3.3. *Let $M \in R^{p\times q}$. If $\operatorname{rank} M = k > 0$, then there exist $L \in R^{p\times k}$ with $\operatorname{rank} L = k$ and a unimodular $Q \in R^{q\times q}$ such that*

$$MQ = \begin{bmatrix} L & 0 \end{bmatrix}. \tag{16}$$

Proof. Although we will make use of row permutations in achieving our result, these may be reversed at the end without disturbing the form (16). Since $M \neq 0$, there exists a row permutation that places a nonzero entry in the first row. Applying Lemma 3.2, we achieve the form (15) with $a \neq 0$. Suppose $\operatorname{rank} C \geq k$. Then C has a kth-order minor $\mu \neq 0$, and $a\mu$ is a $(k+1)$th-order minor of MQ. Since R is an integral domain, $a\mu \neq 0$, which contradicts $\operatorname{rank} M = k$. Thus $\operatorname{rank} C \leq k-1$.

If $k > 1$, the same arguments may then be applied to C, yielding a matrix of the form

$$\begin{bmatrix} a & 0 & 0 \\ d & e & 0 \\ f & g & H \end{bmatrix},$$

where $e \in R - \{0\}$ and $\operatorname{rank} H \leq k-2$. Proceeding inductively, we eventually achieve (16) with k nonzero columns. Since Q is unimodular, $\operatorname{rank} L = \operatorname{rank} M = k$. □

COROLLARY 3.4. *Let $M \in R^{p\times q}$.*

(1) *If $\rho M = p$, then there exists a unimodular Q such that*

$$MQ = \begin{bmatrix} I & 0 \end{bmatrix}.$$

(2) *If $\operatorname{rank} M = k$, then there exist $L \in R^{k\times k}$ with $\operatorname{rank} L = k$ and unimodular P and Q such that*

$$PMQ = \begin{bmatrix} L & 0 \\ 0 & 0 \end{bmatrix}.$$

(3) *If* $\operatorname{rank} M = \rho M$, *then there exist unimodular* P *and* Q *such that*

$$PMQ = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}.$$

(4) *If* $\operatorname{rank} M = \rho M = p$, *then there exists* $L \in R^{(q-p)\times q}$ *such that* $\begin{bmatrix} M \\ L \end{bmatrix}$ *is unimodular.*

Proof. (1) From Theorem 3.3, there exists Q_1 and L such that

$$MQ_1 = \begin{bmatrix} L & 0 \end{bmatrix}.$$

But $\rho L = \rho M = p$, so L is unimodular. Let

$$Q = Q_1 \begin{bmatrix} L^{-1} & 0 \\ 0 & I \end{bmatrix}.$$

(2) From Theorem 3.3, there exist $L_1 \in R^{k\times k}$and $L_2 \in R^{(p-k)\times k}$ with

$$\operatorname{rank} \begin{bmatrix} L_1 \\ L_2 \end{bmatrix} = k$$

and a unimodular Q such that

$$MQ = \begin{bmatrix} L_1 & 0 \\ L_2 & 0 \end{bmatrix}.$$

Also, there exist $L \in R^{k\times k}$ with $\operatorname{rank} L = k$ and a unimodular P such that

$$\begin{bmatrix} L_1^T & L_2^T \end{bmatrix} P^T = \begin{bmatrix} L^T & 0 \end{bmatrix}.$$

Hence,

$$PMQ = \left((MQ)^T P^T\right)^T = \left(\begin{bmatrix} L_1^T & L_2^T \\ 0 & 0 \end{bmatrix} P^T\right)^T = \begin{bmatrix} L^T & 0 \\ 0 & 0 \end{bmatrix}^T.$$

(3) Suppose $\operatorname{rank} M = k$. From part (2), there exist L, P_1, Q such that

$$P_1 MQ = \begin{bmatrix} L & 0 \\ 0 & 0 \end{bmatrix},$$

where $L \in R^{k\times k}$. Since $\rho M = k$, L is unimodular. Let

$$P = \begin{bmatrix} L^{-1} & 0 \\ 0 & I \end{bmatrix} P_1.$$

(4) From part (1), there exists a unimodular Q such that

$$MQ = \begin{bmatrix} I & 0 \end{bmatrix}.$$

Then

$$\begin{bmatrix} J \\ L \end{bmatrix} = Q^{-1}$$

is unimodular, and

$$J = \begin{bmatrix} I & 0 \end{bmatrix} Q^{-1} = M. \qquad \square$$

Corollary 3.4, part (4) is contained in Lemma 59, p. 345 of [6]. However, our proof is more directly applicable to our development.

Another advantage of working in an integral domain is that, if $M \in R^{p\times p}$, $x \in R^p$, and $Mx = 0$, then either $x = 0$ or $\det M = 0$, since

$$(\det M)\, x = (\operatorname{adj} M)\, Mx = 0.$$

We will make frequent use of this fact in developing our main results.

The next result is complementary to Theorem 2.5, part (2).

THEOREM 3.5. *If* $\operatorname{Im} E + A \operatorname{Ker} E = R^n$, *then* (E, A) *has unit index.*

Proof. If $E = 0$, then $\operatorname{Im} A = R^n$. From Lemma 2.8, A is unimodular. Then the standard canonical form with $N = 0$ is achieved by letting $P = A^{-1}$, $Q = I$, and $D = 0$. If $E \neq 0$, we apply Corollary 3.4, part (2) to obtain

$$PEQ = \begin{bmatrix} E_{11} & 0 \\ 0 & 0 \end{bmatrix}, \quad PAQ = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

with $\det E_{11} \neq 0$. Let

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x \in R^n.$$

Since R is an integral domain, $PEQx = 0$ implies $x_1 = 0$, so

$$\operatorname{Ker} PEQ = \operatorname{Im} \begin{bmatrix} 0 \\ I \end{bmatrix},$$

$$\operatorname{Im} \begin{bmatrix} E_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} = \operatorname{Im} PEQ + PAQ \operatorname{Ker} PEQ = P(\operatorname{Im} E + A \operatorname{Ker} E) = R^n.$$

From Lemma 2.8, E_{11} and A_{22} are unimodular. From Lemma 2.9, part (1), (E, A) has a unit index. $\square$

The next theorem, complementary to Theorem 2.7, is our main result.

THEOREM 3.6. *If*

(1) $\operatorname{rank} E = \rho E$,

(2) $\operatorname{Im} E + A \operatorname{Ker} E + \operatorname{Im} B = R^n$,

(3) (E, A) *is presolvable,*

then (E, A, B) *is impulse controllable.*

Proof. Presolvability of (E, A) admits three cases. If (PS1) holds, (E, A) has unit index from Theorem 3.5. Setting $K = 0$, $(E, A + BK)$ has unit index and (E, A, B) is impulse controllable. To analyze the remaining cases, we invoke Corollary 3.4, part (3). Let

$$\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} = PEQ, \quad \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = PAQ, \quad \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = PB,$$

where partitioning conforms to $n = n_1 + n_2$. If (PS2) holds,

$$\operatorname{Im} \begin{bmatrix} I \\ 0 \end{bmatrix} \cap \operatorname{Im} \begin{bmatrix} A_{12} \\ A_{22} \end{bmatrix} = \operatorname{Im} PEQ \cap PAQ \operatorname{Ker} PEQ = P(\operatorname{Im} E \cap A \operatorname{Ker} E) \neq 0,$$

so there exist x and y such that

$$\begin{bmatrix} y \\ 0 \end{bmatrix} = \begin{bmatrix} A_{12} \\ A_{22} \end{bmatrix} x \neq 0.$$

Hence, $x \neq 0$ and $A_{22}x = 0$. Since R is an integral domain, $\det A_{22} = 0$. Similarly, if (PS3) holds,

$$\operatorname{Ker} \begin{bmatrix} A_{12} \\ A_{22} \end{bmatrix} = \operatorname{Ker} PEQ \cap \operatorname{Ker} PAQ = Q^{-1} (\operatorname{Ker} E \cap \operatorname{Ker} A) \neq 0,$$

so there exists $x \neq 0$ such that

$$\begin{bmatrix} A_{12} \\ A_{22} \end{bmatrix} x = 0.$$

Hence, $A_{22}x = 0$ and $\det A_{22} = 0$. In either case, we need consider only singular A_{22}.

Note that

$$\begin{aligned} \operatorname{Im} \begin{bmatrix} I & A_{12} & B_1 \\ 0 & A_{22} & B_2 \end{bmatrix} &= \operatorname{Im} PEQ + PAQ \operatorname{Ker} PEQ + \operatorname{Im} PB \\ &= P (\operatorname{Im} E + A \operatorname{Ker} E + \operatorname{Im} B) \\ &= R^n, \end{aligned}$$

so $\operatorname{Im} \begin{bmatrix} A_{22} & B_2 \end{bmatrix} = R^{n_2}$. Let $r = \operatorname{rank} A_{22}$. From Corollary 3.4, part (2), there exist P_1 and Q_1 such that

$$P_1 A_{22} Q_1 = \begin{bmatrix} \widehat{A} & 0 \\ 0 & 0 \end{bmatrix},$$

where $\widehat{A} \in R^{r \times r}$. Let

$$\begin{bmatrix} \overline{B} \\ \overline{C} \end{bmatrix} = P_1 B_2.$$

Then

$$\operatorname{Im} \begin{bmatrix} \widehat{A} & \overline{B} \\ 0 & \overline{C} \end{bmatrix} = \operatorname{Im} \begin{bmatrix} P_1 A_{22} Q_1 & P_1 B_2 \end{bmatrix} = P_1 \operatorname{Im} \begin{bmatrix} A_{22} & B_2 \end{bmatrix} = R^{n_2}$$

and $\operatorname{Im} \overline{C} = R^{n_2 - r}$. From Corollary 3.4, part (1) and Lemma 2.8, there exists a unimodular Q_2 such that

$$\overline{C} Q_2 = \begin{bmatrix} I & 0 \end{bmatrix}.$$

Let

$$\begin{bmatrix} \widetilde{B} & \widehat{B} \end{bmatrix} = \overline{B} Q_2$$

and

$$P_2 = \begin{bmatrix} I & -\widetilde{B} \\ 0 & I \end{bmatrix}.$$

Then

$$\operatorname{Im}\begin{bmatrix} \widehat{A} & 0 & \widehat{B} \\ 0 & I & 0 \end{bmatrix} = \operatorname{Im}\left(P_2 \begin{bmatrix} \widehat{A} & \overline{B} \\ 0 & \overline{C} \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & Q_2 \end{bmatrix} \right) = P_2 \operatorname{Im}\begin{bmatrix} \widehat{A} & \overline{B} \\ 0 & \overline{C} \end{bmatrix} = R^{n_2},$$

and $\operatorname{Im}[\widehat{A} \quad \widehat{B}] = R^r$, so Lemma 2.8 guarantees the existence of a right inverse. From Corollary 3.4, part (4), there are W and Y such that

$$U = \begin{bmatrix} \widehat{A} & \widehat{B} \\ W & Y \end{bmatrix}$$

is unimodular. Let $K_1 \in R^{m \times n_1}$ be arbitrary,

$$K_2 = Q_2 \begin{bmatrix} W & Y \\ 0 & I \end{bmatrix} Q_1^{-1}, \quad K = \begin{bmatrix} K_1 & K_2 \end{bmatrix} Q^{-1}.$$

Then

$$\begin{aligned} P_2 P_1 \left(A_{22} + B_2 K_2 \right) Q_1 &= P_2 P_1 A_{22} Q_1 + \left(P_2 P_1 B_2 Q_2 \right) \left(Q_2^{-1} K_2 Q_1 \right) \\ &= \begin{bmatrix} \widehat{A} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & \widehat{B} \\ I & 0 \end{bmatrix} \begin{bmatrix} W & Y \\ 0 & I \end{bmatrix} \\ &= U, \end{aligned}$$

so $A_{22} + B_2 K_2$ is unimodular. From Lemma 2.9, part (1),

$$\left(PEQ, P\left(A + BK \right) Q \right) = \left(\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} + B_1 K_1 & A_{12} + B_1 K_2 \\ A_{21} + B_2 K_1 & A_{22} + B_2 K_2 \end{bmatrix} \right)$$

has a unit index, so $(E, A + BK)$ has a unit index and (E, A, B) is impulse controllable. □

Let $\mathcal{I}$ be the set of all impulse eliminating K. The arguments used in the proof of Theorem 3.6 can be generalized to construct a large subset of $\mathcal{I}$. We begin by fixing $P_1, P_2, Q, Q_1, Q_2, A_{22}, B_2, \widehat{A}, \widehat{B}$ as above. Then, for any K_1, W, Y, T, V with V and

$$U = \begin{bmatrix} \widehat{A} & \widehat{B} \\ W & Y \end{bmatrix}$$

unimodular, we set

$$K_2 = Q_2 \begin{bmatrix} W & Y \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ T & V \end{bmatrix} Q_1^{-1}.$$

It follows that

$$P_2 P_1 \left(A_{22} + B_2 K_2 \right) Q_1 = \begin{bmatrix} \widehat{A} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & \widehat{B} \\ I & 0 \end{bmatrix} \begin{bmatrix} W & Y \\ 0 & I \end{bmatrix} = U \begin{bmatrix} I & 0 \\ T & V \end{bmatrix}$$

is unimodular. Setting $K = \begin{bmatrix} K_1 & K_2 \end{bmatrix} Q^{-1}$ guarantees that $(E, A + BK)$ has unit index.

We note that the map $\pi\left(K_1, W, Y, T, V \right) = K$ is one-to-one. Indeed, if we choose K in the range of π, then K_1 is uniquely determined, and setting $L = Q_2^{-1} K_2 Q_1$ yields

$$\begin{bmatrix} W & Y \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ T & V \end{bmatrix} = \begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix},$$

so

$$T = L_{21}, \quad V = L_{22}, \quad Y = L_{12}L_{22}^{-1}, \quad W = L_{11} - L_{12}L_{22}^{-1}L_{21}.$$

Hence, π may be considered a parametrization of the set of all impulse eliminating K with unimodular V (i.e., the 2,2 block of $Q_2^{-1}K_2Q_1$). Unfortunately, this may not be a complete parametrization of $\mathcal{I}$, as the example

$$E = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$$

illustrates. Here, direct calculation shows that $\mathcal{I}$ consists of all matrices of the form

$$K = \begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix}$$

with k_{12} a unit. However, π only yields those matrices of the form

$$K = \begin{bmatrix} W + YT & YV \\ T & V \end{bmatrix}$$

with V and Y units. Although π does predict that $k_{12} = YV$ must be a unit, it does not allow $k_{22} = V$ to be a nonunit, in spite the admissibility of such values. Hence, the range of π is a proper subset of $\mathcal{I}$.

4. Applications to time-varying singular systems. In this section, we consider time-varying differential equations

$$E(t)\,\dot{x} = A(t)\,x + B(t)\,u, \tag{17}$$

where the entries of E, A, and B belong to a ring of real-valued functions on $\mathbb{R}$. We assume $E(t), A(t) \in \mathbb{R}^{n\times n}$ and $B(t) \in \mathbb{R}^{n\times m}$. The interesting case occurs when $E(t)$ is singular on a subset of $\mathbb{R}$. Such systems have been studied at length under the assumption that E, A, and B are either constant [7] or real analytic [3], [4]. We will show that these cases fit into our algebraic framework and examine certain additional classes of functions that can be treated in our setting. Our work does not apply to problems where E, A, B, and K are allowed to have arbitrary entries in C^n (as in [17] and [18]), since C^n is not Hermite.

In studying (17), it is useful to consider a change of variables of the form $x = Q(t)\,z$, where $Q(t)$ is everywhere nonsingular and where both Q and Q^{-1} belong to a given class of functions. Assuming differentiability of Q, direct substitution yields the equivalent system

$$P(t)\,E(t)\,Q(t)\,\dot{z} = P(t)\left(A(t)\,Q(t) - E(t)\,\dot{Q}(t)\right)z + P(t)\,B(t)\,u, \tag{18}$$

where $P(t)$ is also nonsingular for every t. (Note the relationship of (18) to the group action (12).)

Another important consideration in working with any kind of differential equation is that of solvability. Roughly, this means that (17) exhibits existence and uniqueness of solutions over a large class of forcing functions u. In the case of equations based on matrices over the real analytic functions $\mathcal{A}(\mathbb{R})$, Campbell and Petzold [3] define (E, A) to be *analytically solvable* if, for every C^n function u, the system

$$E(t)\,\dot{x} = A(t)\,x + u \tag{19}$$

has at least one C^1 solution x on $\mathbb{R}$ and no two distinct solutions coincide for any t. They then proceed to show that analytic solvability is equivalent to the existence of analytic nonsingular matrices P and Q that put (18) into standard canonical form. Hence, analytic solvability is equivalent to algebraic solvability.

In the time-invariant setting, analytic solvability of (17) reduces to the condition that the matrix pencil (E, A) be *regular*, i.e.,

$$\det(sE - A) \not\equiv 0. \tag{20}$$

(See [8, pp. 45–49].) From [8, Thm. 3, p. 28], (20) is equivalent to the existence of nonsingular $P, Q \in \mathbb{R}^{n\times n}$ that put the pencil into *Weierstrass canonical form*:

$$PEQ = \begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix}, \quad PAQ = \begin{bmatrix} X & 0 \\ 0 & I \end{bmatrix}, \tag{21}$$

where N is nilpotent. Since $\dot{Q} = 0$, (21) and (7) are the same, so (20) is equivalent to algebraic solvability.

In addition to solvability, we note that the unit index property is a natural concept in both the constant and real analytic settings, occurring iff $N \equiv 0$.

In order to study the impulsive behavior of singular systems, we must adopt a more sophisticated viewpoint based on distribution theory. In (19) we may investigate the consequences of applying an input u, which is arbitrary C^1 up to time $t = t_0$ and drops abruptly to 0 at t_0. As discussed in [15, Ch. 22], the resulting solution exists as a distribution and is, in fact, the unique distribution x satisfying $x(t) = 0$ for $t < t_0$ and

$$E(t)\,\dot{x} = A(t)\,x + \delta_{t_0} E(t_0)\,x_0, \tag{22}$$

where δ_{t_0} is the unit impulse and $x_0 = \lim_{t\to t_0^-} x(t)$. Equation (22) gives a precise meaning to the natural response of (17) with arbitrary initial conditions.

Our principal objective is to find a matrix $K(t)$, whose entries reside in the *same ring of functions as the entries of* E, A, *and* B, and such that the state feedback law $u = K(t)\,x$ yields a unit index closed-loop system

$$E(t)\,\dot{x} = (A(t) + B(t)\,K(t))\,x + \delta_{t_0} E(t_0)\,x_0. \tag{23}$$

Thus we are simultaneously treating a wide variety of constrained feedback problems, which have not been considered in the literature.

In order to apply our results to (17), we first need to identify a function ring R that satisfies the conditions that (1) R is an Hermite domain, (2) R is closed under differentiation, (3) solvability in the classical sense implies presolvability, and (4) the analytic and algebraic notions of the unit index property coincide. Note that it follows from (4) that the analytic and algebraic notions of impulse controllability must also coincide. Once these conditions are established, we are guaranteed that the results of sections 2 and 3 apply to systems over R. In particular, Theorems 2.7 and 3.6 give necessary and sufficient algebraic conditions under which (17) is impulse controllable. It remains only to translate conditions (1) and (2) from Theorems 2.7 and 3.6 into analytic terms.

For the remainder of this paper, we restrict ourselves to subrings R (with identity) of $\mathcal{A}(\mathbb{R})$. Properties (1) and (2) will have to be established case by case. On the other hand, (3) and (4) hold automatically for $\mathcal{A}(\mathbb{R})$ as a consequence of previous results.

Indeed, condition (3) may be established by examining the proof of Theorem 2 in [3]. In light of our Theorem 3.3 and its corollary, the arguments used by Campbell and Petzold carry over verbatim to R, demonstrating that analytic solvability of (E, A) guarantees algebraic solvability and, therefore, presolvability. To establish (4), suppose (E, A) is analytically (and algebraically) solvable. If $N \equiv 0$, then (E, A) has unit index in the algebraic sense with $D = -\dot{Q}$. Conversely, suppose (E, A) has an algebraic unit index. Then, from Theorem 2.3, we may choose Q such that setting $D = -\dot{Q}$ yields P that achieves (21) with $N = 0$. Hence, the two notions of unit index coincide. This establishes that our algebraic theory applies to any Hermite subring of $\mathcal{A}(\mathbb{R})$ which is closed under differentiation.

Time-invariant systems. To treat time-invariant systems

$$E\dot{x} = Ax + Bu,$$

set $R = \mathbb{R}$. Since $\mathbb{R}$ is a field, it is Hermite. Viewing $\mathbb{R}$ as the set of constant functions, it is closed under differentiation. We therefore conclude that Theorems 2.7 and 3.6 specialize to the characterization of time-invariant impulse controllability first established in [16]. The proofs of Theorems 2.7 and 3.6 thus constitute an alternative to the known proofs of this result, as presented in [10, Thm. 6], [12, Thm. 2.5.1], and [13, Thm. 3-2.1].

General analytic systems. For $R = \mathcal{A}(\mathbb{R})$, [5, Lem. 1] shows that $\mathcal{A}(\mathbb{R})$ is Hermite. (In fact, it is shown in [11, Thm. 1.19], that $\mathcal{A}(\mathbb{R})$ is a Bezout domain.) R is closed under differentiation, so conditions (1) and (2) of Theorems 2.7 and 3.6 are necessary and sufficient for impulse controllability. It remains to link the algebraic conditions to analytic conditions on $E(t)$, $A(t)$, and $B(t)$.

THEOREM 4.1. *Conditions* (1) *and* (2) *of Theorems* 2.7 *and* 3.6 *hold for* $R = \mathcal{A}(\mathbb{R})$ *iff* $\operatorname{rank}_{\mathbb{R}} E(t)$ *is constant and* $\operatorname{Im} E(t) + A(t) \operatorname{Ker} E(t) + \operatorname{Im} B(t) = \mathbb{R}^n$ *for every* $t \in \mathbb{R}$.

Proof. (Sufficient) Suppose $\operatorname{rank}_{\mathbb{R}} E(t) = k$. Then $\operatorname{rank} E = k$ and, from Corollary 3.4, part (2), there exist unimodular P and Q such that

$$PEQ = \begin{bmatrix} E_{11} & 0 \\ 0 & 0 \end{bmatrix},$$

where $E_{11} \in R^{k \times k}$ and $\operatorname{rank} E_{11} = k$. But $\operatorname{rank}_{\mathbb{R}} E_{11}(t)$ must also be constant, so E_{11} is unimodular. Let

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = PAQ, \quad \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = PB.$$

Then

$$\operatorname{Im} \begin{bmatrix} E_{11}(t) & A_{12}(t) & B_1(t) \\ 0 & A_{22}(t) & B_2(t) \end{bmatrix} = \operatorname{Im} E(t) + A(t) \operatorname{Ker} E(t) + \operatorname{Im} B(t) = \mathbb{R}^n$$

for every t, so $\operatorname{rank}_{\mathbb{R}} \begin{bmatrix} A_{22}(t) & B_2(t) \end{bmatrix} = n - k$. Let $\{\mu_i(t)\}$ be the $(n-k)$th-order minors of $\begin{bmatrix} A_{22}(t) & B_2(t) \end{bmatrix}$. Each μ_i is an analytic function and the μ_i have no common zero. Hence, $u = \sum \mu_1^2$ has no zero and is therefore a unit of R. Also,

$$\sum \left(\frac{\mu_i}{u} \right) \mu_i = 1,$$

so $\rho \left[\begin{array}{cc} A_{22} & B_2 \end{array} \right] = n-k$. From Corollary 3.4, part (1), there exists a unimodular Q_1 such that

$$\left[\begin{array}{cc} A_{22} & B_2 \end{array} \right] Q_1 = \left[\begin{array}{cc} I & 0 \end{array} \right].$$

If $x \in R^{n-k}$, then

$$\left[\begin{array}{cc} A_{22} & B_2 \end{array} \right] Q_1 \left[\begin{array}{c} x \\ 0 \end{array} \right] = x,$$

so $x \in \operatorname{Im} \left[\begin{array}{cc} A_{22} & B_2 \end{array} \right]$. But x is arbitrary, so $\operatorname{Im} \left[\begin{array}{cc} A_{22} & B_2 \end{array} \right] = R^{n-k}$. The theorem follows from Lemma 2.8 and Lemma 2.9, part (2).

(Necessary) From Corollary 3.4, part (3), there exist unimodular P and Q such that

$$P(t) E(t) Q(t) = \left[\begin{array}{cc} I & 0 \\ 0 & 0 \end{array} \right]$$

for every t. Hence, $\operatorname{rank}_{\mathbb{R}} E(t)$ is constant. Let $x \in \mathbb{R}^n$. Viewing x as a constant function, it follows from $\operatorname{Im} E + A \operatorname{Ker} E + \operatorname{Im} B = R^n$ that there exist $u \in R^m$ and $y, z \in R^n$ such that $Ez = 0$ and $Ey + Az + Bu = x$. But this means $E(t) z(t) = 0$ and $E(t) y(t) + A(t) z(t) + B(t) u(t) = x$ for every t, so $\operatorname{Im} E(t) + A(t) \operatorname{Ker} E(t) + \operatorname{Im} B(t) = \mathbb{R}^n$. □

Theorem 4.1 shows that Theorems 2.7 and 3.6 specialize to Theorem 4.1 of [4] for systems over the real analytic functions.

Now we apply our theory to classes of time-varying singular systems (17) which have not been previously studied.

Polynomial systems. Let $R = \mathbb{R}[t]$ be the polynomials on $\mathbb{R}$ with real coefficients. $\mathbb{R}[t]$ is a subring of $\mathcal{A}(\mathbb{R})$ containing 1 and a principal ideal domain, so it is Hermite. $\mathbb{R}[t]$ is closed under differentiation. Theorem 4.1 applies to $\mathbb{R}[t]$ without modification.

Periodic systems. Let $\mathcal{P}(\tau)$ be the analytic functions on $\mathbb{R}$ with period $\tau > 0$. (τ need not be the fundamental period.) $\mathcal{P}(\tau)$ is a subring of $\mathcal{A}(\mathbb{R})$ containing 1 and is closed under differentiation.

THEOREM 4.2. *$\mathcal{P}(\tau)$ is a Bezout domain.*

Proof. We need to show that every finitely generated ideal in $\mathcal{P}(\tau)$ is principal. It suffices to show that, for every $a, b \in \mathcal{P}(\tau)$, there exists $c \in \mathcal{P}(\tau)$ such that $cR = aR + bR$. In view of [9, Thm. 3.7, p. 78], a and b have finitely many zeros in any bounded interval. Let $\{z_1, \ldots, z_q\}$ be the common zeros of a and b in the interval $[0, \tau)$, counting multiplicities, and define

$$c(t) = \prod_k \left(e^{2\pi i \frac{t}{\tau}} - e^{2\pi i \frac{z_k}{\tau}} \right).$$

Then $c \in \mathcal{P}(\tau)$ with zeros $\{z_k\}$, and c is a common divisor of a and b. Let $\overline{a} = a/c$ and $\overline{b} = b/c$. If $x, y \in R$, then

$$ax + by = c\left(\overline{a}x + \overline{b}y\right) \in cR,$$

so $aR + bR \subset cR$. To prove the converse, note that $\overline{a}$ and $\overline{b}$ have no common zero, so $u = \overline{a}^2 + \overline{b}^2$ has no zero and is, therefore, a unit of R. For any $r \in R$, set $x = \overline{a}r/u$ and $y = \overline{b}r/u$. Then

$$cr = cr\frac{\overline{a}^2 + \overline{b}^2}{u} = ax + by \in aR + bR,$$

so $cR \subset aR + bR$. □

It follows from Theorem 4.2 that $\mathcal{P}(\tau)$ is a Hermite domain. It can be further shown that $\mathcal{P}(\tau)$ is a principal ideal domain. Theorem 4.1 applies to $\mathcal{P}(\tau)$ without modification.

Systems analytic at ∞. Let $\mathcal{A}_\infty(\mathbb{R})$ be the subring of $\mathcal{A}(\mathbb{R})$ consisting of all functions analytic at ∞. (x analytic at ∞ means that $x\left(\frac{1}{t}\right)$ is analytic at 0.) From the chain rule,

$$\dot{x}\left(\frac{1}{t}\right) = -t^2 \frac{d}{dt}\left(x\left(\frac{1}{t}\right)\right),$$

so $\mathcal{A}_\infty(\mathbb{R})$ is closed under differentiation.

THEOREM 4.3. *$\mathcal{A}_\infty(\mathbb{R})$ and $\mathcal{P}(\tau)$ are isomorphic.*

Proof. Let

$$\phi(t) = \begin{cases} \tan\left(\pi \frac{t}{\tau}\right), & t \neq \left(k+\frac{1}{2}\right)\tau, \\ \infty, & t = \left(k+\frac{1}{2}\right)\tau. \end{cases}$$

ϕ has period τ and is analytic, except for poles at $\left(k+\frac{1}{2}\right)\tau$. $1/\phi$ is analytic about $\left(k+\frac{1}{2}\right)\tau$, where it has a zero. For any $x \in \mathcal{A}_\infty(\mathbb{R})$, define $x_p(t) = x(\phi(t))$. Then x_p has period τ. Since $x\left(\frac{1}{t}\right)$ is analytic about 0, x_p is analytic about $\left(k+\frac{1}{2}\right)\tau$ and therefore on all of $\mathbb{R}$. Hence, the map $h : x \to x_p$ takes $\mathcal{A}_\infty(\mathbb{R})$ into $\mathcal{P}(\tau)$ and is obviously a ring homomorphism. Since the range of ϕ is $\mathbb{R}$, $x_p \equiv 0$ implies $x \equiv 0$, and h is $1-1$. Given any $x_p \in \mathcal{P}(\tau)$, $x(t) = x_p\left(\frac{\tau}{\pi}\arctan(t)\right)$ defines a function in $\mathcal{A}_\infty(\mathbb{R})$. But $\frac{\tau}{\pi}\arctan(\phi(t)) = t$, so $h(x) = x_p$ and h is onto. □

It follows from Theorems 4.2 and 4.3 that $\mathcal{A}_\infty(\mathbb{R})$ is a Hermite domain.

The conditions of Theorem 4.1 must be augmented to handle analyticity at ∞.

THEOREM 4.4. *Conditions* (1) *and* (2) *of Theorems* 2.7 *and* 3.6 *hold for $R = \mathcal{A}_\infty(\mathbb{R})$ iff*

$$\operatorname*{rank}_{\mathbb{R}} E(t) = \operatorname*{rank}_{\mathbb{R}} E(\infty),$$

$$\operatorname{Im} E(t) + A(t)\operatorname{Ker} E(t) + \operatorname{Im} B(t) = \operatorname{Im} E(\infty) + A(\infty)\operatorname{Ker} E(\infty) + \operatorname{Im} B(\infty) = \mathbb{R}^n$$

for every $t \in \mathbb{R}$.

Proof. (Sufficient) Suppose $\operatorname{rank}_{\mathbb{R}} E(t) = k$. As in the proof of Theorem 4.1, there exist unimodular P and Q such that

$$PEQ = \begin{bmatrix} E_{11} & 0 \\ 0 & 0 \end{bmatrix},$$

where $E_{11} \in R^{k\times k}$ and $\operatorname{rank} E_{11} = k$. But $\operatorname{rank}_{\mathbb{R}} E_{11}(t) = \operatorname{rank}_{\mathbb{R}} E_{11}(\infty) = k$, so E_{11} is unimodular. Then

$$\operatorname{Im}\begin{bmatrix} E_{11}(t) & A_{12}(t) & B_1(t) \\ 0 & A_{22}(t) & B_2(t) \end{bmatrix} = \operatorname{Im} E(t) + A(t)\operatorname{Ker} E(t) + \operatorname{Im} B(t) = \mathbb{R}^n$$

for every t (including $t = \infty$), so $\operatorname{rank}_{\mathbb{R}}\begin{bmatrix} A_{22}(t) & B_2(t) \end{bmatrix} = n - k$ and the minors $\{\mu_i\}$ have no common finite or infinite zero. The remainder of the sufficiency proof proceeds without modification.

(Necessary) From Corollary 3.4, part (3), there exist unimodular P and Q such that

$$P(t)E(t)Q(t) = P(\infty)E(\infty)Q(\infty) = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}.$$

Hence, $\operatorname{rank}_{\mathbb{R}} E(t) = \operatorname{rank}_{\mathbb{R}} E(\infty)$. Let $x \in \mathbb{R}^n$. Viewing x as a constant function, there exist $u \in R^m$ and $y, z \in R^n$ such that $Ez = 0$ and $Ey + Az + Bu = x$. But this means

$$E(t) z(t) = E(\infty) z(\infty) = 0,$$

$$E(t) y(t) + A(t) z(t) + B(t) u(t) = E(\infty) y(\infty) + A(\infty) z(\infty) + B(\infty) u(\infty) = x.$$

Since x is arbitrary, the theorem follows. □

Example. We close this section with a simple example illustrating how our results may be applied to periodic systems. Let

$$T(t) = \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix},$$

and note that T is unimodular over $\mathcal{P}(2\pi)$. Consider the singular system with

$$E = \begin{bmatrix} 0 & T \\ 0 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ I \end{bmatrix}.$$

(E, A, B) is already in standard canonical form, so it is analytically and algebraically solvable. A simple calculation shows that $u \equiv 0$ leads to

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -\delta_{t_0} x_{03} \\ -\delta_{t_0} x_{04} \\ 0 \\ 0 \end{bmatrix}.$$

We wish to find an *analytic periodic* state feedback matrix $K(t)$ to eliminate impulses in the closed-loop system.

Note that $\operatorname{rank}_{\mathbb{R}} E(t) = 2$,

$$\operatorname{Ker} E(t) = \operatorname{Im} \begin{bmatrix} I \\ 0 \end{bmatrix},$$

$$\operatorname{Im} E(t) + A(t) \operatorname{Ker} E(t) + \operatorname{Im} B(t) = \operatorname{Im} \begin{bmatrix} T(t) & I & 0 \\ 0 & 0 & I \end{bmatrix} = \mathbb{R}^4$$

for every t. Theorems 3.6 and 4.1 guarantee that (E, A, B) is impulse controllable.

As in the proof of Theorem 3.6, we obtain the unimodular matrices

$$P = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}, \quad Q = \begin{bmatrix} 0 & I \\ T^T & 0 \end{bmatrix}.$$

Then

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = PAQ = Q,$$

so $A_{22} = 0$, $r = 0$, and $P_1 = Q_1 = P_2 = Q_2 = I$. Let $K_1, V, Y \in (\mathcal{P}(2\pi))^{2\times 2}$ with V, Y unimodular, and apply the parametrization π, as described at the end of section 3. This yields the state feedback matrix

$$K = \begin{bmatrix} K_1 & YV \end{bmatrix} Q^{-1} = \begin{bmatrix} K_1 & YV \end{bmatrix} \begin{bmatrix} 0 & T \\ I & 0 \end{bmatrix} = \begin{bmatrix} YV & K_1 T \end{bmatrix} \in (\mathcal{P}(2\pi))^{2\times 4}$$

and the periodic closed-loop system

$$\begin{bmatrix} 0 & T(t) \\ 0 & 0 \end{bmatrix} \dot{x} = \begin{bmatrix} I & 0 \\ Y(t)V(t) & I + K_1(t)T(t) \end{bmatrix} x. \tag{24}$$

Our theory guarantees that (24) has a unit index. This can be verified directly by interchanging the block columns of (24) and applying Lemma 2.9, part (1).

5. Conclusion. Our work demonstrates that the solutions of the state feedback impulse elimination problem, as originally developed for the time-invariant and time-varying cases in [10] and [4], share a common algebraic basis. Once exposed, this structure lends itself naturally to numerous generalizations, requiring only a small amount of analytic effort to turn the problem into algebra. The rings discussed in this paper are only a few of the many possibilities. For example, it is easy to show that similar conclusions hold for the real analytic functions with an isolated singularity at ∞, those with a pole or removable singularity at ∞, those with a zero of order at least k at a fixed point in $\mathbb{R} \cup \{\infty\}$, rational functions with no pole in $\mathbb{R}$, etc. Perhaps the greatest challenge is to fully exploit our theory by proposing a Hermite domain which is not principal ideal domain, Bezout, etc. We leave this challenge for further research.

Acknowledgement. The author wishes to thank Nigel Boston for his many helpful suggestions during the course of this research.

REFERENCES

[1] B. R. McDonald, *Linear Algebra over Commutative Rings*, Marcel Dekker, New York, 1984.
[2] I. Kaplansky, *Elementary divisors and modules*, Trans. Amer. Math. Soc., 66 (1949), pp. 464–491.
[3] S. L. Campbell and L. R. Petzold, *Canonical forms and solvable singular systems of differential equations*, SIAM J. Algebraic Discrete Methods, 4 (1983), pp. 517–521.
[4] C.-J. Wang, *State feedback impulse elimination of linear time-varying singular systems*, Automatica, 32 (1996), pp. 133–136.
[5] L. M. Silverman and R. S. Bucy, *Generalizations of a theorem of Dolezal*, Math. Systems Theory, 4 (1970), pp. 334–339.
[6] M. Vidyasagar, *Control System Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985.
[7] F. L. Lewis, *A Survey of Linear Singular Systems*, Circuits Systems Signal Process., 5 (1986), pp. 3–36.
[8] F. R. Gantmacher, *The Theory of Matrices*, Vol. II, Chelsea, New York, 1960.
[9] J. B. Conway, *Functions of One Complex Variable*, 2nd ed., Springer-Verlag, New York, 1978.
[10] J. D. Cobb, *Feedback and pole placement in descriptor variable systems*, Internat. J. Control, 33 (1981), pp. 1135–1146.
[11] J. W. Brewer, J. W. Bunce, and F. S. Van Vleck, *Linear Systems over Commutative Rings*, Marcel Dekker, New York, 1986.
[12] S. L. Campbell, *Singular Systems of Differential Equations* II, Pitman, Boston, 1982.
[13] L. Dai, *Singular Control Systems*, Lecture Notes in Control and Inform. Sci. 118, Springer-Verlag, Berlin, 1989.
[14] G. C. Verghese, B. Lévy, and T. Kailath, *A generalized state-space for singular systems*, IEEE Trans. Automat. Control, 25 (1981), pp. 811–831.
[15] G. Doetsch, *Introduction to the Theory and Application of the Laplace Transform*, Springer-Verlag, New York, 1974.
[16] J. D. Cobb, *Controllability, observability, and duality in singular systems*, IEEE Trans. Automat. Control, 29 (1984), pp. 1076–1082.
[17] R. Byers, P. Kunkel, and V. Mehrmann, *Regularization of linear descriptor systems with variable coefficients*, SIAM J. Control Optim., 35 (1997), pp. 117–133.
[18] P. Kunkel, V. Mehrmann, and W. Rath, *Analysis and numerical solution of control problems in descriptor form*, Math. Control Signals Systems, 14 (2001), pp. 29–61.
[19] J. D. Cobb, *A further interpretation of inconsistent initial conditions in descriptor variable systems*, IEEE Trans. Automat. Control, 28 (1983), pp. 920–922.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 6, pp. 2210–2232

HIGH-GAIN STATE FEEDBACK ANALYSIS BASED ON SINGULAR SYSTEM THEORY*

DANIEL COBB† AND JACOB EAPEN†

Abstract. We consider linear, time-invariant state-space systems under high-gain state feedback. The analysis is couched in terms of singular system theory and Grassman manifolds. Our work is distinguished from that of other authors by the fact that we do not allow a gain-dependent state coordinate change. Simple necessary and sufficient conditions are proven under which a singular system is a high-gain limit of a given state-space system. It is shown that the feedback matrix achieves a limit on an appropriate Grassmanian, so infinite gains constitute well-defined mathematical objects. The special cases of minimum-order stable and zeroth-order limits are studied in depth, including an analysis of solution behavior. Finally, the classical "cheap control" problem is interpreted within the context of our results.

Key words. high-gain feedback, state feedback, singular systems

AMS subject classifications. 93B52, 93B55, 93B05

DOI. 10.1137/040620060

1. Introduction.

Consider the linear, time-invariant state-space system

$$\dot{x} = Ax + Bu, \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. For any $K \in \mathbb{R}^{m\times n}$, we may apply state feedback

$$u = -Kx + v, \tag{2}$$

yielding the closed-loop system

$$\dot{x} = (A - BK)x + Bv. \tag{3}$$

In this paper, we are interested in the "high-gain limits" of (3) as $\|K\| \to \infty$. We seek a characterization of all such limits for a given system (1). In addition, we will specialize our results to certain important classes of limits and develop conditions under which a limit of (2) constitutes a well-defined system in its own right. We will then apply our results to the classical "cheap control" problem.

Numerous references deal with the issue of high-gain limits under state feedback. For example, early papers such as [1] treat high gain in a classical singular perturbation context. Much of this work can be viewed largely as a special case of our results. The details will be provided in sections 4–6.

More recent efforts, such as [2], [3], and [4], study high-gain limits in great depth. However, this body of work is fundamentally different from ours in that a K-dependent coordinate change is allowed, while our approach admits no coordinate change. The consequences of the two approaches are strikingly different. Indeed, consider the 1st-order system

$$\dot{x} = u$$

*Received by the editors December 2, 2004; accepted for publication August 5, 2005; published electronically January 26, 2006.
http://www.siam.org/journals/sicon/44-6/62006.html
†Department of Electrical and Computer Engineering, University of Wisconsin, 1415 Engineering Drive, Madison, WI 53706-1691 (cobb@engr.wisc.edu, eapen@cae.wisc.edu).

with feedback

$$u = -kx + v.$$

Our analysis (and that of [1]) dictates that the closed-loop system be written

$$-\frac{1}{k}\dot{x} = x - \frac{1}{k}v,$$

yielding $x = 0$ in the limit. Note that controllability is progressively weakened as k increases, and lost entirely for $k = \infty$. This is precisely the effect one would observe in practice, with the variable x representing the fixed (i.e., *K-independent*) state of the plant.

On the other hand, the analyses in [2], [3], and [4] allow a *K-dependent* coordinate change. In this case, the kth closed-loop system becomes

$$p_k q_k \dot{z} = -p_k k q_k z + p_k v,$$

where $x = q_k z$, and p_k, q_k are arbitrary nonzero sequences. For any $g \neq 0$, setting

$$p_k = 1, \quad q_k = \frac{1}{kg}$$

yields the controllable limit $z = gv$. The problem here is that the loss of controllability is masked by the coordinate change $z = kgx$, which scales the physical state x progressively higher as $k \to \infty$.

Another phenomenon that can occur with a K-dependent coordinate change is illustrated by the example

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u, \tag{4}$$
$$u = -\begin{bmatrix} k^2 & 1 \end{bmatrix} x.$$

Let $x = Q_k z$ and premultiply (4) by P_k, where P_k, Q_k are nonsingular. Then

$$P_k Q_k \dot{z} = P_k \begin{bmatrix} 0 & 1 \\ -k^2 & -1 \end{bmatrix} Q_k z, \tag{5}$$

which is equivalent to a system of the form

$$X_k \dot{z} = z. \tag{6}$$

If $Q_k = I$,

$$X_k = \begin{bmatrix} -\frac{1}{k^2} & -\frac{1}{k^2} \\ 1 & 0 \end{bmatrix} \to \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \tag{7}$$

irrespective of P_k. On the other hand, setting $P_k = I$ and

$$Q_k = \begin{bmatrix} \frac{1}{k} & 0 \\ 0 & 1 \end{bmatrix}$$

yields

$$X_k = \begin{bmatrix} -\frac{1}{k^2} & -\frac{1}{k} \\ \frac{1}{k} & 0 \end{bmatrix} \to 0. \tag{8}$$

Substituting (7) and (8) into (6) produces vastly different results. In particular, (7) produces impulses, while (8) does not. (See [9, Ch. 22].) Losing track of the impulsive behavior in (6) and (8) is again due to the progressive redefinition of the state.

Our approach disallows coordinate changes of the state x. A moment's reflection indicates that, in our setting, the high-gain limits of (3) form a subset of those in [2], [3], and [4]. Nevertheless, characterization of these "fixed coordinate" limits requires an independent analysis. Although the limits we obtain must satisfy the necessary conditions proven in [2] and [3], we will establish alternative conditions, which are arguably simpler and both necessary and sufficient. We will also conduct a careful analysis of stable and "zeroth-order" limits, which heretofore have not been explicitly studied in the literature, at least at this level of generality.

One of our objectives is to establish results which are dual to those we developed for observers in [6]. To this end, much of our work relies on the theory of differentiable manifolds. (See, e.g., [10].)

Throughout the paper, we assume for convenience that $\operatorname{rank} B = m$. For a system where this is not the case, an input coordinate change $\widehat{u} = Tu$ can be used to reduce the problem to our framework.

2. Preliminaries. Before we can talk about the limits of (1), we need some elementary results from singular system theory. Consider the matrix differential equation

$$E\dot{x} = Fx + Gu, \tag{9}$$

where $E, F \in \mathbb{R}^{n\times n}$ and $G \in \mathbb{R}^{n\times m}$. We assume the matrix pencil (E, F) is *regular*, i.e.,

$$\Delta(s) = |sE - F| \not\equiv 0.$$

The roots of Δ are the *eigenvalues* of the system. Consider the *Stiefel manifold* $\mathcal{V}_n\left(\mathbb{R}^{n\times(2n+m)}\right)$ of all $\begin{bmatrix} E & F & G \end{bmatrix} \in \mathbb{R}^{n\times(2n+m)}$ with full rank. Also, let

$$\Sigma(n,m) = \left\{ \begin{bmatrix} E & F & G \end{bmatrix} \;\middle|\; \Delta \not\equiv 0 \right\}.$$

Since $\Delta \not\equiv 0$ implies $\begin{bmatrix} E & F \end{bmatrix}$ has full rank, $\Sigma(n,m) \subset \mathcal{V}_n\left(\mathbb{R}^{n\times(2n+m)}\right)$. Both $\Sigma(n,m)$ and $\mathcal{V}_n\left(\mathbb{R}^{n\times(2n+m)}\right)$ are complementary to algebraic varieties in $\mathbb{R}^{n\times(2n+m)}$ and are, therefore, open and dense in $\mathbb{R}^{n\times(2n+m)}$.

Since premultiplication of (9) by a nonsingular matrix M does not affect the dynamics of (9), it is natural to identify systems of the form (9), which are related by such a transformation. On the other hand, right multiplication of E and A amounts to a coordinate change, so we avoid such transformations, retaining the coordinate-dependent nature of conventional state-space theory. We claim that this approach leads to a simpler theory overall.

With these ideas in mind, we couch our problem in terms of the *Grassman manifold* $\mathcal{G}_n\left(\mathbb{R}^{2n+m}\right)$. A Grassmanian is obtained by applying the equivalence relation

$$\begin{bmatrix} E_1 & F_1 & G_1 \end{bmatrix} \approx \begin{bmatrix} E_2 & F_2 & G_2 \end{bmatrix} \quad \text{iff} \quad \exists \text{ nonsingular } M \ni \begin{bmatrix} E_1 & F_1 & G_1 \end{bmatrix} = M \begin{bmatrix} E_2 & F_2 & G_2 \end{bmatrix} \tag{10}$$

to $\mathcal{V}_n\left(\mathbb{R}^{n\times(2n+m)}\right)$ and forming the quotient manifold $\mathcal{G}_n\left(\mathbb{R}^{2n+m}\right)$ with dimension $n(n+m)$. Charts on $\mathcal{G}_n\left(\mathbb{R}^{2n+m}\right)$ may be constructed by setting n columns of

$[E \quad F \quad G]$ to the $n \times n$ identity matrix and varying the remaining entries. Doing this in all $\binom{2n+m}{n}$ ways generates an atlas for $\mathcal{G}_n\left(\mathbb{R}^{2n+m}\right)$. We denote points in $\mathcal{G}_n\left(\mathbb{R}^{2n+m}\right)$ by $[E,F,G]$. Setting

$$\mathcal{L}(n,m) = \left\{ [E,F,G] \in \mathcal{G}_n\left(\mathbb{R}^{2n+m}\right) \;\middle|\; \Delta \not\equiv 0 \right\}$$

is consistent with the quotient structure of $\mathcal{G}_n\left(\mathbb{R}^{2n+m}\right)$, since premultiplication of $\begin{bmatrix} E & F & G \end{bmatrix}$ by a nonsingular M scales Δ by a nonzero constant. Let

$$\mu : \mathcal{V}_n\left(\mathbb{R}^{2n+m}\right) \to \mathcal{G}_n\left(\mathbb{R}^{2n+m}\right)$$

be the submersion defined by $\begin{bmatrix} E & F & G \end{bmatrix} \to [E,F,G]$. Then μ is continuous and open (see [10, Prop. 6.1.5]). Hence, $\mathcal{L}(n,m) = \mu\left(\Sigma(n,m)\right)$ is an open, dense submanifold of $\mathcal{G}_n\left(\mathbb{R}^{2n+m}\right)$. This makes $\mathcal{L}(n,m)$ an analytic manifold of dimension $n(n+m)$. We studied $\mathcal{L}(n,m)$ in [5].

We will make frequent use of the Weierstrass decomposition ([8, pp. 24–28]): For any regular pencil (E,F), there exists nonsingular M,N such that

$$MEN = \begin{bmatrix} I & 0 \\ 0 & E_f \end{bmatrix}, \qquad MFN = \begin{bmatrix} F_s & 0 \\ 0 & I \end{bmatrix}, \tag{11}$$

where E_f is nilpotent. E_f and F_s are unique up to similarity. Define the *order* of (E,F) to be $\operatorname{ord}(E,F) = \deg \Delta$ (i.e., the dimension of F_s) and the *index* $\operatorname{ind}(E,F)$ to be the smallest integer $q \geq 1$ such that $E_f^q = 0$. The functions ord and ind are uniquely defined on $\Sigma(n,m)$. In fact, both are invariant under the equivalence (11), so we may apply them to points in $\mathcal{L}(n,m)$:

$$\begin{aligned} \operatorname{ord}[E,F,G] &= \operatorname{ord}(E,F), \\ \operatorname{ind}[E,F,G] &= \operatorname{ind}(E,F). \end{aligned}$$

Eigenvalues are also invariant over orbits in $\mathcal{V}_n(\mathbb{R}^{n\times(2n+m)})$, so we may refer to a point $[E,F,G]$ as being *stable* if all its eigenvalues λ satisfy $\operatorname{Re}\lambda < 0$ and $\operatorname{ind}[E,F,G] = 1$.

We will need to consider solutions of (9). To this end, we review some basic facts from the theory of distributions. (See, e.g., [11].) Let $\mathcal{D}$ be the space of C^∞ functions $\phi : \mathbb{R} \to \mathbb{R}$ with bounded support, and let $\mathcal{D}'$ denote the dual space of $\mathcal{D}$. A *distribution* f is any member of $\mathcal{D}'$. Each locally L^1 function f (i.e., L^1 on bounded intervals) may be considered a distribution, since it determines a functional $\phi \to \int f\phi$. The unit impulse δ is defined to be the evaluation functional $\langle \delta, \phi \rangle = \phi(0)$. Every distribution has a derivative defined by $\langle \dot{f}, \phi \rangle = -\langle f, \dot{\phi} \rangle$; thus $\langle \delta^{(i)}, \phi \rangle = (-1)^i \phi^{(i)}(0)$. A sequence of distributions f_k is said to converge *weak** to f if $\langle f_k, \phi \rangle \to \langle f, \phi \rangle$ for every $\phi \in \mathcal{D}$. One advantage of working with distributions is that differentiation is a weak*-continuous operation. Besides weak* convergence, we will sometimes refer to uniform convergence $f_k \to f$ on an interval in $\mathcal{I} \subset \mathbb{R}$. This simply means that there exist locally L^1 functions g_k, g defined on $\mathcal{I}$ such that $\langle f_k, \phi \rangle = \langle g_k, \phi \rangle$, $\langle f, \phi \rangle = \langle g, \phi \rangle$ for all ϕ with support in $\mathcal{I}$ and $g_k \to g$ uniformly. Let $U \subset \mathbb{R}$ be the largest open set such that $\operatorname{supp}\phi \subset U$ implies $\langle f, \phi \rangle = 0$. The *support* of f is $\operatorname{supp} f = U^c$. Let $\mathcal{D}'_+$ be the distributions with support in $[0,\infty)$.

In order to apply arbitrary initial conditions x_0 to (9), it is convenient to consider the augmented system

$$E\dot{x} = Fx + Gu + \delta E x_0, \tag{12}$$

which yields a unique solution $x \in \mathcal{D}'_+$. (See [9, Ch. 22] for details.). Let

$$\begin{bmatrix} G_s \\ G_f \end{bmatrix} = MG, \quad \begin{bmatrix} x_s \\ x_f \end{bmatrix} = N^{-1}x, \quad \begin{bmatrix} x_{0s} \\ x_{0f} \end{bmatrix} = N^{-1}x_0 \tag{13}$$

and $\exp(F_s) : \mathbb{R} \to \mathbb{R}^{\deg \Delta \times \deg \Delta}$ be given by

$$\exp(F_s)t = \begin{cases} e^{tF_s}, & t \geq 0, \\ 0, & t < 0. \end{cases}$$

Define the *state-transition matrix*

$$\Phi = N \begin{bmatrix} \exp(F_s) & 0 \\ 0 & -\sum_{i=0}^{q-1} \delta^{(i)} E_f^i \end{bmatrix} M. \tag{14}$$

Direct calculation shows that Φ is the inverse Laplace transform of $(sE - A)^{-1}$, so Φ may be viewed as a map on $\Sigma(n, 0)$, obviously $1 - 1$. Since Φ is $1 - 1$, it varies over each orbit in $\Sigma(n, 0)$, so Φ cannot be defined consistently on $\mathcal{L}(n, 0)$. Φ may be extended trivially to $\Sigma(n, m)$, with similar consequences. The state transition matrix relates to the system (12) as follows.

THEOREM 1.

(1) $E\dot{\Phi} = A\Phi + \delta I$.

(2) *The solution of* (12) *is* $x = \Phi E x_0 + \Phi * Gu$.

(3) *The system* (12) *is asymptotically stable iff* ΦE *is bounded and decays asymptotically to* 0.

Proof. (1) and (2) follow by direct calculation.

(3) By asymptotic stability, we mean that, for $u \equiv 0$, we have conditions (a) $x(t) \to 0$ as $t \to \infty$ for every x_0, and (b) $\sup_t |x(t)| \to 0$ as $x_0 \to 0$. Boundedness and decay of ΦE are equivalent to the eigenvalues λ of F_s satisfying $\operatorname{Re} \lambda < 0$ and $E_f = 0$.

(Sufficient) From (11) and (14),

$$\Phi E = N \begin{bmatrix} \exp(F_s) & 0 \\ 0 & 0 \end{bmatrix} N^{-1},$$

so conditions (a) and (b) are met relative to $\Phi E x_0$.

(Necessary) We have $\Phi(t) E x_0 \to 0$ for every x_0, so

$$\Phi E = N \begin{bmatrix} \exp(F_s) & 0 \\ 0 & -\sum_{i=0}^{q-1} \delta^{(i)} E_f^{i+1} \end{bmatrix} N^{-1} \to 0,$$

which implies F_s is stable. Furthermore, $\Phi E x_0$ is bounded for every x_0, so ΦE is bounded, which implies it contains no impulses, i.e., $E_f = 0$. ∎

3. The manifold of closed-loop systems. The present paper closely follows the development of [6], where the dual problem of the limiting behavior of state observers under high-gain feedback was studied. One might speculate that the state feedback case should be obtained from [6] merely be taking the "transpose" of all theorems. While some theorems do transfer over in this way, much of the state feedback theory is different. One way to see that this must be true is to observe that, in both cases, systems are identified when they are related by left multiplication by a

nonsingular M. In contrast, pure transposition of the observer problem would require *right* multiplication by M, leading to a K-dependent coordinate change, which we explicitly avoid.

The closed-loop system (3) for a given plant (1) imbeds naturally into $\mathcal{L}(n,m)$ via the map $K \to [I, A - BK, B]$. We denote the image of $\mathbb{R}^{m\times n}$ under this map by $\mathcal{C}_r$. We further denote the closure of $\mathcal{C}_r$ in $\mathcal{L}(n,m)$ by $\mathcal{C}$ and consider the set $\mathcal{C}_s = \mathcal{C} - \mathcal{C}_r$. $\mathcal{C}$ may be regarded as the set of all limits of (3), $\mathcal{C}_r$ the full-order limits (i.e., ordinary state-space systems) and $\mathcal{C}_s$ the singular limits (i.e., generalized state-space systems). Another way to define $\mathcal{C}$, $\mathcal{C}_r$, and $\mathcal{C}_s$ is via the submersion μ. Let

$$\Omega_r = \left\{ \begin{bmatrix} M & M(A-BK) & MB \end{bmatrix} \;\middle|\; M \text{ nonsingular} \right\}.$$

Obviously, $\Omega_r \subset \Sigma(n,m)$. Let Ω be the closure of Ω_r in $\Sigma(n,m)$, and $\Omega_s = \Omega - \Omega_r$. It is easy to see that $\mathcal{C} = \mu(\Omega)$, $\mathcal{C}_r = \mu(\Omega_r)$, and $\mathcal{C}_s = \mu(\Omega_s)$.

We need the following lemma to prove Theorem **??**, which establishes the basic structure of $\mathcal{C}$.

LEMMA 2. *Let $X, Y \in \mathbb{R}^{n\times n}$ with $\operatorname{rank}[X \quad Y] = n$. There exist $K_k \in \mathbb{R}^{m\times n}$ and nonsingular $X_k \in \mathbb{R}^{n\times n}$ such that $X_k \to X$ and $X_k B K_k \to Y$ iff $\operatorname{rank}[XB \quad Y] = m$.*

Proof. (Necessary) For large k,

$$\operatorname{rank}\begin{bmatrix} XB & Y \end{bmatrix} \le \operatorname{rank}\begin{bmatrix} X_k B & X_k B K_k \end{bmatrix} = \operatorname{rank}\begin{bmatrix} B & BK_k \end{bmatrix} = m.$$

Suppose

$$\operatorname{rank}\begin{bmatrix} XB & Y \end{bmatrix} < m,$$

and let $R \subset \mathbb{R}^n$ be a subspace such that

$$\operatorname{Im} B \oplus R = \mathbb{R}^n.$$

Then $\dim R = n - m$, and

$$\begin{aligned} \operatorname{rank}\begin{bmatrix} X & Y \end{bmatrix} &= \dim(\operatorname{Im} X + \operatorname{Im} Y) \\ &= \dim(XR + \operatorname{Im} XB + \operatorname{Im} Y) \\ &\le \dim XR + \dim(\operatorname{Im} XB + \operatorname{Im} Y) \\ &\le \dim R + \operatorname{rank}\begin{bmatrix} XB & Y \end{bmatrix} \\ &< n. \end{aligned}$$

From this contradiction, we conclude

$$\operatorname{rank}\begin{bmatrix} XB & Y \end{bmatrix} = m.$$

(Sufficient) Let

$$\begin{aligned} R &= \operatorname{Im} XB \cap \operatorname{Im} Y, \\ S &= \operatorname{Ker} X \cap \operatorname{Im} B, \end{aligned}$$

$p = \dim R$, and $q = \dim S$. Then

$$m = q + \operatorname{rank} XB,$$

and there exists a nonsingular $T \in \mathbb{R}^{n\times n}$ such that

$$YT = \begin{bmatrix} Y_1 & Y_2 \end{bmatrix}$$

with $\operatorname{Im} Y_1 = R$ and

$$\operatorname{Im} XB \cap \operatorname{Im} Y_2 = 0.$$

Hence, we may select $H \in \mathbb{R}^{m\times p}$ such that $XBH = Y_1$. Also,

$$\operatorname{rank} XB + \operatorname{rank} Y_2 = \operatorname{rank} \begin{bmatrix} XB & Y_2 \end{bmatrix} \leq \operatorname{rank} \begin{bmatrix} XB & Y \end{bmatrix} = m,$$

so

$$\operatorname{rank} Y_2 \leq q.$$

We may choose $J \in \mathbb{R}^{m\times q}$ such that $\operatorname{Im} BJ = S$. Then $XBJ = 0$, and

$$\operatorname{rank} BJ = q \geq \operatorname{rank} Y_2.$$

Thus there exists $Z \in \mathbb{R}^{n\times n}$ such that $ZBJ = Y_2$.

Let $Z_k = X + \frac{1}{k}Z$ and, for each k, select nonsingular $Z_{kj} \to Z_k$ as $j \to \infty$. We may select a sequence $j_k \uparrow \infty$ such that

$$\|Z_{kj_k} - Z_k\| < \frac{1}{k^2}$$

for every k. Setting $X_k = Z_{kj_k}$, we have

$$\|X_k - X\| \leq \|X_k - Z_k\| + \|Z_k - X\| = \frac{1}{k^2} + \frac{1}{k}\|Z\|,$$

so $X_k \to X$. Let $K_k = \begin{bmatrix} H & kJ \end{bmatrix} T^{-1}$. Then

$$\begin{aligned} X_k B K_k &= \begin{bmatrix} X_k BH & k(X_k - Z_k)BJ + kZ_k BJ \end{bmatrix} T^{-1} \\ &= \begin{bmatrix} X_k BH & k(Z_{kj_k} - Z_k)BJ + kXBJ + ZBJ \end{bmatrix} T^{-1} \\ &\to \begin{bmatrix} Y_1 & Y_2 \end{bmatrix} T^{-1} \\ &= Y. \qquad \square \end{aligned}$$

THEOREM 3.
(1) $\mathcal{C} = \{[X, XA - Y, XB] \in \mathcal{L}(n,m) \mid \operatorname{rank} \begin{bmatrix} XB & Y \end{bmatrix} = m\}$.
(2) *$\mathcal{C}$ is a regular submanifold of $\mathcal{L}(n,m)$ with dimension nm.*
(3) *$\mathcal{C}_r$ is a (relatively) open, dense submanifold of $\mathcal{C}$*
(4) *$[X, XA - Y, XB] \in \mathcal{C}_s$ iff $\operatorname{rank} \begin{bmatrix} XB & Y \end{bmatrix} = m$ with X singular.*

Proof. (1) Let

$$\Omega_e = \left\{ \begin{bmatrix} X & XA - Y & XB \end{bmatrix} \in \mathcal{V}_n\left(\mathbb{R}^{2n+m}\right) \;\middle|\; \operatorname{rank} \begin{bmatrix} XB & Y \end{bmatrix} = m \right\}.$$

Setting $X = M$ and $Y = MBK$ yields

$$\begin{bmatrix} X & XA - Y & XB \end{bmatrix} = \begin{bmatrix} M & M(A - BK) & MB \end{bmatrix},$$
$$\operatorname{rank} \begin{bmatrix} XB & Y \end{bmatrix} = \operatorname{rank} \begin{bmatrix} MB & MBK \end{bmatrix} = \operatorname{rank} \begin{bmatrix} B & BK \end{bmatrix} = m,$$

so $\Omega_r \subset \Omega_e \cap \Sigma(n,m)$. It suffices to show that the closure of Ω_r in $\mathcal{V}_n\left(\mathbb{R}^{2n+m}\right)$ is Ω_e, because then the closure of Ω_r in $\Sigma(n,m)$ is $\Omega = \Omega_e \cap \Sigma(n,m)$, and part (1) follows from $\mu(\Omega_r) = \mathcal{C}_r$, $\mu(\Omega) = \mathcal{C}$.

For any nonsingular $T \in \mathbb{R}^{n\times n}$, let

$$L_T = \begin{bmatrix} T^{-1} & T^{-1}A & T^{-1}B \\ 0 & -I & 0 \end{bmatrix}.$$

Choose X, Y such that

$$\begin{bmatrix} X & Y \end{bmatrix} L_I = \begin{bmatrix} X & XA - Y & XB \end{bmatrix} \in \Omega_e.$$

L_T has independent rows, so $\operatorname{rank}\begin{bmatrix} X & Y \end{bmatrix} = n$. From Lemma 2, there exist sequences X_k and K_k, with X_k nonsingular, such that $X_k \to X$ and $X_k B K_k \to Y$. Hence,

$$\begin{bmatrix} X_k & X_k(A - BK_k) & X_k B \end{bmatrix} \to \begin{bmatrix} X & XA - Y & XB \end{bmatrix},$$

and the closure of Ω_r contains Ω_e. Conversely, if

$$\begin{bmatrix} X_k & X_k(A - BK_k) & X_k B \end{bmatrix} \to \begin{bmatrix} X & F & G \end{bmatrix} \in \mathcal{V}_n\left(\mathbb{R}^{2n+m}\right),$$

then $X_k \to X$ and $G = XB$. Let $Y = XA - F$. Then

$$\begin{aligned} \operatorname{rank}\begin{bmatrix} X & Y \end{bmatrix} &= \operatorname{rank}\begin{bmatrix} X & Y \end{bmatrix}\begin{bmatrix} I & A \\ 0 & -I \end{bmatrix} \\ &= \operatorname{rank}\begin{bmatrix} X & F \end{bmatrix} \\ &= \operatorname{rank}\begin{bmatrix} X & F & G \end{bmatrix} \\ &= n, \end{aligned}$$

$$X_k B K_k = X_k A - X_k(A - BK_k) \to Y,$$

so, from Lemma 2,

$$\operatorname{rank}\begin{bmatrix} XB & Y \end{bmatrix} = m.$$

Hence, $\begin{bmatrix} X & F & G \end{bmatrix} \in \Omega_e$, and Ω_e contains the closure of Ω_r.

(2) This part of the proof will be based on the following construction. Choose a nonsingular T such that

$$T^{-1}B = \begin{bmatrix} 0 \\ I \end{bmatrix},$$

and consider the diagram

$$\begin{array}{ccccc} & \widehat{h} & & \widehat{g} & \\ \mathcal{V}_m\left(\mathbb{R}^{m+n}\right) & \to & \mathcal{V}_n\left(\mathbb{R}^{2n}\right) & \to & \mathcal{V}_n\left(\mathbb{R}^{2n+m}\right) \\ \downarrow \pi & & \downarrow \nu & & \downarrow \mu \\ \mathcal{G}_m\left(\mathbb{R}^{m+n}\right) & \to & \mathcal{G}_n\left(\mathbb{R}^{2n}\right) & \to & \mathcal{G}_n\left(\mathbb{R}^{2n+m}\right), \\ & h & & g & \end{array}$$

where

$$\begin{aligned} \widehat{g}\left(\begin{bmatrix} \widetilde{X} & Y \end{bmatrix}\right) &= \begin{bmatrix} \widetilde{X} & Y \end{bmatrix} L_T, \\ \widehat{h}(Z) &= \begin{bmatrix} I & 0 \\ 0 & Z \end{bmatrix}, \end{aligned}$$

and μ, ν, and π are the standard submersions. We note that

$$\widehat{g}\left(M\begin{bmatrix}\widetilde{X} & Y\end{bmatrix}\right) = M\widehat{g}\left(\begin{bmatrix}\widetilde{X} & Y\end{bmatrix}\right),$$

and

$$\widehat{h}\,(NZ) = \begin{bmatrix}I & 0\\ 0 & N\end{bmatrix}\widehat{h}\,(Z)$$

for any nonsingular M, N, so g and h may be defined to make the diagram commute. We are mainly interested in the compositions $f = g \circ h$ and $\widehat{f} = \widehat{g} \circ \widehat{h}$. Note that $\widehat{g}$, $\widehat{h}$, and hence $\widehat{f}$ are $1-1$. Furthermore, if

$$\widehat{g}\left(\begin{bmatrix}\widetilde{X}_a & Y_a\end{bmatrix}\right) = M\widehat{g}\left(\begin{bmatrix}\widetilde{X} & Y\end{bmatrix}\right),$$

we obtain

$$\widehat{g}\left(\begin{bmatrix}\widetilde{X}_a & Y_a\end{bmatrix}\right) = \widehat{g}\left(M\begin{bmatrix}\widetilde{X} & Y\end{bmatrix}\right),$$

so

$$\begin{bmatrix}\widetilde{X}_a & Y_a\end{bmatrix} = M\begin{bmatrix}\widetilde{X} & Y\end{bmatrix}$$

and g is $1-1$. Now suppose

$$\widehat{h}\,(Z_a) = M\widehat{h}\,(Z)\,.$$

Then

$$\begin{bmatrix}I & 0\\ 0 & Z_a\end{bmatrix} = \begin{bmatrix}M_{11} & M_{12}\\ M_{21} & M_{22}\end{bmatrix}\begin{bmatrix}I & 0\\ 0 & Z\end{bmatrix}.$$

Inspection of the block matrix equations yields $Z_a = M_{22}Z$ with M_{22} nonsingular, so h and f are $1-1$.

Let $\mathcal{C}_e = \mu\,(\Omega_e)$. Since $\mathcal{L}(n,m)$ is open in $\mathcal{G}_n\left(\mathbb{R}^{2n+m}\right)$, it suffices to demonstrate that $\mathcal{C}_e$ satisfies (2), because then $\mathcal{C} = \mathcal{C}_e \cap \mathcal{L}(n,m)$ inherits the same properties. We begin by showing that $f\left(\mathcal{G}_m\left(\mathbb{R}^{m+n}\right)\right) = \mathcal{C}_e$. Consider any point $[X, XA - Y, XB] \in \mathcal{C}_e$. Setting $\widetilde{X} = XT$ and partitioning

$$\begin{bmatrix}\widetilde{X}_1 & \widetilde{X}_2\end{bmatrix} = \widetilde{X},$$

with $\widetilde{X}_1 \in \mathbb{R}^{n\times(n-m)}$, $\widetilde{X}_2 \in \mathbb{R}^{n\times m}$, we obtain

$$\operatorname{rank}\begin{bmatrix}\widetilde{X}_2 & Y\end{bmatrix} = \operatorname{rank}\begin{bmatrix}\widetilde{X}T^{-1}B & Y\end{bmatrix} = \operatorname{rank}\begin{bmatrix}XB & Y\end{bmatrix} = m,$$

$$\operatorname{rank}\begin{bmatrix}\widetilde{X}_1 & \widetilde{X}_2 & Y\end{bmatrix} = \operatorname{rank}\begin{bmatrix}XT & Y\end{bmatrix} = \operatorname{rank}\begin{bmatrix}X & Y\end{bmatrix} = n,$$

so $\operatorname{rank}\widetilde{X}_1 = n - m$. Hence, there exists $Z_1 \in \mathbb{R}^{m\times m}$, $Z_2 \in \mathbb{R}^{m\times n}$, and a nonsingular M such that

$$M\begin{bmatrix}\widetilde{X}_1 & \widetilde{X}_2 & Y\end{bmatrix} = \begin{bmatrix}I & 0 & 0\\ 0 & Z_1 & Z_2\end{bmatrix},$$

and

$$\operatorname{rank}\begin{bmatrix} Z_1 & Z_2 \end{bmatrix} = m.$$

It follows that

$$f\left([Z_1, Z_2]\right) = g\left(\left[\widetilde{X}, Y\right]\right) = [X, XA - Y, XB],$$

which yields the desired result. In fact, letting functions ϕ range over an atlas of $\mathcal{G}_m\left(\mathbb{R}^{n+m}\right)$, $\left\{\phi \circ f^{-1}\right\}$ becomes an atlas for $\mathcal{C}_e$, making f an analytic diffeomorphism between $\mathcal{G}_m\left(\mathbb{R}^{n+m}\right)$ and $\mathcal{C}_e$.

As a map into $\mathcal{G}_n\left(\mathbb{R}^{2n+m}\right)$, we can prove that f is analytic by showing that g and h are analytic. Let $\xi \in \mathcal{G}_n\left(\mathbb{R}^{2n}\right)$, and choose charts ϕ on $\mathcal{G}_n\left(\mathbb{R}^{2n}\right)$ and $\psi \in \mathcal{G}_n\left(\mathbb{R}^{2n+m}\right)$ such that ξ and $g(\xi)$ lie in the domains of ϕ and ψ, respectively. Then $\psi \circ g \circ \phi^{-1}$ is a rational function, where the denominator has no zero, and is thus analytic. Since ϕ, ψ are arbitrary, g is analytic. Analyticity of h is proved similarly.

To show that $\mathcal{C}_e$ is a submanifold of $\mathcal{G}_n\left(\mathbb{R}^{2n+m}\right)$, we must also prove that f has full rank. We need to show that the derived linear function f_* at each point of $\mathcal{G}_m\left(\mathbb{R}^{n+m}\right)$ is $1-1$. From [10, Prop. 4.3.1], $f_* = g_* \circ h_*$, so it suffices to prove that g_* and h_* are $1-1$. Since $g \circ \nu = \mu \circ \widehat{g}$, the same theorem guarantees

$$g_* \circ \nu_* = \mu_* \circ \widehat{g}_*.$$

Since $\widehat{g}$ is $1-1$ and μ_*, ν_* are onto,

$$\begin{aligned} \operatorname{rank} g_* &= \operatorname{rank}\left(\mu_* \circ \widehat{g}_*\right) \\ &\geq \operatorname{rank} \widehat{g}_* - \left(\dim \mathcal{V}_n\left(\mathbb{R}^{2n+m}\right) - \operatorname{rank} \mu_*\right) \\ &= 2n^2 - \left(n(2n+m) - n(n+m)\right) \\ &= n^2, \end{aligned}$$

so g_* is $1-1$. Unfortunately, this calculation does not work for h_*. To prove h_* is $1-1$, consider any point $\xi \in \mathcal{G}_m\left(\mathbb{R}^{n+m}\right)$ and a chart ϕ whose coordinate domain contains ξ. Applying ϕ amounts to choosing m columns $\{c_i\}$ of $\begin{bmatrix} Z_1 & Z_2 \end{bmatrix}$, setting them equal to the $m \times m$ identity matrix, and allowing the remaining entries to vary, forming an $m \times n$ matrix $\widetilde{Z}$. In a neighborhood of $h(\xi)$, each point of $\mathcal{G}_n(\mathbb{R}^{2n})$ may be represented as $\begin{bmatrix} I & S_1 \\ 0 & S_2 \end{bmatrix}$, where the columns $\{c_i\}$ of $\begin{bmatrix} S_1 \\ S_2 \end{bmatrix}$ are $\begin{bmatrix} 0 \\ I \end{bmatrix}$. This generates a chart ψ of $\mathcal{G}_n\left(\mathbb{R}^{2n}\right)$, whose coordinate domain contains $h(\xi)$. It is easy to see that

$$\psi\left(h\left(\phi^{-1}\left(\widetilde{Z}\right)\right)\right) = \begin{bmatrix} 0 \\ \widetilde{Z} \end{bmatrix}.$$

From [10, p. 58], h_* has matrix representation $\frac{\partial\left(\psi \circ h \circ \phi^{-1}\right)}{\partial \widetilde{Z}}$. But $\psi \circ h \circ \phi^{-1}$ is linear, so

$$h_*\left(\widetilde{Z}\right) = \begin{bmatrix} 0 \\ \widetilde{Z} \end{bmatrix},$$

which is $1-1$. Hence, we conclude that $\mathcal{C}_e$ is an nm-dimensional submanifold of $\mathcal{G}_n\left(\mathbb{R}^{2n+m}\right)$.

Finally we prove regularity of $\mathcal{C}_e$. We need to show that the topologies that $\mathcal{C}_e$ inherits from $\mathcal{G}_m\left(\mathbb{R}^{m+n}\right)$ (through f) and from $\mathcal{G}_n\left(\mathbb{R}^{2n+m}\right)$ (as a subset) coincide.

Since f is analytic, it is continuous, and $f^{-1}\left(W \cap f\left(\mathcal{G}_m\left(\mathbb{R}^{m+n}\right)\right)\right) = f^{-1}\left(W\right)$ is open in $\mathcal{G}_m\left(\mathbb{R}^{m+n}\right)$ for every open $W \subset \mathcal{G}_n\left(\mathbb{R}^{2n+m}\right)$. To prove the converse, let $U \subset \mathcal{G}_m\left(\mathbb{R}^{m+n}\right)$ be open. Then $\pi^{-1}\left(U\right)$ is open. For any $Z \in \pi^{-1}\left(U\right)$ there exists $\varepsilon > 0$ such that the ball $B\left(Z, \varepsilon\right) \subset \pi^{-1}\left(U\right)$. Then $NB\left(Z, \varepsilon\right) \subset \pi^{-1}\left(U\right)$ for every nonsingular N. Let

$$\widetilde{L} = \begin{bmatrix} T & A \\ 0 & -I \\ 0 & 0 \end{bmatrix},$$

and define

$$\begin{aligned} W_Z &= \left\{ M^{-1} B\left(\widehat{f}\left(Z\right), \frac{\varepsilon}{2\left\|\widetilde{L}\right\|}\right) \;\middle|\; M \text{ nonsingular} \right\} \\ &= \mu^{-1}\left(\mu\left(B\left(\widehat{f}\left(Z\right), \frac{\varepsilon}{2\left\|\widetilde{L}\right\|}\right)\right)\right), \end{aligned}$$

$$W = \bigcup_{Z \in \pi^{-1}(U)} W_Z.$$

Since μ is open, each W_Z and, therefore, W are open. It suffices to show that

$$\widehat{f}\left(\pi^{-1}\left(U\right)\right) = W \cap \widehat{f}\left(\mathcal{V}_m\left(\mathbb{R}^{m+n}\right)\right). \tag{15}$$

Indeed, since W is a union of orbits in $\mathcal{V}_n(\mathbb{R}^{2n+m}), \mu(W \cap A) = \mu(W) \cap \mu(A)$ for any A, from which it follows that

$$\begin{aligned} f\left(U\right) &= \mu\left(\widehat{f}\left(\pi^{-1}\left(U\right)\right)\right) \\ &= \mu\left(W \cap \widehat{f}\left(\mathcal{V}_m\left(\mathbb{R}^{m+n}\right)\right)\right) \\ &= \mu\left(W\right) \cap \mu\left(\widehat{f}\left(\mathcal{V}_m\left(\mathbb{R}^{m+n}\right)\right)\right) \\ &= \mu\left(W\right) \cap \mathcal{C}_e, \end{aligned}$$

so $f\left(U\right)$ is (relatively) open in $\mathcal{C}_e$.

To prove (15), first note that $Z \in \pi^{-1}\left(U\right)$ implies $\widehat{f}\left(Z\right) \in W_Z$, so

$$\widehat{f}\left(\pi^{-1}\left(U\right)\right) \subset W \cap \widehat{f}\left(\mathcal{V}_m\left(\mathbb{R}^{m+n}\right)\right).$$

Conversely, suppose $Z_a \in \mathcal{V}_m(\mathbb{R}^{m+n})$, $\Delta \in B\big(0, \frac{\varepsilon}{2\|\widetilde{L}\|}\big)$, and nonsingular M satisfy

$$M^{-1}\left(\widehat{f}\left(Z\right) + \Delta\right) = \widehat{f}\left(Z_a\right).$$

Then

$$\left\|M\widehat{f}\left(Z_a\right) - \widehat{f}\left(Z\right)\right\| < \frac{\varepsilon}{2\left\|\widetilde{L}\right\|}.$$

But

$$M\widehat{f}(Z_a) - \widehat{f}(Z) = \left(M \begin{bmatrix} I & 0 \\ 0 & Z_a \end{bmatrix} - \begin{bmatrix} I & 0 \\ 0 & Z \end{bmatrix} \right) L_T$$

and $L_T \widetilde{L} = I$, so

$$\left\| M \begin{bmatrix} I & 0 \\ 0 & Z_a \end{bmatrix} - \begin{bmatrix} I & 0 \\ 0 & Z \end{bmatrix} \right\| \leq \left\| M\widehat{f}(Z_a) - \widehat{f}(Z) \right\| \left\| \widetilde{L} \right\| < \frac{\varepsilon}{2}.$$

Partitioning M, we obtain

$$\|M_{22} Z_a - Z\| < \frac{\varepsilon}{2}$$

(assuming an appropriate norm). Let N be a nonsingular matrix such that

$$\left\| N^{-1} - M_{22} \right\| < \frac{\varepsilon}{2\|Z_a\|}.$$

Then

$$\left\| N^{-1} Z_a - Z \right\| \leq \left\| N^{-1} - M_{22} \right\| \|Z_a\| + \|M_{22} Z_a - Z\| < \varepsilon,$$

so $Z_a \in NB(Z, \varepsilon)$. Hence,

$$\widehat{f}\left(\pi^{-1}(U)\right) \supset M^{-1} B \left(\widehat{f}(Z), \frac{\varepsilon}{2\left\| \widetilde{L} \right\|} \right) \cap \widehat{f}\left(\mathcal{V}_m\left(\mathbb{R}^{m+n}\right)\right)$$

for every nonsingular M, and

$$\widehat{f}\left(\pi^{-1}(U)\right) \supset W \cap \widehat{f}\left(\mathcal{V}_m\left(\mathbb{R}^{m+n}\right)\right).$$

(3) Density of $\mathcal{C}_r$ follows from the definition of $\mathcal{C}$. To show $\mathcal{C}_r$ is open in $\mathcal{C}$, it suffices to show that Ω_r is open in Ω. Let $\sigma \in \Omega_r$ and $\sigma_k \in \Omega$ with $\sigma_k \to \sigma$. Then there exist $M, X_k, Y_k \in \mathbb{R}^{n \times n}$, and $K \in \mathbb{R}^{m \times n}$, with M nonsingular and

$$\operatorname{rank} \begin{bmatrix} X_k B & Y_k \end{bmatrix} = m,$$

such that

$$\begin{aligned} \sigma &= \begin{bmatrix} M & M(A - BK) & MB \end{bmatrix}, \\ \sigma_k &= \begin{bmatrix} X_k & X_k A - Y_k & X_k B \end{bmatrix}. \end{aligned}$$

Since $\sigma_k \to \sigma$, $X_k \to M$, thus X_k is nonsingular for large k and

$$\operatorname{rank} \begin{bmatrix} B & X_k^{-1} Y_k \end{bmatrix} = m.$$

Then $\operatorname{Im} X_k^{-1} Y_k \subset \operatorname{Im} B$, so there exists $K_k \in \mathbb{R}^{m \times n}$ such that $X_k^{-1} Y_k = B K_k$. Therefore,

$$\sigma_k = \begin{bmatrix} X_k & X_k(A - BK_k) & X_k B \end{bmatrix} \in \Omega_r,$$

and Ω_r is open in Ω.

(4) (Sufficient) This follows from the definition of Ω_s and $\mathcal{C}_s = \mu(\Omega_s)$.
(Necessary) Assume X is nonsingular. From part (1),

$$\operatorname{rank}\begin{bmatrix} B & X^{-1}Y \end{bmatrix} = \operatorname{rank}\begin{bmatrix} XB & Y \end{bmatrix} = m,$$

so $\operatorname{Im} X^{-1}Y \subset \operatorname{Im} B$, and there exists $K \in \mathbb{R}^{m\times n}$ such that $X^{-1}Y = BK$. It follows that

$$[X, XA - Y, XB] = [X, X(A - BK), XB] \in \mathcal{C}_r,$$

which is a contradiction. ∎

Theorem 3, part (4) characterizes all degenerate closed-loop systems $\mathcal{C}_s$. This corresponds to applying a sequence of feedback matrices K_k such that $\|K_k\| \to \infty$, driving some or all eigenvalues to ∞ in magnitude. Since $\mathcal{C}_s$ is obtained with no state coordinate change, $\mathcal{C}_s$ must be a subset of the high-gain limits considered in [2] and [3]. In particular, each point in $\mathcal{C}_s$ must satisfy the necessary conditions established in [2, Thm. 1] and [3, Cor. 4.3]. Compared with these results, our characterization of $\mathcal{C}_s$ has a very different form, is necessary *and* sufficient, and is arguably simpler.

4. Stable and zeroth-order limits. In this section, we study certain subsets of $\mathcal{C}$ which have special significance. In particular, we examine those systems in $\mathcal{C}$ which are stable (i.e., all eigenvalues satisfy $\operatorname{Re}\lambda < 0$) and those with order 0. We begin with a discussion of an important submanifold of $\mathcal{C}$, which will help simplify the development. Let

$$\mathcal{C}_I = \{[X, I, XB] \in \mathcal{C}\}.$$

$\mathcal{C}_I$ is simply the set of points in $\mathcal{C}$ with no eigenvalue at 0. Each point in $\mathcal{C}_I$ corresponds to a system

$$X\dot{x} = x + XBv + \delta X x_0 \tag{16}$$

with state transition matrix determined by

$$X\dot{\Phi} = \Phi + \delta I.$$

From Theorem 3, part (1), we obtain

$$\mathcal{C}_I = \left\{ [X, I, XB] \in \mathcal{G}_n\left(\mathbb{R}^{2n+m}\right) \;\middle|\; \operatorname{rank}\begin{bmatrix} XB & XA - I \end{bmatrix} = m \right\}.$$

The next result gives several alternative characterizations of $\mathcal{C}_I$.

THEOREM 4. *For any $X \in \mathbb{R}^{n\times n}$, the following are equivalent:*
(1) $\operatorname{rank}[XB \quad XA - I] = m$.
(2) $\operatorname{Ker}[X \quad I] \subset \operatorname{Im}\left[\begin{smallmatrix} B & A \\ 0 & -1 \end{smallmatrix}\right]$.
(3) $\operatorname{Im}(AX - I) \subset \operatorname{Im} B$.
(4) *There exists $U \in \mathbb{R}^{m\times n}$ such that $AX + BU = I$.*

Proof. (1 $\iff$ 2) From elementary linear algebra,

$$\begin{aligned} \operatorname{rank}\begin{bmatrix} XB & XA - I \end{bmatrix} &= \operatorname{rank}\begin{bmatrix} X & I \end{bmatrix}\begin{bmatrix} B & A \\ 0 & -I \end{bmatrix} \\ &\geq \operatorname{rank}\begin{bmatrix} B & A \\ 0 & -I \end{bmatrix} - \left(2n - \operatorname{rank}\begin{bmatrix} X & I \end{bmatrix}\right) \\ &= (n + m) - (2n - n) \\ &= m \end{aligned} \tag{17}$$

with equality iff

$$\operatorname{Ker}\begin{bmatrix} X & I \end{bmatrix} \subset \operatorname{Im}\begin{bmatrix} B & A \\ 0 & -I \end{bmatrix}.$$

(2 $\Longleftrightarrow$ 3) Condition (2) is equivalent to saying that, for each x, there exist y, z such that

$$\begin{bmatrix} B & A \\ 0 & -I \end{bmatrix} \begin{bmatrix} y \\ z \end{bmatrix} = \begin{bmatrix} -x \\ Xx \end{bmatrix}. \tag{18}$$

Writing out the equations, (18) is the same as $By = (AX - I)\,x$, which is a restatement of (3).

(3 $\Longleftrightarrow$ 4) Condition (3) says that there exists U such that $AX - I = -BU$, which is the same as (4). □

Theorem 4, part (4) indicates that $\mathcal{C}_I$ is nonempty iff $\begin{bmatrix} A & B \end{bmatrix}$ has full rank; i.e., iff 0 is a controllable mode of (1). In this case, the affine set

$$\mathcal{W} = \left\{ \begin{bmatrix} X \\ U \end{bmatrix} \in \mathbb{R}^{2n \times n} \;\middle|\; AX + BU = I \right\}$$

will prove central to our theory. The next result gives a precise relationship between $\mathcal{C}_I$ and $\mathcal{W}$.

Theorem 5.

(1) $[X, I, XB] \in \mathcal{C}_I$ *iff there exists* $U \in \mathbb{R}^{m \times n}$ *such that* $\begin{bmatrix} X \\ U \end{bmatrix} \in \mathcal{W}$. *In this case,* U *is unique.*

(2) *Let* $K_k \in \mathbb{R}^{m \times n}$. *Then* $[I, A - BK_k, B] \to [X, I, XB] \in \mathcal{C}_I$ *as* $k \to \infty$ *iff* $A - BK_k$ *is nonsingular for large* k *and* $(A - BK_k)^{-1} \to X$. *In this case,* $-K_k(A - BK_k)^{-1} \to U$.

(3) $\mathcal{C}_I$ *is a (relatively) open, dense submanifold of* $\mathcal{C}$, *diffeomorphic to* $\mathcal{W}$.

Proof.

(1) All but uniqueness is a restatement of Theorem 4, part (4). Uniqueness follows from $BU = I - AX$ and $\operatorname{rank} B = m$.

(2) If $(A - BK_k)^{-1} \to X$,

$$[I, A - BK_k, B] = \left[(A - BK_k)^{-1}, I, (A - BK_k)^{-1} B \right] \to [X, I, XB]. \tag{19}$$

To prove the converse, we note that μ is a submersion, so there exist nonsingular M_k such that

$$M_k \begin{bmatrix} I & A - BK_k & B \end{bmatrix} \to \begin{bmatrix} X & I & XB \end{bmatrix}.$$

Hence, $M_k \to X$ and $M_k (A - BK_k) \to I$, so $A - BK_k$ is nonsingular for large k, and

$$(A - BK_k)^{-1} = (M_k (A - BK_k))^{-1} M_k \to X.$$

If $[X, I, XB] \in \mathcal{C}_I$, part (1) indicates that there exists a unique U such that $AX + BU = I$. Then

$$BK_k (A - BK_k)^{-1} = A (A - BK_k)^{-1} - I \to AX - I = -BU,$$

$$K_k (A - BK_k)^{-1} \to -U.$$

(3) Consider the open, dense subset

$$\Omega_I = \left\{ \begin{bmatrix} X & Y \end{bmatrix} L_I \in \mathcal{V}_n\left(\mathbb{R}^{2n+m}\right) \;\middle|\; \operatorname{rank}\begin{bmatrix} XB & Y \end{bmatrix} = m, \quad \det\left(AX - Y\right) \neq 0 \right\}$$

of Ω. Since μ is a submersion, $\mathcal{C}_I = \mu\left(\Omega_I\right)$ is open and dense in $\mathcal{C}$. The map

$$f : \begin{bmatrix} X \\ U \end{bmatrix} \to [X, I, XB]$$

takes $\mathcal{W}$ onto $\mathcal{C}_I$ by (1). Since U is uniquely determined by X, f is $1-1$. Both $\mathcal{W}$ and $\mathcal{C}_I$ are covered by single coordinate domains. One may construct an affine chart ϕ for $\mathcal{W}$ and apply the chart

$$\psi : [X, I, XB] \to X$$

to $\mathcal{C}_I$. Then $\psi \circ f \circ \phi^{-1}$ is an affine diffeomorphism, so f is a diffeomorphism. ∎

Since closed-loop systems in $\mathcal{C}_I$ (or, alternatively, $\mathcal{W}$) have no eigenvalue at 0, $\mathcal{C}_I$ contains all stable limits and all zeroth-order limits. The structure of $\mathcal{W}$ is dual to the structure of the manifold $\mathcal{V}$ we studied in [5].

Restricting to $\mathcal{C}_I$ yields a surprising result related to controllability of the closed-loop system (16).

THEOREM 6. *Let* $[X, I, XB] \in \mathcal{C}_I$. *Then* $\operatorname{rank} X \geq n-m$ *with equality iff* $XB = 0$.

Proof. $\begin{bmatrix} A & B \end{bmatrix}$ has full rank, so we may choose nonsingular M, N such that

$$MB = \begin{bmatrix} 0 \\ I \end{bmatrix}, \quad MAN = \begin{bmatrix} I & 0 \\ \widetilde{A}_{21} & \widetilde{A}_{22} \end{bmatrix}. \tag{20}$$

Let

$$\begin{bmatrix} \widetilde{X}_{11} & \widetilde{X}_{12} \\ \widetilde{X}_{21} & \widetilde{X}_{22} \end{bmatrix} = N^{-1} X M^{-1}, \quad \begin{bmatrix} \widetilde{U}_1 & \widetilde{U}_2 \end{bmatrix} = U M^{-1}.$$

Then

$$\begin{bmatrix} \widetilde{X}_{11} & \widetilde{X}_{12} \\ A_{21}\widetilde{X}_{11} + A_{22}\widetilde{X}_{21} + \widetilde{U}_1 & A_{21}\widetilde{X}_{12} + A_{22}\widetilde{X}_{22} + \widetilde{U}_2 \end{bmatrix} = M\left(AX + BU\right)M^{-1} = I,$$

so X and XB have the form

$$X = N \begin{bmatrix} I & 0 \\ X_{21} & X_{22} \end{bmatrix} M, \quad XB = N \begin{bmatrix} 0 \\ X_{22} \end{bmatrix}.$$

Hence, $\operatorname{rank} X \geq n - m$ with equality iff $X_{22} = 0$. ∎

Theorem 6 states that high-gain limits of (3), where the rank of X degenerates maximally, have the unfortunate property that the input v exerts no control whatsoever on the system. This is undoubtedly a limitation for control problems where closed-loop tracking to a reference input is required.

Now we consider the special cases of minimum-order stable and zeroth-order limits. By applying essentially the same arguments as in [5], several results are obtained immediately. These are summarized in Theorems 7 and 8. The first is based on the following construction. Choose any nonsingular matrix T such that

$$T^{-1}B = \begin{bmatrix} 0 \\ I \end{bmatrix}, \tag{21}$$

and let

$$\begin{bmatrix} \widetilde{A}_{11} & \widetilde{A}_{12} \\ \widetilde{A}_{21} & \widetilde{A}_{22} \end{bmatrix} = T^{-1}AT, \tag{22}$$

where $\widetilde{A}_{22} \in \mathbb{R}^{m\times m}$. If (A, B) is stabilizable,

$$\operatorname{rank} \begin{bmatrix} \lambda I - \widetilde{A}_{11} & -\widetilde{A}_{12} & 0 \\ -\widetilde{A}_{21} & \lambda I - \widetilde{A}_{22} & I \end{bmatrix} = n$$

for every λ with $\operatorname{Re}\lambda \geq 0$. Hence, $\operatorname{rank} \begin{bmatrix} \lambda I - \widetilde{A}_{11} & \widetilde{A}_{12} \end{bmatrix} = n - m$ (i.e., $(\widetilde{A}_{11}, \widetilde{A}_{12})$ is stabilizable). We may thus choose Λ such that $\widetilde{A}_{11} - \widetilde{A}_{12}\Lambda$ is stable and set

$$X = T \begin{bmatrix} \left(\widetilde{A}_{11} - \widetilde{A}_{12}\Lambda\right)^{-1} & 0 \\ -\Lambda\left(\widetilde{A}_{11} - \widetilde{A}_{12}\Lambda\right)^{-1} & 0 \end{bmatrix} T^{-1}, \tag{23}$$

$$U = \begin{bmatrix} -\left(\widetilde{A}_{21} - \widetilde{A}_{22}\Lambda\right)\left(\widetilde{A}_{11} - \widetilde{A}_{12}\Lambda\right)^{-1} & I \end{bmatrix} T^{-1}. \tag{24}$$

By direct calculation, $AX + BU = I$, so $\begin{bmatrix} X \\ U \end{bmatrix} \in \mathcal{W}$ and $\xi = [X, I, 0] \in \mathcal{C}_I$. Note that $\operatorname{ind}\xi = 1$ and $(\widetilde{A}_{11} - \widetilde{A}_{12}\Lambda)^{-1}$ is stable, so ξ is stable. From Theorem 1, part (1), the state transition matrix is

$$\Phi = T \begin{bmatrix} \left(\widetilde{A}_{11} - \widetilde{A}_{12}\Lambda\right)\exp\left(\widetilde{A}_{11} - \widetilde{A}_{12}\Lambda\right) & 0 \\ -\Lambda\left(\left(\widetilde{A}_{11} - \widetilde{A}_{12}\Lambda\right)\exp\left(\widetilde{A}_{11} - \widetilde{A}_{12}\Lambda\right) + \delta I\right) & -\delta I \end{bmatrix} T^{-1}, \tag{25}$$

so

$$\Phi X = T \begin{bmatrix} \exp\left(\widetilde{A}_{11} - \widetilde{A}_{12}\Lambda\right) & 0 \\ -\Lambda\exp\left(\widetilde{A}_{11} - \widetilde{A}_{12}\Lambda\right) & 0 \end{bmatrix} T^{-1}. \tag{26}$$

Letting

$$\begin{bmatrix} \widetilde{x}_{01} \\ \widetilde{x}_{02} \end{bmatrix} = T^{-1}x_0,$$

we obtain the solution of (16):

$$x = T \begin{bmatrix} I \\ -\Lambda \end{bmatrix} \exp\left(\widetilde{A}_{11} - \widetilde{A}_{12}\Lambda\right)\widetilde{x}_{01}.$$

THEOREM 7.

(1) *$\mathcal{C}_s$ contains a stable point iff (A, B) is stabilizable.*

(2) *If $\xi \in \mathcal{C}_s$ is stable, then $\operatorname{ord}\xi \geq n - m$ with equality iff $\xi = [X, I, 0]$, where X has the structure* (23).

Proof. See [6, Thms. 4.2 and 4.3]. ∎

We are also interested in the zeroth-order closed-loop limits

$$\mathcal{C}_0 = \left\{ \xi \in \mathcal{C} \;\middle|\; \operatorname{ord} \xi = 0 \right\}.$$

$\mathcal{C}_0$ corresponds precisely to those $\xi = [X, I, XB] \in \mathcal{C}_I$ with X nilpotent. From Theorem 1, part (1), the state transition matrix is

$$\Phi = -\sum_{i=0}^{q-1} \delta^{(i)} X^i, \tag{27}$$

so the solution of (16) is

$$x = \Phi X x_0 + \Phi * v = -\sum_{i=0}^{n-1} X^{i+1} B v^{(i)} - \sum_{i=1}^{n-1} \delta^{(i-1)} X^i x_0.$$

The system corresponds to successive differentiation of the input v plus a "noise" term.

THEOREM 8.

(1) *$\mathcal{C}_0$ is nonempty iff (A, B) is controllable.*

(2) *If (A, B) is controllable and $m = 1$, $\mathcal{C}_0$ is a singleton.*

(3) *If (A, B) is controllable, $m = 1$, $\xi_k \in \mathcal{C}_r$, and all eigenvalues λ_{ik} of ξ_k satisfy $|\lambda_{ik}| \to \infty$, then ξ_k converges to the unique point in $\mathcal{C}_0$.*

(4) *If (A, B) is controllable and $m > 1$, $\mathcal{C}_0$ is uncountable and unbounded (as a subset of $\mathcal{W}$).*

(5) *Every $\xi \in \mathcal{C}_0$ satisfies* $\operatorname{ind} \xi \geq \frac{n}{m}$.

Proof. See [6, Thms. 5.1–5.3]. □

Next, we consider $\mathcal{C}_r$ approximations $[I, A - BK_k, B]$ to certain points in $\mathcal{C}_s$. This is important in applications, since points with singular X can be achieved as limits only as $\|K_k\| \to \infty$ in (3). In view of (12), the closed-loop system (3) can be written equivalently as

$$(A - BK_k)^{-1} \dot{x} = x + (A - BK_k)^{-1} B v + \delta (A - BK_k)^{-1} x_0, \tag{28}$$

yielding state transition matrix

$$\Phi_k = (A - BK_k) \exp (A - BK_k) \tag{29}$$

and solution

$$x_k = \Phi_k (A - BK_k)^{-1} x_0 + \Phi_k * Bv. \tag{30}$$

We are interested in finding a sequence $\{K_k\}$ that yields not only convergence of $[I, A - BK_k, B]$ in $\mathcal{C}$, but also the strongest possible convergence of the forced and natural response in (30).

We begin by considering stable systems.

THEOREM 9. *Let $\xi \in \mathcal{C}_s$ be stable with* $\operatorname{ord} \xi = n - m$, *and let*

$$K_k = \begin{bmatrix} \widetilde{A}_{21} + k\Lambda & \widetilde{A}_{22} + kI \end{bmatrix} T^{-1}, \tag{31}$$

$$\xi_k = [I, A - BK_k, B].$$

Then

(1) $\xi_k \to \xi$,

(2) $\Phi_k (A - BK_k)^{-1}$ *is uniformly bounded,*

(3) $\Phi_k \to \Phi$ *uniformly on* $[\varepsilon, \infty)$ *for every* $\varepsilon > 0$,

(4) $\Phi_k \to \Phi$ *weak*,*

where Φ *is given by* (25).

Proof.

(1)–(3) See [6, Thm. 6.2].

(4) From (2), (3), $(A - BK_k)^{-1} \Phi_k \to X\Phi$ weak*. Since differentiation is weak* continuous,

$$\Phi_k = (A - BK_k)^{-1} \dot{\Phi}_k - \delta I \to X\dot{\Phi} - \delta I = X. \qquad \square$$

The results of [1] can be interpreted in terms of Theorems 7 and 9. In [1], the special case

$$K_\mu = -\frac{1}{\mu} K \tag{32}$$

is considered, where K is a fixed matrix and $\mu > 0$ is small. Adopting (21) and (22) and setting

$$\begin{bmatrix} \widetilde{K}_1 & \widetilde{K}_2 \end{bmatrix} = KT,$$

it is assumed in [1, equations (32) and (33)] that $\widetilde{K}_2$ and $\widetilde{A}_{11} - \widetilde{A}_{12}\widetilde{K}_2^{-1}\widetilde{K}_1$ are stable. Under these conditions, (32) constitutes an alternative to (31). Indeed, define

$$\Gamma_\mu = \mu\widetilde{A}_{22} + \widetilde{K}_2, \quad \Delta_\mu = \widetilde{A}_{11} - \widetilde{A}_{12}\Gamma_\mu^{-1}\left(\mu\widetilde{A}_{21} + \widetilde{K}_1\right),$$

and note that Γ_μ and Δ_μ are stable for small $\mu > 0$. Block matrix inversion reveals

$$\begin{aligned} X &= T \begin{bmatrix} \Delta_\mu^{-1} & -\mu\Delta_\mu^{-1}\widetilde{A}_{12}\Gamma_\mu^{-1} \\ -\Gamma_\mu^{-1}\left(\mu\widetilde{A}_{21} + \widetilde{K}_1\right)\Delta_\mu^{-1} & \mu\left(\Gamma_\mu^{-1} + \Gamma_\mu^{-1}\left(\mu\widetilde{A}_{21} + \widetilde{K}_1\right)\Delta_\mu^{-1}\widetilde{A}_{12}\Gamma_\mu^{-1}\right) \end{bmatrix} T^{-1} \\ &\to T \begin{bmatrix} \left(\widetilde{A}_{11} - \widetilde{A}_{12}\widetilde{K}_2^{-1}\widetilde{K}_1\right)^{-1} & 0 \\ -\widetilde{K}_2^{-1}\widetilde{K}_1\left(\widetilde{A}_{11} - \widetilde{A}_{12}\widetilde{K}_2^{-1}\widetilde{K}_1\right)^{-1} & 0 \end{bmatrix} T^{-1}, \end{aligned}$$

which is the same as (23) with $\Lambda = \widetilde{K}_2^{-1}\widetilde{K}_1$. Although the structures (31) and (32) are slightly different, the methods of [6, Thm. 6.2] can easily be modified to prove Theorem 9 relative to (32). Note that, in [1], only asymptotic stability for each $\mu > 0$ is actually proven.

Now consider zeroth-order systems $\xi \in \mathcal{C}_0$. Theorem 8, part (1), guarantees that (A, B) is controllable. From [15, pp. 342–343], there exist $\widetilde{K} \in \mathbb{R}^{m\times n}, w \in \mathbb{R}^m$ such that $(A - B\widetilde{K}, Bw)$ is controllable with $A - B\widetilde{K}$ nilpotent. Thus there exists a nonsingular N such that

$$N^{-1}\left(A - B\widetilde{K}\right)N = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix}, \quad N^{-1}Bw = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

THEOREM 10. *Let*

$$\beta_{ik} = \binom{n}{i} k^{n-i}, \quad \widehat{K}_k = \begin{bmatrix} \beta_{0k} & \cdots & \beta_{n-1,k} \end{bmatrix},$$

$$K_k = \widetilde{K} + w\widehat{K}_k N^{-1},$$

$$\xi_k = [I, A - BK_k, B].$$

Then

(1) ξ_k *converges to a point in* $\mathcal{C}_0$,

(2) $\Phi_k \to \Phi$ *uniformly on* $[\varepsilon, \infty)$ *for every* $\varepsilon > 0$,

(3) $\Phi_k \to \Phi$ *weak*,*

where Φ *is given by* (27).

Proof.

(1) From Theorem 5, part (2), it suffices to prove that $(A - BK_k)^{-1} \to X$ for some nilpotent X. This follows by the same arguments as in [6, Thm. 6.3].

(2), (3) See [6, Thm. 6.3]. ∎

Note that, in Theorem 10, boundedness of the natural response matrix $\Phi_k(A - BK_k)^{-1}$ was dropped. This is a consequence of the appearance of impulses in Φ when $\xi \in \mathcal{C}_0$ and $X \neq 0$. We can, in fact, prove a stronger result, which demonstrates the disastrous effect of driving the system to a limit with $\operatorname{ord} \xi < n - m$.

THEOREM 11. *Let* $m < n$, $1 < p \leq \infty$, *and* $\xi_k \in \mathcal{C}$ *be stable for all* k. *If the eigenvalues* λ_{ik} *of* ξ_k *satisfy* $\max_i \{|\lambda_{ik}|\} \to \infty$ *as* $k \to \infty$, *then* $\|\Phi_k X_k\|_p \to \infty$.

Proof. See [6, Thm. 6.4]. ∎

5. The limiting compensator. The state feedback law (2) may be written

$$\begin{bmatrix} I & K \end{bmatrix} \begin{bmatrix} u \\ x \end{bmatrix} = v. \tag{33}$$

This suggests that compensators of the form (2) are naturally identified with points $[I, K]$ in the Grassmanian $\mathcal{G}_m(\mathbb{R}^{m+n})$. In the proof of Theorem **??**, we considered the maps $g : \mathcal{G}_n(\mathbb{R}^{2n}) \to \mathcal{G}_n(\mathbb{R}^{2n+m})$ and $h : \mathcal{G}_m(\mathbb{R}^{m+n}) \to \mathcal{G}_n(\mathbb{R}^{2n})$ defined by

$$g\left(\left[\widetilde{X}, Y\right]\right) = \left[\widetilde{X}T^{-1}, \widetilde{X}T^{-1}A - Y, \widetilde{X}T^{-1}B\right], \tag{34}$$

$$h([Z_1, Z_2]) = \left(\begin{bmatrix} I & 0 \\ 0 & Z_1 \end{bmatrix}, \begin{bmatrix} 0 \\ Z_2 \end{bmatrix}\right), \tag{35}$$

where T is given by (21). The composition $f = g \circ h$ was shown to be an analytic diffeomorphism between the manifolds $\mathcal{G}_m(\mathbb{R}^{m+n})$ and $\mathcal{C}_e = f(\mathcal{G}_m(\mathbb{R}^{m+n}))$, with $\mathcal{C}_e$ regular in $\mathcal{G}_n(\mathbb{R}^{2n+m})$. Consider the open, dense submanifolds $\mathcal{F} = f^{-1}(\mathcal{C})$ and $\mathcal{F}_r = f^{-1}(\mathcal{C}_r)$ of $\mathcal{G}_m(\mathbb{R}^{m+n})$, and let $\mathcal{F}_s = f^{-1}(\mathcal{C}_s)$. The next result establishes basic properties of state feedback (33).

THEOREM 12.

(1) $\mathcal{F}_r = \{[I, K] \in \mathcal{G}_m(\mathbb{R}^{m+n}) \mid K \in \mathbb{R}^{m \times n}\}$.

(2) $\mathcal{F}_s = \{[Z_1, Z_2] \in \mathcal{F} \mid \det Z_1 = 0\}$.

Proof. (1) The result follows by the calculation

$$\begin{aligned} f([I,K]) &= g(h([I,K])) \\ &= g\left(\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}, \begin{bmatrix} 0 \\ K \end{bmatrix}\right) \\ &= \left[T^{-1}, T^{-1}A - T^{-1}BK, T^{-1}B\right] \\ &= [I, A - BK, B]. \end{aligned}$$

(2) This follows from $\mathcal{F}_r = \{[Z_1, Z_2] \in \mathcal{G}_m(\mathbb{R}^{m+n}) \mid \det Z_1 \neq 0\}$ and $\mathcal{F}_s = \mathcal{F} - \mathcal{F}_r$. ∎

The properties of f guarantee that, if K_k is any sequence of feedback matrices such that the closed-loop system (3) converges in $\mathcal{C}$, then the sequence $[I, K_k]$ also converges in $\mathcal{G}_m(\mathbb{R}^{m+n})$. By Theorem 12, degeneration of (3) to a point in $\mathcal{C}_s$ occurs iff $[I, K_k]$ converges to a point in $\mathcal{F}_s$. In other words, the limiting compensator always exists, and it is singular iff the limiting closed-loop system is singular. Compensators in $\mathcal{F}_s$ are not physically realizable, since they correspond to feedback laws of the form

$$Z_1 u = -Z_2 x + v$$

with Z_1 singular. Yet, as a mathematical object, each compensator in $\mathcal{F}$ determines a well-defined closed-loop system.

For the special case of minimum-order stable limits, as in Theorem 7, we can obtain the form of Z_1 and Z_2 explicitly.

THEOREM 13. *If* $\xi = [X, I, XB]$ *is given by* (23), *then* $f^{-1}(\xi) = \left[0, \begin{bmatrix} \Lambda & I \end{bmatrix}\right]$.

Proof. Choose a representative $\begin{bmatrix} Z_1 & Z_2 \end{bmatrix}$ for $f^{-1}(\xi)$. From (23), (34), and (35),

$$\begin{bmatrix} I & 0 \\ 0 & Z_1 \end{bmatrix} T^{-1} = MT \begin{bmatrix} \left(\widetilde{A}_{11} - \widetilde{A}_{12}\Lambda\right)^{-1} & 0 \\ -\Lambda\left(\widetilde{A}_{11} - \widetilde{A}_{12}\Lambda\right)^{-1} & 0 \end{bmatrix} T^{-1}$$

for some nonsingular M. Hence, $Z_1 = 0$ and

$$\begin{bmatrix} \widetilde{A}_{11} - \widetilde{A}_{12}\Lambda \\ 0 \end{bmatrix} = MT \begin{bmatrix} I \\ -\Lambda \end{bmatrix}.$$

Letting

$$\begin{bmatrix} \widetilde{M}_{11} & \widetilde{M}_{12} \\ \widetilde{M}_{21} & \widetilde{M}_{22} \end{bmatrix} = MT,$$

we obtain $\widetilde{M}_{21} = \widetilde{M}_{22}\Lambda$. Also, from (34) and (35),

$$\begin{bmatrix} 0 \\ Z_2 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} T^{-1}A - M,$$

so

$$Z_2 = -\widetilde{M}_{22}\begin{bmatrix} \Lambda & I \end{bmatrix}.$$

Since $\operatorname{rank}\begin{bmatrix} Z_1 & Z_2 \end{bmatrix} = m$, $\widetilde{M}_{22}$ is nonsingular. Premultiplication of $\begin{bmatrix} Z_1 & Z_2 \end{bmatrix}$ by $-\widetilde{M}_{22}^{-1}$ yields the desired result. ∎

We conclude this section by examining behavior of the input function u under high-gain feedback. For simplicity, we will consider only the case where $v = 0$. If we apply the feedback gains K_k to (3), then both u and x depend on k and are related by the feedback law

$$u_k = K_k x_k.$$

In Theorems 9 and 10, we established cases under which the state-transition matrix Φ_k converges in two different topologies. More generally, consider the linear subspace

$$\mathcal{D}_0' = C[0,\infty) + \operatorname{span}\left\{\delta, \dot{\delta}, \ddot{\delta}, \dots\right\} \subset \mathcal{D}_+',$$

where $C[0,\infty)$ is the set of continuous functions on $\mathbb{R}$ with support in $[0,\infty)$. Both weak* convergence and uniform convergence on every $[\varepsilon,\infty)$ correspond to specific topologies on $\mathcal{D}_0$. It is easy to show that both make $\mathcal{D}_0$ a topological vector space.

THEOREM 14. *Suppose $\mathcal{D}_0$ is given a topology that makes it a topological vector space. If $[I, A - BK_k, B] \to [X, I, XB] \in \mathcal{C}_I$ and $\Phi_k \to \Phi$ in $\mathcal{D}_0$, then $u_k \to U\Phi x_0$ in $\mathcal{D}_0$.*

Proof. From (29) and Theorem 5, part (2),

$$\begin{aligned} U\Phi + K_k (A - BK_k)^{-1} \Phi_k &= \left(U + K_k (A - BK_k)^{-1}\right)\Phi + K_k (A - BK_k)^{-1} (\Phi_k - \Phi) \\ &\to 0, \end{aligned}$$

so

$$u_k = -K_k (A - BK_k)^{-1} \Phi_k x_0 \to U\Phi x_0. \qquad \square$$

Theorem 14 can be extended to $v \neq 0$ through the choice of an appropriate space of inputs v and exploiting the properties of the convolution operator. We leave the details to the reader.

6. Application to cheap control. A classical problem in the theory of linear-quadratic optimal control is the "cheap control" problem, where an input function $u^*(t)$ is sought to minimize the cost

$$J(\varepsilon) = \int_0^\infty x^T x + \varepsilon u^T u dt$$

subject to (1), with fixed initial condition x_0 and small $\varepsilon \geq 0$. For $\varepsilon > 0$, this problem has been extensively studied (e.g., see [13], [7], [12], [14]). The solution is obtained by constructing the unique positive definite symmetric solution $P(\varepsilon)$ of the algebraic Riccati equation

$$P(\varepsilon)A + A^T P(\varepsilon) - \frac{1}{\varepsilon} P(\varepsilon) B B^T P(\varepsilon) + I = 0.$$

Then, for each x_0, the optimal u and x are related by the feedback law

$$u^* = -\frac{1}{\varepsilon} B^T P(\varepsilon) x^*,$$

yielding the closed-loop system

$$\left(A - \frac{1}{\varepsilon} B B^T P(\varepsilon)\right)^{-1} \dot{x}^* = x^* + \delta \left(A - \frac{1}{\varepsilon} B B^T P(\varepsilon)\right)^{-1} x_0$$

(cf. (28)).

For $\varepsilon = 0$, we adopt (21) and (22), let

$$\begin{bmatrix} \widetilde{Q}_{11} & \widetilde{Q}_{12} \\ \widetilde{Q}_{12}^T & \widetilde{Q}_{22} \end{bmatrix} = T^T T,$$

and let Γ be the unique positive definite symmetric solution of the reduced Riccati equation

$$\Gamma\left(\widetilde{A}_{11} - \widetilde{A}_{12}\widetilde{Q}_{22}^{-1}\widetilde{Q}_{12}^T\right) + \left(\widetilde{A}_{11} - \widetilde{A}_{12}\widetilde{Q}_{22}^{-1}\widetilde{Q}_{12}^T\right)^T \Gamma - \Gamma\widetilde{A}_{12}\widetilde{Q}_{22}^{-1}\widetilde{A}_{12}^T\Gamma + \widetilde{Q}_{11} - \widetilde{Q}_{12}\widetilde{Q}_{22}^{-1}\widetilde{Q}_{12}^T = 0.$$

Setting

$$\Lambda = \widetilde{Q}_{22}^{-1}\left(A_{12}^T\Gamma + \widetilde{Q}_{12}^T\right) \tag{36}$$

leads to values of X, U, and Φ according to (23), (24), and (25). It is shown in [14, Cor. 2.6.1] that $J(0)$ is minimized, subject to (1), by $x^* = \Phi x_0$ and $u^* = U\Phi x_0$. Furthermore, [14, Thm. 2.7.1] indicates that

$$\left(A - \frac{1}{\varepsilon}BB^T P(\varepsilon)\right)^{-1} \to X$$

as $\varepsilon \to 0^+$. These facts are now interpreted in the context of the present paper.

THEOREM 15. *For each $\varepsilon \geq 0$, let $\xi_\varepsilon^* \in \mathcal{C}_r$ be the optimal closed-loop system in the cheap control problem. Then $\xi_\varepsilon^* \to \xi_0^*$ in $\mathcal{C}$ as $\varepsilon \to 0^+$, where ξ_0^* is stable and $\operatorname{ord}\xi_0^* = n - m$. The limiting system ξ_0^* is determined uniquely by the singular compensator $\left[0, \begin{bmatrix}\Lambda & I\end{bmatrix}\right] \in \mathcal{G}_m(\mathbb{R}^{m+n})$ as in Theorem 13, where Λ is given by (36).*

7. Conclusions. In this paper, we have developed a general theory of high-gain state feedback, retaining a fixed state coordinate system. For many control problems, this approach lends itself to a more natural interpretation of results than if the coordinates were allowed to vary with the feedback gain. Relationships to other seminal work in the area have been drawn. As in our earlier similar work on high-gain observers, the present paper has focused primarily on system parameter convergence and behavior of solutions, particularly in the cases of stable and zeroth-order limits. A unique aspect of our results is that even infinite state feedback gains are identified with specific mathematical objects. As future work, we hope to be able to extend our results to observer-based output feedback and, ultimately, to general output feedback.

REFERENCES

[1] K. D. YOUNG, P. KOKOTOVIC, AND V. UTKIN, *A singular perturbation analysis of high-gain feedback systems*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 931–938.

[2] D. HINRICHSEN AND J. O'HALLORAN, *Orbit closures of matrix pencils and system limits under high gain feedback*, in Proceedings of the 29th IEEE Conference on Decision and Control, Honolulu, HI, 1990, pp. 550–560.

[3] D. HINRICHSEN AND J. O'HALLORAN, *A pencil approach to high gain feedback and generalized state space systems*, Kybernetika, 31 (1995), pp. 109–139.

[4] D. HINRICHSEN AND J. O'HALLORAN, *Limits of generalized state space systems under proportional and derivative feedback*, Math. Control Signals Systems, 10 (1997), pp. 97–124.

[5] J. D. COBB, *Fundamental properties of the manifold of singular and regular systems*, J. Math. Anal. Appl., 120 (1986), pp. 328–353.

[6] J. D. Cobb, *A unified theory of full-order and low-order observers based on singular system theory*, IEEE Trans. Automat. Control, 39 (1994), pp. 2497–2502.
[7] B. A. Francis and K. Glover, *Bounded peaking in the optimal linear regulator with cheap control*, IEEE Trans. Automat. Control, 23 (1978), pp. 608–617.
[8] F. R. Gantmacher, *The Theory of Matrices*, Vol. II, Chelsea, New York, 1959.
[9] G. Doetsch, *Introduction to the Theory and Application of the Laplace Transformation*, Springer-Verlag, New York, Heidelberg, 1974.
[10] F. Brickell and R. S. Clark, *Differentiable Manifolds*, Van Nostrand Reinhold, London, 1970.
[11] I. M. Gelfand and G. E. Shilov, *Generalized Functions*, Vol. I, Academic Press, New York, 1964.
[12] B. A. Francis, *The optimal linear-quadratic time-invariant regulator with cheap control*, IEEE Trans. Automat. Control, 24 (1979), pp. 616–621.
[13] H. Kwakernaak and R. Sivan, *The maximally achievable accuracy of linear optimal regulators and linear optimal filters*, IEEE Trans. Automat. Control, AC-17 (1972), pp. 79–86.
[14] N. Safari-Shad, *Deterministic L^2/H^2 Optimization of Linear Dynamical Systems*, Ph.D. Thesis, University of Wisconsin-Madison, Madison, WI, 1992.
[15] C.-T. Chen, *Linear System Theory and Design*, Oxford University Press, Oxford, UK, 1984.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 6, pp. 2233–2252

RENDEZVOUS IN HIGHER DIMENSIONS*

STEVE ALPERN† AND VIC BASTON‡

Abstract. Two players are placed on the integer lattice Z^n (consisting of points in n-dimensional space with all coordinates integers) so that their vector difference is of length 2 and parallel to one of the axes. Their aim is to move to an adjacent node in each period, so that they meet (occupy same node) in least expected time $R(n)$, called the rendezvous value. We assume they have no common notion of directions or orientations (i.e., no common notion of "clockwise"). We extend the known result $R(1) = 3.25$ of Alpern and Gal to obtain $R(2) = 197/32 = 6.16$, and the bounds $2n \leq R(n) \leq \left(32n^3 + 12n^2 - 2n - 3\right)/12n^2$. For $n = 2$ we characterize the set of all optimal strategies and show that none of them simultaneously maximizes the probability of meeting by time t for all t. This behavior differs from that found by Anderson and Fekete, and the authors, for the related problem where the players are initially placed at diagonals of one of the squares of the lattice Z^2.

Key words. rendezvous, search, game, plane

AMS subject classifications. 90B40, 90D99

DOI. 10.1137/S0363012904443531

1. Introduction. This paper generalizes the (player-asymmetric) rendezvous problem on the line [6] to higher dimensions. That problem, which we call here $\Gamma(1)$, asks how two unit-speed players initially placed a fixed distance D apart on the line, and faced in random directions (which they each call "forward"), can meet in minimum expected time. Neither knows the direction to the other, nor do they have a common notion of a "positive" or "forward" direction. By taking $D = 2$ we may model the line as the lattice $Z = Z^1$ of integers, with the players moving to adjacent nodes (integers) in each integer period, until the first (meeting) time that they occupy the same node. (Taking D even ensures they will not pass each other on the line without meeting.) A strategy for each player can be taken as a sequence of F's (for forward) and B's (for backward), which determines a player's motion relative to his starting node and his initial direction. Adopting Player I's coordinate system, with his initial node as 0 and his forward direction to the right, we equiprobably place Player II at the nodes -2 and +2. We may consider that there are four *agents* of Player II, two starting at -2 (facing in either direction) and two at +2. It was shown in [6] that the strategy pair

$$(1) \qquad (F, F, B, B, B, B) \text{ (for Player I) and } (F, B, B, B, F, F) \text{ (for II)}$$

is optimal, having minimum expected meeting time 13/4 (called the *rendezvous value* and, since the line has dimension $n = 1$, denoted here by $R(1) = 13/4 = 3.25$). It is easy to verify the expected meeting time 13/4 for this strategy pair (though establishing optimality is obviously harder). If the two players start facing each

*Received by the editors May 6, 2004; accepted for publication (in revised form) May 25, 2005; published electronically February 3, 2006.
http://www.siam.org/journals/sicon/44-6/44353.html
†Department of Mathematics, The London School of Economics, Houghton Street, WC2A London, UK (alpern@lse.ac.uk).
‡University of Southampton, SO17 1BJ Southampton, UK (vjdb@uk.ac.soton.maths) and Department of Mathematics, The London School of Economics, Houghton Street, WC2A 2AE London, UK.

other (that is, Player II starts at +2 facing left), then since they both start with a forward move they meet at time $t = 1$ at node +1. This corresponds to the upper right entry in the *meeting time matrix* (2) for the four agents of Player II, and the reader should verify the other three. The expected meeting time is therefore $(2 + 1 + 6 + 4)/4 = 13/4$.

(2)

location\direction	$\rightarrow$	$\leftarrow$
+2	2	1
−2	6	4

The optimal strategy (1) for $\Gamma(1)$ is almost a *wait for mommy* (WFM) strategy in which Player I (Mommy) optimally searches out the possible initial starting points of II (Child), while II stays still. In fact (F, F, B, B, B, B) is such a Mommy strategy, though II does not wait. However, note that II is indeed back at his starting node at the two times $t = 2, 6$ that I might search it. At the intermediate time $t = 4$, Player II searches out a possible starting point of I. So we may call this strategy an *alternating wait for mommy* (AWFM) strategy, in which the players alternate taking the role of Mommy and searching out starting points of the other. Such strategies will play a role in the higher dimensional games. (The analysis of more general linear rendezvous problems can be found in [11] and [5].)

To generalize this problem to n dimensions, we consider the game $\Gamma(n)$ in which the two players are initially placed in the lattice Z^n (n-vectors of integers, with nodes adjacent if they differ by 1 in one coordinate and are identical in the others) so that their vector difference is of length 2 and parallel to one of the coordinate axes. We assume that they do not know the initial location of the other, and that they have no common ordering of the axes (in particular, no common notion of clockwise for the planar problem, $n = 2$). Note that while restricting the players to move in the lattice Z^n rather than the Euclidean space R^n does not essentially change the problem when $n = 1$, it certainly does so for $n > 1$.

The organization and main results of the paper are as follows. In section 2, we give a rigorous definition of the rendezvous problem on the planar lattice Z^2, for a general distribution of the initial vector between the players. The presentation is similar but simpler than that given in [3] and contains some general optimality conditions.

In sections 3 and 4 we solve the planar rendezvous problem $\Gamma(2)$, establishing (Theorem 14) that the rendezvous value is $R(2) = 197/32$. Although the AWFM strategy is not optimal, a variation of it called the nearly alternating wait for mommy (NAWFM) strategy (drawn in Figure 4) is optimal. We also determine the full set of optimal strategies (Theorem 15). We show that, unlike the situation in $\Gamma(1)$, no strategy in $\Gamma(2)$ is *uniformly optimal*, in the sense that it simultaneously maximizes the probability of meeting by time t, for all t.

The results on planar rendezvous which we establish in section 4 may be contrasted with related results with different assumptions. It is shown in [4] that the players can do better (lower expected meeting time) if they have a common notion of clockwise. Our results may also be contrasted with those for the planar rendezvous model introduced by Anderson and Fekete [9] and extended by the authors [3], for the *diagonal start problem* (players start at opposite diagonals of a square in the Z^2 lattice). In that model all optimal strategies are uniformly optimal and having a common notion of clockwise does not help the players.

In section 5, we obtain (Theorem 16) bounds on the rendezvous value $R(n)$, showing that it is asymptotically bounded above by $8n/3$. These results are the first obtained for rendezvous in more than two dimensions, and are reported in the book [7],

which attributes them to an earlier version of this paper. Section 6 gives a discussion of some of our results.

The version of rendezvous search adopted in this paper is the so-called *player-asymmetric* (or just asymmetric) version, in which the players can have distinct strategies. For example, they may have mobile phones and agree which role each will take, e.g., one Mommy and one Child in the WFM strategy pair. The other, *player-symmetric* version (where they both must adopt the same mixed strategy) will not be discussed here, but can be found in [1], [8], and [10].

A general survey of the rendezvous search problem can be found in [2] and a unified presentation of results up to 2003 can be found in Book II of [7]. The main works on planar rendezvous are [9] and [13], together with our work cited above.

2. Rendezvous in the plane: Strategies and agents. In this section we give a formal presentation of rendezvous on the planar lattice Z^2 without a common notion of clockwise. A similar but more general presentation is given in [3]. The (lattice) distance d between two nodes is defined as the sum of the edges in a shortest connecting path, or equivalently $d\left((z_1, z_2), (w_1, w_2)\right) = |z_1 - w_1| + |z_2 - w_2|$. At time $t = 0$, nature places the two players on even nodes with the vector v from I to II drawn from a given distribution. (A node $z \in Z^2$ is called *even* if the sum of its coordinates is even; otherwise it is called *odd*.) In every time period each player must move to an adjacent node. This restriction, combined with the "even distance" initial placement (originating in the interval network of Howard [12]), ensures that the two players will always have the same parity and that they cannot pass each other on an edge without meeting at a node.

We analyze the progress of the game in terms of Player I's coordinate system (and sense of clockwise). In this perspective, the initial random placement is achieved by nature placing I at the origin facing north ($\mathcal{N}$) and placing II at the even node v_i, with probability p_i, $i = 1, \ldots, K$, facing equiprobably in any of the four possible directions, and with either the same or opposite notion of clockwise as Player I. Aside from this section, dealing with arbitrary initial locations v_i, we will be mainly concerned with the game $\Gamma(2)$ described in the Introduction. We denote this general planar game by $\Gamma = \Gamma(v_1, \ldots, v_k; p_1, \ldots, p_K)$. In terms of this more general framework, the game $\Gamma(2)$ can be defined in the following way.

DEFINITION 1. *The game* $\Gamma(2)$ *begins with Player I initially placed at the origin* $(0, 0)$ *and II initially placed equiprobably at one of the four nodes* $v_1 = (0, 2)$, $v_2 = (2, 0)$, $v_3 = (0, -2)$, $v_4 = (-2, 0)$.

When the game begins, the players have no common notion of locations or directions in the plane, and no common notion of clockwise. As observers, we adopt I's coordinate system. The orientations of Player II can be seen as transformations (or rigid motions, or symmetries) of the "standard orientation" of Player I. Figure 1 shows the eight equiprobable orientations.

The orientations in the top row ($k = 0$) all have the same notion of clockwise as Player I (the usual one), with the upper left orientation the standard one, and the one in column j obtained from it by applying the 90° clockwise rotation $\mathcal{R}$ to it j times. The orientations in the bottom row ($k = 1$) have the other (reversed) notion of clockwise, and can be obtained from the one above by a single application of the reflection ϕ about the vertical axis. Formally,

$$\mathcal{R}^j = \text{clockwise rotation by angle } j\ \pi/2,\ j = 0, 1, 2, 3, \tag{3}$$

$$\phi(w_1, w_2) = (-w_1, w_2), \text{ and the eight orientations of II are} \tag{4}$$

$$\phi^k \mathcal{R}^j,\ j = 0, 1, 2, 3;\ k = 0, 1. \tag{5}$$

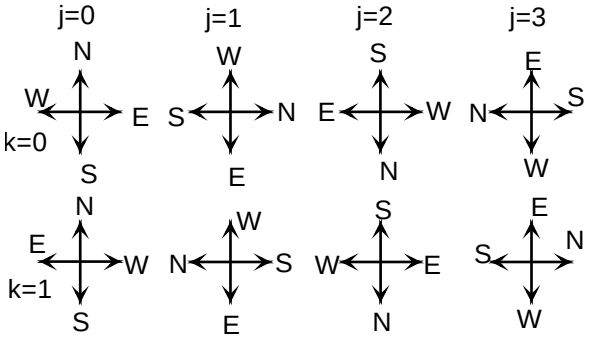


FIG. 1. *Eight orientations in the plane.*

The orientations determine how Player II will move when given an instruction (strategy). For example, if the instruction is for II to go E (east), then (in Player I's coordinate system, which we also adopt as observers) he will go S (south) if he has orientation $j = 1, k = 0$; he will go N if he has orientation $j = 3$, $k = 1$.

We can now define a strategy and show how a pair of strategies determines a set of meeting times for the two players, one for each initial configuration (initial location v_i and initial orientation j, k).

DEFINITION 2. *A* strategy *for a player (in* Γ *or* $\Gamma(2)$*) is a sequence of directions* $D_t \in \{N = (0,1), E = (1,0), S = (0,-1), W = (-1,0)\}$, $t = 1, 2, \ldots$. *A player pursuing this strategy moves successively one unit in his direction* $D_0, D_1, \ldots$, *according to his initial orientation. Equivalently, it can be seen as his net displacement* $f(t)$ *at time* t *from his initial location, given by* $f(0) = (0,0)$ *and for* $t \geq 1$,

$$f(t) = \sum_{m=1}^{t} D_m. \tag{6}$$

So for example the strategy beginning $N, E, E,$ (I's strategy, thick grey line in Figure 2) corresponds to a net displacement function f with

$$[f(0), f(1), f(2), f(3)] = [(0,0), (0,1), (1,1), (2,1)]. \tag{7}$$

We shall deal with strategy pairs (f, g) where Player I adopts f and II adopts g. Sometimes we will use the symmetric notation (f_1, f_2). In this setting, the location of Player I at time t is simply $f(t)$, while the location of II (in I's coordinate system) depends on his initial configuration, as described below. If the initial configuration (i, j, k) gives Player II initial location v_i and orientation $\phi^k \mathcal{R}^j$, then the location of Player II at time t under strategy g is given by

$$g_{i,j,k}(t) = v_i + \phi^k \mathcal{R}^j (g(t)). \tag{8}$$

It is useful to note that if $g(t) = (x, y)$, then the displacements of the eight agents starting at v_i at time t are of the form $(\pm x, \pm y)$ or $(\pm y, \pm x)$. If none of the conditions $xy = 0$ (at least one 0) or $|x| = |y|$ hold, there are eight distinct displacements; if exactly one of these conditions holds there are four distinct displacements; if both hold ($x = y = 0$), all agents are back at their starting node v_i.

DEFINITION 3. *The* $8K$ *(in our example,* 32*) paths* $g_{i,j,k}$ *are called the* agents *of Player II. We call* $g_{i,j,k}$ *the* agent starting at v_i in direction j, with the same (if $k = 0$)

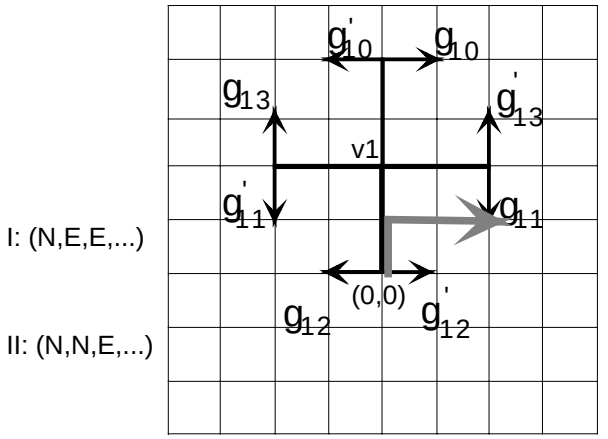


FIG. 2. *Meetings at $t \leq 3$ for (N,E,E,...), (N,N,E,...).*

or different (if $k = 1$) notion of clockwise as I. *Each agent from v_i is the actual path of Player II with probability $p_i/8$. We will sometimes use the simpler two subscript agent notation $g_{i,j} = g_{i,j,0}$ and $g'_{i,j} = g_{i,j,1}$.*

The time taken for agent $g_{i,j,k}$ to be met by Player I is called its *meeting time* and denoted

$$\omega_{i,j,k}(f,g) = \min\{t : f(t) = g_{i,j,k}(t)\}, \tag{9}$$

and the time required to meet *all* the agents is called *maximum time* $M(f,g)$, where

$$M(f,g) = \max_{i,j,k} \omega_{i,j,k}(f,g). \tag{10}$$

For each time t we calculate the number of agents x_t that Player I meets (for the first time) at time t as

$$x_t = x_t(f,g) = \#\{(i,j,k) : \omega_{i,j,k}(f,g) = t\}, \tag{11}$$

and we call the sequence $[x_1, x_2, \ldots, x_M]$ the *agent number profile*, or sometimes just *agent profile*, or just *profile*. Figure 2 illustrates several important concepts for the case where the strategy pair begins with (N,E,E,...) for I and (N,N,E,...) for II. Player I's path $(0,0),(0,1),(1,1),(2,1)$ for $t \leq 3$ is shown in a thick grey line. The possible paths taken by II if he starts at $v_1 = (0,2)$, that is, the 8 paths $g_{1,j,k}(t)$, $j = 0,\ldots 3$, $k = 0,1$, are drawn in medium black lines starting at $v_1 = (0,2)$. Note that all the paths $g_{1,j,0}(t) = g_{1,j}(t)$ in our notation make *right* turns at move 3, while the $g'_{1,j} = g_{1,j,1}$ make *left* turns. Note that Player I meets agents $g'_{1,2}$ and $g_{1,2}$ at node $(0,1)$ at time 1, and meets agent $g_{1,1}$ at node $(2,1)$ at time 3. Thus in our notation we have $\omega_{1,2,0} = \omega_{1,2,1} = 1$ and $\omega_{1,1,0} = 3$. Note that when there are no turns since the last return to a starting point (a single direction is repeated), agents with a common notion of that direction will be at the same location, so if one is met, the other will be also. Since Player I will not meet agents from any nodes other than v_1 by time 3, it can be seen that he meets two agents at time $t = 1$, no agents at time $t = 2$, and one agent at time $t = 3$, so that $x_1 = 2, x_2 = 0$, $x_3 = 1$, and the agent number profile begins $[2,0,1,\ldots]$. For comparison with a later table (18) we write this as a table of

$\omega_{i,j,k}$, with blank entries depending on the strategy for $t > 3$.

(12)

$\omega_{i,j,k}$	$k=0$				$k=1$			
j	0	1	2	3	0	1	2	3
$i=1$		3	1				1	

Given a strategy pair (f, g), the expected value of the meeting times $\omega_{i,j,k}$ is called the *expected meeting time* and denoted $T(f, g)$. Thus

$$T(f,g) = \frac{1}{8}\sum_{i,j,k} p_i \omega_{i,j,k}(f,g), \quad \text{or simply} \tag{13}$$

$$= \frac{1}{32}\sum_{i,j,k} \omega_{i,j,k}(f,g) \text{ for } \Gamma(2). \tag{14}$$

If $M = M(f, g)$ is finite, we may also calculate $T(f, g)$ for $\Gamma(2)$ by the meeting number sequence $x = [x_1, x_2, \ldots, x_M]$ as

$$T(f,g) = \frac{1}{32}\sum_{t=1}^{M} t\ x_t. \tag{15}$$

The rendezvous value R for the game Γ is the least expected time,

$$R(\Gamma) = \min_{f,g} T(f,g), \tag{16}$$

and any pair f, g achieving the minimum is called *optimal.*

We know that $R(\Gamma)$ is finite because M, and hence T, is finite for the WFM strategy.

A stronger notion of optimality is the following.

DEFINITION 4. *A strategy pair is called* uniformly optimal *if for all t it maximizes the probability that the players have met by time t. (Note that if there is a uniformly optimal strategy, then all optimal strategies must be uniformly optimal.)*

A uniformly optimal strategy maximizes the expected utility $U(\omega)$ of the meeting time ω as long as the utility function U is nonincreasing in ω (earlier meetings are preferred to later ones); an optimal strategy is only required to accomplish this for the particular utility function $U(\omega) = -\omega$.

The authors have shown that, given the order in which the agents are met, any optimal strategy pair is greedy in the sense that the next agent is met as quickly as possible, with both players taking geodesics to the next meeting point. The following result was proved without the assumption that the players must move in each period (that is, staying still was allowed), and hence shows that our assumption here that they must move does not take away any optimal strategies.

THEOREM 5 (see [3]). *Let (f_1, f_2) be an optimal strategy pair. Define $\omega^0 = 0$ and let $\omega^1 < \omega^2 < \cdots < \omega^N$ denote the associated set of meeting times with the $8K$ agents, listed in increasing order. Let d denote the graph distance on the lattice Z^2. Then*

$$d\left(f_i\left(\omega^{m-1}\right), f_i\left(\omega^m\right)\right) = \omega^m - \omega^{m-1}, \text{ for } i = 1, 2 \text{ and } m = 1, \ldots, N. \tag{17}$$

In other words, both players move in time-minimizing paths between consecutive meeting points. In particular, neither player ever stays still, and consequently both players are at even (odd) nodes at all even (odd) times. Furthermore, if the agent $g_{i,j,k}$ is met at time ω^m, then that meeting point $f_1(\omega^m) = g_{i,j,k}(\omega^m)$ is a lattice midpoint of the locations of I and agent $g_{i,j,k}$ of II at time ω^{m-1}, and hence occurs at the earliest possible time (given their locations at time ω^{m-1}).

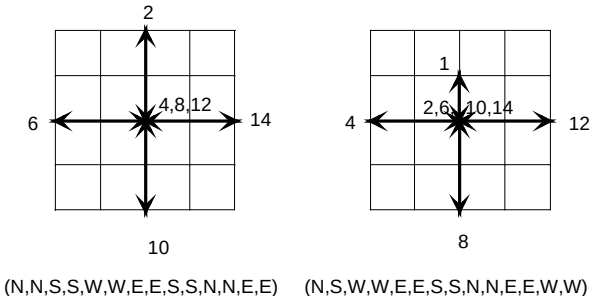


FIG. 3. *An AWFM strategy* $(\breve{f}, \breve{g})$ *for* $n = 2$.

3. Analysis of Γ(2). In this section we develop some fundamental lemmas for the problem $\Gamma(2)$. In the following section these lemmas will be applied to solve $\Gamma(2)$, in the sense of finding all optimal strategies and the rendezvous value. It will be convenient to introduce the notation $\mathcal{S} = \{(0,2),(2,0),(0,-2),(-2,0)\}$ for the set of Player II starting nodes, and $\mathcal{S}_0$ for the set $\mathcal{S}$ augmented by Player I's starting point $(0,0)$. Note that all nodes in $\mathcal{S}_0$ have even parity.

The main results of this section are Theorems 13, 14, and 15, which together determine the rendezvous value $R(\Gamma(2)) = 197/32$ and the complete set of optimal strategies for $\Gamma(2)$. Since we know players following optimal strategies will move in every period (never stay still), we will henceforth only consider strategies with this property. Since the players both start (in $\mathcal{S}_0$) at nodes of even parity, it follows that both are at even nodes at even times and at odd nodes at odd times.

3.1. Special strategies. We now define some particular strategies, AWFM (valid for all $\Gamma(n)$) and NAWFM (for $\Gamma(2)$). In [7], a family of strategies called AWFM was proposed for the general parallel start game $\Gamma(n)$. The players alternate taking the role of Mommy (searching out the starting points of the other) and Child (coming back to one's starting point to be found there when the other comes looking). We will use these strategies to analyze the $n-$dimensional case in the final section.

DEFINITION 6. *A strategy in* $\Gamma(n)$ *is called AWFM if* (a) *Player I successively visits the* $2n$ *possible starting locations of II (in any order) at times* $\mathcal{T}_1 = \{2, 6, \ldots, 2+4(2n-1)\}$, *while returning to his start* $(0,0)$ *at the intermediate times* $\mathcal{T}_2 = \{4, 8, \ldots, 4(2n-1)\}$, *and* (b) *Player II makes his first move a single unit in any direction, is back at his start at times* $\mathcal{T}_1$, *and visits all but one of the possible initial locations of I at times* $\mathcal{T}_2$. *The maximum time for this strategy is clearly* $M = 2 + 4(2n-1)$, *or* $M = 14$ *for* $n = 2$.

It was shown in [7] that, for $n = 2$, any AWFM strategy (as illustrated in Figure 3) gives the maximal probability of meeting by time t for any $t \leq 7$, and suggested that this strategy might in fact be optimal (like the $n = 1$ version, where optimality was established in [6]), or even uniformly optimal.

The meeting time sequence x for $(\breve{f}, \breve{g})$ (and indeed for any AWFM strategy) is given by

$$x = [2,6,0,6,0,6,0,4,0,4,0,2,0,2] \text{ with } T\left(\breve{f}, \breve{g}\right) = \frac{198}{32}.$$

The strategy $(\tilde{f}, \tilde{g})$ drawn in Figure 4 has a lower expected meeting time T than AWFM. Note that I's strategy $\tilde{f}$ is the same as that for AWFM $(\breve{f})$ up to time 10,

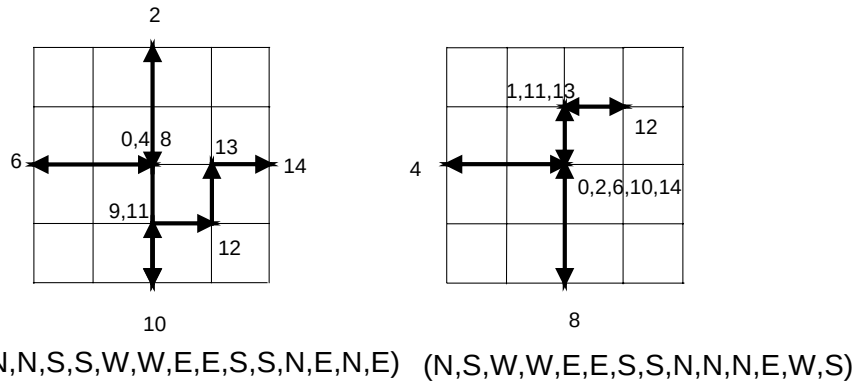


FIG. 4. *The NAWFM strategy* $(\tilde{f}, \tilde{g})$.

but then I goes from $(0,-2)$ to $(2,0)$ without going back through his starting point $(0,0)$. For this reason we call $(\tilde{f}, \tilde{g})$ the NAWFM strategy.

The expected meeting time for the NAWFM strategy pair $(\tilde{f}, \tilde{g})$ can be evaluated by considering the following table of $\omega_{i,j,k}(\tilde{f}, \tilde{g})$ and setting x_t to be the number of t's found in the table. The entries $\omega_{i,j,k}$ in the following table are calculated as in the earlier simpler table (12) which was based on the strategy of Figure 2.

(18)

$\omega_{i,j,k}\left(\tilde{f}, \tilde{g}\right)$	$k = 0$				$k = 1$			
j	0	1	2	3	0	1	2	3
$i = 1, (0,2)$	2	2	1	2	2	2	1	2
$i = 2, (2,0)$	4	8	12	13	14	12	4	8
$i = 3, (0,-2)$	10	4	8	10	9	10	8	10
$i = 4, (-2,0)$	6	6	4	6	4	6	6	6

This gives the agent number profile of

$$x = x\left(\tilde{f}, \tilde{g}\right) = [2,6,0,6,0,6,0,4,0,4,0,2,1,1] \tag{19}$$

(e.g., there are six 2's, so $x_2 = 6$), and hence by (14) or (15),

$$T\left(\tilde{f}, \tilde{g}\right) = \frac{197}{32}. \tag{20}$$

3.2. Nodes with multiple agents. It will turn out from our subsequent analysis that optimal strategies necessarily involve some simultaneous meetings of Player I with two or more agents of Player II. Since this necessitates several agents being simultaneously at a common node, we now analyze how this can occur (both for agents from the same starting node and from different starting nodes).

LEMMA 7 (same starting point). *Suppose that there are $m > 1$ distinct agents from a common starting node v who are all at the same node $z \neq v$ at the same time t. Then*

(i) *$m = 2$ and the agents have distinct notions of clockwise.*
(ii) *The line between v and z makes an angle $r\pi/4$ with the vertical, $r \in \{0,1,2,3\}$.*
(iii) *If r is odd, then the two agents have distinct notions of north (distinct j) and will move to distinct locations at the next move.*

Proof. (i) The four rotations $\mathcal{R}^j$ take any nonzero vector into four distinct vectors, so agents at the same location must have different notions of clockwise, and hence there can be at most two of them.

(ii) By part (i), the $m = 2$ agents at z have distinct notions of clockwise (k equal to 0 and 1) and so by (8) have respective locations

$$v + \mathcal{R}^j(g(t)) \text{ and } v + \phi\mathcal{R}^{j'}(g(t)), \text{ both equal to } z.$$

(Recall that ϕ is the reflection $(w_1, w_2) \to (-w_1, w_2)$ about the vertical axis, and $\mathcal{R}^j$ is the clockwise rotation by $j\pi/2$, $j = 0, 1, 2, 3$). Consequently both $\mathcal{R}^j$ and $\phi\mathcal{R}^{j'}$ take $g(t)$ into $w = z - v$, and hence w is a nonzero fixed point of $(\phi\mathcal{R}^{j'})(\mathcal{R}^j)^{-1} = \phi\mathcal{R}^{j'-j} = \phi\mathcal{R}^n$, $n = j' - j$. If $n = 0$, w lies on the vertical axis (invariant set for ϕ), so $r = 0$; if $n = 1$, $\phi\mathcal{R}^n(w_1, w_2) = (-w_2, -w_1)$, so the fixed points form the line $w_2 = -w_1$, an angle $3\pi/4$ with the vertical; if $n = 2$, then $\phi\mathcal{R}^n(w) = (w_1, -w_2)$ with the fixed point line $w_2 = -w_2$ (the horizontal axis), so $r = 2$; if $n = 3$, then the fixed set w of $\phi\mathcal{R}^n$ is the line $w_1 = w_2$, so $r = 1$. Note that in all cases, n and r have the same parity.

(iii) By the previous remark, if r is odd, then $n = j' - j$ is odd, so the rotations $\mathcal{R}^j$ and $\mathcal{R}^{j'}$ are not the same or opposite ($j \neq -j'$), and hence the moves of the two agents (namely $\mathcal{R}^j(g(t+1) - g(t))$ and $\phi\mathcal{R}^{j'}(g(t+1) - g(t))$) will be distinct, in fact at right angles to each other. (This can also be seen from Figure 1, which shows that two such agents will have no direction notion in common.) ∎

LEMMA 8 (different starting points). *Suppose that agents from $m > 1$ distinct starting points are simultaneously at a common node z. Then*

(i) *If $m = 2$, z lies on the (Euclidean) perpendicular bisector $B(a, b)$ of the two starting points a, b.*

(ii) *If $m > 2$, z is the origin $(0, 0)$.*

Proof. The vectors $z - v$, for the distinct starting points v, are equivalent under rotations, and in particular all have the same Euclidean length (moreover, the set of the absolute values of their two coordinates are the same). Since the Euclidean equidistant set of two points a, b at even lattice distance is their perpendicular bisector $B(a, b)$, this gives (i). The only point equidistant from three starting points is the origin (in either metric), giving (ii). ∎

LEMMA 9. *There can be at most two agents at a common location z at any odd time t. Hence $x_t \leq 2$ for odd t.*

Proof. Assume $x_t \geq 3$. Since t is odd, z has odd parity, and hence $z \neq (0, 0)$. Hence by Lemma 8(ii) , the agents come from at most two distinct starting nodes $a, b \in \mathcal{S}$. Since by Lemma 7(i) at most two agents can come from a single starting point, we conclude that $a \neq b$ and $x_t \leq 4$. By Lemma 8(i), z belongs to the perpendicular bisector $B(a, b)$ of a and b, which can only contain odd parity nodes z if $a = -b$, and either z_1 or z_2 is 0. Without loss of generality we may assume $a = (2, 0)$, $b = (-2, 0)$, and $z_1 = 0$. Since by Lemma 7(ii) the vectors $z - a$ and $z - b$ make angles $r\pi/4$ with the vertical, the coordinate z_2 must be 2, 0, or -2, but all these possibilities result in an even parity z, whereas z is odd, contradicting the assumption. ∎

3.3. Starting point meetings (SPMs). Lemma 7(i) says that the only place to simultaneously meet more than two agents from a common starting node v is at v itself. Consequently, meetings of this type are important and will be given a special name.

DEFINITION 10. *We say there is an* SPM *at time t and node z if either I meets an agent of II at $z = (0, 0)$ (called* Type I*) or Player I meets an agent starting at $z \in \mathcal{S}$ at that node z (called* Type II*). We denote this (either case) as* SPM(t).

Note that the *type* of the SPM is the name of the Player who is back at his starting point at the time of the meeting. Some elementary properties of SPMs are given below.

LEMMA 11 (starting point meetings).

(i) *SPM(t) implies t is even,* $x_{t+1} = 0$, *and* $x_t \leq 6$.

(ii) $x_t > 4$ *implies SPM(t), and hence for even* $t, x_t + x_{t+1} \leq 6$ *(so* $x_i \leq 6$ *for all* i*).*

(iii) $x_t + x_{t+1} \leq 4$ *if no SPM at even time* t.

(iv) *If SPM(t) and SPM(t+2), then the corresponding types are distinct.*

Proof. (i) Since an SPM must occur at a node z in $\mathcal{S}_0$, which contains only even nodes, the time t must be even. Note that at an SPM, one of the players is at his start, while the other is some vector $\mathcal{R}^j(0,2)$ away from his start. Since the difference between their starts has the form $\mathcal{R}^k(0,2)$, their difference at an SPM is of the form $\mathcal{R}^k(0,2) + \mathcal{R}^j(0,2)$, and their distance is 4. So the earliest time for the next meeting is $t+2$, and hence $x_{t+1} = 0$. Suppose the SPM is of Type II. At time $t-1$, Player I will be at some node z' adjacent to z, and there will also be exactly two agents (counting both met and unmet agents) at z' from the starting point z. Since there are only 8 agents who start at z, I can meet at most $8 - 2 = 6$ new agents at time t at z. If the SPM is of Type I, there will be 8 agents at the origin at time t (including ones already met) since $g(t)$ will be a rotation of $(2,0)$. As in the previous argument, I will have met two of them on the previous move. So in either case we have $x_t \leq 6$.

(ii) If the first part of (ii) is true, the second part (the "hence") follows from $x_t + x_{t+1} \leq 6 + 0 = 6$, by part (i) and Lemma 9. So we need only prove the first part.

Suppose I meets $x_t > 4$ agents at some location z at some (even) time t, and there is no SPM. So $z \neq (0,0)$ (otherwise we have a Type I SPM). Suppose $z \notin \mathcal{S}$. Since at most two agents can come from any starting point (Lemma 7(i)), the agents who met at time t must come from at least three distinct starting nodes, and hence (Lemma 8(ii)) $z = (0,0)$, again contradicting our assumption. So we may assume that $z \in \mathcal{S}$ and II is not back at his starting point (otherwise we have a Type II SPM). Since we know $z \neq (0,0)$, Lemma 8(ii) says I can meet agents from at most two starting points. Since II is not back at his start at time t, Lemma 7(i) says there are at most two from each, so in total at most four, contradicting the assumption $x_t > 4$, and we are done.

(iii) Assuming there is no SPM at time t, part (ii) implies that $x_t \leq 4$. If $x_t \leq 2$, the result follows from Lemma 9 (without any other assumptions), so we can assume that $x_t \geq 3$. Hence by Lemma 7(i), agents from at least two starting points a and b must be met at time t at a node z, and by Lemma 8(i), $z \in B(a,b)$.

Suppose $z \in \mathcal{S}$, in which case we can assume without loss of generality that $z = (0,2)$, with $a = (2,0)$ and $b = (-2,0)$. In this case $g(t)$ is a rotation of $(2,2)$, which means the unmet agents at time t are at lattice distance 4 from z and cannot be met before time $t+2$. So $x_t + x_{t+1} \leq x_t + 0 \leq 4$, as required.

Suppose $z \notin \mathcal{S}$. Since $z \neq (0,0)$ (this would be an SPM), we can assume that I meets two agents from a and at least one from b. By Lemma 7(ii), we know that $z - a$ makes an angle $r\pi/4$ with the vertical. If $a = -b$ ("opposite" starting points) then the only nodes z with this property on $B(a,b)$ are in $\mathcal{S}_0$, which we have ruled out. So assume without loss of generality that $a = (2,0)$ and $b = (0,2)$, in which case the only z's in $B(a,b)$ with the required angle property are $(1,1)$ and $(2,2)$. If $z = (2,2)$, then $g(t)$ is a rotation of $(2,0)$, and the same observation as in the previous paragraph gives $x_{t+1} = 0$, and we are done. If $z = (1,1)$, then $g(t) = (1,1)$ and there will be four met and unmet agents at z at time t. However, by Lemma 7(iii) these agents will occupy all four nodes adjacent to z at time $t-1$, and again at time $t+1$. By the $t-1$

observation, we have $x_t \leq 3$, and by the $t+1$ observation and Lemma 9 we know that of the maximum of two agents at I's location z' at $t+1$, one has already been met, so $x_{t+1} \leq 1$. Hence $x_t + x_{t+1} \leq 3+1 = 4$, as required.

(iv) If SPM(t) is of Type II and occurs at some $v \in \mathcal{S}$, then SPM(t+2) cannot occur at another $v' \in \mathcal{S}$ (also be of Type II), because $d(v, v') = 4$. If SPM(t) is of Type I, then at time t all unmet agents are at distance 4 from the origin, and hence cannot get there for another Type I meeting before time $t+4$. So consecutive SPMs cannot be of the same type. □

LEMMA 12. *If SPM(t) but not SPM(t + 2), then $x_{t+2} \leq 2$ and $x_{t+2} + x_{t+3} \leq 3$.*

Proof. Since SPM(t), one of the players is at his start at time t. By renaming them, if necessary, we may assume that Player I is at the origin at time t (Type I). The proof now divides into cases according to the location z of Player I at time $t+2$.

$z = (0,0)$ At time t, unmet agents are at distance 4 from z, so at time $t+2$ they are at distance at least 2 from z, and hence $x_{t+2} = 0$. Since $t+3$ is odd, $x_{t+3} \leq 2$ by Lemma 9 and the result follows.

$z \in S$ By symmetry, we may assume $z = (0,2)$. We may assume that at time $t+2$ Player II is not at his starting point ($g(t+2) \neq (0,0)$), since otherwise we have SPM$(t+2)$. If $x_{t+2} \leq 1$, the result follows from Lemma 9, so we may assume $x_{t+2} \geq 2$. The agents who met at time $t+2$ must have starting points $(-2,0)$ or $(2,0)$, since those from $(0,-2)$ were at distance at least 4 from z at time t. The agents who met at time $t+2$ must be at locations $(-2,2)$ and $(2,2)$ at time t. Although there may be two agents at each of these nodes at time t, at most one of these (from each) can get to z by time $t+2$, since they have different notions of clockwise. Thus $x_{t+2} \leq 2$. In this case we have $g(t+2) = (2,2)$, so all unmet agents at time $t+2$ are at distance from z of at least 4 and cannot be met at time $t+3$. Hence $x_{t+2} + x_{t+3} \leq 2$.

$z = (\pm 1, \pm 1)$ By symmetry we take $z = (1,1)$. It is easy to check that the only new agents met at time $t+2$ were at $(2,2)$ at time t. There are at most four of these, two each starting from $(0,2)$ and $(2,0)$. Those from the same starting point have different notions of clockwise (by Lemma 7(i)), and since $g(t+2)$ is a rotation of $(1,1)$, they will be at different locations at time $t+2$. Hence at most one from each of $(0,2)$ and $(2,0)$ will meet I at $(1,1)$ at time $t+2$, and so $x_{t+2} \leq 2$. If $x_{t+2} < 2$, then the result follows from Lemma 9. Assuming $x_{t+2} = 2$, the situation is as stated above, and the only nodes with unmet agents within distance 2 of $(1,1)$ at time $t+2$ are $(3,1), (-1,1), (1,3)$, and $(1,-1)$, that is, the four starting points added to $(1,1)$. Player I can meet agents from only one of these nodes at time $t+3$, at respective nodes $z' = (2,1), (0,1), (1,2)$, and $(1,0)$. If z' is $(2,1)$ (the analysis for $(1,2)$ is similar), the only agents there come from $w = (3,1)$ at $t+2$ and (by Lemma 7 (iii)) go to different locations (namely z' and $(3,0)$) at time $t+3$. So in this case $x_{t+3} = 1$ and we are done. If z' is $(0,1)$ (the case $(1,0)$ is similar), then any agents he meets there were at $w = (-1,1)$ at time $t+2$ and come from one of the starting points ($(-2,0)$ and $(0,2)$). By Lemma 7(i), there can be at most one of these. □

4. Optimal strategies. In this section we first obtain a partial characterization of optimal strategies, show that the NAWFM strategy $(\tilde{f}, \tilde{g})$ is optimal, and finally give a complete characterization of optimality. We can always rename the players so that the first SPM, if there are any, is of Type II. *We adopt this labeling throughout this section.* The following important result determines the existence and type of the

potential SPMs at all even times except for $t = 14$. We shall see later that optimal strategies may or may not have an SPM at time 14.

THEOREM 13. *Any optimal strategy pair in* $\Gamma(2)$ *has SPMs of alternating Types II, I, II, I, II, at times* $2, 4, 6, 8,$ *and* 10*, and no SPM at time* 12.

Proof. To any meeting profile $[x_1, x_2, \dots]$, where x_t denotes the number of new agents met at time t, we associate the integer $T^* = \sum_t t x_t$, which is 32 times the expected meeting time. To aid the reader, we will highlight the numbers x_t in the profile relevant to the argument by grouping them with a $\overbrace{\ \ }$. The NAWFM strategy has the agent number profile $[2, 6, 0, 6, 0, 6, 0, 4, 0, 4, 0, 2, 1, 1]$, for which $T^* = 197$ (expected meeting time 197/32), and hence any optimal strategy has $T^* \le 197$. We will establish the SPM claims successively for $t = 2, 4, 6, 8, 10,$ and 12 by rejecting as optimal any strategy which produces a partial agent number sequence $[x_1, \dots, x_k]$ for which the smallest possible value of T^* exceeds 197. Note that for any strategy we have $x_1 = 2$.

SPM(2) Suppose that $x_2 + x_3 \le 4$. Then by Lemma 11(ii) the meeting profile minimizing T^* is $[2, \overbrace{4, 0}, 6, 0, 6, 0, 6, 0, 6, 0, 2]$, with $T^* = 202 > 197$. Hence any optimal strategy has $x_2 + x_3 > 4$, which by Lemma 11(iii) implies the required result SPM(2).

SPM(4) If there is no SPM at time 4, then by Lemma 12 and the established result SPM(2), we have $x_4 + x_5 \le 3$. So by Lemma 11(ii), the minimizing profile is $[2, 6, 0, \overbrace{3, 0}, 6, 0, 6, 0, 6, 0, 3]$, with $T^* = 206 > 197$, contradicting optimality.

SPM(6) If there is no SPM at time 6, then by SPM(4) and Lemma 12 we have $x_6 + x_7 \le 3$. Hence by Lemma 11(ii), the minimizing profile is $[2, 6, 0, 6, 0, \overbrace{3, 0}, 6, 0, 6, 0, 3]$, with $T^* = 200 > 197$.

We now summarize our findings for $t \le 7$. Since the SPMs at $t = 2, 4, 6$ are of alternating type, it is easy to verify that the meeting profile for any optimal strategy starts with

$$[2, 6, 0, 6, 0, 6, 0], \text{ with contribution } T^{**} \text{ to } T^* \text{ of} \tag{21}$$

$$T^{**} = (2 \times 1) + (6 \times 2) + (6 \times 4) + (6 \times 6) = 74. \tag{22}$$

Furthermore, there are precisely two starting nodes a, $b \in \mathcal{S}$, which have not been Type II SPMs (not been visited by I) by time 6. From each of these nodes, two agents have been met at time 4 (the Type I SPM at the origin), and the remaining six agents have not been met by time 6.

Suppose $x_8 + x_9 \le 2$. Then by (21) and Lemma 11(ii), the best profile is $[2, 6, 0, 6, 0, 6, 0, 2, 0, 6, 0, 4]$, with $T^* = 198 > 197$. Hence

$$x_8 + x_9 \ge 3 \text{ and (since } x_9 \le 2 \text{ by Lemma 9)} \tag{23}$$

$$x_8 > 0. \tag{24}$$

We now continue to establish further SPMs.

SPM(8) Suppose there is no SPM at time 8. Then by the established SPM at time 6 and Lemma 12, we have $x_8 \le 2$ and $x_8 + x_9 \le 3$. Hence by (23)

$$x_8 + x_9 = 3 \text{ and } x_9 \ge 1. \tag{25}$$

If we *also* have $x_{10} + x_{11} \le 4$, then $[2, 6, 0, 6, 0, 6, 0, \overbrace{2, 1}, \overbrace{4, 0}, 5]$ is the best profile, with $T^* = 74 + 125 = 199 > 197$. Hence we have $x_{10} + x_{11} \ge 5$.

So by Lemma 11(iii) we have SPM(10), and by Lemma 11(i) we have that $x_{11} = 0$. Hence $x_{10} \geq 5$. The SPM at time 10 cannot be Type I, since in that case Player I could meet at most two agents each from the unvisited nodes a and b, contradicting the result $x_{10} \geq 5$. Thus it is of Type II, with Player I meeting at least five agents at the node (say) a. Since there were six agents from a unmet at time 7, he can have met at most $6 - 5 = 1$ of these at times 8 and 9. Since (25) $x_9 \geq 1$ and the only agents that Player I can meet at time 9 come from a (since they are back at a at the Type II SPM at time 10), we must have $x_9 = 1$, and hence by (25), $x_8 = 2$. Thus the other agent met by I at some node c at time 8 must come from b. Since the SPM at time 6 was of Type II, this agent was back at b then, so $d(b, c) \leq 8 - 6 = 2$. Now Player I goes from some node $e \in \mathcal{S}$ at time 6 (since there was a Type II SPM then), to c at time 8, and to $a \in \mathcal{S}$ at time 10. Since $d(e, a) = 4$, it follows that the distance from c to all three nodes $a, b, e \in \mathcal{S}$ is no more than 2. Hence c must be the origin, and since I met an agent at c at time 8, there was an SPM at time 8, contradicting our initial assumption.

By alternation of types for consecutive SPMs, the SPM at time 8 must be of Type I, so at most two agents from each of the unvisited starting points a and b can be met at time 8, and (by Lemma 11(i)) none at time 9, meaning $x_8 \leq 4$ and $x_9 = 0$.

SPM(10) Suppose there is no SPM at time 10. Then, since we have established SPM(8), it follows from Lemma 12 that $x_{10} \leq 2$ and $x_{10} + x_{11} \leq 3$. If $x_8 + x_9 \leq 3$, it follows from the previous remark and (21) that the best profile satisfying the known constraints is $[\overbrace{2, 6, 0, 6, 0, 6, 0}, \overbrace{3, 0}, \overbrace{2, 1}, 6]$. But this has $T^* = 201 > 197$, so $x_8 + x_9 \geq 4$, and by the remarks above this paragraph $x_8 = 4$. Thus, after time 8, I can meet at most four agents at any SPM of Type II, and hence (by Lemma 7(i)) at most four agents at any time after 8. Hence the best profile is $[\overbrace{2, 6, 0, 6, 0, 6, 0}, \overbrace{4, 0}, \overbrace{2, 1}, 4, 1]$, with $T^* = 198 > 197$. Hence there is an SPM at time 10, which must be of Type II, at node (say) a.

Not SPM(12) Suppose there is an SPM at time 12, which by alternation would have to be of Type I. By Lemma 7(i), I can meet at most two agents from the sole unvisited node b at time 12, so $x_{12} \leq 2$ and (by Lemma 11(i)) $x_{13} = 0$. Hence the best profile would be $[2, 6, 0, 6, 0, 6, 0, 4, 0, 4, 0, 2, 0, 2]$, with $T^* = 198 > 197$. So there is no SPM at time 12. □

We can now use the pattern of SPMs established for optimal strategies in Theorem 13 to determine the unique optimal agent number profile and thus the rendezvous value of $\Gamma(2)$.

THEOREM 14. *The rendezvous value for* $\Gamma(2)$ *is* $197/32$ *and the NAWFM strategy* $(\tilde{f}, \tilde{g})$ *is optimal. Furthermore, every optimal strategy has the agent number profile* $[2, 6, 0, 6, 0, 6, 0, 4, 0, 4, 0, 2, 1, 1]$.

Proof. Consider an agent number profile x corresponding to an optimal strategy. According to (21) the profile must begin with $[2, 6, 0, 6, 0, 6, 0]$. Theorem 13 says that at time $t = 8$ there is a Type I SPM (at the origin), so I will meet there two agents from each of the two starting nodes he has not visited. Hence $x_8 = 4$ and $x_9 = 0$ (by Lemma 11(i)). At time $t = 10$, Theorem 13 says that I will visit his third Player II starting point. Of the eight agents who started there, two were met at the origin at each of the Type I SPMs at $t = 4$ and 8, so $x_{10} = 8 - 2 - 2 = 4$, and again $x_{11} = 0$. So the agent number profile x must be of the form

$$[2, 6, 0, 6, 0, 6, 0, 4, 0, 4, 0, x_{12}, x_{13}, x_{14}].$$

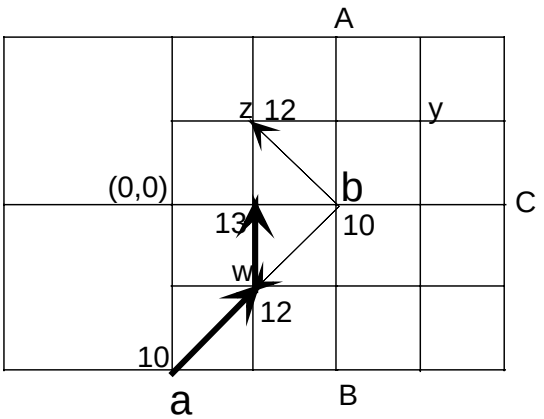


FIG. 5. *Analysis for* $t = 10$ *to* 14.

Since there is no SPM at time 12, it follows from Lemma 12 that $x_{12} \leq 2$ and $x_{12} + x_{13} \leq 3$. So the best profile satisfying all known constraints is

$$\left[2, 6, 0, 6, 0, 6, 0, 4, 0, 4, 0, \overbrace{2, 1}, 1\right],$$

which is (19) the profile of the NAWFM strategy $(\tilde{f}, \tilde{g})$ and has $T^* = 197$. Hence NAWFM is optimal, and the rendezvous value is 197/32. □

We now use the last two theorems to derive the full set of optimal strategy pairs for Γ. We describe the situation in I's coordinate system, but we will not use the agent analysis for II. We begin our analysis at the end of move 10 ($t = 10$). According to Theorem 13, I is at one of the previously unvisited nodes of $\mathcal{S}$, say a, and II (all four remaining agents of) is at the other one, b. Player I can deduce that II is at b, while II knows only that the vector to I is one of the four rotations of $(2, 2)$. We know from the optimal profile of Theorem 14 that they must meet with conditional probability 1/2 at times 12 (two agents of the remaining 4) and 13 (one agent of the remaining 2) and conditional probability 1 at time 14 (the remaining agent must be met). We will find it easier to use a conditional probability analysis, rather than an agent analysis.

If a and b are opposite starting points ($a = -b$) in $\mathcal{S}$, then the only possible location for a meeting at time 12 is the unique midpoint of a and b, namely the origin. But this would constitute an SPM at time 12, which we have ruled out in Theorem 13. The situation from times 10 to 14 is illustrated in Figure 5, with $a = (0, -2)$ and $b = (2, 0)$. Let w denote I's location at time 12. For a meeting, w has to be at one of the three midpoints of a and b : the origin, $(1, -1)$, or $(2, -2)$. The first would imply SPM(12), so can be excluded (Theorem 13). The last would require each player to move 2 units in the same direction on moves 11 and 12. This would leave the players 4 units apart if they fail to meet at time 12, in which case there could not be a meeting at time 13. So w must be the unique Euclidean midpoint of a and b. In our drawing, this gives $w = (1, -1)$. Since the optimal profile has no meeting at time 11, I may get to w in either of the two possible ways (indicated by a diagonal arrow). To achieve a possible meeting at w at time 12, II must move to a diagonally opposite node from his location b at time 10 (must make a turn at time 12). Since he does not know where w is, such a move sequence might lead him to any of the four nodes w, z, y, and

$(3,-1)$). How can II choose among these four directions so as to reach w at time 12 with conditional probability 1/2?

To answer this question, recall (from Theorem 13) that there were Type I SPMs at times 4 and 8. This means Player II has visited two of the possible Player I starting points, that is, two of the three nodes A, B, and C. (He has not visited the actual starting point at the origin or the game would be over.) Now there are two cases. Either (1) he did not visit C (visited two *opposite* starting nodes of I, A and B) or (2) he visited C (two *adjacent* possible starting nodes of I, C and either A or B). He will know which case he is in according to whether he searched adjacent or opposite nodes.

Suppose (1) that Player II did not visit C, and so visited A and B. In this case he can distinguish the vertical coordinate (directions of his searches) from the horizontal (unsearched directions), and for example could determine the origin (I's starting point) with probability 1/2 (he could equiprobably reach the origin or C in two moves). However, the node w (which he needs to reach at time 12 with conditional probability 1/2) is indistinguishable from three other nodes z, y, and $(3,-1)$, and so could not be reached with conditional probability greater than 1/4. Hence he cannot ensure a meeting at time 12 with the required conditional probability 1/2, and so a strategy that did not visit C cannot be optimal.

So an optimal strategy for II must have visited C, that is, previously searched two *adjacent* possible starting nodes of I. In this case II can determine w with the required conditional probability 1/2 as *the node e in the direction opposite the midpoint of the two Player I starting nodes he has visited.* (If he visited A and C, then $e = w$; if B and C, then $e = z$.) Optimal play must bring him to w with conditional probability 1/2, so he must move to e (in either of the two possible ways, since the optimal profile has no meeting at $t = 11$) at time 12. This is indicated by the thin lines going out from b. (This is not a mixed strategy; II goes for his uniquely defined node e.) Assuming II does not meet I at time 12, his location at that time must be $e = z$, and I can deduce this. So to achieve a possible meeting at time 13 (required by the optimal profile), I must move to the unique lattice midpoint of w and z (the node (1,0) in our figure). Player II can conclude at time 12 that Player I is either at w or y. So he must move back toward b equiprobably in either way (toward w or toward y). (Again, this is not a mixed strategy—in agent formulation, one of the two remaining agents will go towards b clockwise, and the other counterclockwise.) If the players have not met by time 13, they both can deduce the location of the other; I is at $(1,0)$ and II is at $(2,1)$. Now they can agree to meet for sure at time 14 at either of their two midpoints: (i) b or (ii) z. This leads to two types of strategy, according to whether they both choose (i) or they both choose (ii).

Hence we have established the following theorem.

THEOREM 15. *A strategy pair is optimal in* $\Gamma(2)$ *if and only if it is one of the following two types,* (i) *and* (ii).

Player I orders the starting points of II as V_1, V_2, V_3, V_4 in any way so long as the first two are adjacent ($V_1 \neq -V_2$). On the first 10 *moves, he goes to V_1 and back, V_2 and back, and to V_3. On moves* 11 *and* 12 *he goes (in either of the two ways) to the unique Euclidean midpoint w of V_3 and V_4. At time* 13 *he goes to the unique midpoint of the origin, w, and V_4. At time* 14, *he either* (i) *goes to V_4 or* (ii) *goes in the same direction as on move* 13.

Player II goes in any direction on move 1, *returning to his start on move* 2. *On moves* 3 *to* 6, *and on moves* 7 *to* 10, *he goes, respectively, to two adjacent possible starting nodes W_1 and W_2 and back to his start. On moves* 11 *and* 12, *he goes (in*

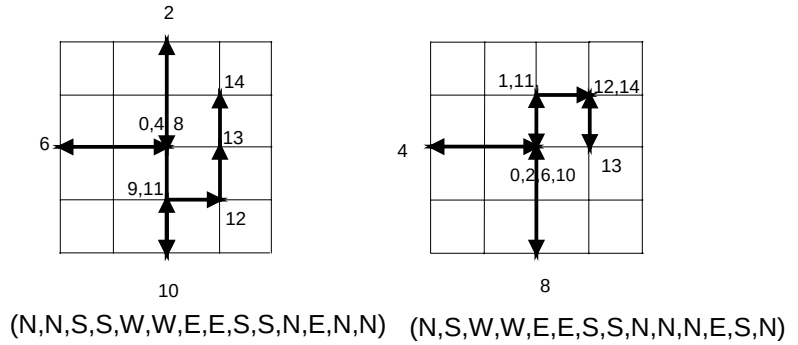


FIG. 6. *Optimal strategy for* Γ *of type* (ii).

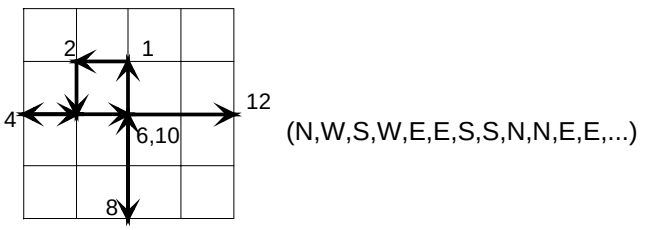


FIG. 7. *Strategy with meeting probability* 31/32 *by time* 12.

either of the two possible ways) to the node e diagonally opposite his start and in the direction opposite to the Euclidean midpoint of W_1 *and* W_2. *On move* 13, *he goes to either of the two midpoints of e and his start. On move* 14, *he goes either to* (i) *his starting point or* (ii) *back to e.*

The strategy $(\tilde{f}, \tilde{g})$ is of type (i). In Figure 6 we give an example of an optimal strategy of type (ii).

Although the NAWFM strategy is optimal, it is not *uniformly optimal*—it does not maximize the probability of meeting by time t for all t. In particular, it does not maximize the probability of meeting by time 12. After move 12, Player I has met 30 of the 32 agents, so the probability that the players have met by time 12 is 30/32. Consider the 12 move strategy for Player I given by $(N, W, S, W, E, E, S, S, N, N, E, E, \dots)$, and drawn in Figure 7. Assuming Player II is back at his start at times 4, 8, and 12, the only agents not met by time 12 have starting point $(0, 2)$. Of the eight agents who started there, two could be met at $(0, 1)$ at time 1, one (only one since this involves a turn) could be met at $(-1, 1)$ at time 2, and two could be met at the origin at both times 6 and 10. This would leave only one agent unmet by time 12, or a 31/32 probability of rendezvous by that time, better than NAWFM. In fact, the agents can be met at the times mentioned above if Player II adopts the move sequence $(N, E, W, S, S, S, N, N, W, W, E, E, \dots)$. Hence the NAWFM strategy is optimal but not uniformly optimal and, as mentioned in Definition 4, there cannot be any uniformly optimal strategy for the game $\Gamma = \Gamma_P$. As mentioned in [3], this is qualitatively distinct from the situation for the diagonal start game Γ_D, where there is a uniformly optimal strategy (and hence all optimal strategies are uniformly optimal). The (uniformly) optimal strategies for the diagonal start game (without common clockwise) are as follows (Corollary 23 of [3]): Player I cyclically orders adjacent (distance 2)

starting nodes as v_1, v_2, v_3, v_4 in any manner. He moves to the midpoint of v_1 and v_4 at time 1, and then cyclically searches the vertices v_i at time $2i$. Player II moves in any direction D_i (of the four possible) at times $i = 1, 3, 5, 7$, going in the opposite direction $-D_i$ at times $i+1$. The only restriction is that $D_7 = \pm D_1$, that is, II must move in the same or opposite direction on move 7 as he did on move 1 (not at right angles). The particular strategy pair of $(N, W, S, S, E, E, N, N), (N, S, N, S, N, S, N, S)$ was earlier shown to be optimal for the common clockwise version of the diagonal start problem by Anderson and Fekete [9], and an easy symmetric argument shows that it does equally well (same expected meeting time) in the no common clockwise version.

5. Rendezvous in higher dimensions. This section determines an upper bound for the rendezvous value $R(n)$ of the game $\Gamma(n)$ which generalizes the game $\Gamma(2)$ of the previous section (as well as the game $\Gamma(1)$ solved in [6]) to n dimensions. The strategy pair giving this bound is the AWFM (Definition 6), which assumes no common ordering of the coordinate axes. Some ideas on lower bounds are also given at the end of the section.

THEOREM 16. *Suppose that two players are initially placed on the n-dimensional integer lattice so that their difference vector is two units long and parallel to one of the coordinate axes. Assume that the players have no common notion of location and no common labeling of the coordinates axes. The rendezvous value $R(n)$ for this game $\Gamma(n)$ satisfies the inequality*

$$2n \leq R(n) \leq \frac{32n^3 + 12n^2 - 2n - 3}{12n^2}. \tag{26}$$

Consequently, we have the asymptotic result

$$2 \leq \lim_{n\to\infty} R(n)/n \leq \frac{8}{3}. \tag{27}$$

Proof. We will need to consider two subsidiary problems $\Gamma_1(m)$ and $\Gamma_2(m)$, for $m = 1, \ldots, 2n$. Both of these problems begin at time $t = 0$ with the placement of Players I and II, respectively, at a pair of nodes A and B which are two units apart along a line parallel to some coordinate axis. Then Player I is displaced to a node A' along a similar two unit line which is not the one leading to B. Player I is told the node A. In the problem $\Gamma_1(m)$, Player I is told $m-1$ directions which are certain to include the direction to B, and Player II is told m such directions. In problem $\Gamma_2(m)$, both players are told m such directions. Special cases of these problems, for $n = 2$, are drawn in Figures 8 and 9.

In order to estimate the rendezvous value $R(n)$ of the original problem $\Gamma(n)$, we must obtain estimates on the respective asymmetric rendezvous values $w_1(m)$ and $w_2(m)$ of $\Gamma_1(m)$ and $\Gamma_2(m)$ for various m (corresponding to the dimension n, which is implicit in our notation).

Suppose that in the problem $\Gamma_1(m)$, the first two moves of the players are as follows: Player I goes to the node A (which he knows) while Player II goes 2 steps randomly in one of the m indicated directions. With probability $1/m$, II will pick the direction to A, and the meeting time will be $t = 2$. Otherwise the two players will be in the initial position of the other problem $\Gamma_2(m-1)$, with the roles (of I and II) reversed. Hence we have

$$w_1(m) \leq \frac{1}{m}(2) + \frac{m-1}{m}(2 + w_2(m-1)). \tag{28}$$

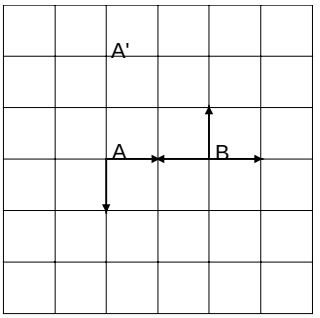


FIG. 8. *Start in* $\Gamma_1(3)$.

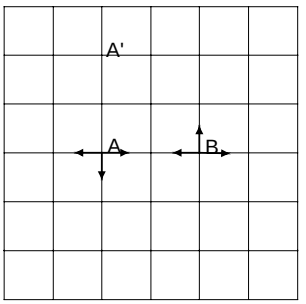


FIG. 9. *Start in* $\Gamma_2(3)$.

Similarly, in the initial position of $\Gamma_2(m)$, the same type of strategy for the first two moves gives

$$w_2(m) \leq \frac{1}{m}(2) + \frac{m-1}{m}(2 + w_1(m)). \tag{29}$$

From these two inequalities we obtain upper bounds $w_i(m) \leq \bar{w}_i(m)$ by solving the associated equalities. This gives us the solutions

$$\bar{w}_1(m) = \frac{4m+1}{3}, \tag{30}$$

$$\bar{w}_2(m) = \frac{-1+3m+4m^2}{3m}. \tag{31}$$

This is consistent with the trivial base case $\Gamma_2(1)$, where I and II start 4 units apart with knowledge of the other's direction, with rendezvous value $w_2(1) = 2$ corresponding to a meeting at A. This case is illustrated in Figure 10.

We now consider the original game $\Gamma(n)$. Suppose that in their first two moves one player (I) goes two units in his forward direction, while the other (II) goes in some direction and then back to his start. With probability $1/(2n)^2$ the two players will go towards each other and meet at time $t = 1$. If Player I goes in the direction of II (probability $1/2n$) and II does not go in the direction of I (probability $(2n-1)/(2n)$), then the two will meet at time $t = 2$ at II's initial location. In the remaining case, I will find himself displaced two units from his start and know of one direction from his start which does not lead to II's starting location. Meanwhile, II will be at his start and not know of any of the $2n$ directions which are not correct. Consequently

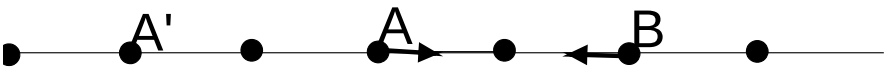


FIG. 10. *Start in* $\Gamma_2(1)$.

the situation at time $t = 2$ will be identical to that of the problem $\Gamma_1(2n)$. Therefore we have the estimate

$$R(n) \leq \left(\frac{1}{(2n)^2}\right) 1 + \left(\frac{1}{2n}\right)\left(\frac{2n-1}{2n}\right) 2 + \left(\frac{2n-1}{2n}\right)(2 + w_1(2n)).$$

Estimating w_1 by the formula (30) for $\bar{w}_1$, and simplifying, we get

$$R(n) \leq \frac{32n^3 + 12n^2 - 2n - 3}{12n^2}, \text{ and hence}$$

$$\lim_{n\to\infty} R(n)/n \leq \frac{8}{3}, \text{ the required right inequality.}$$

We note again that the value of the right-hand side of the top inequality is 13/4 for $n = 1$, which is the *exact* asymmetric rendezvous value for the line as derived in [6] for known initial distance $D = 2$. The strategy pair which gives this expected meeting time is the one which converts $\Gamma(n)$ into $\Gamma_1(2n)$ and thereafter converts each problem $\Gamma_1(m)$ into $\Gamma_2(m-1)$ and each problem $\Gamma_2(m)$ into $\Gamma_1(m)$, $m = 2n, 2n-1, \ldots, 1$ (assuming the players don't meet earlier). It is an AWFM strategy.

We can obtain a lower bound on $R(n)$ by giving the players some additional information and determining their optimal strategy pair in that situation. The simplest way to do this is to give the players common notions of directions. In this case it is well known (see [6] or [9] for one- or two-dimensional arguments) that the players should always move in opposite directions, and that the rendezvous problem is consequently equivalent to one where Player II is stationary and Player I moves with their combined speeds, here equal to 2. The least expected time for Player I to reach the $2n$ possible locations of the stationary Player II, moving with speed 2, is obtained by a path which reaches these locations at the $2n$ times $1, 3, 5, \ldots, 2i-1, \ldots, 2(2n)-1$, while returning to his start in between. The expected time is consequently given by

$$\frac{1}{2n}\sum_{i=1}^{2n}(2i-1) = 2n.$$

Thus the asymptotic value of $R^a(n)/n$ lies between 2 and 8/3, completing the proof of Theorem 16. □

A player-symmetric version of this n-dimensional problem, where both players must follow the same mixed strategy, is analyzed in [7, section 18.2.2]. An alternative treatment of the player-asymmetric problem is outlined in section 18.2.1, based on a direct evaluation of the expected meeting meeting time for the NAWFM strategy. An anonymous referee has observed that this can be simply evaluated by defining random variables X and Y to be, respectively, the first time that I reaches the starting point of II and the first time that II reaches the starting point of I. The meeting time $\min(X, Y)$ can then be calculated from standard results on random variables.

6. Discussion of results. A natural question that arises from our analysis is whether one might be able to obtain a relatively simple analytical expression for the rendezvous value $R(n)$ of $\Gamma(n)$ for $n \geq 3$. The indications from the results on $\Gamma(2)$ are somewhat equivocal. Theorem 14 states that every optimal strategy pair in $\Gamma(2)$ has the same agent number profile and the same holds for $\Gamma(1)$. Furthermore, Theorem 13 tells us that, in all optimal strategy pairs of $\Gamma(2)$, the players initially move to an SPM position and then go as quickly as possible from one SPM position to another until a player has met most of the other player's agents. The reason for this is that a player can meet comparatively few agents of the other player at points which are not either possible starting points of that player or his own starting point. It is intuitively clear that this property holds for $\Gamma(n)$ with $n \geq 3$ so we would expect there would be a similar emphasis on SPM positions for these problems. It also suggests that $R(n)$ is likely to be much nearer the upper bound than the lower bound in Theorem 16, particularly when account is taken of the fact that the lower bound is obtained by considering a corresponding problem in which the players are given important extra information. On the other hand, Theorem 15 detailing the optimal strategy pairs for $\Gamma(2)$ demonstrates that although the players have some freedom in the order in which the SPM positions are visited, it is not unrestricted. For $\Gamma(2)$, by looking at the moves which enabled the final agents to be met, it was comparatively easy to see what modifications to AWFM were needed if better expected meeting times were to be achieved and the restrictions on the SPM positions arose from these modifications. However, even if optimal strategy pairs in $\Gamma(n)$ are modifications of AWFM, the number of possible modifications is likely to increase substantially as n increases and, in addition, there is the difficulty of picturing the situation, particularly with regard to agents.

REFERENCES

[1] S. Alpern, *The rendezvous search problem*, SIAM J. Control Optim., 33 (1995), pp. 673–683.
[2] S. Alpern, *Rendezvous search: A personal perspective*, Oper. Res., 50 (2002), pp. 772–795.
[3] S. Alpern and V. Baston, *Rendezvous on a planar lattice*, Oper. Res., 53 (2005).
[4] S. Alpern and V. Baston, *A common notion of clockwise helps in planar rendezvous*, European J. Oper. Res., 2006, to appear.
[5] S. Alpern and A. Beck, *Asymmetric rendezvous on the line is a double linear search problem*, Math. Oper. Res., 24 (1999), pp. 604–618.
[6] S. Alpern and S. Gal, *Rendezvous search on the line with distinguishable players*, SIAM J. Control Optim., 33 (1995), pp. 1270–1276.
[7] S. Alpern and S. Gal, *The Theory of Search Games and Rendezvous*, Internat. Ser. Oper. Res. Management Sci. 55, Kluwer Academic Publishers, Norwell, MA, 2003.
[8] E. J. Anderson and S. Essegaier, *Rendezvous search on the line with indistinguishable players*, SIAM J. Control Optim., 33 (1995), pp. 1637–1642.
[9] E. J. Anderson and S. Fekete, *Two dimensional rendezvous search*, Oper. Res., 49 (2001), pp. 107–118.
[10] V. J. Baston, *Two rendezvous search problems on the line*, Naval Res. Logist., 46 (1999), pp. 335–340.
[11] S. Gal, *Rendezvous search on the line*, Oper. Res., 47 (1999), pp. 974–976.
[12] J. V. Howard, *Rendezvous search on the interval and the circle*, Oper. Res., 47 (1999), pp. 550–558.
[13] L. C. Thomas and P. B. Hulme, *Searching for targets who want to be found*, J. Oper. Res. Soc., 48 (1997), pp. 44–50.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 6, pp. 2253–2268

WHAT PERIODIC SIGNALS CAN AN EXPONENTIALLY STABILIZABLE LINEAR FEEDFORWARD CONTROL SYSTEM ASYMPTOTICALLY TRACK?*

EERO IMMONEN† AND SEPPO POHJOLAINEN†

Abstract. We study asymptotic tracking and rejection of continuous periodic signals in the context of exponentially stabilizable linear infinite-dimensional systems. Our reference signals are in Sobolev-type spaces $H(\omega_n, f_n)$ and they (as well as the disturbance signals) are generated by an infinite-dimensional exogenous system. We show that there exists a feedforward controller which achieves output regulation if and only if the so-called regulator equations are satisfied and a decomposability condition holds. For SISO systems this result allows us to completely answer the question posed in the title: We show that if the stabilized plant does not have transmission zeros at the frequencies $i\omega_n$ of the reference signals, then all reference signals in $H(\omega_n, f_n)$ can be asymptotically tracked in the presence of disturbances if and only if

$$\left(H_K(i\omega_n)^{-1}[1 - H_d(n)]f_n^{-1}\right)_{n\in I} \in \ell^2.$$

Here $H_K(i\omega_n)$, $n \in I$, is the transfer function of the stabilized plant evaluated at $i\omega_n$, and $(H_d(n))_{n\in I}$ is a sequence of disturbance coefficients for the stabilized plant. Moreover, the sequence $(f_n)_{n\in I}$ consists of weights for the Fourier coefficients of the reference signals. We give four examples to illustrate the theory.

Key words. output regulation, infinite-dimensional systems, regulator equations, Fourier series, Sobolev spaces, periodic signals

AMS subject classifications. 93B99, 93C25, 93D99

DOI. 10.1137/040613093

1. Introduction.

In this paper we study regulation of periodic signals in the context of exponentially stabilizable linear infinite-dimensional systems. By regulation we mean that the output of the system asymptotically tracks given suitably smooth periodic reference signals generated by an infinite-dimensional exogenous system and asymptotically rejects disturbance signals generated by the same exogenous system.

Periodic signals are often encountered, e.g., in acoustics, electric motors, and mechanical systems [12], and various forms of this regulation problem have been intensively studied for several decades. The regulation problem for finite-dimensional signals (generated by finite-dimensional exosystems) and finite-dimensional linear systems was solved quite completely in the 1970s by Francis, Wonham, Davison, and others [7, 9, 10]. Many authors have since generalized these results for infinite-dimensional systems and finite-dimensional reference/disturbance signals: Pohjolainen [17, 18], Hämäläinen and Pohjolainen [11], Logemann and Owens [16], Byrnes et al. [3], and others.

For finite-dimensional reference and disturbance signals it is well known that there are two basic control configurations that guarantee regulation: the feedforward and feedback controllers [7]. The feedforward controller does not lead to a robust design. Robustness means that the control configuration can achieve regulation in spite of variations in the system's and controller's parameters.

*Received by the editors August 9, 2004; accepted for publication (in revised form) June 1, 2005; published electronically February 3, 2006.
http://www.siam.org/journals/sicon/44-6/61309.html
†Institute of Mathematics, Tampere University of Technology, PL 553, 33101 Tampere, Finland (Eero.Immonen@tut.fi, Seppo.Pohjolainen@tut.fi).

In the feedback control design the controller must incorporate a model of the exogenous system according to the internal model principle. This introduces additional dynamics in the feedback loop and thus makes stabilization of the closed loop system a more difficult task, especially in the infinite-dimensional case. The advantage of the feedback controller is that it provides robustness. This is often stated in the following form: Stability of the closed loop system implies regulation.

In this paper we consider the problem of constructing a feedforward controller for exponentially stabilizable infinite-dimensional systems and uniformly continuous periodic reference and disturbance signals. Since such periodic signals may contain an infinite number of distinct frequency components, they must be generated by an infinite-dimensional exogenous system. Hence our results extend the infinite-dimensional feedforward controllers of [3, 16] to allow for infinite-dimensional reference and disturbance signals.

Our approach has been inspired by [4, 5], where the regulation problem with an infinite-dimensional exogenous system was formulated but not rigorously solved. The solution presented in this paper relies on a construction of a scale of Sobolev-type spaces of periodic uniformly continuous functions. This construction makes it possible to completely resolve the existence of a regulating feedforward controller in the single input single output (SISO) case; a characterization in terms of solvability of the so-called regulator equations [3, 9] is given in Theorem 3.1.

In Theorem 4.5 and Corollary 4.7 we assume that the transfer function of the stabilized plant is invertible on the spectrum of the exosystem. We prove that in this case a necessary and sufficient condition for the solvability of the regulation problem is that the transfer function does not approach zero asymptotically too rapidly on that spectrum. In the engineering language this condition says that the amount of damping on the high frequency components determines precisely how nonsmooth the reference/disturbance signals are that the system can asymptotically track/reject. The condition given in terms of weighting coefficients in the aforementioned Sobolev-type spaces makes it possible to single out those spaces which contain signals that can be tracked/rejected by a given system.

To the authors' knowledge the results of this article are new even for finite-dimensional systems. Moreover, they complement and improve those in [14], where the exosystem was built so that it can generate (at least) one given scalar-valued reference signal with ℓ^1 Fourier coefficients. In [14] the authors found a sufficient condition for the asymptotic tracking of this reference signal to occur, under the assumption of exponential stabilizability of the plant. In this paper, the exosystem is constructed so that precisely all reference signals in a Sobolev-type space can be generated; we pose—and completely solve—the problem of asymptotic tracking of all such reference signals and construct the actual control law which achieves output regulation. The construction of this paper also shows that the exogenous signal generator can, in a certain sense, be made isomorphic to a function space (Theorem 2.8). This feature yields two important observations which cannot be made from the results of [14]: (i) the approach of this paper suggests a direct generalization via considering reference signals in more general function spaces (e.g., vector-valued continuous periodic functions in the MIMO case), and (ii) the exosystem becomes an artificial device—we do not need to know its dynamical behavior in order to achieve output regulation (see Theorem 4.5 and Corollary 4.6). More profoundly, as opposed to [14], the results of this paper yield (for SISO systems) a complete answer, via the new condition (4.17), to the question posed in the title.

We conclude this section with a brief outline of the contents of the article. In

section 2 we define the plant and the reference signals that we want to asymptotically track; for this we introduce the Sobolev-type spaces $H(f_n, \omega_n)$. Furthermore, we construct the exogenous system that we assume is generating the reference and disturbance signals. We also formulate the output regulation problem (ω_n, f_n)-RP in section 2. In section 3 we show that for exponentially stabilizable plants the solvability of the (ω_n, f_n)-RP is equivalent to the solvability of the regulator equations and a decomposability condition. This result is then applied in section 4, where the solution of the regulator equations is considered. We obtain an explicit expression for a candidate operator L in the solution of the (ω_n, f_n)-RP. The continuity of L determines whether or not this regulation problem is solvable, and it can be verified by checking condition (4.17). Furthermore, assuming solvability of the output regulation problem, we explicitly write out, in terms of the reference signals, the control law achieving output regulation in Corollary 4.6. We conclude section 4 with two results which show that the capability of output regulation is in some cases an intrinsic property of the plant—different stabilizing feedbacks result in the same capability of output regulation. Finally, in the four examples of section 5 we solve the (ω_n, f_n)-RP in several situations: We consider a finite-dimensional plant, a delay-differential equation, and a heat equation with Neumann boundary conditions. In the last example we show that there are infinite-dimensional systems which cannot track all reference signals in any standard Sobolev space $H^{\alpha}_{per}(0, p)$ of periodic functions [15], even if there are no transmission zeros at the frequencies of the reference signals. These examples demonstrate an important new result of the paper: Transmission zeros are not the only cause of output regulation problems, even for finite-dimensional systems. The intrinsic "smoothness" of the plant in part determines which signals can be regulated. For certain finite-dimensional systems this smoothness is characterized by the relative degree of the plant, as shown in section 5.

1.1. Notation and conventions. For complex separable Hilbert spaces E and F, $\mathcal{L}(E, F)$ denotes the space of bounded linear operators $E \to F$, and E' denotes the space of bounded linear functionals on E. Inner product on E is denoted by $\langle \cdot, \cdot \rangle_E$ (the subscript E is omitted if no confusion can arise). Norm on E is denoted by $\|\cdot\|_E$ (the subscript E is omitted if no confusion can arise). The product space $E \times F$ is endowed with the norm $\sqrt{\|\cdot\|_E^2 + \|\cdot\|_F^2}$ which makes it a Hilbert space. The resolvent set of a closed linear operator $S : E \to F$ is denoted by $\rho(S)$. $R(\lambda, S)$ denotes (whenever it exists) the resolvent operator $(\lambda I - S)^{-1}$. If $\widetilde{E}$ is a subspace of E, then $S|_{\widetilde{E}}$ denotes the restriction of S to $\widetilde{E}$. For $s > \frac{1}{2}$, $H^s_{per}(0, p)$ denotes the Sobolev space of those p-periodic functions f (see [15]) satisfying $\sum_{k=-\infty}^{\infty} (1 + |\frac{2\pi k}{p}|^2)^s |\hat{f}(k)|^2 < \infty$, where $\hat{f}(k)$, $k \in \mathbb{Z}$, denotes a Fourier coefficient. $\Re(z)$ denotes the real part of a complex number z. A sequence $(f_n) \subset \mathbb{C}$ is $\mathcal{O}(g_n)$ for some positive sequence $(g_n) \subset \mathbb{R}$ as $n \to \infty$ if $|f_n| \leq M g_n$ for some $M > 0$ and all sufficiently large $n \geq 0$.

2. Formulation of the problem. In this section, we define the plant and the reference signals that we want to asymptotically track. Furthermore, we construct the exogenous systems that we assume are generating the reference signals. We also formulate the output regulation problem (ω_n, f_n)-RP.

2.1. The plant. We consider a plant described by the following (possibly infinite-dimensional) control system for $t \geq 0$:

$$\dot{z}(t) = Az(t) + Bu(t) + \mathcal{U}_{dist}(t), \tag{2.1a}$$

$$y(t) = Cz(t) + Du(t), \tag{2.1b}$$

$$z(0) = z_0 \in Z, \tag{2.1c}$$

where $z(t) \in Z$ is the state of the system (Z is a separable complex Hilbert space), A generates a C_0-semigroup of linear operators $T_A(t)$ on Z, $u(t) \in U$ is an input, and $y(t) \in Y$ is the output. The input space U is a complex separable Hilbert space and the output space $Y = \mathbb{C}$. The control operator $B \in \mathcal{L}(U, Z)$, the observation operator $C \in \mathcal{L}(Z, Y) = Z'$, and the feedthrough operator $D \in \mathcal{L}(U, Y)$. The pair (A, B) is assumed to be exponentially stabilizable; i.e., we assume that there exists $K \in \mathcal{L}(Z, U)$ such that $A+BK$ generates an exponentially stable C_0-semigroup on Z. The term $\mathcal{U}_{dist}(t)$ represents a disturbance (to be defined shortly). Finally, since we allow z_0 to be outside of $\mathcal{D}(A)$, (2.1a) is to be considered in the mild sense [6].

2.2. Sobolev spaces $H(f_n, \omega_n)$ and the exogenous system. Throughout this article, we assume that the periodic reference signals are in Sobolev spaces $H(\omega_n, f_n)$.

DEFINITION 2.1. *Let $I \subset \mathbb{Z}$, let $p > 0$, and let $\omega_n = \frac{2\pi n}{p}$ for every $n \in I$. Let $(f_n)_{n\in I} \subset \mathbb{R}$ such that $f_n \geq 1$ for each $n \in I$ and $(f_n^{-1})_{n\in I} \in \ell^2$. The Sobolev space $H(f_n, \omega_n)$ is defined as $\{\, u : \mathbb{R} \to \mathbb{C} \mid u(t) = \sum_{n\in I} a_n e^{i\omega_n t}$ for each $t \in \mathbb{R}$, $\sum_{n\in I} |f_n|^2|a_n|^2 < \infty$, and $(a_n)_{n\in I} \subset \mathbb{C}\,\}$.*

PROPOSITION 2.2. *Each $u \in H(f_n, \omega_n)$ is uniformly continuous and p-periodic. Moreover, $H(f_n, \omega_n)$ is a Hilbert space with respect to the inner product $\langle u, v\rangle_f = \sum_{n\in I} a_n\overline{b_n}|f_n|^2$. Here $u(t) = \sum_{n\in I} a_n e^{i\omega_n t}$ and $v(t) = \sum_{n\in I} b_n e^{i\omega_n t}$ for every $t \in \mathbb{R}$. Moreover, $\overline{b_n}$ denotes the complex conjugate of b_n.*

Proof. The uniform continuity and p-periodicity of $u(t)$ are evident. It is also easy to see that $H(f_n, \omega_n)$ is an inner product space. The completeness arguments follow those in [2, pp. 124–125]. ∎

These Sobolev spaces generalize certain Sobolev spaces of periodic functions [15] in the following way.

PROPOSITION 2.3. *Let $I = \mathbb{Z}$, $\gamma > \frac{1}{2}$, and $f_n = \sqrt{1+\omega_n^2}^{\,\gamma}$ for each $n \in \mathbb{Z}$. Then $H(f_n, \omega_n) = H^\gamma_{per}(0, p)$.*

Proof. By definition, $H^\gamma_{per}(0, p)$ is the Sobolev space of p-periodic functions u satisfying $\sum_{k=-\infty}^{\infty}(1 + |\frac{2\pi k}{p}|^2)^\gamma|\hat{u}(k)|^2 < \infty$, where $\hat{u}(k)$, $k \in \mathbb{Z}$, denotes a Fourier coefficient (see [15]). Hence $H^\gamma_{per}(0, p) \subset H(f_n, \omega_n)$. To prove the converse inclusion, we need only observe that absolutely and uniformly convergent trigonometric series of periodic functions are Fourier series (see, e.g., [2, p. 202]). ∎

Let the sequences $(\omega_n)_{n\in I}$ and $(f_n)_{n\in I}$ be fixed in the remainder of this subsection. We next construct an exogenous system which generates the functions in $H(f_n, \omega_n)$. To this end, let W be a Hilbert space with inner product $\langle\cdot,\cdot\rangle$ and an orthonormal basis $(\phi_n)_{n\in I}$. Set $S = \sum_{n\in I} i\omega_n\langle\cdot, \phi_n\rangle\phi_n$ with $\mathcal{D}(S) = \{\, w \in W \mid \sum_{n\in I} |\omega_n|^2|\langle w, \phi_n\rangle|^2 < \infty\,\}$. It is clear that S generates the C_0-semigroup $T_S(t) = \sum_{n\in I} e^{i\omega_n t}\langle\cdot, \phi_n\rangle\phi_n$ in W.

LEMMA 2.4. *Consider the sequence $(f_n)_{n\in I}$. Let $\mathcal{F} : \mathcal{D}(\mathcal{F}) \subset W \to W$ be such that*

$$\mathcal{F}w = \sum_{n\in I} f_n\langle w, \phi_n\rangle\phi_n, \quad \mathcal{D}(\mathcal{F}) = \left\{ w \in W \;\middle|\; \sum_{n\in I} |f_n|^2|\langle w, \phi_n\rangle|^2 < \infty \right\}. \tag{2.2}$$

Then the operator $\mathcal{F}$ is linear, closed, and densely defined in W. Moreover, $\mathcal{F}^{-1} \in \mathcal{L}(W)$.

Proof. The linearity of $\mathcal{F}$ is evident. Clearly all elements $w \in W$ for which $\langle w, \phi_n \rangle = 0$ for $|n| \geq M$ lie in $\mathcal{D}(\mathcal{F})$ and form a dense set in W. Hence $\mathcal{F}$ is densely defined in W. Let $(w_k)_{k\geq 0} \subset \mathcal{D}(\mathcal{F})$ be a sequence such that $w_k \to w$ as $k \to \infty$ and $\mathcal{F}w_k \to y \in W$ as $k \to \infty$. Since the sequence $(\mathcal{F}w_k)_{k\geq 0}$ converges, it is bounded, and $\sup_{k\geq 0} \sum_{n\in I} |f_n|^2 |\langle w_k, \phi_n \rangle|^2 < \infty$. This implies that $\sum_{n\in I} |f_n|^2 |\langle w, \phi_n \rangle|^2 < \infty$, i.e., $w \in \mathcal{D}(\mathcal{F})$. Now for each $n \in I$, $\langle y, \phi_n \rangle = \lim_{k\to\infty} \langle \mathcal{F}w_k, \phi_n \rangle = \lim_{k\to\infty} f_n \langle w_k, \phi_n \rangle = f_n \langle w, \phi_n \rangle = \langle \mathcal{F}w, \phi_n \rangle$, and so $\mathcal{F}w = y$. This proves that $\mathcal{F}$ is a closed operator. It is plain to see that $\mathcal{F}^{-1} = \sum_{n\in I} f_n^{-1} \langle \cdot, \phi_n \rangle \phi_n$ and that $\mathcal{F}^{-1}$ is bounded in W since $(f_n^{-1})_{n\in I} \in \ell^2$. ∎

THEOREM 2.5. *With $\mathcal{F}$ as in (2.2), define $\langle x, y \rangle_{\mathcal{F}} = \langle \mathcal{F}x, \mathcal{F}y \rangle$ for every $x, y \in \mathcal{D}(\mathcal{F})$. Then $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ is an inner product on $\mathcal{D}(\mathcal{F})$, and the space $W_{\mathcal{F}} = (\mathcal{D}(\mathcal{F}), \langle \cdot, \cdot \rangle_{\mathcal{F}})$ is a Hilbert space such that*

1. *$W_{\mathcal{F}} \hookrightarrow W$;*
2. *the restriction $T_S(t)|_{W_{\mathcal{F}}}$ of $T_S(t)$ to $W_{\mathcal{F}}$ is a C_0-semigroup on $W_{\mathcal{F}}$.*

Proof. Let $\|\cdot\|_{\mathcal{F}}$ denote the norm induced by $\langle \cdot, \cdot \rangle_{\mathcal{F}}$. It is readily verified that $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ satisfies the axioms of inner product because $f_n \geq 1$ for every $n \in I$. By Lemma 2.4, $\mathcal{F}$ is a closed operator with a bounded inverse. Let $|\cdot|_{\mathcal{F}}$ denote the graph norm in $\mathcal{D}(\mathcal{F})$. That $W_{\mathcal{F}}$ is a Hilbert space now follows from the estimates (valid for every $w \in \mathcal{D}(\mathcal{F})$)

$$|w|_{\mathcal{F}} = \|w\|_W + \|\mathcal{F}w\|_W = \|\mathcal{F}^{-1}\mathcal{F}w\|_W + \|\mathcal{F}w\|_W \tag{2.3}$$

$$\leq \big[\|\mathcal{F}^{-1}\|_{\mathcal{L}(W)} + 1\big] \|\mathcal{F}w\|_W = \big[\|\mathcal{F}^{-1}\|_{\mathcal{L}(W)} + 1\big] \|w\|_{\mathcal{F}} \tag{2.4}$$

$$\leq \big[\|\mathcal{F}^{-1}\|_{\mathcal{L}(W)} + 1\big] (\|\mathcal{F}w\|_W + \|w\|_W) \tag{2.5}$$

$$= \big[\|\mathcal{F}^{-1}\|_{\mathcal{L}(W)} + 1\big] |w|_{\mathcal{F}} \tag{2.6}$$

and the fact that $\mathcal{D}(\mathcal{F})$ is a Banach space when endowed with the graph norm. Additionally, since $f_n \geq 1$ for every $n \in I$, we have that

$$\|w\|_W^2 = \sum_{n\in I} |\langle w, \phi_n \rangle|^2 \leq \sum_{n\in I} |f_n|^2 |\langle w, \phi_n \rangle|^2 = \|w\|_{\mathcal{F}}^2 \quad \forall w \in \mathcal{D}(\mathcal{F}), \tag{2.7}$$

which shows that $W_{\mathcal{F}} \hookrightarrow W$.

It remains to show that the restriction $T_S(t)|_{W_{\mathcal{F}}}$ of $T_S(t)$ to $W_{\mathcal{F}}$ is a C_0-semigroup. For arbitrary $w \in \mathcal{D}(\mathcal{F})$,

$$\|T_S(t)w\|_{\mathcal{F}}^2 = \sum_{n\in I} |f_n|^2 |e^{i\omega_n t}|^2 |\langle w, \phi_n \rangle|^2 = \sum_{n\in I} |f_n|^2 |\langle w, \phi_n \rangle|^2 = \|w\|_{\mathcal{F}}^2, \tag{2.8}$$

which shows that $W_{\mathcal{F}}$ is $T_S(t)$-invariant and that $T_S(t)|_{W_{\mathcal{F}}}$ is an isometry (and hence bounded) for each $t \geq 0$. The semigroup property of $T_S(t)|_{W_{\mathcal{F}}}$ is easy to verify using the semigroup property of $T_S(t)$ in W and the $T_S(t)$-invariance of $W_{\mathcal{F}}$. We show the strong continuity of $T_S(t)|_{W_{\mathcal{F}}}$. Let $w \in \mathcal{D}(\mathcal{F})$. Then

$$\|T_S(t)|_{W_{\mathcal{F}}} w - w\|_{\mathcal{F}}^2 = \sum_{n\in I} |f_n|^2 |e^{i\omega_n t} - 1|^2 |\langle w, \phi_n \rangle|^2 = \sum_{n\in I} |(e^{i\omega_n t} - 1) f_n \langle w, \phi_n \rangle|^2 \tag{2.9}$$

$$= \|T_S(t)\mathcal{F}w - \mathcal{F}w\|_W^2 \to 0 \quad \text{as } t \to 0^+ \tag{2.10}$$

because $T_S(t)$ is strongly continuous on W. This completes the proof. ∎

LEMMA 2.6. *Let $Q : \mathcal{D}(Q) \subset W \to \mathbb{C}$ be such that*

$$Qw = \sum_{n\in I} \langle w, \phi_n\rangle \quad \forall w \in \mathcal{D}(Q) = \left\{ w \in W \;\middle|\; \left|\sum_{n\in I} \langle w, \phi_n\rangle\right| < \infty \right\}. \tag{2.11}$$

Then $Q \in \mathcal{L}(W_{\mathcal{F}}, \mathbb{C})$, where $W_{\mathcal{F}}$ is defined as in Theorem 2.5.

Proof. Since $(f_n^{-1})_{n\in I} \in \ell^2$, we have for some $M > 0$ by the Schwarz inequality that

$$|Qw| \le \sum_{n\in I} |\langle w, \phi_n\rangle||f_n f_n^{-1}| \le M\|w\|_{\mathcal{F}} \quad \forall w \in W_{\mathcal{F}}, \tag{2.12}$$

and so $Q \in \mathcal{L}(W_{\mathcal{F}}, \mathbb{C})$. ∎

DEFINITION 2.7 (the exogenous system). *Let $S_{\mathcal{F}}$ denote the generator of $T_S(t)|_{W_{\mathcal{F}}}$ on $W_{\mathcal{F}}$ (cf. Theorem 2.5). With the above definitions and notation, the exogenous system is given as*

$$\dot{w}(t) = S_{\mathcal{F}} w(t), \quad w(0) = w_0 \in W_{\mathcal{F}}, \tag{2.13a}$$

$$y_{ref}(t) = Qw(t), \quad t \ge 0, \tag{2.13b}$$

$$\mathcal{U}_{dist}(t) = Pw(t) \tag{2.13c}$$

on the state space $W_{\mathcal{F}}$. Here $P \in \mathcal{L}(W_{\mathcal{F}}, Z)$ is some known disturbance operator and (2.13a) *is to be considered in the mild sense.*

The exosystem (2.13) is capable of generating precisely the reference signals in the Sobolev space $H(f_n, \omega_n)$. Moreover, for each $y_{ref} \in H(f_n, \omega_n)$, there is exactly one initial state $w_0 \in W_{\mathcal{F}}$ such that $QT_S(t)|_{W_{\mathcal{F}}} w_0 = y_{ref}(t)$ for $t \ge 0$. These facts follow from the next theorem.

THEOREM 2.8. *There exists a bounded linear bijection $\mathcal{T} : H(f_n, \omega_n) \to W_{\mathcal{F}}$ such that for $y_{ref} \in H(f_n, \omega_n)$ we have that $\mathcal{T} y_{ref} = w_0$ with $QT_S(t)|_{W_{\mathcal{F}}} w_0 = y_{ref}(t)$ for each $t \ge 0$.*

Proof. Let $y_{ref} \in H(f_n, \omega_n)$ be such that $y_{ref}(t) = \sum_{n\in I} a_n e^{i\omega_n t}$. By our assumption, the estimate $\sum_{n\in I} |f_n|^2 |a_n|^2 < \infty$ holds. Define a mapping $\mathcal{T} : H(f_n, \omega_n) \to W_{\mathcal{F}}$ as

$$\mathcal{T} y_{ref} = \sum_{n\in I} a_n \phi_n = w_0. \tag{2.14}$$

Then $\mathcal{T}$ is linear and bounded. By the definition of $W_{\mathcal{F}}$, it is clear that $\mathcal{T}$ is surjective. If $\mathcal{T} y_{ref} = 0$, then $a_n = 0$ for every $n \in I$ by the orthonormality of $(\phi_n)_{n\in I}$ in W. Consequently $y_{ref} = 0$, and hence $\mathcal{T}$ is injective. Moreover, $T_S(t)|_{W_{\mathcal{F}}} w_0 = \sum_{n\in I} e^{i\omega_n t} a_n \phi_n$ for each $t \ge 0$, and so

$$QT_S(t)|_{W_{\mathcal{F}}} w_0 = \sum_{n\in I} a_n e^{i\omega_n t} = y_{ref}(t) \quad \forall t \ge 0. \tag{2.15}$$

The proof is complete. ∎

2.3. The output regulation problem (ω_n, f_n)-RP. Let $(\omega_n)_{n\in I}$ and $(f_n)_{n\in I}$ be as in Definition 2.1. Let $W_{\mathcal{F}}$ be as in Theorem 2.5 and let $S_{\mathcal{F}}$ be as in (2.13a). The task is to find a feedback control law

$$u(t) = Kz(t) + Lw(t) \tag{2.16}$$

such that $K \in \mathcal{L}(Z, U)$, $L \in \mathcal{L}(W_{\mathcal{F}}, U)$, and

- $A+BK$ is the generator of an exponentially stable C_0-semigroup $T_{A+BK}(t)$ on Z;
- for the closed loop system on $Z \times W_{\mathcal{F}}$ given by

$$\dot{z}(t) = (A+BK)z(t) + (BL+P)w(t), \tag{2.17a}$$

$$\dot{w}(t) = S_{\mathcal{F}}w(t), \tag{2.17b}$$

the tracking error

$$e(t) = y(t) - y_{ref}(t) = (C+DK)z(t) + (DL-Q)w(t) \to 0 \text{ as } t \to \infty \tag{2.18}$$

for all initial conditions $z(0) = z_0 \in Z$ and $w(0) = w_0 \in W_{\mathcal{F}}$.

3. A characterization of the solvability of the (ω_n, f_n)-RP. In Theorem 3.1 below we present a characterization for the solvability of the output regulation problem (ω_n, f_n)-RP in terms of the solvability of the so-called regulator equations, a decomposition property, and two continuity conditions. The result is an extension of Theorem IV.1 in [3] and Theorem 3.1 in [14]. In [3] Byrnes et al. proved this result for finite-dimensional exogenous systems, while in [14] the authors generalized Theorem IV.1 in [3] to cover output regulation of at least one given periodic reference signal (generated by an infinite-dimensional exosystem) in the case in which the feedthrough operator $D = 0$. In the exogenous system of [14] both the observation operator Q and the initial state $w(0)$ depend on the signal to be regulated; Theorem 3.1 in [14] provides a necessary and sufficient condition that this reference signal can be regulated. In Theorem 3.1 below we obtain a complete characterization for output regulation of all reference signals in a Sobolev-type space.

THEOREM 3.1. *Let $I \subset \mathbb{Z}$, $(\omega_n)_{n\in I}$, and $(f_n)_{n\in I}$ be fixed. Let the pair (A,B) be exponentially stabilizable with $K \in \mathcal{L}(Z,U)$. Then there exists an $L \in \mathcal{L}(W_{\mathcal{F}},U)$ such that the (ω_n, f_n)-RP is solvable using the control law $u(t) = Kz(t)+Lw(t)$ if and only if there exists a decomposition $L = \Gamma - K\Pi$, where $\Gamma \in \mathcal{L}(W_{\mathcal{F}},U)$ and $\Pi \in \mathcal{L}(W_{\mathcal{F}},Z)$ satisfy the following regulator equations for every $n \in I$:*

$$A\Pi\phi_n + B\Gamma\phi_n + P\phi_n = \Pi S_{\mathcal{F}}\phi_n, \tag{3.1a}$$

$$C\Pi\phi_n + D\Gamma\phi_n = 1. \tag{3.1b}$$

Proof (necessity). By the assumptions, the control law $u(t) = Kz(t) + Lw(t)$ solves the (ω_n, f_n)-RP. Hence by definition $L \in \mathcal{L}(W_{\mathcal{F}},U)$. Since $A+BK$ generates the exponentially stable C_0-semigroup $T_{A+BK}(t)$ on Z, the growth bound $\omega(T_{A+BK}) = \inf_{t>0}(\frac{1}{t}\log\|T_{A+BK}(t)\|) < 0$. On the other hand, $T_S(t)|_{W_{\mathcal{F}}}$ is an isometric (semi)group on each of the spaces $W_{\mathcal{F}}$, so its growth bound is 0. Corollary 8 in [21] then guarantees that the linear operator $\Pi : W_{\mathcal{F}} \to Z$ defined as

$$\Pi w = \int_0^\infty T_{A+BK}(t)(BL+P)T_S(-t)|_{W_{\mathcal{F}}} w dt \quad \forall w \in W_{\mathcal{F}} \tag{3.2}$$

is the unique bounded (i.e., $\mathcal{L}(W_{\mathcal{F}},Z)$) solution of the Sylvester-type operator equation $\Pi S_{\mathcal{F}} = (A+BK)\Pi + BL + P$ in $\mathcal{D}(S_{\mathcal{F}})$. Consequently, if we choose $\Gamma = L + K\Pi \in \mathcal{L}(W_{\mathcal{F}},U)$, it is clear that $\Pi S_{\mathcal{F}}\phi_n = A\Pi\phi_n + B\Gamma\phi_n + P\phi_n$ for each $n \in \mathbb{Z}$. We also have the decomposition $L = \Gamma - K\Pi$.

We next show that also the second regulator equation (3.1b) is satisfied with these choices of Π and Γ. To this end, consider the composite operator $\mathcal{A}$ on the composite state space $Z \times W_{\mathcal{F}}$ (see (2.17)) defined as

$$\mathcal{A} = \begin{pmatrix} A+BK & BL+P \\ 0 & S_{\mathcal{F}} \end{pmatrix}. \tag{3.3}$$

Since $A+BK$ generates the C_0-semigroup $T_{A+BK}(t)$ on Z and $S_{\mathcal{F}}$ generates the C_0-semigroup $T_S(t)|_{W_{\mathcal{F}}}$ on $W_{\mathcal{F}}$, it is clear that $\mathcal{A}$ generates a C_0-semigroup $T_{\mathcal{A}}(t)$ on $Z \times W_{\mathcal{F}}$ because $BL+P \in \mathcal{L}(W_{\mathcal{F}}, Z)$ (see also [6, Lemma 3.2.2]). An easy calculation reveals that this semigroup is given by

$$T_{\mathcal{A}}(t) = \begin{pmatrix} T_{A+BK}(t) & \int_0^t T_{A+BK}(\tau)(BL+P)T_S(t-\tau)|_{W_{\mathcal{F}}} d\tau \\ 0 & T_S(t)|_{W_{\mathcal{F}}} \end{pmatrix}. \tag{3.4}$$

Now choose an arbitrary eigenvector ϕ_n, $n \in I$, of $S_{\mathcal{F}}$. Then

$$T_{\mathcal{A}}(t)\left(\begin{smallmatrix} \Pi\phi_n \\ \phi_n \end{smallmatrix}\right) = \begin{pmatrix} T_{A+BK}(t)\Pi\phi_n + \int_0^t T_{A+BK}(\tau)(BL+P)e^{i\omega_n(t-\tau)}\phi_n d\tau \\ e^{i\omega_n t}\phi_n \end{pmatrix}. \tag{3.5}$$

According to the first regulator equation (3.1a), for each $t \geq \tau \geq 0$ we have that

$$(BL+P)e^{i\omega_n(t-\tau)}\phi_n = \big[\Pi S_{\mathcal{F}} - (A+BK)\Pi\big]e^{i\omega_n(t-\tau)}\phi_n. \tag{3.6}$$

A direct calculation then shows that for $0 < \tau < t$ (recall that $\Pi(\mathcal{D}(S_{\mathcal{F}})) \subset \mathcal{D}(A) = \mathcal{D}(A+BK)$)

$$\frac{d}{d\tau}\Big(T_{A+BK}(\tau)\Pi e^{i\omega_n(t-\tau)}\phi_n\Big) = T_{A+BK}(\tau)(A+BK)\Pi e^{i\omega_n(t-\tau)}\phi_n \tag{3.7}$$

$$- T_{A+BK}(\tau)\Pi i\omega_n e^{i\omega_n(t-\tau)}\phi_n \tag{3.8}$$

$$= T_{A+BK}(\tau)\big[(A+BK)\Pi - \Pi S_{\mathcal{F}}\big]e^{i\omega_n(t-\tau)}\phi_n. \tag{3.9}$$

Hence for $t \geq 0$,

$$\int_0^t T_{A+BK}(\tau)(BL+P)e^{i\omega_n(t-\tau)}\phi_n d\tau \tag{3.10}$$

$$= \int_0^t T_{A+BK}(\tau)\big[\Pi S_{\mathcal{F}} - (A+BK)\Pi\big]e^{i\omega_n(t-\tau)}\phi_n d\tau \tag{3.11}$$

$$= -\int_0^t \frac{d}{d\tau}\Big(T_{A+BK}(\tau)\Pi e^{i\omega_n(t-\tau)}\phi_n\Big)d\tau \tag{3.12}$$

$$= -T_{A+BK}(t)\Pi\phi_n + \Pi e^{i\omega_n t}\phi_n, \tag{3.13}$$

and so

$$T_{\mathcal{A}}(t)\left(\begin{smallmatrix} \Pi\phi_n \\ \phi_n \end{smallmatrix}\right) = \begin{pmatrix} T_{A+BK}(t)\Pi\phi_n - T_{A+BK}(t)\Pi\phi_n + \Pi e^{i\omega_n t}\phi_n \\ e^{i\omega_n t}\phi_n \end{pmatrix} = \begin{pmatrix} \Pi e^{i\omega_n t}\phi_n \\ e^{i\omega_n t}\phi_n \end{pmatrix}. \tag{3.14}$$

Since the (ω_n, f_n)-RP is solvable, the tracking error corresponding to the particular initial states $z(0) = \Pi\phi_n \in Z$ and $w(0) = \phi_n \in W_{\mathcal{F}}$ (for arbitrary $n \in I$) satisfies

$$e(t) = (C+DK, DL-Q)T_{\mathcal{A}}(t)\left(\begin{smallmatrix} \Pi\phi_n \\ \phi_n \end{smallmatrix}\right) = (C\Pi + DK\Pi + DL - Q)e^{i\omega_n t}\phi_n \tag{3.15}$$

$$= (C\Pi + D\Gamma - Q)e^{i\omega_n t}\phi_n, \tag{3.16}$$

which tends to 0 as $t \to \infty$. But this is possible only if $C\Pi\phi_n + D\Gamma\phi_n = Q\phi_n = 1$ for every $n \in I$. Hence also the second regulator equation (3.1b) is satisfied.

(Sufficiency) Choose $L = \Gamma - K\Pi \in \mathcal{L}(W_{\mathcal{F}}, U)$, where $\Pi \in \mathcal{L}(W_{\mathcal{F}}, Z)$ and $\Gamma \in \mathcal{L}(W_{\mathcal{F}}, U)$ satisfy the regulator equations (3.1) for every $n \in I$. Consider the control law $u(t) = Kz(t) + Lw(t)$. Since by the assumptions the C_0-semigroup $T_{A+BK}(t)$ generated by $A + BK$ is exponentially stable, it remains to show that the error term $e(t) \to 0$ as $t \to \infty$ for any $z(0) = z_0 \in Z$ and $w(0) = w_0 \in W_{\mathcal{F}}$.

Let $z_0 \in Z$ and $w_0 \in W_{\mathcal{F}}$ be arbitrary. Consider the semigroup $T_{\mathcal{A}}(t)$ (see (3.4)) generated by $\mathcal{A}$ (see (3.3)) on $Z \times W_{\mathcal{F}}$. Then

$$T_{\mathcal{A}}(t)\left(\begin{smallmatrix} z_0 \\ w_0 \end{smallmatrix}\right) = \begin{pmatrix} T_{A+BK}(t)z_0 + \int_0^t T_{A+BK}(\tau)(BL+P)T_S(t-\tau)|_{W_{\mathcal{F}}} w_0 d\tau \\ T_S(t)|_{W_{\mathcal{F}}} w_0 \end{pmatrix}. \tag{3.17}$$

Now an application of the Lebesgue dominated convergence theorem and relations (3.11)–(3.13) yields

$$\int_0^t T_{A+BK}(\tau)(BL+P)T_S(t-\tau)|_{W_{\mathcal{F}}} w_0 d\tau \tag{3.18}$$

$$= \int_0^t T_{A+BK}(\tau)(BL+P) \sum_{n\in I} e^{i\omega_n(t-\tau)} \langle w_0, \phi_n \rangle \phi_n d\tau \tag{3.19}$$

$$= \sum_{n\in I} \langle w_0, \phi_n \rangle \int_0^t T_{A+BK}(\tau)(BL+P) e^{i\omega_n(t-\tau)} \phi_n d\tau \tag{3.20}$$

$$= \sum_{n\in I} \langle w_0, \phi_n \rangle \left[-T_{A+BK}(t)\Pi\phi_n + \Pi e^{i\omega_n t}\phi_n \right] \tag{3.21}$$

$$= \Pi T_S(t)|_{W_{\mathcal{F}}} w_0 - T_{A+BK}(t)\Pi w_0 \quad \forall t \geq 0, \tag{3.22}$$

since $\Pi \in \mathcal{L}(W_{\mathcal{F}}, Z)$ and $w_0 \in W_{\mathcal{F}}$.

Using this information we can work out the explicit expression for the tracking error $e(t)$ as follows.

(3.23)
$$e(t) = (C + DK, DL - Q) T_{\mathcal{A}}(t)\left(\begin{smallmatrix} z_0 \\ w_0 \end{smallmatrix}\right)$$

(3.24)
$$= (C+DK)T_{A+BK}(t)z_0 + (C+DK)\int_0^t T_{A+BK}(\tau)(BL+P)T_S(t-\tau)|_{W_{\mathcal{F}}} w_0 d\tau$$

(3.25)
$$+ (DL-Q)T_S(t)|_{W_{\mathcal{F}}} w_0$$

(3.26)
$$= (C+DK)T_{A+BK}(t)(z_0 - \Pi w_0) + (C\Pi + DK\Pi + DL - Q)T_S(t)|_{W_{\mathcal{F}}} w_0$$

(3.27)
$$= (C+DK)T_{A+BK}(t)(z_0 - \Pi w_0) + (C\Pi + D\Gamma - Q)T_S(t)|_{W_{\mathcal{F}}} w_0$$

(3.28)
$$= (C+DK)T_{A+BK}(t)(z_0 - \Pi w_0) \quad \forall t \geq 0$$

by the second regulator equation (3.1b) and the fact that the operator $C\Pi + D\Gamma - Q \in \mathcal{L}(W_{\mathcal{F}}, Y)$.

By our assumption, $T_{A+BK}(t)$ is exponentially stable. Consequently, $e(t) \to 0$ for every $z_0 \in Z$ and $w_0 \in W_{\mathcal{F}}$. This shows that the control law $u(t) = Kz(t) + Lw(t)$ solves the (ω_n, f_n)-RP. The proof is complete. ∎

Remark 3.2. In the sufficiency part of Theorem 3.1, mere weak stability of the semigroup $T_{A+BK}(t)$ (i.e., that $f(T_{A+BK}(t)z_0) \to 0$ for every $z_0 \in Z$ and every $f \in Z'$ as $t \to \infty$) would guarantee that $|e(t)| \to 0$ as $t \to \infty$. Whenever the semigroup $T_{A+BK}(t)$ is exponentially stable, we in fact obtain exponentially fast decay of $|e(t)|$.

Remark 3.3. Theorem 3.1 shows that if $u(t) = Kz(t) + Lw(t)$ solves the (ω_n, f_n)-RP, then *necessarily* $L = \Gamma - K\Pi$, i.e., the operators K and L cannot be independent of each other. This shows that although small additive bounded perturbations to K do not affect exponential stability of $T_{A+BK}(t)$ [8], such perturbations do in general destroy output regulation. To the authors' knowledge this fact has not been explicitly stated before in related earlier work, e.g., [3, 14].

4. Solution of the regulator equations—SISO systems. In this section we solve the regulator equations (3.1) for SISO systems under the assumption that the stabilized plant does not have transmission zeros at the Fourier frequencies $i\omega_n$ of the reference signals. These solutions are then used to construct a candidate operator L for the solution of the output regulation problem (ω_n, f_n)-RP; its continuity determines whether or not the problem is solvable. Using this explicit series representation for L, in condition (4.17) we completely characterize the solvability of the (ω_n, f_n)-RP by the growth of the transfer function of the stabilized plant on the imaginary axis. Furthermore, we obtain an explicit expression, in terms of the reference signals, for the control law which achieves output regulation. Our arguments are similar to those in [14]; however, our results are more complete here. In particular, for SISO systems we derive the verifiable condition (4.17) which completely answers the question posed in the title.

Throughout this section, we assume that the plant is a SISO system and that the sequences $(\omega_n)_{n\in I}$ and $(f_n)_{n\in I}$ (see Definition 2.1) are fixed. Moreover, we assume that the stabilizing feedback $K \in \mathcal{L}(Z, U)$ for the pair (A, B) is fixed.

DEFINITION 4.1. *The transfer function $H(s)$ of the plant is defined as $H(s) = CR(s, A)B + D$ for every $s \in \rho(A)$. The transfer function $H_K(s)$ of the stabilized plant is defined as $H_K(s) = (C + DK)R(s, A + BK)B + D$ for $s \in \rho(A + BK)$. The sequence of disturbance coefficients for the stabilized plant is defined as $(H_d(n))_{n\in I} = ((C + DK)R(i\omega_n, A + BK)P\phi_n)_{n\in I} \subset \mathbb{C}$.*

DEFINITION 4.2. *The plant (respectively, stabilized plant) has a transmission zero at $s = s_0$ if $H(s_0) = 0$ (respectively, $H_K(s_0) = 0$).*

The next lemma shows that in $\rho(A)$ the concept of transmission zero does not depend on K; for the case $D = 0$ the result was stated in Lemma V.2 of [3].

LEMMA 4.3. *Let $s_0 \in \rho(A) \cap \rho(A + BK)$. Then the plant has a transmission zero at $s = s_0$ if and only if the stabilized plant has a transmission zero at $s = s_0$.*

Proof. Let $s = s_0$ be a transmission zero of the plant. Clearly $CR(s_0, A)B + D = 0$ if and only if

$$\ker \begin{pmatrix} C & D \\ s_0 I - A & -B \end{pmatrix} \neq \{0\}, \tag{4.1}$$

where the domain of definition of the operator $\mathcal{R} = \left(\begin{smallmatrix} C & D \\ s_0 I - A & -B \end{smallmatrix}\right)$ is $\mathcal{D}(A) \times \mathbb{C}$. Let

$0 \neq \left(\begin{smallmatrix} x \\ u \end{smallmatrix}\right) \in \ker \mathcal{R}$. Then since $x = R(s_0, A)Bu$, we must have that $u \neq 0$. Moreover,

$$\begin{pmatrix} C & D \\ s_0 I - A & -B \end{pmatrix} \begin{pmatrix} I & 0 \\ K & I \end{pmatrix} \begin{pmatrix} I & 0 \\ -K & I \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix} = 0, \tag{4.2}$$

which implies

$$\begin{pmatrix} C + DK & D \\ s_0 I - A - BK & -B \end{pmatrix} \begin{pmatrix} x \\ u - Kx \end{pmatrix} = 0. \tag{4.3}$$

Let $\mathcal{R}_K = \left(\begin{smallmatrix} C+DK & D \\ s_0 I - A - BK & -B \end{smallmatrix}\right)$ with $\mathcal{D}(\mathcal{R}_K) = \mathcal{D}(\mathcal{R})$. If $x = 0$, then $0 \neq \left(\begin{smallmatrix} 0 \\ u \end{smallmatrix}\right) \in \ker \mathcal{R}_K$. On the other hand, if $x \neq 0$, then $0 \neq \left(\begin{smallmatrix} x \\ u - Kx \end{smallmatrix}\right) \in \ker \mathcal{R}_K$. In any case $\ker \mathcal{R}_K \neq \{0\}$. The above means that $(C + DK)R(s_0, A + BK)B + D = 0$, i.e., that the stabilized plant has a transmission zero at $s = s_0$.

Similar arguments show that the converse also holds. We omit the details. ∎

PROPOSITION 4.4. *Let $n \in I$. If the stabilized plant does not have a transmission zero at $s = i\omega_n$, and if we define $L\phi_n = H_K(i\omega_n)^{-1}[1 - H_d(n)]$, then*

$$\Pi\phi_n = R(i\omega_n, A + BK)\big[BL\phi_n + P\phi_n\big], \tag{4.4}$$

$$\Gamma\phi_n = L\phi_n + K\Pi\phi_n \tag{4.5}$$

are solutions of the regulator equations (3.1).

Proof. Since $S\phi_n = i\omega_n\phi_n$, it is clear that if we solve the equations

$$(A + BK)\Pi\phi_n + BL\phi_n + P\phi_n = i\omega_n\Pi\phi_n, \tag{4.6a}$$

$$(C + DK)\Pi\phi_n + DL\phi_n = 1 \tag{4.6b}$$

for $\Pi\phi_n$ and $L\phi_n$, and then set $\Gamma\phi_n = L\phi_n + K\Pi\phi_n$, we simultaneously solve regulator equations (3.1) for $\Pi\phi_n$ and $\Gamma\phi_n$. From (4.6a) we obtain

$$(A + BK)\Pi\phi_n + BL\phi_n + P\phi_n = i\omega_n\Pi\phi_n \Leftrightarrow \Pi\phi_n = R(i\omega_n, A + BK)(BL\phi_n + P\phi_n) \tag{4.7}$$

because $A + BK$ generates an exponentially stable C_0-semigroup. Applying this expression for $\Pi\phi_n$ to (4.6b) yields

$$(C + DK)R(i\omega_n, A + BK)(BL\phi_n + P\phi_n) + DL\phi_n = 1 \tag{4.8}$$

$$\Leftrightarrow \big[(C + DK)R(i\omega_n, A + BK)B + D\big]L\phi_n + (C + DK)R(i\omega_n, A + BK)P\phi_n = 1 \tag{4.9}$$

$$\Leftrightarrow H_K(i\omega_n)L\phi_n + H_d(n) = 1 \tag{4.10}$$

$$\Leftrightarrow H_K(i\omega_n)^{-1}\big[1 - H_d(n)\big] = L\phi_n \tag{4.11}$$

by the assumption that $H_K(i\omega_n) \neq 0$. Hence equations (4.6) have a unique solution which is also a solution of the regulator equations (3.1). ∎

THEOREM 4.5. *Suppose that for every $n \in I$, $s = i\omega_n$ is not a transmission zero of the stabilized plant. Define*

$$L = \sum_{n \in I} H_K(i\omega_n)^{-1}[1 - H_d(n)]\langle \cdot, \phi_n \rangle. \tag{4.12}$$

Then the (ω_n, f_n)-RP is solvable using the control law $u(t) = Kz(t) + Lw(t)$ if and only if $L \in \mathcal{L}(W_{\mathcal{F}}, U)$.

Proof (necessity). If the control law $u(t) = Kz(t) + Lw(t)$ solves the (ω_n, f_n)-RP, then by definition L must be in $\mathcal{L}(W_{\mathcal{F}}, U)$.

(Sufficiency) Suppose that $L \in \mathcal{L}(W_{\mathcal{F}}, U)$. Then the linear operator $\Pi : W_{\mathcal{F}} \to Z$ defined by $\Pi w = \int_0^\infty T_{A+BK}(\tau)(BL+P)T_S(-\tau)|_{W_{\mathcal{F}}} w d\tau$ for each $w \in W_{\mathcal{F}}$ is in $\mathcal{L}(W_{\mathcal{F}}, Z)$. Since $T_S(-t)|_{W_{\mathcal{F}}}\phi_n = e^{-i\omega_n t}\phi_n$ and $R(i\omega_n, A+BK)z = \int_0^\infty e^{-i\omega_n t} T_{A+BK}(t) z dt$ for every $z \in Z$ (Proposition 5.1.5 in [1]), we have that $\Pi\phi_n = R(i\omega_n, A+BK)(BL+P)\phi_n$ for each n. Consequently, by Proposition 4.4 the operators Π and $\Gamma = K\Pi + L \in \mathcal{L}(W_{\mathcal{F}}, U)$ solve the regulator equations (3.1) for every $n \in I$. Theorem 3.1 then guarantees that the (ω_n, f_n)-RP is solvable with the control law $u(t) = Kz(t) + Lw(t)$. ∎

COROLLARY 4.6. *Suppose that the assumptions of Theorem* 4.5 *are satisfied and that* L *defined in* (4.12) *is in* $\mathcal{L}(W_{\mathcal{F}}, U)$*, so that* (ω_n, f_n)*-RP is solvable using* $u(t) = Kz(t) + Lw(t)$*. Then for every* $y_{ref} \in H(f_n, \omega_n)$ *the corresponding control law* $u_{y_{ref}}(t)$ *which achieves output regulation of* $y_{ref}(t)$ *is given by*

$$u_{y_{ref}}(t) = Kz(t) + \sum_{n\in I} H_K(i\omega_n)^{-1}[1 - H_d(n)] y_n e^{i\omega_n t} \quad \forall t \geq 0, \tag{4.13}$$

where $y_{ref}(t) = \sum_{n\in I} y_n e^{i\omega_n t}$ *for each* t.

Proof. Let $y_{ref} \in H(f_n, \omega_n)$ be fixed, and let $y_{ref}(t) = \sum_{n\in I} y_n e^{i\omega_n t}$ for each t. By Theorem 2.8 the corresponding initial state of the exosystem $w(0) = \sum_{n\in I} y_n \phi_n \in W_{\mathcal{F}}$. Using continuity we work out $Lw(t) = LT_S(t)|_{W_{\mathcal{F}}} w(0)$ as

$$LT_S(t)|_{W_{\mathcal{F}}} w(0) = \sum_{n\in I} y_n LT_S(t)|_{W_{\mathcal{F}}}\phi_n \tag{4.14}$$

$$= \sum_{n\in I} y_n L e^{i\omega_n t}\phi_n \tag{4.15}$$

$$= \sum_{n\in I} y_n e^{i\omega_n t} H_K(i\omega_n)^{-1}[1 - H_d(n)] \quad \forall t \geq 0 \tag{4.16}$$

because $\phi_n \in W_{\mathcal{F}}$ for every $n \in I$ and $T_S(t)|_{W_{\mathcal{F}}}\phi_n = T_S(t)\phi_n = e^{i\omega_n t}\phi_n$ for each $n \in I$. The proof is completed by the observation that in our construction the control law $u(t) = Kz(t) + LT_S(t)|_{W_{\mathcal{F}}} w(0)$ achieves asymptotic tracking of $QT_S(t)|_{W_{\mathcal{F}}} w(0) = y_{ref}(t)$. ∎

In particular, if there are no disturbances and the plant is already exponentially stable, then the control law (4.13) reduces to the remarkably simple $u_{y_{ref}}(t) = \sum_{n\in I} H(i\omega_n)^{-1} y_n e^{i\omega_n t}$. Corollary 4.6 shows, under certain assumptions, that once we know that (ω_n, f_n)-RP is solvable, knowledge of the dynamical behavior of the exogenous system is irrelevant for asymptotic tracking of the reference signals (we need to have knowledge of the sequence $(P\phi_n)_{n\in I} \subset Z$ though): It is irrelevant *how* the control signal (4.13) is generated.

The following corollary characterizes the solvability of the (ω_n, f_n)-RP by the asymptotic behavior of $H_K(i\omega_n)^{-1}[1 - H_d(n)]$ as $n \to \pm\infty$.

COROLLARY 4.7. *Suppose that the assumptions of Theorem* 4.5 *are satisfied. Let* L *be defined as in* (4.12)*. Then the control law* $u(t) = Kz(t) + Lw(t)$ *solves the* (ω_n, f_n)*-RP if and only if*

$$\left(H_K(i\omega_n)^{-1}[1 - H_d(n)] f_n^{-1}\right)_{n\in I} \in \ell^2. \tag{4.17}$$

In the disturbance-free case (i.e., whenever $P = 0$*), the above condition reduces to* $(H_K(i\omega_n)^{-1} f_n^{-1})_{n\in I} \in \ell^2$.

Proof. Since $W_{\mathcal{F}}$ is a Hilbert space, by the Riesz representation theorem $L \in \mathcal{L}(W_{\mathcal{F}}, U) = \mathcal{L}(W_{\mathcal{F}}, \mathbb{C})$ if and only if there exists a unique element $l \in W_{\mathcal{F}}$ such that $Lw = \langle w, l\rangle_{\mathcal{F}}$ for every $w \in W_{\mathcal{F}}$. Then we must have that $\langle \phi_n, l\rangle |f_n|^2 = H_K(i\omega_n)^{-1}[1 - H_d(n)]$, or $\langle l, \phi_n\rangle = \overline{H_K(i\omega_n)^{-1}[1 - H_d(n)]}|f_n|^{-2}$ for every $n \in I$. But the element l thus defined is in $W_{\mathcal{F}}$ if and only if

$$\sum_{n\in I} |\langle l, \phi_n\rangle|^2 |f_n|^2 = \sum_{n\in I} |H_K(i\omega_n)^{-1}[1 - H_d(n)]|^2 |f_n|^{-2} < \infty. \tag{4.18}$$

This and Theorem 4.5 give the desired result. ∎

The above results formalize the intuitive idea that in order to be able to track a periodic reference signal, the stabilized plant should not attenuate high frequency oscillations too drastically and at the same time the reference signal should be smooth enough. We conclude this section with some results which in some cases simplify the verification of condition (4.17).

THEOREM 4.8. *Let A generate an exponentially stable C_0-semigroup and let $A + BK$, for $K \in \mathcal{L}(Z, U)$, also generate an exponentially stable C_0-semigroup. Then there exist $m, M \geq 0$ (which do not depend on $n \in I$) such that $\|CR(i\omega_n, A)B\| \leq m\|CR(i\omega_n, A + BK)B\| \leq M\|CR(i\omega_n, A)B\|$ for each $n \in I$.*

Proof. By an elementary calculation, we have that $CR(i\omega_n, A)B[I + KR(i\omega_n, A + BK)B] = CR(i\omega_n, A + BK)B$ and that $CR(i\omega_n, A)B = CR(i\omega_n, A + BK)B[I - KR(i\omega_n, A)B]$ for every $n \in I$. Since A and $A + BK$ generate exponentially stable C_0-semigroups, $\|R(i\omega_n, A)\|$ and $\|R(i\omega_n, A + BK)\|$ are uniformly bounded in n, according to the Riemann–Lebesgue lemma [6]. The desired conclusion now follows by some obvious norm estimates. ∎

According to Theorem 4.8, if $D = 0$, if there are no disturbances, and if both A and $A + BK$ generate exponentially stable C_0-semigroups, then $(H(i\omega_n)f_n^{-1})_{n\in I} \in \ell^2$ if and only if $(H_K(i\omega_n)f_n^{-1})_{n\in I} \in \ell^2$. In particular, the capability of output regulation is an intrinsic property of the plant which is independent of the stabilizing feedback K.

COROLLARY 4.9. *Let A and $A + BK$, where $K \in \mathcal{L}(Z, U)$, generate exponentially stable analytic C_0-semigroups.*

1. *For $D = 0$, condition (4.17) holds if $(H(i\omega_n)^{-1}[1 - H_d(n)]f_n^{-1})_{n\in I} \in \ell^2$.*
2. *For $D \neq 0$, condition (4.17) holds if $H(i\omega_n) \neq 0$ for each $n \in I$ and $([1 - H_d(n)]f_n^{-1})_{n\in I} \in \ell^2$.*

Proof. The case $D = 0$ is settled by Theorem 4.8.

Let $D \neq 0$. By exponential stability and Lemma 4.3, for all $n \in I$ we have $H_K(i\omega_n) \neq 0$. Since A and $A + BK$ generate exponentially stable analytic semigroups, we have $\lim_{n\to\pm\infty} H_K(i\omega_n) = D \neq 0$. Consequently, for some $\delta > 0$ we have $\delta < \inf_{n\in I} |H_K(i\omega_n)| < \sup_{n\in I} |H_K(i\omega_n)| < \infty$, and so

$$\left|H_K(i\omega_n)^{-1}[1 - H_d(n)]f_n^{-1}\right| \leq \frac{1}{\delta}\left|[1 - H_d(n)]f_n^{-1}\right| \quad \forall n \in I. \tag{4.19}$$

This shows that condition (4.17) holds if $([1 - H_d(n)]f_n^{-1})_{n\in I} \in \ell^2$. ∎

5. Examples.

Example 5.1. Consider a finite-dimensional exponentially stable SISO plant that is not subject to any disturbances (i.e., $P = 0$). Consider reference signals in the Sobolev space $H^{\gamma}_{per}(0, p)$, $\gamma > \frac{1}{2}$; i.e., set $I = \mathbb{Z}$ and $f_n = \sqrt{1 + \omega_n^2}^{\,\gamma}$ for each $n \in \mathbb{Z}$. Let N denote the relative degree of the transfer function $H(s)$ of the plant, and assume that there are no transmission zeros in the set of Fourier frequencies $\{\, i\omega_n \mid n \in \mathbb{Z} \,\}$ of the reference signals.

By the relative degree condition, $H(i\omega_n)^{-1}$ is $\mathcal{O}(|\omega_n|^N)$ as $n \to \pm\infty$. If we define L as in (4.12) (with $K = 0$ since the plant is already stable), then it is easy to see that $L \in \mathcal{L}(W_{\mathcal{F}}, U)$ if $\gamma > N + \frac{1}{2}$. This implies that for such γ, all reference signals $y_{ref} \in H^\gamma_{per}(0,p)$ can be asymptotically tracked using the control law $u(t) = Lw(t)$ by Theorem 4.5. On the other hand, for $\gamma \leq N + \frac{1}{2}$, there are reference signals in $H^\gamma_{per}(0,p)$ which cannot be asymptotically tracked by Corollary 4.7.

Example 5.2. Consider the following scalar delay differential equation [19] with control and observation. Let $a > 0$, $r \neq 0$, $\tau_1 > \tau_2 > 0$, and

$$\dot{x}(t) = -ax(t) - b\big[x(t-\tau_1) + x(t-\tau_2)\big] + u(t), \tag{5.1a}$$

$$y(t) = rx(t), \quad t \geq 0. \tag{5.1b}$$

Taking initial conditions for $x(\cdot)$ into account, the pair (5.1) can be formulated as a plant of the form (2.1) in which $D = 0$ and $\mathcal{U}_{dist} = 0$ [6]. Moreover, it can be shown (see, e.g., [6, Lemma 4.3.9]) that the transfer function $H(s) = CR(s,A)B$ of the plant is given by

$$H(s) = \frac{r}{s + a + b(e^{-s\tau_1} + e^{-s\tau_2})} \tag{5.2}$$

for those $s \in \mathbb{C}$ at which the denominator is not equal to zero.

The semigroup generated by A is exponentially stable if and only if $s + a + b(e^{-s\tau_1} + e^{-s\tau_2}) \neq 0$ for all $s \in \{\, z \in \mathbb{C} \mid \Re(z) \geq 0 \,\}$ [6, Theorem 5.1.7]. Ruan and Wei [19] give a complete characterization (in terms of a, b, τ_1, and τ_2) of those instances in which all roots of equation $s + a + b(e^{-s\tau_1} + e^{-s\tau_2}) = 0$ have negative real parts. In their characterization, the parameter b lies on an interval (b_0^-, b_0^+). We assume that the semigroup generated by A is exponentially stable. By the above discussion, then $i\omega_n \in \rho(A)$ and $H(i\omega_n) \neq 0$ for every $n \in \mathbb{Z}$.

It is evident that for every $\gamma > \frac{3}{2}$, $\sum_{n=-\infty}^{\infty} |H(i\omega_n)^{-1}|^2 (1+\omega_n^2)^{-\gamma} < \infty$, and that for every $\gamma \leq \frac{3}{2}$, $\sum_{n=-\infty}^{\infty} |H(i\omega_n)^{-1}|^2 (1+\omega_n^2)^{-\gamma} = \infty$. Consequently, by Corollary 4.7 the system can track all reference signals in $H^\gamma_{per}(0,p)$ for $\gamma > \frac{3}{2}$. On the other hand, for $\gamma \leq \frac{3}{2}$ in every Sobolev space $H^\gamma_{per}(0,p)$ there are reference signals which cannot be asymptotically tracked.

Example 5.3. Consider a disturbance-free controlled one-dimensional heat equation on the interval $[0,1]$ with Neumann boundary conditions $\frac{\partial z(x,t)}{\partial t} = \frac{\partial^2 z(x,t)}{\partial^2 x} + Bu(t)$, $\frac{\partial z(0,t)}{\partial t} = \frac{\partial z(1,t)}{\partial t} = 0$, $z(x,0) = \psi(x)$. The output is given as $y(t) = Cz(t)$. The bounded control operator $B : \mathbb{C} \to L^2(0,1)$ is defined by $Bu = b(x)u$, with $b(x) = 2\chi_{[\frac{1}{2},1]}(x)$. Here $\chi_{[\epsilon,\delta]}(x)$ denotes the characteristic function of the interval $[\epsilon,\delta]$. The bounded observation operator $C : L^2(0,1) \to \mathbb{C}$ is defined by $C\psi = \int_0^1 c(x)\psi(x)dx$, with $c(x) = 2\chi_{[0,\frac{1}{2}]}(x)$.

It is well known how to put this system in the form (2.1) [3, 6, 14]. It can also be shown [3] that the transfer function of this heat plant is $H(s) = \frac{2\sinh(\sqrt{s}/2)}{s\sqrt{s}\cosh(\sqrt{s}/2)}$ for $s \in \rho(A)$. Now $i\omega_n = i\frac{2\pi n}{p} \in \rho(A)$ for $n \neq 0$ [3], and $i\omega_n$ is not a transmission zero of this plant for $n \neq 0$.

Let $I = \mathbb{Z} \setminus \{0\}$, and let $f_n = \sqrt{1+\omega_n^2}^{\,\gamma}$. Let K be any bounded exponentially stabilizing feedback for the pair (A,B) (such a K is of course known to exist [3]). Then by Lemma 4.3, $H_K(i\omega_n)^{-1}$ exists for $n \neq 0$. Let us define L as in (4.12) (with $P = 0$ and $D = 0$). By some elementary calculations [14], in this case $H_K(i\omega_n)^{-1} = H(i\omega_n)^{-1}[I - KR(i\omega_n, A)B]$ for every $n \in I$. It is easy to see that $\|I - KR(i\omega_n, A)B\|$

is uniformly bounded for $n \neq 0$. Moreover, $H(i\omega_n)^{-1} = \mathcal{O}(|\omega_n|^{\frac{3}{2}})$ as $n \to \pm\infty$. Consequently $(H_K(i\omega_n)^{-1}\sqrt{1+\omega_n^2}^{-\gamma})_{n\in I} \in \ell^2$ if $\gamma > 2$. By Corollary 4.7, this system is capable of asymptotically tracking those periodic reference signals in $H^\gamma_{per}(0,p)$, with $\gamma > 2$, that lack the constant term in the Fourier series description. More accurate information on the signals which can be asymptotically tracked could be obtained by working out the explicit expression for $H_K(s)$.

We remark that Byrnes et al. (see sections III and VI of [3]) have thoroughly studied and simulated output regulation problems for the above heat plant in the case of constant and sinusoidal reference/disturbance signals. On the other hand, while in [14] the above system was used to track one p-periodic reference signal, here we may use Corollary 4.6 to track all sufficiently smooth p-periodic signals which lack the constant term in the Fourier series description.

Example 5.4. In this example we show that there exist infinite-dimensional systems which cannot track all reference signals in $H^\gamma_{per}(0,p)$ for any $\gamma > \frac{1}{2}$, even if there are no transmission zeros in the set of Fourier frequencies of the reference signals. This is in strong contrast to the finite-dimensional case, as is seen from Example 5.1 above. We refer the reader to [6, 20] for relevant notation and definitions.

Let $f \in \mathcal{D}(\mathbb{R})$ be a test function such that $\operatorname{supp}(f) \subset [0,a]$, where $0 < a < \infty$. Let $Z = \{\, g \in H^1(0,a) \mid g(a) = 0 \,\}$, where $H^1(0,a)$ denotes the standard Sobolev space. Since Z is the null space of a continuous linear functional, it is a closed subspace of $H^1(0,a)$. Let A be the generator of the left shift semigroup $T_A(t)$ on Z defined as $(T_A(t)g)(x) = g(x+t)$ for $x + t \leq a$, and $(T_A(t)g)(x) = 0$ otherwise, for every $g \in Z$. Clearly $T_A(t)$ is exponentially stable [8]. Let C be the point evaluation at the origin, i.e., $Cg = g(0)$ for every $g \in Z$. It is easy to show (see, e.g., [13]) that $C \in \mathcal{L}(Z,\mathbb{C})$. Finally, let $Bu = fu$ for $u \in \mathbb{C}$. Then evidently $B \in \mathcal{L}(\mathbb{C},Z)$. Moreover, the system (2.1) (with $D = 0$ and $\mathcal{U}_{dist} = 0$) has $f(t)$ as its impulse response [6]. In fact, $CT_A(t)B = [f(x+t)]_{x=0} = f(t)$ for every $t \geq 0$.

By applying Fourier transforms, we see that $H(i\omega) = \mathcal{F}(f)(i\omega)$ is a rapidly decreasing function. Hence $\sup_{\omega\in\mathbb{R}}(1+\omega^2)^N|H(i\omega)| < \infty$ for every $N \in \mathbb{N}$. For the purpose of output regulation, we may assume that $H(i\omega_n) \neq 0$ for every $n \in \mathbb{Z}$. Otherwise there would exist $m \in \mathbb{Z}$ such that the regulator equations (3.1) are not solvable for this m, and hence there would exist an infinite-dimensional system which cannot track all reference signals in $H^\gamma_{per}(0,p)$ for any $\gamma > \frac{1}{2}$. But by the above, $H(i\omega_n)^{-1}$ grows faster than every polynomial in n as $n \to \pm\infty$. By Corollary 4.7 and the fact that L (if it exists in $\mathcal{L}(W_{\mathcal{F}},\mathbb{C})$) is unique in this case, for arbitrary $\gamma > \frac{1}{2}$ there always exists $y_{ref} \in H^\gamma_{per}(0,p)$ which this system cannot asymptotically track using the control law $u(t) = Lw(t)$. In other words, if the system can asymptotically track all signals in $H(\omega_n, f_n)$, then $H(\omega_n, f_n) \subset C^\infty(\mathbb{R})$ (the space of infinitely smooth functions on $\mathbb{R}$). Moreover, the situation cannot be remedied using an auxiliary stabilizing feedback $Kz(t)$ by Theorem 4.8.

6. Conclusions. In this article we have studied regulation of periodic signals with a feedforward controller. We have constructed a scale of Sobolev-type spaces $H(\omega_n, f_n)$ for the reference signals. This construction played a key role as we showed that solvability of the regulation problem is equivalent to solvability of the regulator equations and a decomposability condition. In the case of SISO systems, assuming that the transfer function of the stabilized plant is invertible on the spectrum of the exosystem, we have completely characterized the reference signals which can be asymptotically tracked in the presence of disturbances.

Possible directions for future research include the design of a feedback controller

with robustness properties and the study of the internal model principle for infinite-dimensional exogenous systems. Additionally, the idea introduced in this article that the exogenous system is in a sense equivalent to the reference function space should be useful in generalizing our results for MIMO systems and reference functions which are not periodic.

Acknowledgments. The authors thank the reviewers for their many useful comments for improving the original manuscript.

REFERENCES

[1] W. Arendt, C. Batty, M. Hieber, and F. Neubrander, *Vector-Valued Laplace Transforms and Cauchy Problems*, Birkhäuser Verlag, Basel, 2001.

[2] G. Bachman, L. Narici, and E. Beckenstein, *Fourier and Wavelet Analysis*, Springer-Verlag, New York, 2000.

[3] C. Byrnes, I. Laukó, D. Gilliam, and V. Shubov, *Output regulation for linear distributed parameter systems*, IEEE Trans. Automat. Control, 45 (2000), pp. 2236–2252.

[4] C. Byrnes, D. Gilliam, V. Shubov, and J. Hood, *An example of output regulation for a distributed parameter system with infinite dimensional exosystem*, in Electronic Proceedings of MTNS 2002, http://www.nd.edu/~mtns/papers/22618_2.pdf.

[5] C. Byrnes, D. Gilliam, V. Shubov, and J. Hood, *Examples of output regulation for distributed parameter systems with infinite dimensional exosystem*, in Proceedings of the 40th IEEE Conference on Decision and Control, IEEE, Piscataway, NJ, 2001, pp. 547–548.

[6] R. Curtain and H. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1995.

[7] E. Davison, *Multivariable tuning regulators: The feedforward and robust control of a general servomechanism problem*, IEEE Trans. Automat. Control, 21 (1976), pp. 35–47.

[8] K.-J. Engel and R. Nagel, *One-Parameter Semigroups for Linear Evolution Equations*, Springer-Verlag, New York, 2000.

[9] B. A. Francis, *The linear multivariable regulator problem*, SIAM J. Control Optim., 15 (1977), pp. 486–505.

[10] B. Francis and W. Wonham, *The internal model principle of control theory*, Automatica J. IFAC, 12 (1976), pp. 457–465.

[11] T. Hämäläinen and S. Pohjolainen, *A finite-dimensional robust controller for systems in the CD-algebra*, IEEE Trans. Automat. Control, 45 (2000), pp. 421–431.

[12] G. Hillerström and K. Walgama, *Repetitive control theory and applications—a survey*, in Proceedings of the 13th IFAC World Congress, Vol. D, San Francisco, CA, 1996, pp. 1–6.

[13] E. Immonen, S. Pohjolainen, and T. Hämäläinen, *On the realization of periodic functions*, Systems Control Lett., 54 (2005), pp. 225–235.

[14] E. Immonen and S. Pohjolainen, *Output regulation of periodic signals for DPS: An infinite-dimensional signal generator*, IEEE Trans. Automat. Control, 50 (2005), pp. 1799–1804.

[15] R. Iorio and V. Iorio, *Fourier Analysis and Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 2001.

[16] H. Logemann and D. Owens, *Low-gain control of an unknown infinite-dimensional system: A frequency domain approach*, Dynam. Stab. Syst., 4 (1989), pp. 13–29.

[17] S. Pohjolainen, *A feedforward controller for distributed parameter systems*, Internat. J. Control, 34 (1981), pp. 173–184.

[18] S. Pohjolainen, *Robust multivariable PI-controller for infinite-dimensional systems*, IEEE Trans. Automat. Control, 27 (1982), pp. 17–31.

[19] S. Ruan and J. Wei, *On the zeros of transcendental functions with applications to stability of delay differential equations with two delays*, Dyn. Contin. Discrete Impuls. Syst. Ser. A Math. Anal., 10 (2003), pp. 863–874.

[20] W. Rudin, *Functional Analysis*, Tata McGraw–Hill, New Delhi, 1979.

[21] Q. Vu, *The operator equation $AX - XB = C$ with unbounded operators A and B and related abstract Cauchy problems*, Math. Z., 208 (1991), pp. 567–588.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 6, pp. 2269–2284

ON ITERATIVE SOLUTIONS OF GENERAL COUPLED MATRIX EQUATIONS*

FENG DING† AND TONGWEN CHEN‡

Abstract. In this paper we study coupled matrix equations, which are encountered in many systems and control applications. First, we extend the well-known Jacobi and Gauss–Seidel iterations and present a large family of iterative methods, which are then applied to develop iterative solutions to coupled Sylvester matrix equations. The basic idea is to regard the unknown matrices to be solved as parameters of a system to be identified and to obtain the iterative solutions by applying a hierarchical identification principle. Next, we generalize the Sylvester equations to general coupled matrix equations, and propose a gradient-based iterative algorithm for the solutions, using a block-matrix inner product—the star ($\star$) product; we prove that the iterative algorithm always converges to the (unique) solutions for any initial values. One advantage of the algorithms proposed is that they require less storage space in implementation than existing numerical methods. Finally, we test the algorithms and show their effectiveness using numerical examples.

Key words. matrix equations, gradient search principle, Jacobi iteration, Gauss–Seidel iteration, Hadamard product, star product, hierarchical identification principle

AMS subject classifications. 15A06, 93B30, 15A24

DOI. 10.1137/S0363012904441350

1. Introduction.

In stability analysis of control systems and robust control [6], we often need to solve (coupled) matrix equations of the following forms:

- Continuous-time (CT) Sylvester equation: $AX + XB = C$.
- Discrete-time (DT) Sylvester equation: $AXB^{\mathrm{T}} + X = C$.
- Generalized Sylvester matrix equation: $AXB^{\mathrm{T}} + CXD^{\mathrm{T}} = F$.
- Coupled Sylvester matrix equations: $AX + YB = C$ and $DX + YE = F$.

Here, X and Y are unknown matrices in $\mathbb{R}^{m\times n}$; A, B, C, etc. represent constant (coefficient) matrices of appropriate dimensions (to be specified later).

The conventional method of solution is to expand the matrix equations to form a set of equations of the form $\mathcal{A}x = b$ by means of the Kronecker product. However, the dimensions of the associated matrix $\mathcal{A}$ are very high when m and n are large. Computational difficulties arise because excessive computer memory is required for computation and inversion of large matrices of size $(mn) \times (mn)$ or even $(2mn) \times (2mn)$.

Other methods are based on matrix transformations into forms for which solutions may be readily computed; examples of such forms include the Jordan canonical form [17], the companion form [4, 3], and the Hessenberg–Schur form [2, 15]. In this area, Chu gave an algorithm for solving generalized/coupled Sylvester equations [7]; Syrmos, Misra, and Aripirala [29], Stykel [28], and Takaba, Morihhira, and Katayama [30] discussed numerical solutions to generalized coupled Lyapunov equations; many authors studied least squares solutions of matrix equations of the form $AXB^* +$

*Received by the editors February 20, 2004; accepted for publication (in revised form) August 7, 2005; published electronically February 3, 2006. This research was supported by the Natural Sciences and Engineering Research Council of Canada and the National Natural Science Foundation of China.
http://www.siam.org/journals/sicon/44-6/44135.html

†Control Science and Engineering Research Center, Southern Yangtze University, Wuxi 214122, China (fding@sytu.edu.cn, fding@ece.ualberta.ca). Corresponding authors are F. Ding and T. Chen.

‡Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada T6G 2V4 (tchen@ece.ualberta.ca).

$CYD^* = E$ based on the singular value decomposition of coefficient matrices [14, 1, 32, 26]; Jonsson and Kågström proposed recursive block algorithms for solving CT Sylvester equations, CT Lyapunov equations, and coupled Sylvester matrix equations [19, 20]; and finally, Borno presented a parallel algorithm for solving the coupled Lyapunov equations [5].

However, the above-mentioned algorithms all require computing some additional matrix transformation/decomposition; moreover, they are not suitable for more general coupled matrix equations of the form

$$\begin{cases} A_{11}X_1B_{11} + A_{12}X_2B_{12} + \cdots + A_{1p}X_pB_{1p} = C_1, \\ A_{21}X_1B_{21} + A_{22}X_2B_{22} + \cdots + A_{2p}X_pB_{2p} = C_2, \\ \quad \vdots \\ A_{p1}X_1B_{p1} + A_{p2}X_2B_{p2} + \cdots + A_{pp}X_pB_{pp} = C_p. \end{cases} \tag{1}$$

Here, $X_j \in \mathbb{R}^{m\times n}$ are the unknown matrices to be solved. For such coupled matrix equations, the above-mentioned methods require dealing with matrices whose dimensions grow quickly as m, n, and p increase—a major disadvantage. We would like to comment that the coupled matrix equations in (1) are quite general and include many matrix equations, e.g., the ones mentioned above [29, 28] as special cases; in particular, they also encompass generalized (coupled) Lyapunov and Sylvester equations which occur in the study of linear jump parameter systems [5].

Iterative algorithms are popular in the areas of matrix algebra and system identification [16, 22, 24, 25]. For example, Starke and Niethammer reported an iterative method for solutions of CT Sylvester equations by using the SOR (successive overrelaxation) technique [27], and Mukaidani, Xu, and Mizukami discussed an iterative algorithm for generalized algebraic Lyapunov equations [24]. In our work, we focus on numerical solutions for coupled Sylvester matrix equations and general coupled matrix equations in (1), and we present the gradient-based iterative algorithms by using the gradient search principle and the hierarchical identification principle. We mention that least squares iterative algorithms were given in [9] for coupled matrix equations and gradient-based iterative algorithms in [10] for noncoupled matrix equations.

For matrix equations, exact solutions are important, but it is often not necessary to compute exact solutions for many applications such as stability analysis in control systems, and approximate solutions are sufficient. Also, if parameters in system matrices contain uncertainty, it is not possible to obtain exact solutions for robust stability analysis [9, 10, 13, 24, 23, 21].

The paper is organized as follows. In section 2, we extend the well-known Jacobi and Gauss–Seidel iterations and present a large family of iterative methods for linear equations. In sections 3 and 4, we derive gradient-based iterative algorithms for, respectively, coupled Sylvester matrix equations and general coupled matrix equations, and we study the convergence properties of the algorithms. In section 5 we present two examples to illustrate the effectiveness of the algorithms proposed in the paper. Finally, we offer some concluding remarks in section 6.

2. A large family of iterative methods. Consider the equation

$$Ax = b, \quad A \in \mathbb{R}^{n\times n}, \quad b \in \mathbb{R}^n. \tag{2}$$

Here, $A = [a_{ij}]$ is a full-rank matrix with nonzero diagonal elements and $x \in \mathbb{R}^n$ is an unknown vector to be computed. Letting D denote the diagonal part, L and U the

strictly lower and upper triangular parts of A, we have

$$D = \operatorname{diag}[a_{11},\ a_{22},\ \ldots,\ a_{nn}] \in \mathbb{R}^{n\times n},$$

$$L = \begin{bmatrix} 0 & 0 & \cdots & \cdots & 0 \\ a_{21} & 0 & \ddots & & \vdots \\ a_{31} & a_{32} & 0 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ a_{n1} & a_{n2} & \cdots & a_{n,n-1} & 0 \end{bmatrix} \in \mathbb{R}^{n\times n},$$

$$U = \begin{bmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & 0 & a_{23} & & a_{2n} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & a_{n-1,n} \\ 0 & \cdots & \cdots & 0 & 0 \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

The matrices just defined satisfy $L + D + U = A$. In this case both the Jacobi and Gauss–Seidel iterations [16] can be applied to give an iterative solution $x(k)$ of x in the form of

$$Mx(k) = Nx(k-1) + b, \quad k = 1, 2, 3, \ldots ;$$

for the Jacobi method, $M = D$ and $N = -(L + U)$; for the Gauss–Seidel method, $M = L + D$ and $N = -U$.

The main drawback of the Jacobi and Gauss–Seidel iterations is that they do not guarantee that $x(k)$ converges to the exact solution $x = A^{-1}b$. This inspires us to study new iterative methods.

Let $G_k \in \mathbb{R}^{n\times n}$ be a matrix to be determined and let $\mu > 0$ be the step size or the convergence factor. We present a large family of iterative methods as follows:

$$x(k) = x(k-1) + \mu G_k[b - Ax(k-1)], \quad k = 1, 2, 3, \ldots . \tag{3}$$

This family includes the Jacobi and Gauss–Seidel iterations as special cases. For example, when $G_k = D^{-1}$ and $\mu = 1$, we get the Jacobi method; when $G_k = (L+D)^{-1}$ and $\mu = 1$, we obtain the Gauss–Seidel method.

THEOREM 1. *For the iterative algorithm in* (3)*, assume system* (2) *has a unique solution. Then the iterative solution* $x(k)$ *in* (3) *converges to the exact solution* x *(i.e.,* $\lim_{k\to\infty} x(k) = x$*) for any initial values* $x(0)$ *if there exists* $\varepsilon > 0$ *independent of* k *such that*

$$\mu(G_kA)^{\mathrm{T}}(G_kA) + \varepsilon I \le (G_kA)^{\mathrm{T}} + (G_kA) \quad \textit{for all } k. \tag{4}$$

Here I *represents an identity matrix of appropriate dimensions. In fact, if* $(G_kA)^{\mathrm{T}} + (G_kA)$ *is positive definite, a conservative choice of the convergence factor can be given by*

$$0 < \mu < \frac{\lambda_{\min}[(G_kA)^{\mathrm{T}} + (G_kA)]}{\lambda_{\max}[(G_kA)^{\mathrm{T}}(G_kA)]} \quad \textit{for all } k,$$

where $\lambda_{\max}(\lambda_{\min})$ *denotes the maximum (resp., minimum) eigenvalue.*

Note that for time-invariant systems of the form $x(k) = Hx(k-1),\ H \in \mathbb{R}^{n\times n}$; the fact that all eigenvalues of H are inside the unit circle guarantees the convergence of $x(k)$ to zero as $k \to \infty$. But for time-varying systems of the form $x(k) = H_k x(k-1),\ H_k \in \mathbb{R}^{n\times n}$, such a conclusion is no longer true because the condition that all the eigenvalues of H_k are inside the unit circle is neither sufficient nor necessary for stability—see the appendix for some stability examples. In the following, we give a proof of Theorem 1 based on the Lyapunov stability theorem.

Proof. Define the error vector

$$\tilde{x}(k) = x(k) - x.$$

Substituting (3) into the above equation and using (2) yield

$$\begin{aligned}\tilde{x}(k) &= \tilde{x}(k-1) + \mu G_k\left[Ax - Ax(k-1)\right]\\ &= \tilde{x}(k-1) - \mu G_k A\tilde{x}(k-1)\\ &= [I - \mu G_k A]\tilde{x}(k-1).\end{aligned}$$

Define a nonnegative definite Lyapunov function

$$S(k) = \tilde{x}^{\mathrm{T}}(k)\tilde{x}(k).$$

Hence

$$\begin{aligned}S(k) &= \tilde{x}^{\mathrm{T}}(k-1)[I - \mu(G_k A)^{\mathrm{T}}][I - \mu G_k A]\tilde{x}(k-1)\\ &= \tilde{x}^{\mathrm{T}}(k-1)\tilde{x}(k-1) - \mu\tilde{x}^{\mathrm{T}}(k-1)[G_k A + (G_k A)^{\mathrm{T}} - \mu(G_k A)^{\mathrm{T}}(G_k A)]\tilde{x}(k-1)\\ &= S(k-1) - \mu\tilde{x}^{\mathrm{T}}(k-1)[G_k A + (G_k A)^{\mathrm{T}} - \mu(G_k A)^{\mathrm{T}}(G_k A)]\tilde{x}(k-1).\end{aligned}$$

Using (4), it is not difficult to get

$$\begin{aligned}\Delta S(k) &:= S(k) - S(k-1)\\ &= -\mu\tilde{x}^{\mathrm{T}}(k-1)[G_k A + (G_k A)^{\mathrm{T}} - \mu(G_k A)^{\mathrm{T}}(G_k A)]\tilde{x}(k-1)\\ &\le -\mu\varepsilon\tilde{x}^{\mathrm{T}}(k-1)\tilde{x}(k-1) \le 0.\end{aligned}$$

So we have $S(k) \to 0$ as $k \to \infty$. This proves Theorem 1. □

It is well known that if all eigenvalues of $(I - D^{-1}A)$ are inside the unit circle, the Jacobi iterative solution will converge to the exact solution, and if all eigenvalues of $[I - (L+D)^{-1}A]$ are inside the unit circle, the Gauss–Seidel solution will converge to the exact solution. Thus introducing a convergence factor μ in the Jacobi and the Gauss–Seidel iterations can relax their convergence conditions because for appropriate values of μ, $(I - \mu D^{-1}A)$ and $[I - \mu(L+D)^{-1}A]$ may have all eigenvalues inside the unit circle. We can also draw the following corollaries from Theorem 1.

COROLLARY 1. *If we take $G_k = D^{-1}$, then the Jacobi iteration with the convergence factor μ,*

$$x(k) = x(k-1) + \mu D^{-1}[b - Ax(k-1)],$$

yields $\lim_{k\to\infty} x(k) = x$ if $A^{\mathrm{T}}D^{-1} + D^{-1}A$ is a positive-definite matrix, and

$$0 < \mu < \frac{\lambda_{\min}[A^{\mathrm{T}}D^{-1} + D^{-1}A]}{\lambda_{\max}[A^{\mathrm{T}}D^{-2}A]}.$$

COROLLARY 2. *If we take* $G_k = (L+D)^{-1}$, *then the Gauss–Seidel iteration with the convergence factor* μ,

$$x(k) = x(k-1) + \mu(L+D)^{-1}[b - Ax(k-1)],$$

also yields $\lim_{k\to\infty} x(k) = x$ *if* $A^{\mathrm{T}}(L+D)^{-\mathrm{T}} + (L+D)^{-1}A > 0$, *and*

$$0 < \mu < \frac{\lambda_{\min}[A^{\mathrm{T}}(L+D)^{-\mathrm{T}} + (L+D)^{-1}A]}{\lambda_{\max}[A^{\mathrm{T}}(L+D)^{-\mathrm{T}}(L+D)^{-1}A]}.$$

COROLLARY 3. *If we take* $G_k = A^{\mathrm{T}}$, *then the gradient iterative algorithm* [9],

$$\begin{cases} x(k) = x(k-1) + \mu A^{\mathrm{T}}[b - Ax(k-1)], \quad k = 1,\ 2,\ 3,\ \ldots, \\ 0 < \mu < \frac{2}{\lambda_{\max}[A^{\mathrm{T}}A]} \quad or \quad 0 < \mu < \frac{2}{\|A\|^2}, \end{cases} \tag{5}$$

yields $\lim_{k\to\infty} x(k) = x$. *Here,* $\|X\|^2 = \mathrm{tr}[XX^{\mathrm{T}}]$.

COROLLARY 4. *If* A *is a nonsquare* $m \times n$ *full column-rank matrix and we take* $G_k = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}$, *then the following least squares iterative algorithm leads to* $\lim_{k\to\infty} x(k) = x$ [9]:

$$x(k) = x(k-1) + \mu(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}[b - Ax(k-1)], \quad 0 < \mu < 2.$$

Later we will use the iterative algorithm in (5) to develop iteration techniques for general coupled matrix equations.

3. Coupled Sylvester equations. In this section, we apply a hierarchical identification principle to solve the coupled Sylvester matrix equation

$$\begin{cases} AX + YB = C, \\ DX + YE = F. \end{cases} \tag{6}$$

Here $A,\ D \in \mathbb{R}^{m\times m}$, $B,\ E \in \mathbb{R}^{n\times n}$, and $C,\ F \in \mathbb{R}^{m\times n}$ are given constant matrices; $X,\ Y \in \mathbb{R}^{m\times n}$ are the unknown matrices to be solved.

First, let us introduce some notation. I_n is the $n \times n$ identity matrix. For two matrices M and N, $M \otimes N$ is their Kronecker product. For two $m \times n$ matrices X and Y with

$$X = [\ x_1, \quad x_2, \quad \ldots, \quad x_n\ ] \in \mathbb{R}^{m\times n},$$

$\mathrm{col}[X]$ is an mn-dimensional vector formed by columns of X,

$$\mathrm{col}[X] = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix},$$

and

$$\mathrm{col}[X, Y] = \begin{bmatrix} \mathrm{col}[X] \\ \mathrm{col}[Y] \end{bmatrix} \in \mathbb{R}^{2mn}.$$

The following result is well known.

LEMMA 1. *System* (6) *has a unique solution*

$$\text{col}[X, Y] = S_2^{-1}\text{col}[C, F] \tag{7}$$

if and only if the matrix

$$S_2 := \begin{bmatrix} I_n \otimes A & B^{\mathrm{T}} \otimes I_m \\ I_n \otimes D & E^{\mathrm{T}} \otimes I_m \end{bmatrix} \in \mathbb{R}^{(2mn)\times(2mn)}$$

is nonsingular and the corresponding homogeneous matrix equation ($AX + YB = \mathbf{0},\ DX + YE = \mathbf{0}$) *has a unique solution:* $X = Y = \mathbf{0}$.

According to the hierarchical identification principle [11, 12], system (6) is decomposed into two subsystems and then, based on the gradient search principle, the parameters of each subsystem are identified. In this way we derive the iterative algorithm. The details are as follows.

Define two matrices

$$b_1 := \begin{bmatrix} C - YB \\ F - YE \end{bmatrix}, \tag{8}$$

$$b_2 := [C - AX,\ F - DX]. \tag{9}$$

Then from (6), we obtain two fictitious subsystems

$$S_1: \quad \begin{bmatrix} A \\ D \end{bmatrix} X = b_1,$$

$$S_2: \quad Y[B,\ E] = b_2.$$

Let $X(k)$ and $Y(k)$ be the estimates or iterative solutions of X and Y, associated with subsystems S_1 and S_2. Then using the gradient search principle or applying Corollary 3 to S_1 and S_2 leads to the following recursive equations:

$$X(k) = X(k-1) + \mu \begin{bmatrix} A \\ D \end{bmatrix}^{\mathrm{T}} \left\{ b_1 - \begin{bmatrix} A \\ D \end{bmatrix} X(k-1) \right\}, \tag{10}$$

$$Y(k) = Y(k-1) + \mu\{b_2 - Y(k-1)[B,\ E]\}[B,\ E]^{\mathrm{T}}. \tag{11}$$

Here, $\mu > 0$ is the iterative step size or convergence factor to be given later. Substituting (8) into (10) and (9) into (11) gives

$$\begin{aligned} X(k) &= X(k-1) + \mu \begin{bmatrix} A \\ D \end{bmatrix}^{\mathrm{T}} \left\{ \begin{bmatrix} C - YB \\ F - YE \end{bmatrix} - \begin{bmatrix} A \\ D \end{bmatrix} X(k-1) \right\} \\ &= X(k-1) + \mu \begin{bmatrix} A \\ D \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} C - YB - AX(k-1) \\ F - YE - DX(k-1) \end{bmatrix}, \end{aligned} \tag{12}$$

$$\begin{aligned} Y(k) &= Y(k-1) + \mu\{[C - AX,\ F - DX] - Y(k-1)[B,\ E]\}[B,\ E]^{\mathrm{T}} \\ &= Y(k-1) + \mu[C - AX - Y(k-1)B,\ F - DX - Y(k-1)E][B,\ E]^{\mathrm{T}}. \end{aligned} \tag{13}$$

Because the expressions on the right-hand sides of (12) and (13) contain the unknown parameter matrices Y and X, it is impossible to realize the algorithm in (12) and (13). According to the hierarchical identification principle, the unknown variables Y in (12) and X in (13) are replaced by their estimates $Y(k-1)$ and $X(k-1)$ at

time $k-1$. Hence, we obtain the iterative solutions $X(k)$ and $Y(k)$ for the coupled Sylvester equation in (6):

$$\begin{aligned} X(k) &= X(k-1) \\ &+\mu \begin{bmatrix} A \\ D \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} C - AX(k-1) - Y(k-1)B \\ F - DX(k-1) - Y(k-1)E \end{bmatrix}, \end{aligned} \tag{14}$$

$$\begin{aligned} Y(k) &= Y(k-1) \\ &+\mu[C - AX(k-1) - Y(k-1)B,\ F - DX(k-1) - Y(k-1)E][B,\ E]^{\mathrm{T}}. \end{aligned} \tag{15}$$

The convergence factor may be taken to satisfy

$$0 < \mu < \frac{2}{\lambda_{\max}[A^{\mathrm{T}}A] + \lambda_{\max}[D^{\mathrm{T}}D] + \lambda_{\max}[BB^{\mathrm{T}}] + \lambda_{\max}[EE^{\mathrm{T}}]} =: \mu_0 \tag{16}$$

or

$$0 < \mu < \frac{2}{\|A\|^2 + \|B\|^2 + \|D\|^2 + \|E\|^2}.$$

To initialize the algorithm, we take $X(0) = Y(0) = \mathbf{0}$ or some small real matrix, e.g., $X(0) = Y(0) = 10^{-6}\mathbf{1}_{m\times n}$ with $\mathbf{1}_{m\times n}$ being an $m \times n$ matrix whose elements are all 1.

THEOREM 2. *If the coupled Sylvester equation in* (6) *has unique solutions* X *and* Y*, then for any initial values, the iterative solutions* $X(k)$ *and* $Y(k)$ *given by the algorithm in* (14)–(15) *converge to the solutions* X *and* Y*:*

$$\lim_{k\to\infty} X(k) = X, \qquad \lim_{k\to\infty} Y(k) = Y.$$

The proof of Theorem 2 is omitted here but can be given later with the proof of Theorem 3.

In order to enhance the convergence properties, normally we should choose a large convergence factor μ which leads to a fast convergence rate of $X(k)$ to X and $Y(k)$ to Y; but too large μ may violate the condition of this theorem (also see (23)). Usually, there exists some best μ so that a fast convergence rate can be achieved—see the examples to be studied later.

4. General coupled matrix equations. In this section, we will generalize the Sylvester system of equations and introduce the block-matrix inner product to develop iterative solutions for a more general coupled matrix equations of the form

$$\begin{cases} A_{11}X_1B_{11} + A_{12}X_2B_{12} + \cdots + A_{1p}X_pB_{1p} = C_1, \\ A_{21}X_1B_{21} + A_{22}X_2B_{22} + \cdots + A_{2p}X_pB_{2p} = C_2, \\ \quad \vdots \\ A_{p1}X_1B_{p1} + A_{p2}X_2B_{p2} + \cdots + A_{pp}X_pB_{pp} = C_p. \end{cases} \tag{17}$$

Here, $A_{ij} \in \mathbb{R}^{m\times m}$, $B_{ij} \in \mathbb{R}^{n\times n}$, and $C_i \in \mathbb{R}^{m\times n}$ are given constant matrices; $X_i \in \mathbb{R}^{m\times n}$ are the unknown matrices to be determined.

Here, we succinctly express the iterative solutions by using the block-matrix inner product [9]—the star product, denoted by $\star$—which differs from the Hadamard (inner)

product [18, 8, 31] or the general matrix multiplication. Let

$$X=\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}\in\mathbb{R}^{(mp)\times n},\quad Y=\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{bmatrix}\in\mathbb{R}^{(np)\times m},\quad X_i,\, Y_i^{\mathrm{T}}\in\mathbb{R}^{m\times n},$$

$$S_A=\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1p} \\ A_{21} & A_{22} & \cdots & A_{2p} \\ \vdots & \vdots & & \vdots \\ A_{p1} & A_{p2} & \cdots & A_{pp} \end{bmatrix},\quad S_B=\begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1p} \\ B_{21} & B_{22} & \cdots & B_{2p} \\ \vdots & \vdots & & \vdots \\ B_{p1} & B_{p2} & \cdots & B_{pp} \end{bmatrix},$$

$$S_{B^{\mathrm{T}}}=\begin{bmatrix} B_{11}^{\mathrm{T}} & B_{12}^{\mathrm{T}} & \cdots & B_{1p}^{\mathrm{T}} \\ B_{21}^{\mathrm{T}} & B_{22}^{\mathrm{T}} & \cdots & B_{2p}^{\mathrm{T}} \\ \vdots & \vdots & & \vdots \\ B_{p1}^{\mathrm{T}} & B_{p2}^{\mathrm{T}} & \cdots & B_{pp}^{\mathrm{T}} \end{bmatrix},$$

$$S_p=\begin{bmatrix} B_{11}^{\mathrm{T}}\otimes A_{11} & B_{12}^{\mathrm{T}}\otimes A_{12} & \cdots & B_{1p}^{\mathrm{T}}\otimes A_{1p} \\ B_{21}^{\mathrm{T}}\otimes A_{21} & B_{22}^{\mathrm{T}}\otimes A_{22} & \cdots & B_{2p}^{\mathrm{T}}\otimes A_{2p} \\ \vdots & \vdots & & \vdots \\ B_{p1}^{\mathrm{T}}\otimes A_{p1} & B_{p2}^{\mathrm{T}}\otimes A_{p2} & \cdots & B_{pp}^{\mathrm{T}}\otimes A_{pp} \end{bmatrix}.$$

Then the block-matrix star product is defined as

$$X\star Y=\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}\star\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{bmatrix}=\begin{bmatrix} X_1Y_1 \\ X_2Y_2 \\ \vdots \\ X_pY_p \end{bmatrix},$$

$$X\star S_B=\begin{bmatrix} X_1B_{11} & X_1B_{12} & \cdots & X_1B_{1p} \\ X_2B_{21} & X_2B_{22} & \cdots & X_2B_{2p} \\ \vdots & \vdots & & \vdots \\ X_pB_{p1} & X_pB_{p2} & \cdots & X_pB_{pp} \end{bmatrix}.$$

The definitions of $S_A\star X$ and $S_A\star S_B$ can be found in [9]. In the above definitions, we assume that the dimensions of matrices are compatible. The block matrix star Kronecker product, denoted by $\circledast$, is defined by

$$S_{B^{\mathrm{T}}}\circledast S_A=S_p\,.$$

For the Hadamard product (denoted by $\circ$), we have $X\circ Y=Y\circ X$, but $X\circ S_A$ is not defined. For the star product, taking into account the dimension compatibility, we have $AB\star C=A(B\star C)\neq(AB)\star C$ (the multiplier and multiplicand matrices are not necessarily of the same size); in general, $A\star B\neq B\star A$, $A\star B\star C=(A\star B)\star C\neq A\star(B\star C)$.

Let $I_{mp\times m}=[I_m,\ I_m,\ \ldots,\ I_m]^{\mathrm{T}}\in\mathbb{R}^{(mp)\times m}$. Then the star product has the following properties:

$$I_{mp\times m}^{\mathrm{T}}X\star Y=[X_1,\ X_2,\ \ldots,\ X_p]Y=\sum_{i=1}^{p}X_iY_i,$$

$$\operatorname{tr}\left\{ X_i^{\mathrm{T}} \begin{bmatrix} A_{1i} \\ A_{2i} \\ \vdots \\ A_{pi} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \tilde{C}_1 \\ \tilde{C}_2 \\ \vdots \\ \tilde{C}_p \end{bmatrix} \star \begin{bmatrix} B_{1i}^{\mathrm{T}} \\ B_{2i}^{\mathrm{T}} \\ \vdots \\ B_{pi}^{\mathrm{T}} \end{bmatrix} \right\} = \operatorname{tr}\left\{ \begin{bmatrix} A_{1i}X_iB_{1i} \\ A_{2i}X_iB_{2i} \\ \vdots \\ A_{pi}X_iB_{pi} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \tilde{C}_1 \\ \tilde{C}_2 \\ \vdots \\ \tilde{C}_p \end{bmatrix} \right\},$$

$$\left\| \begin{bmatrix} A_{1i} \\ A_{2i} \\ \vdots \\ A_{pi} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \tilde{C}_1 \\ \tilde{C}_2 \\ \vdots \\ \tilde{C}_p \end{bmatrix} \star \begin{bmatrix} B_{1i}^{\mathrm{T}} \\ B_{2i}^{\mathrm{T}} \\ \vdots \\ B_{pi}^{\mathrm{T}} \end{bmatrix} \right\|^2 \le \sum_{j=1}^{p} \|A_{ji}B_{ji}\|^2 \left\| \begin{bmatrix} \tilde{C}_1 \\ \tilde{C}_2 \\ \vdots \\ \tilde{C}_p \end{bmatrix} \right\|^2.$$

LEMMA 2. *Equation* (17) *has unique solutions* X_i *if and only if the matrix* S_p *is nonsingular; in this case, the unique solutions are given by*

$$\operatorname{col}[X_1, X_2, \ldots, X_p] = S_p^{-1}\operatorname{col}[C_1, C_2, \ldots, C_p].$$

and if $C_i = \mathbf{0}$ $(i = 1,\ 2,\ \cdots,\ p)$*, the corresponding homogeneous equation in* (17) *has unique solutions* $X_i = \mathbf{0}$ $(i = 1,\ 2,\ \ldots,\ p)$.

The iterative solution for the general coupled matrix equation in (17) is obtained by generalizing that for the coupled Sylvester equation in (6); to do this, we first consider the coupled Sylvester equation in (6) in the more general form

$$\begin{cases} AXI_B + I_AYB = C, \\ DXI_E + I_DYE = F, \end{cases}$$

whose iterative solution in (14)–(15) can be expressed as

$$X(k) = X(k-1)$$

$$+\mu \begin{bmatrix} A \\ D \end{bmatrix}^{\mathrm{T}} \left\{ \begin{bmatrix} C - AX(k-1)I_B - I_AY(k-1)B \\ F - DX(k-1)I_E - I_DY(k-1)E \end{bmatrix} \star [I_B,\ I_E]^{\mathrm{T}} \right\}, \tag{18}$$

$$Y(k) = Y(k-1)$$

$$+\mu \begin{bmatrix} I_A \\ I_D \end{bmatrix}^{\mathrm{T}} \left\{ \begin{bmatrix} C - AX(k-1)I_E - I_DY(k-1)B \\ F - DX(k-1)I_E - I_DY(k-1)E \end{bmatrix} \star [B,\ E]^{\mathrm{T}} \right\}. \tag{19}$$

If I_A, I_B, I_D, and I_E are identity matrices of appropriate dimensions, then the algorithm in (18) and (19) is equivalent to the algorithm given in (14) and (15).

Let $X_i(k)$ be the estimates or iterative solutions of X_i. For (17), we propose the following iterative algorithm to compute the solutions X_i $(i = 1,\ 2,\ \ldots,\ p)$:

$$X_i(k) = X_i(k-1)$$

$$+\mu \begin{bmatrix} A_{1i} \\ A_{2i} \\ \vdots \\ A_{pi} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} C_1 - \sum_{j=1}^{p} A_{1j}X_j(k-1)B_{1j} \\ C_2 - \sum_{j=1}^{p} A_{2j}X_j(k-1)B_{2j} \\ \vdots \\ C_p - \sum_{j=1}^{p} A_{pj}X_j(k-1)B_{pj} \end{bmatrix} \star [B_{1i},\ B_{2i},\ \ldots,\ B_{pi}]^{\mathrm{T}}, \tag{20}$$

$$0 < \mu < 2\left(\sum_{i=1}^{p}\sum_{j=1}^{p} \|A_{ij}B_{ij}\|^2 \right)^{-1} =: \mu_0. \tag{21}$$

THEOREM 3. *If the matrix equation in* (17) *has unique solutions* X_i, $i = 1,\ 2,\ \ldots,\ p$, *then the iterative solutions* $X_i(k)$ *given by the algorithm in* (20) *and* (21) *converge to the solutions* X_i *for any initial value:*

$$\lim_{k\to\infty} X_i(k) = X_i, \quad i = 1,\ 2,\ \ldots,\ p.$$

Proof. Define the estimation error matrices

$$\tilde{X}_i(k) = X_i(k) - X_i.$$

Let

$$\begin{bmatrix} \tilde{C}_1(k) \\ \tilde{C}_2(k) \\ \vdots \\ \tilde{C}_p(k) \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{p} A_{1j}\tilde{X}_j(k-1)B_{1j} \\ \sum_{j=1}^{p} A_{2j}\tilde{X}_j(k-1)B_{2j} \\ \vdots \\ \sum_{j=1}^{p} A_{pj}\tilde{X}_j(k-1)B_{pj} \end{bmatrix}.$$

By using (17) and (20), it is not difficult to get

$$\tilde{X}_i(k) = \tilde{X}_i(k-1) - \mu \begin{bmatrix} A_{1i} \\ A_{2i} \\ \vdots \\ A_{pi} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \tilde{C}_1(k) \\ \tilde{C}_2(k) \\ \vdots \\ \tilde{C}_p(k) \end{bmatrix} \star [B_{1i},\ B_{2i},\ \ldots,\ B_{pi}]^{\mathrm{T}}.$$

Taking the norm on the above equation and using the star product properties, we have

$$\begin{aligned} \|\tilde{X}_i(k)\|^2 \le \|\tilde{X}_i(k-1)\|^2 &- 2\mu \operatorname{tr}\left\{ \begin{bmatrix} A_{1i}\tilde{X}_i(k-1)B_{1i} \\ A_{2i}\tilde{X}_i(k-1)B_{2i} \\ \vdots \\ A_{pi}\tilde{X}_i(k-1)B_{pi} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \tilde{C}_1(k) \\ \tilde{C}_2(k) \\ \vdots \\ \tilde{C}_p(k) \end{bmatrix} \right\} \\ &+ \mu^2 \sum_{j=1}^{p} \|A_{ji}B_{ji}\|^2 \left\| \begin{bmatrix} \tilde{C}_1(k) \\ \tilde{C}_2(k) \\ \vdots \\ \tilde{C}_p(k) \end{bmatrix} \right\|^2. \end{aligned} \tag{22}$$

Define a nonnegative definite function:

$$V(k) = \sum_{i=1}^{p} \|\tilde{X}_i(k)\|^2.$$

By using (22), it follows that

$$\begin{aligned} V(k) &\le V(k-1) - 2\mu \left\| \begin{bmatrix} \tilde{C}_1(k) \\ \tilde{C}_2(k) \\ \vdots \\ \tilde{C}_p(k) \end{bmatrix} \right\|^2 + \mu^2 \sum_{i=1}^{p}\sum_{j=1}^{p} \|A_{ji}B_{ji}\|^2 \left\| \begin{bmatrix} \tilde{C}_1(k) \\ \tilde{C}_2(k) \\ \vdots \\ \tilde{C}_p(k) \end{bmatrix} \right\|^2 \\ &= V(k-1) - \mu \left(2 - \mu \sum_{i=1}^{p}\sum_{j=1}^{p} \|A_{ij}B_{ij}\|^2 \right) \sum_{i=1}^{p} \|\tilde{C}_i(k)\|^2 \\ &\le V(0) - \mu \left(2 - \mu \sum_{i=1}^{p}\sum_{j=1}^{p} \|A_{ij}B_{ij}\|^2 \right) \sum_{l=1}^{\infty}\sum_{i=1}^{p} \|\tilde{C}_i(l)\|^2. \end{aligned} \tag{23}$$

If the convergence factor μ is chosen to satisfy

$$0 < \mu < 2 \left(\sum_{i=1}^{p}\sum_{j=1}^{p} \|A_{ij}B_{ij}\|^2 \right)^{-1},$$

then

$$\sum_{k=1}^{\infty}\sum_{i=1}^{p} \|\tilde{C}_i(k)\|^2 < \infty.$$

It follows that as $k \to \infty$,

$$\sum_{i=1}^{p} \|\tilde{C}_i(k)\|^2 = \sum_{j=1}^{p} \|A_{ij}\tilde{X}_j(k-1)B_{ij}\|^2 = 0$$

or

$$\sum_{j=1}^{p} A_{ij}\tilde{X}_j(k-1)B_{ij} \to \mathbf{0}, \quad i = 1,\ 2,\ \ldots,\ p.$$

According to Lemma 2, this complete the proof of Theorem 3. □

Let

$$X(k) = \begin{bmatrix} X_1(k) \\ X_2(k) \\ \vdots \\ X_p(k) \end{bmatrix} \in \mathbb{R}^{(mp)\times n}, \quad C = \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_p \end{bmatrix} \in \mathbb{R}^{(mp)\times n}.$$

Then (17) can be simply expressed as

$$S_A \star X \star S_B I_{np\times n} = C.$$

By using the star product properties, (20) can be written in the following more compact form:

$$X(k) = X(k-1) + \mu S_A^{\mathrm{T}} \begin{bmatrix} C_1 - \sum_{j=1}^{p} A_{1j} X_j(k-1) B_{1j} \\ C_2 - \sum_{j=1}^{p} A_{2j} X_j(k-1) B_{2j} \\ \vdots \\ C_p - \sum_{j=1}^{p} A_{pj} X_j(k-1) B_{pj} \end{bmatrix} \star S_{B^{\mathrm{T}}}$$

$$= X(k-1) + \mu S_A^{\mathrm{T}} [C - S_A \star X(k-1) \star S_B I_{np \times n}] \star S_{B^{\mathrm{T}}}. \tag{24}$$

The convergence factor in the algorithm in (20) or (24) may also be taken as

$$0 < \mu < 2 \left\{ \sum_{i=1}^{p} \sum_{j=1}^{p} \lambda_{\max}[A_{ij} A_{ij}^{\mathrm{T}}] \lambda_{\max}[B_{ij} B_{ij}^{\mathrm{T}}] \right\}^{-1}.$$

5. Examples. In this section, we present two examples to illustrate the performance of the proposed algorithms.

Example 1. Suppose that the coupled Sylvester matrix equations are $AX + YB = C$ and $DX + YE = F$ with

$$A = \begin{bmatrix} 2.00 & 1.00 \\ -1.00 & 2.00 \end{bmatrix}, \quad B = \begin{bmatrix} 1.00 & -0.20 \\ 0.20 & 1.00 \end{bmatrix}, \quad C = \begin{bmatrix} 13.20 & 10.60 \\ 0.60 & 8.40 \end{bmatrix},$$

$$D = \begin{bmatrix} -2.00 & -0.50 \\ 0.50 & 2.00 \end{bmatrix}, \quad E = \begin{bmatrix} -1.00 & -3.00 \\ 2.00 & -4.00 \end{bmatrix}, \quad F = \begin{bmatrix} -9.50 & -18.00 \\ 16.00 & 3.50 \end{bmatrix}.$$

Then the unique solutions of X and Y from (7) are

$$X = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} = \begin{bmatrix} 4.00 & 3.00 \\ 3.00 & 4.00 \end{bmatrix},$$

$$Y = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} = \begin{bmatrix} 2.00 & 1.00 \\ -2.00 & 3.00 \end{bmatrix}.$$

Taking $X(0) = Y(0) = 10^{-6}\mathbf{1}_{2\times 2}$, we apply the algorithm in (14) and (15) to compute $X(k)$ and $Y(k)$. The iterative solutions $X(k)$ and $Y(k)$ are shown in Table 1, where

$$\delta = \sqrt{\frac{\|X(k) - X\|^2 + \|Y(k) - Y\|^2}{\|X\|^2 + \|Y\|^2}}$$

is the relative iteration error. The errors δ with different convergence factors μ are shown in Figure 1. From Table 1 and Figure 1, it is clear that the errors are becoming smaller and smaller and go to zero as k increases. This indicates that the proposed algorithm is effective. The effect of changing the convergence factor μ is also illustrated in Figure 1. We see that for $\mu = 1/101.16, 1/38.47, 1/30, 1/20$, the

TABLE 1
The iterative solutions ($\mu = 1/20.00$).

k	x_{11}	x_{12}	x_{21}	x_{22}	y_{11}	y_{12}	y_{21}	y_{22}	δ (%)
2	3.49715	1.07818	3.22924	2.30096	2.17503	0.26353	−1.60310	2.30842	34.50146582
4	3.74044	1.70293	3.36547	3.08953	2.26460	0.70926	−1.88191	2.79275	20.73545753
6	3.81220	2.11537	3.25459	3.46987	2.21517	0.90624	−1.98444	2.91727	13.42666154
8	3.86856	2.39509	3.15907	3.67547	2.16522	0.98697	−2.02647	2.95577	8.94369709
10	3.91045	2.58580	3.09671	3.79372	2.12397	1.01811	−2.04035	2.97133	6.05894632
12	3.93954	2.71596	3.05859	3.86493	2.09134	1.02761	−2.04137	2.97985	4.15211551
14	3.95922	2.80489	3.03571	3.90943	2.06629	1.02779	−2.03709	2.98556	2.86862687
16	3.97245	2.86574	3.02197	3.93810	2.04752	1.02436	−2.03112	2.98970	1.99358825
18	3.98135	2.90743	3.01365	3.95705	2.03373	1.01995	−2.02514	2.99275	1.39157217
20	3.98735	2.93606	3.00856	3.96983	2.02375	1.01570	−2.01982	2.99498	0.97468241
22	3.99142	2.95576	3.00542	3.97860	2.01661	1.01205	−2.01537	2.99658	0.68460255
24	3.99418	2.96934	3.00345	3.98471	2.01155	1.00910	−2.01178	2.99770	0.48201790
26	3.99605	2.97871	3.00220	3.98901	2.00799	1.00679	−2.00895	2.99849	0.34012071
28	3.99733	2.98520	3.00141	3.99207	2.00550	1.00503	−2.00676	2.99902	0.24048339
30	3.99820	2.98969	3.00091	3.99426	2.00377	1.00371	−2.00507	2.99938	0.17036544
Solution	4.00000	3.00000	3.00000	4.00000	2.00000	1.00000	−2.00000	3.00000	

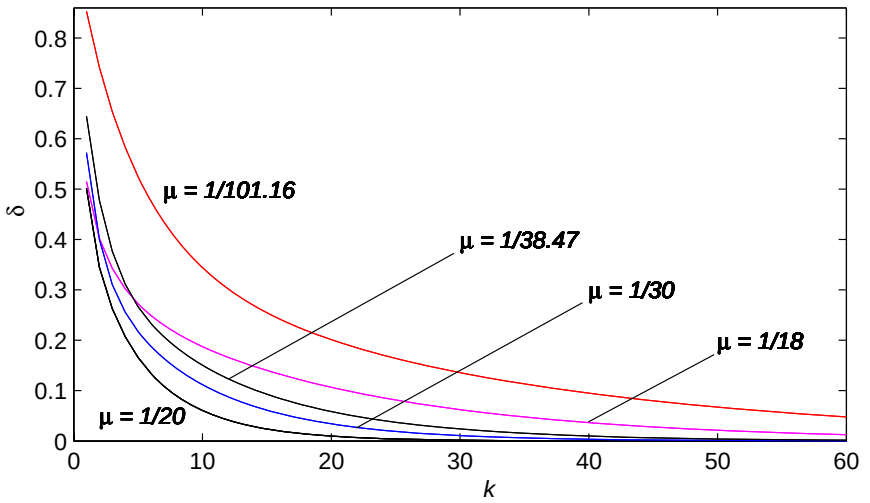


FIG. 1. *The relative error* δ *versus* k. $\mu = [\|A\|^2 + \|B\|^2 + \|D\|^2 + \|E\|^2]^{-1} = 1/101.16$, $\mu_0 = 2/38.47$, $\mu = \{\lambda_{\max}[A^{\mathrm{T}}A] + \lambda_{\max}[D^{\mathrm{T}}D] + \lambda_{\max}[BB^{\mathrm{T}}] + \lambda_{\max}[EE^{\mathrm{T}}]\}^{-1} = 1/38.47$.

larger the convergence factor μ is, the faster the convergence rate of the algorithm (or, the smaller the iteration error). However, if we keep enlarging μ, e.g., $\mu = 1/18$, the iteration error will become larger. Thus, there exists a best convergence factor such that the fastest convergence rate is obtained. This is still a topic left for the future.

Example 2. Suppose that the coupled matrix equations are

$$A_{11}X_1B_{11} + A_{12}X_2B_{12} = C_1,$$
$$A_{21}X_1B_{21} + A_{22}X_2B_{22} = C_2,$$

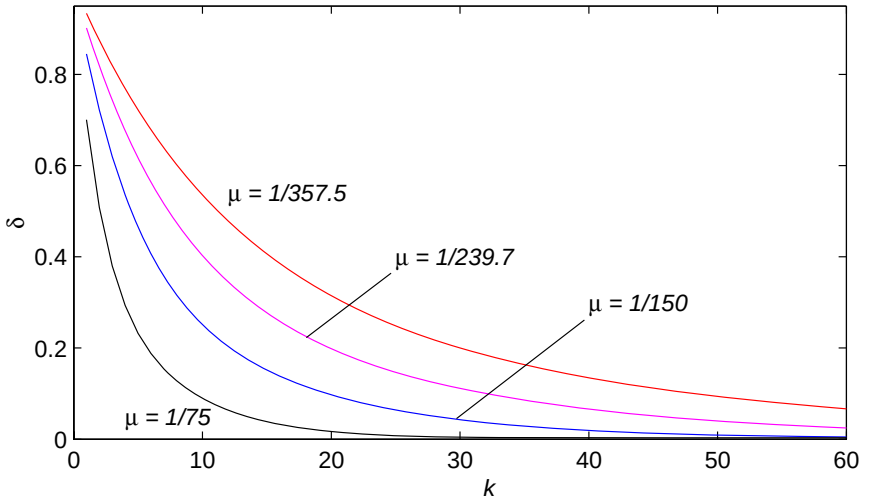


FIG. 2. *The relative error δ versus k.* $\mu = (\sum_{i=1}^{2}\sum_{j=1}^{2}\|A_{ij}B_{ij}\|^2)^{-1} = 1/357.5$, $\mu_0 = 2/239.7$, $\mu = \{\sum_{i=1}^{2}\sum_{j=1}^{2}\lambda_{\max}[A_{ij}^{\mathrm{T}}A_{ij}]\lambda_{\max}[B_{ij}^{\mathrm{T}}B_{ij}]\}^{-1} = 1/239.7$.

where

$$A_{11} = \begin{bmatrix} 3.00 & -2.00 \\ -1.00 & 1.00 \end{bmatrix}, \quad B_{11} = \begin{bmatrix} 1.00 & 1.00 \\ -1.00 & -2.00 \end{bmatrix}, \quad A_{12} = \begin{bmatrix} 2.00 & 1.00 \\ 1.00 & -2.00 \end{bmatrix},$$

$$B_{12} = \begin{bmatrix} 1.00 & -2.00 \\ -1.00 & 2.00 \end{bmatrix}, \quad C_1 = \begin{bmatrix} 1.30 & -3.60 \\ -2.10 & -1.30 \end{bmatrix},$$

$$A_{21} = \begin{bmatrix} 1.00 & 2.00 \\ 1.50 & -1.00 \end{bmatrix}, \quad B_{21} = \begin{bmatrix} 2.00 & -1.00 \\ 1.00 & 2.00 \end{bmatrix}, \quad A_{22} = \begin{bmatrix} 1.00 & -2.00 \\ 2.00 & -1.00 \end{bmatrix},$$

$$B_{22} = \begin{bmatrix} 1.00 & -1.00 \\ -2.00 & 1.00 \end{bmatrix}, \quad C_2 = \begin{bmatrix} 17.40 & 24.10 \\ 12.55 & 2.20 \end{bmatrix}.$$

Taking $X_1(0) = X_2(0) = 10^{-6}\mathbf{1}_{2\times 2}$, we apply the algorithm in (20)–(21) to compute $X_1(k)$ and $X_2(k)$. The iterative errors δ with different convergence factors μ are shown in Figure 2.

From Figure 2, we conclude that as μ increases from $\mu = 1/357.5, 1/239.7, 1/150$ to $1/75$, the iteration errors become smaller and smaller and eventually go to zero. This confirms the proposed theorems. In simulation, as we gradually enlarge μ, e.g., when $\mu > 1/70$, the errors become larger, and the algorithm diverges for $\mu > 1/65$.

6. Conclusions. Gradient iterative algorithms for solving Sylvester coupled matrix equations and general coupled matrix equations are studied by using the gradient search principle. The analysis indicates that, as in the least squares iterative algorithms in [9], the gradient iterative algorithms can achieve good convergence properties for any initial values. The family of iterative methods proposed for linear (coupled) matrix equations can be extended to study iterative solutions of other linear or nonlinear matrix equations, e.g., Riccati equations.

Appendix. Examples on stability. For a time-varying system,

$$x(k) = H_k x(k-1),\ H_k \in \mathbb{R}^{n\times n},\ x(0) = x_0 \neq 0, \tag{25}$$

even if H_k is stable for any k, i.e., all eigenvalues of H_k are inside the unit circle, there is no guarantee that system (25) is stable, implying $x(k) \to 0$. For example, take

$$H_k = \frac{1}{8}\begin{bmatrix} 0 & 9+(-1)^k 7 \\ 9-(-1)^k 7 & 0 \end{bmatrix}.$$

The two eigenvalues of H_k, $\pm\frac{1}{\sqrt{2}}$, are inside the unit circle, but the transition matrix of the system is given by

$$L_k = \begin{cases} \begin{bmatrix} 0 & 2^{-2k} \\ 2^k & 0 \end{bmatrix}, & k \text{ is odd}, \\ \begin{bmatrix} 2^k & 0 \\ 0 & 2^{-2k} \end{bmatrix}, & k \text{ is even}. \end{cases}$$

Clearly, the system $x(k) = H_k x(k-1)$ is unstable.

For another example, take

$$H_k = \begin{bmatrix} 1-(-1)^k & 0 \\ 0 & 1+(-1)^k \end{bmatrix}$$

or

$$H_k = \begin{cases} \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}, & k \text{ is odd}, \\ \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}, & k \text{ is even}. \end{cases}$$

One eigenvalue of H_k is 2, outside the unit circle, but the system $x(k) = H_k x(k-1)$ is exponentially stable.

REFERENCES

[1] J. K. Baksalary and R. Kala, *The matrix equation $AXB+CYD = E$*, Linear Algebra Appl., 30 (1980), pp. 141–147.

[2] A. Barraud, *A numerical algorithm to solve $A^{\mathrm{T}}XA-X = Q$*, IEEE Trans. Automat. Control, 22 (1977), pp. 883–885.

[3] R. Bitmead, *Explicit solutions of the discrete-time Lyapunov matrix equation and Kalman–Yakubovich equations*, IEEE Trans. Automat. Control, 26 (1981), pp. 1291–1294.

[4] R. Bitmead and H. Weiss, *On the solution of the discrete-time Lyapunov matrix equation in controllable canonical form*, IEEE Trans. Automat. Control, 24 (1979), pp. 481–482.

[5] I. Borno, *Parallel computation of the solutions of coupled algebraic Lyapunov equations*, Automatica, 31 (1995), pp. 1345–1347.

[6] T. Chen and B. A. Francis, *Optimal Sampled-data Control Systems*, Springer, London, 1995.

[7] K. E. Chu, *The solution of the matrix equations $AXB-CXD = E$ and $(YA-DZ, YC-BZ) = (E, F)$*, Linear Algebra Appl., 93 (1987), pp. 93–105.

[8] G. Corach and D. Stojanoff, *Index of Hadamard multiplication by positive matrices* II, Linear Algebra Appl., 332/334 (2001), pp. 503–517.

[9] F. Ding and T. Chen, *Iterative least squares solutions of coupled Sylvester matrix equations*, Systems Control Lett., 54 (2005), pp. 95–107.

[10] F. Ding and T. Chen, *Gradient based iterative algorithms for solving a class of matrix equations*, IEEE Trans. Automat. Control, 50 (2005), pp. 1216–1221.

[11] F. Ding and T. Chen, *Hierarchical gradient-based identification of multivariable discrete-time systems*, Automatica, 41 (2005), pp. 315–325.

[12] F. Ding and T. Chen, *Hierarchical least squares identification methods for multivariable systems*, IEEE Trans. Automat. Control, 50 (2005), pp. 397–402.

[13] Y. Fang, K. A. Loparo, and X. Feng, *New estimates for solutions of Lyapunov equations*, IEEE Trans. Automat. Control, 42 (1997), pp. 408–411.

[14] J. D. Gardiner, A. J. Laub, J. J. Amato, and C. B. Moler, *Solution of the Sylvester matrix equation* $AXB^{\mathrm{T}} + CXD^{T} = E$, ACM Trans. Math. Software, 18 (1992), pp. 223–231.

[15] G. H. Golub, S. Nash, and C. F. Van Loan, *A Hessenberg–Schur method for the matrix problem* $AX + XB = C$, IEEE Trans. Automat. Control, 24 (1979), pp. 909–913.

[16] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.

[17] J. Heinen, *A technique for solving the extended discrete Lyapunov matrix equation*, IEEE Trans. Automat. Control, 17 (1972), pp. 156–157.

[18] C. R. Johnson and L. Elsner, *The relationship between Hadamard and conventional multiplication for positive definite matrices*, Linear Algebra Appl., 92 (1987), pp. 231–240.

[19] I. Jonsson and B. Kågström, *Recursive blocked algorithm for solving triangular systems.* I. *One-sided and coupled Sylvester-type matrix equations*, ACM Trans. Math. Software, 28 (2002), pp. 392–415.

[20] I. Jonsson and B. Kågsrtöm, *Recursive blocked algorithm for solving triangular systems.* II. *Two-sided and generalized Sylvester and Lyapunov matrix equations*, ACM Trans. Math. Software, 28 (2002), pp. 416–435.

[21] W. H. Kwon, Y. S. Moon, and S. C. Ahn, *Bounds in algebraic Riccati and Lyapunov equations: A survey and some new results*, Internat. J. Control, 64 (1996), pp. 377–389.

[22] L. Ljung, *System Identification: Theory for the User*, 2nd ed., Prentice–Hall, Englewood Cliffs, NJ, 1999.

[23] T. Mori and A. Derese, *A brief summary of the bounds on the solution of the algebraic matrix equations in control theory*, Internat. J. Control, 39 (1984), pp. 247–256.

[24] H. Mukaidani, H. Xu, and K. Mizukami, *New iterative algorithm for algebraic Riccati equation related to* H_∞ *control problem of singularly perturbed systems*, IEEE Trans. Automat. Control, 46 (2001), pp. 1659–1666.

[25] L. Qiu and T. Chen, *Unitary dilation approach to contractive matrix completion*, Linear Algebra Appl., 379 (2004), pp. 345–352.

[26] S.-Y. Shim and Y. Chen, *Least squares solution of matrix equation* $AXB^* + CYD^* = E$, SIAM J. Matrix Anal. Appl., 24 (2003), pp. 802–808.

[27] G. Starke and W. Niethammer, *SOR for* $AX - XB = C$, Linear Algebra Appl., 154 (1991), pp. 355–375.

[28] T. Stykel, *Numerical solution and perturbation theory for generalized Lyapunov equations*, Linear Algebra Appl., 349 (2002), pp. 155–185.

[29] V. L. Syrmos, P. Misra, and R. Aripirala, *On the discrete generalized Lyapunov equation*, Automatica, 31 (1995), pp. 297–301.

[30] K. Takaba, N. Morihhira, and T. Katayama, *A generalized Lyapunov theorem for descriptor system*, Systems Control Lett., 24 (1995), pp. 49–51.

[31] S. Xiang, *On an inequality for the Hadamard product of an M-matrix or an H-matrix and its inverse*, Linear Algebra Appl., 367 (2003), pp. 17–27.

[32] G. Xu, M. Wei, and D. Zheng, *On solutions of matrix equation* $AXB + CYD = F$, Linear Algebra Appl., 279 (1998), pp. 93–109.

SIAM J. CONTROL OPTIM.
Vol. 44, No. 6, pp. 2285–2299

SOME NEW REGULARITY PROPERTIES FOR THE MINIMAL TIME FUNCTION*

GIOVANNI COLOMBO†, ANTONIO MARIGONDA†, AND PETER R. WOLENSKI‡

Abstract. A minimal time problem with linear dynamics and convex target is considered. It is shown, essentially, that the epigraph of the minimal time function $T(\cdot)$ is φ-convex (i.e., it satisfies a kind of exterior sphere condition with locally uniform radius), provided $T(\cdot)$ is continuous. Several regularity properties are derived from results in [G. Colombo and A. Marigonda, Calc. Var. Partial Differential Equations, 25 (2005), pp. 1–31], including twice a.e. differentiability of $T(\cdot)$ and local estimates on the total variation of DT.

Key words. nonsmooth analysis, proximally smooth and φ-convex sets, small time controllability, functions with φ-convex epigraph

AMS subject classifications. 49N05, 49J52

DOI. 10.1137/050630076

1. Introduction.

The regularity of the minimal time function $T(\cdot)$ is a widely studied topic (see, e.g., [5, 24, 6, 7, 8, 25, 3] and references therein), under different viewpoints. In particular, it is proved in [7] that with linear dynamics and convex target, $T(\cdot)$ is semiconvex provided the Petrov condition holds. The latter is equivalent to the Lipschitz continuity of $T(\cdot)$ near the target and thus is a type of *strong* local controllability condition. Since $T(\cdot)$ is not necessarily convex (see p. 100 in [14]) even for a point-target, this is a natural regularity class for a linear minimum time problem.

Classical examples, however, exhibit minimal time functions that are not locally Lipschitz even though the system is small time locally controllable (see, e.g., [3, Example 2.7, p. 242]). Therefore, it is natural to seek conditions that identify regularity properties of $T(\cdot)$ in situations where $T(\cdot)$ is not locally Lipschitz. This motivated the results in [12], where a class of lower semicontinuous functions was studied whose epigraph satisfy an external sphere condition with locally uniform radius; this property, for general sets, is often referred to as *positive reach* [16], *φ-convexity* [15], or *proximal smoothness* [11]. Such functions are semiconvex if and only if they are locally Lipschitz and therefore are a good candidate to extend the result in [7] under more general controllability conditions. In [12], functions with φ-convex epigraph were shown to have several fine properties. In particular, a function in this class is of locally bounded variation; moreover, a.e. x admits a neighborhood where the function is indeed semiconvex, and as a consequence it is twice differentiable almost everywhere.

It will be shown below that the epigraph of $T(\cdot)$ is φ-convex, under suitable controllability assumptions. More precisely, we prove that for a linear control problem with a convex target S, the epigraph of $T(\cdot)$ is φ-convex (Theorem 3.7), provided T is continuous. Our assumptions are satisfied in several situations, including, e.g., the

*Received by the editors April 27, 2005; accepted for publication (in revised form) September 21, 2005; published electronically February 3, 2006. Work partially supported by MIUR project "Viscosity, metric, and control theoretic methods for nonlinear partial differential equations."
http://www.siam.org/journals/sicon/44-6/63007.html

†Dipartimento di Matematica Pura e Applicata, Università di Padova, via Belzoni 7, 35131 Padova, Italy (colombo@math.unipd.it, amarigo@math.unipd.it).

‡Department of Mathematics, Louisiana State University, 326 Lockett Hall, Baton Rouge, LA 70803-4918 (wolenski@math.lsu.edu).

case where the system fulfils the Kalman rank condition and the target is the origin. An example where small time controllability does not hold, yet is covered by Theorem 3.7, is presented in section 2.4.

Our analysis depends on a representation formula for the normal cone to sublevel sets of T, which is proved using simple tools of convex analysis together with Pontryagin's maximum principle. The techniques used here are essentially linear, due to the repeated use of explicit formulas. The main difficulty to handle is the possibility of having points where both the subdifferential and the superdifferential of T are empty, due to the lack of Lipschitz continuity. Finally, the regularity results in [12] are applied to $T(\cdot)$, and the corresponding properties of T are listed in Corollary 3.8.

We recall that for nonlinear dynamics, the semiconvexity of $T(\cdot)$ is generally not present (see, e.g., [6, Example 4.3]). However, in analogy with [6] and [8], one may expect regularity results of a similar nature under more restrictive assumptions on the target and dynamics. We mention that proving such a nonlinear result by methods analogous to ours must overcome two main difficulties: first, the existing nonlinear results rely either on the Lipschitz continuity of $T(\cdot)$ (see [7]) or are rather general, but provide substantially weaker estimates (see [8]); second, weaker controllability conditions lead to singularities of $T(\cdot)$ that are of both semiconvex and semiconcave type (see [4]) together with cusp points. Hence it is not clear how to obtain a nonlinear version of our Theorem 3.1, and this will be a topic of future research.

2. Preliminaries. This section briefly introduces concepts from nonsmooth analysis, geometric measure theory, and control theory.

2.1. Nonsmooth analysis. A standard reference for the nonsmooth concepts introduced here is [10]. Let $K \subseteq \mathbb{R}^n$ be closed. We denote, for $x \in \mathbb{R}^n$,

$$\begin{array}{rcll} d_K(x) & = & \min\{\|y-x\| : y \in K\} & \text{(the distance of } x \text{ from } K\text{)}, \\ \pi_K(x) & = & \{y \in K : \|y-x\| = d_K(x)\} & \text{(the projections of } x \text{ onto } K\text{)}, \\ B(K,\rho) & = & \{y \in \mathbb{R}^n : d_K(y) \le \rho\}. & \end{array}$$

A vector v is a *proximal normal* to K at $x \in K$ (notated by $v \in N_K^P(x)$) if there exists $\sigma = \sigma(v,x) \ge 0$ such that

$$\langle v, y-x\rangle \le \sigma\|y-x\|^2 \quad \text{for all } y \in K. \tag{2.1}$$

For $v \neq 0$, then $v \in N_K^P(x)$ if and only if this there exists $\lambda > 0$ such that $\pi_K(x+\lambda v) = \{x\}$. If K is convex, then $N_K^P(x)$ equals the normal cone $N_K(x)$ to K at x as defined in convex analysis, namely, the set of vectors $v \in \mathbb{R}^n$ for which

$$\langle v, y-x\rangle \le 0 \quad \text{for all } y \in K.$$

Suppose $f : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ is lower semicontinuous and $\mathrm{epi}(f) := \{(x,\xi) : \xi \ge f(x)\}$ and $\mathrm{dom}(f) = \{x \in \mathbb{R}^n : f(x) < +\infty\}$ are its epigraph and (effective) domain, respectively. Let $x \in \mathrm{dom}(f)$. A vector $\zeta \in \mathbb{R}^n$ is a *proximal subgradient* of f at x (notated by $\zeta \in \partial_P f(x)$) if $(\xi,-1) \in N^P_{\mathrm{epi}(f)}(x, f(x))$; equivalently (see [10, Theorem 1.2.5]), $\xi \in \partial_P f(x)$ if and only if there exist $\sigma, \eta > 0$ such that

$$f(y) \ge f(x) + \langle \zeta, y-x\rangle - \sigma\|y-x\|^2 \quad \text{for all } y \in B(x,\eta). \tag{2.2}$$

The following class of sets (see [11, section 4]) will play a major role in our analysis.

Definition 2.1. *Suppose $K \subseteq \mathbb{R}^n$ is closed and $r > 0$. Then K is* r-proximally smooth *if the distance function d_K is continuously differentiable on $B(K,r)\setminus K$.*

Geometrically, in virtue of [11, Theorem 4.1], this means that every nonzero proximal normal to K is realized by an r-ball, i.e.,

$$\langle v, y - x\rangle \le \frac{1}{2r}\|y - x\|^2 \tag{2.3}$$

for all $x, y \in K$ and $v \in N_K^P(x)$, $\|v\| = 1$. Moreover, if K is proximally smooth, then the Clarke normal cone to K at x coincides with $N_K^P(x)$ for all $x \in K$, and in particular $N_K^P(x)$ is nontrivial (see [11]) at all points x on the boundary of K.

Proximal smoothness is rather restrictive for noncompact sets such as epigraphs. The following generalization allows for the constant in (2.3) to depend on x.

DEFINITION 2.2. *Suppose $K \subseteq \mathbb{R}^n$ is closed and $\varphi : K \to [0, +\infty)$ is continuous. We say that K is φ-convex if*

$$\langle v, y - x\rangle \le \varphi(x)\|y - x\|^2 \tag{2.4}$$

for all $x, y \in K$ and $v \in N_K^P(x)$ with $\|v\| = 1$.

Comparing (2.3) and (2.4) reveals that K is r-proximally smooth if and only if it is φ-convex with $\varphi(x) = \frac{1}{2r}$ for all $x \in K$. Such sets are also referred to as prox-regular in [22], and several characterizations are known (see [16, 11, 22]). However, they will not be used here. We recall that, in particular, convex sets, or sets with a $\mathcal{C}^{1,1}$-boundary, are φ-convex.

If K is the epigraph of a continuous function $T(\cdot)$, then the φ-convexity condition (2.4) takes the form

$$\langle (\zeta, \xi), (y, \beta) - (x, \alpha)\rangle \le \varphi(x, \alpha)(\|\zeta\| + |\xi|)\,(\|y - x\|^2 + |\beta - \alpha|^2) \tag{2.5}$$

for all $x, y \in \mathrm{dom}(T)$, $\alpha \ge T(x)$, $\beta \ge T(y)$, $(\zeta, \xi) \in N^P_{\mathrm{epi}(T)}(x, \alpha)$, with $\varphi : \mathrm{epi}(T) \to [0, +\infty)$ continuous.

2.2. Geometric measure theory. The study of some fine regularity properties of φ-convex sets and functions with φ-convex epigraph is taken up in [12] and will be quoted here. Stating these requires concepts from geometric measure theory [1, 20].

For $0 \le k \le n$, the k-dimensional Hausdorff measure in $\mathbb{R}^n$ is denoted by $\mathcal{H}^k$. The Hausdorff dimension of a set E is $\mathcal{H} - \dim(E) := \inf\{k \ge 0 : \mathcal{H}^k(E) = 0\}$. A set $E \subseteq \mathbb{R}^n$ is *countably k-rectifiable* if there exist countably many Lipschitz functions $f_i : \mathbb{R}^k \to \mathbb{R}^n$ such that

$$\mathcal{H}^k\left(E \setminus \bigcup_{i=0}^{+\infty} f_i(\mathbb{R}^k)\right) = 0.$$

Let $\Omega \subset \mathbb{R}^n$ be open, and $u \in L^1(\Omega)$; we say that u is a *function of bounded variation in Ω* ($u \in BV(\Omega)$) if the distributional derivative of u is representable by a finite Radon measure in Ω, i.e., if

$$\int_\Omega u \frac{\partial \varphi}{\partial x_i}\, dx = -\int_\Omega \varphi\, dD_i u \text{ for all } \varphi \in C_c^\infty(\Omega),\ i = 1, \dots, n,$$

for some finite Radon measure $Du = (D_1 u, \dots, D_n u)$.

2.3. Control theory: Generalities. We consider throughout the paper a linear control system of the form

$$\left\{\begin{array}{lcl} \dot{y}(t) & = & Ay(t)+u(t) \quad \text{a.e.},\\ u(t) & \in & \mathcal{U} \ \text{a.e.},\\ y(0) & = & x, \end{array}\right. \tag{2.6}$$

where $A \in \mathrm{Mat}_{n\times n}(\mathbb{R})$. The control set $\mathcal{U} \subset \mathbb{R}^n$ is compact and convex, and the control function $u(\cdot)$ is measurable. For all $t > 0$, we denote by $\mathcal{U}^t_{\mathrm{ad}}$ the set of admissible controls, i.e., the measurable functions $u : [0,t] \to \mathbb{R}^n$, such that $u(t) \in \mathcal{U}$ a.e. on $[0,t]$. For any $u(\cdot) \in \mathcal{U}^t_{\mathrm{ad}}$, the unique Carathéodory solution of (2.6) is denoted by $y^{x,u}(\cdot)$.

Suppose we are now given a closed nonempty set $S \subset \Omega$, which is called the *target set*. For fixed $x \notin S$, the *minimal time* $T(x)$ to reach S from x is defined by

$$T(x) := \inf\{T \geq 0 \,:\, \exists\, u(\cdot) \text{ such that } y^{x,u}(T) \in S\}.$$

When the set of controls $u(\cdot)$ steering x to S is empty, then $T(x) = +\infty$. Since the velocity sets $F(y) := \{Ay + u \,:\, u \in \mathcal{U}\}$ are convex, then standard arguments (see [9, Theorem 9.2.i, p. 311]) show the infimum is actually a minimum (provided it is finite); that is, there exists an optimal control steering x to S in the minimal time.

The reachable set from a point $x \in \Omega$ in time T is the set

$$R^T(x) = \{y(T) : y(\cdot) \text{ satisfies (2.6)}\}.$$

If $\bar{x} \in R^T(x)$, then $\bar{x}$ is *realized* by the control function $\bar{u}(\cdot)$ if $\bar{x} = y^{x,\bar{u}}(T)$. Note that $\bar{x} \in R^T(x)$ is realized by $\bar{u}(\cdot)$ if and only if the (equivalent) formulas

$$\bar{x} = e^{AT}x + \int_0^T e^{A(T-t)}\bar{u}(t)\,dt \quad \text{and} \quad x = e^{-AT}\bar{x} - \int_0^T e^{-At}\bar{u}(t)\,dt \tag{2.7}$$

hold. It is well known that $R^T(x)$ is convex and compact. It is convenient to also notate as $R^T_-(\bar{x})$ the reversed-time reachable set from a point $\bar{x}$, which is the reachable set associated to the dynamics $\dot{y} = -Ay - u$. Namely,

$$R^T_-(\bar{x}) = \{y(T) : \dot{y}(t) = -Ay(t) - u(t),\ u(\cdot) \in \mathcal{U}^T_{\mathrm{ad}} \text{ a.e.},\ y(0) = \bar{x}\}.$$

It is clear that $\bar{x} \in R^T(x)$ if and only if $x \in R^T_-(\bar{x})$. For $r > 0$, let

$$\begin{aligned} S(r) &= \{x \in \mathbb{R}^n : T(x) \leq r\},\\ \mathcal{R} &= \{x \in \mathbb{R}^n : T(x) < +\infty\}, \end{aligned}$$

and observe

$$S(r) = \bigcup_{\bar{x}\in S,\, 0\leq T\leq r} R^T_-(\bar{x}).$$

Recall that a closed set $S \subseteq \Omega$ is *strongly invariant* for the system (2.6) if for all $x \in S$ and $T > 0$, one has $R^T(x) \subseteq S$. Analogously, S is *weakly invariant* (or *viable*) if for all $x \in S$ and all small $T > 0$, there exists a trajectory of (2.6) which remains in S for all $t \in [0,T]$.

A major tool in our analysis is the minimized Hamiltonian $h : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, given by

$$h(x,\zeta) = \langle Ax,\zeta\rangle + \min_{u\in\mathcal{U}}\langle u,\zeta\rangle. \tag{2.8}$$

It is known that a set S is weakly invariant for the dynamics (2.6) if $h(x,\zeta)\le 0$ for all $x\in S$ and $\zeta\in N_S^P(x)$ (see [24], [10, Theorem 2.10]).

The *adjoint equation* associated with (2.6) is

$$\begin{cases} \dot p(t) = -A^\top p(t), \\ p(T) = \bar p, \end{cases} \tag{2.9}$$

and an adjoint arc is

$$p(t) = e^{A^\top(T-t)}\bar p, \tag{2.10}$$

which is the solution of (2.9). Pontryagin's maximum principle is stated next.

PROPOSITION 2.3 (maximum principle). *Suppose $\bar x\in R^T(x)$ is realized by $\bar u(\cdot)$. Then $\bar x\in$ bdry $R^T(x)$ (= the boundary of $R^T(x)$) if and only if there exists $\bar p\neq 0$ so that the solution $\bar p(\cdot)$ of* (2.9) *satisfies*

$$\langle \bar p(t),\bar u(t)\rangle = \max_{u\in U}\langle \bar p(t),u\rangle \tag{2.11}$$

for almost all $t\in[0,T]$. Moreover, in this case, $p(t)\in N_{R^t(x)}\big(y^{x,\bar u}(t)\big)$ for each $t\in[0,T]$.

A standard reference for the proof is [17, section 13].

2.4. Continuity of the minimal time function. Continuity properties of the minimal time function is a widely studied topic, mainly in connection with controllability. We refer to Chapter IV in [3] and references therein for an introduction to the subject.

DEFINITION 2.4. *The control system* (2.6) *is* small time controllable *(STC) near the target S if $S\subseteq$ int $S(r)$ for all small $r>0$.*

We collect some known results relating STC to continuity of $T(\cdot)$, with main emphasis on a target more general than a singleton, in the following theorem.

THEOREM 2.5. *Assume (for simplicity) that S is compact.*

(1) *Suppose $S=\{0\}$ and $0\in$ rel int $S(r)$ for all $r>0$. Then $T(\cdot)$ is continuous on $\mathcal{R}$.*

(2) (Generalized Petrov condition.) *Suppose there exist $\delta>0$ and a continuous nondecreasing function $\mu:[0,\delta]\to[0,+\infty)$ with the properties*

(a) *$\mu(0)=0$, $\mu(\rho)>0$ for $\rho>0$, and $\int_0^\delta \frac{d\rho}{\mu(\rho)}<+\infty$;*

(b) *for all $x\in B(S,\delta)\setminus S$ there exists $\bar s\in\pi_S(x)$ such that*

$$h(x,x-\bar s)\le -\mu(\|x-\bar s\|)\|x-\bar s\|. \tag{2.12}$$

Then the system (2.6) *is STC near S and the minimal time function is continuous in a neighborhood of S.*

(3) (Second order Petrov condition.) *Suppose that S is the closure of an open set with $\mathcal{C}^2$-boundary, and assume that there exist $\delta>0$ and $\eta>0$ such that for all $x\in B(S,\delta)\setminus S$,*

(a) *$h(x,Dd_S(x))\le 0$,*

(b) $\langle Dd_S(x), A^2x\rangle + 2\left\langle \langle D^2 d_S(x), Ax\rangle, Ax\right\rangle \leq -\eta$.

Then the system (2.6) *is STC near* S *and the minimal time function is Hölder continuous with exponent* $1/2$ *in a neighborhood of* S.

(4) *Suppose* $S = \{0\}$ *and* $\mathcal{U} = \{Bu : u \in \mathbb{R}^m,\ u \in [-1,1]^m\}$, $B \in Mat_{n\times m}(\mathbb{R})$. *The following are equivalent for a fixed integer* k, $k = 0, 1, \ldots, n-1$.

(a) $T(\cdot)$ *is Hölder continuous in* $\mathbb{R}^n$ *with exponent* $1/(k+1)$;

(b) (Kalman rank condition)

$$\operatorname{rank}[B, AB, \ldots, A^kB] = n.$$

Proof. The proof of (1) is in [14, Theorem II.4.3]. Various versions of (2), obtained with different methods, can be found, e.g., in [24], [6], [7, Chapter 8, section 8.2], [23], [18], [21], [19]. Condition (3) is a particular case of a controllability result contained in [19]. The proof of (4) can be found in [2, Chapter 2, section 6]. □

We will consider a slightly more general situation, where the continuity of the minimal time function is not directly linked to an STC condition. We illustrate this with a simple example.

Example 1. *Let* $\alpha > 1$ *and* $S = \{(x,y) \in \mathbb{R}^2 : y \geq |x|^\alpha\}$. Let $\mathcal{U} = [-1,1]$ and consider the linear control system

$$\left\{ \begin{array}{rcl} \dot{x} &=& u \in \mathcal{U}, \\ \dot{y} &=& 0. \end{array} \right. \tag{2.13}$$

None of the conditions listed in Theorem 2.5 is satisfied in a neighborhood of S, and actually $\mathcal{R} = \mathbb{R} \times [0, +\infty)$ is not a neighborhood of S. Let $x > 0$, $0 \leq y < x^\alpha$. Then $T(x,y) = x - y^{1/\alpha}$, which is continuous on $\mathcal{R} \setminus S$ but not locally Lipschitz. We observe that for all (x,y), there exists a control $u(x,y)$ (actually $u(x,y) = -\operatorname{sgn}(x)$) such that $A(x,y) + u(x,y) = (u(x,y), 0)$ points toward S. However, the angle between the vector pointing to S and the external normal to S is not uniformly bounded away from 0, and in fact this angle tends to 0 as $(x,y) \to (0,0)$. We estimate its rate of convergence to 0 along the x-axis. Let $\xi(x) = x + \alpha x^{2\alpha-1}$. Observe that the segment joining $(\xi(x), 0)$ and (x, x^α) is orthogonal to the graph of $y = x^\alpha$ at (x, x^α). Moreover, $\xi(x) \sim x$ for $x \to 0$ and

$$d_S((\xi(x), 0)) = x^\alpha \sqrt{1 + \alpha^2 x^{2(\alpha-1)}} \sim x^\alpha \text{ for } x \to 0.$$

Finally,

$$\begin{aligned} \min_{u\in[-1,1]} \left\langle (u,0), \frac{(\xi(x),0) - (x,x^\alpha)}{d_S(\xi(x),0)} \right\rangle &= \frac{x - \xi(x)}{d_S(\xi(x),0)} \\ &= -\frac{\alpha x^{\alpha-1}}{\sqrt{1 + \alpha^2 x^{2(\alpha-1)}}} \\ &\leq -\text{const}\,(d_S(\xi(x),0))^{\frac{\alpha-1}{\alpha}}. \end{aligned}$$

□

In this example, the angle satisfies an estimate of the type (2.12). However, this estimate does not hold in an entire neighborhood of S, and the continuity of T in $\mathcal{R}$ is not covered by any of the statements in Theorem 2.5. The forthcoming paper [19] contains a result covering the Hölder continuity of T in $\mathcal{R}$ also in the above example.

3. The epigraph of the minimal time function, and differentiability properties. We repeat the setting we are concerned with. We consider the linear system

$$\left\{\begin{array}{lcll} \dot{y}(t) & = & Ay(t)+u(t) & \text{a.e.,} \\ y(0) & = & x, & \\ u(t) & \in & \mathcal{U} & \text{a.e.} \end{array}\right. \tag{3.1}$$

with $\mathcal{U} \subseteq \mathbb{R}^n$ compact and convex. Let $S \neq \emptyset$ be the target set.

Let $\delta > 0$ be given, and set $\mathcal{R}_\delta = S(\delta) \setminus S$. We make the following further assumptions:

(H1) S is closed and convex, and $h(x,\zeta) \leq 0$ for all $x \in S$ and $\zeta \in N_S(x)$;

(H2) $T(\cdot)$ is continuous in $S(\delta)$.

Observe that (H1) and (H2) do not imply STC, because $S(\delta)$ is not required to be a neighborhood of S. Such a situation is illustrated by Example 1.

The following result is an easy consequence of (H1).

PROPOSITION 3.1. *Under the above assumption* (H1)*, the sets $S(r)$ are compact and convex, and if $r_1 \leq r_2$ we have $S(r_1) \subseteq S(r_2)$. Therefore $\mathcal{R}$ is convex.*

We need a few technical lemmas. A version of Lemmas 3.2 and 3.3 already appeared in [13, section 2]. We repeat the proofs here, in order to make this paper more self-contained. The first two concern a representation of the normal cone to the level sets of T and of the proximal subdifferential of T.

LEMMA 3.2. *Let* (H1) *hold, and let $r \geq 0$, $x \in \mathbb{R}^n$ with $T(x) = r$, and $\bar{x} \in S \cap R^r(x)$. Then*

$$N_{S(r)}(x) = \Big\{ -e^{A^\top r}\bar{p} : \bar{p} \in [-N_S(\bar{x})] \cap N_{R^r(x)}(\bar{x}) \Big\}, \tag{3.2}$$

and therefore the right-hand side is independent of $\bar{x} \in S \cap R^r(x)$.

Proof (see also [13, Theorems 4 and 8]).

"$\subseteq$." Let $\zeta \in N_{S(r)}(x)$. Then, by convexity,

$$\langle \zeta, y - x \rangle \leq 0 \quad \text{for all } y \in S(r). \tag{3.3}$$

Let $\bar{u}(\cdot) \in \mathcal{U}_{ad}^r$ be an admissible control that realizes $\bar{x}$, and thus (2.7) holds with $T = r$. The rest of the proof is broken into two claims. □

Claim 1. $e^{-A^\top r}\zeta \in N_S(\bar{x})$.

Proof of Claim 1. Let $\bar{y} \in S$, and define

$$y = e^{-Ar}\bar{y} - \int_0^r e^{-At}\bar{u}(t)\,dt, \tag{3.4}$$

which therefore belongs to $S(r)$. We have

$$\begin{aligned} \langle e^{-A^\top r}\zeta, \bar{y} - \bar{x} \rangle &= \langle \zeta, e^{-Ar}\bar{y} - e^{-Ar}\bar{x} \rangle \\ &= \langle \zeta, y - x \rangle \quad \text{(by (2.7) and (3.4))} \\ &\leq 0 \quad \text{(by (3.3) and since } y \in S(r)\text{)}. \end{aligned}$$

It follows that $e^{-A^\top r}\zeta \in N_S(\bar{x})$.

Claim 2. $-e^{-A^\top r}\zeta \in N_{R^r(x)}(\bar{x})$.

Proof. First note that $x \in R^r_-(\bar{x}) \subseteq S(r)$, and therefore $\zeta \in N_{R^r_-(\bar{x})}(x)$. By Proposition 2.3 applied to the reversed time data $-A$ and $-\mathcal{U}$, we have that for all $t \in [0, r]$,

$$\left\langle -e^{-A^\top t}\zeta, \bar{u}(t) \right\rangle = \max_{u \in U} \left\langle -e^{-A^\top t}\zeta, u \right\rangle. \tag{3.5}$$

Now suppose $\bar{y} \in R^r(x)$, so that

$$\bar{y} = e^{Ar}x + \int_0^r e^{A(r-t)}u(t)\,dt$$

for some $u(\cdot) \in \mathcal{U}^r_{ad}$. We have

$$\begin{aligned}\left\langle -e^{-A^\top r}\zeta, \bar{y} - \bar{x} \right\rangle &= \left\langle -e^{-A^\top r}\zeta, \int_0^r e^{A(r-t)}\big(u(t) - \bar{u}(t)\big)\,dt \right\rangle \\ &= \int_0^r \left\langle -e^{-A^\top t}\zeta, u(t) - \bar{u}(t) \right\rangle dt \\ &\le 0,\end{aligned}$$

where the last inequality follows from (3.5). The validity of Claim 2 is now established.

It is clear that the "$\subseteq$" inclusion in (3.2) follows from Claims 1 and 2.

"$\supseteq$." Let $\bar{x} \in S \cap R^r(x)$, and let $\bar{p} \in [-N_S(\bar{x})] \cap N_{R^r(x)}(\bar{x})$. Let $y \in S(r)$ and $\bar{y} \in S \cap R^r(y)$. Respectively, let $u(\cdot), \bar{u}(\cdot) \in \mathcal{U}^T_{ad}$ realize $\bar{y}$, $\bar{x}$, and thus

$$\begin{aligned} y &= e^{-Ar}\bar{y} - \int_0^r e^{-At}u(t)\,dt, \\ x &= e^{-Ar}\bar{x} - \int_0^r e^{-At}\bar{u}(t)\,dt. \end{aligned} \tag{3.6}$$

We have

$$\begin{aligned}\left\langle -e^{A^\top r}\bar{p}, y - x \right\rangle &= \left\langle -\bar{p}, e^{Ar}\big(y - x\big) \right\rangle \\ &= \left\langle -\bar{p}, \bar{y} - \bar{x} \right\rangle + \int_0^r \left\langle -p(t), \bar{u}(t) - u(t) \right\rangle dt\end{aligned}$$

by (3.6). Since $-\bar{p} \in N_S(\bar{x})$, the first term on the right-hand side of the previous expression is nonpositive. By the maximum principle, the second term is also nonpositive. Hence the assertion $-e^{A^\top r}\bar{p} \in N_{S(r)}(x)$ follows, and the proof is concluded. ∎

LEMMA 3.3. *Let the assumption* (H1) *hold. Let $x \in S(r)$, $T(x) = r > 0$ and let $\bar{x} \in S \cap R^r(x)$. Then a vector ζ belongs to $\partial_P T(x)$ if and only if*

$$h(x, \zeta) = -1$$

and

$$-e^{-A^\top r}\zeta \in \big[-N_S(\bar{x})\big] \cap N_{R^r(x)}(\bar{x}).$$

Proof. By Theorem 5.1 in [25],

$$\partial_P T(x) = N_{S(r)}(x) \bigcap \big\{\zeta : h(x, \zeta) = -1\big\}. \tag{3.7}$$

Then the statement follows from Lemma 3.2. ∎

The next three lemmas concern the Hamiltonian, mainly in connection with normal vectors to the epigraph of T.

LEMMA 3.4. *Let $r > 0$, $x_0 \in S(r)$. If $\zeta \in N_{S(r)}(x_0)$, then $h(x_0, \zeta) \leq 0$.*

Proof. By contradiction, let $\zeta \in N_{S(r)}(x_0)$ be such that $h(x_0, \zeta) > 0$. By definition of Hamiltonian, we have $\zeta \neq 0$. Let $x(\cdot)$ be an optimal trajectory starting from $x(0) = x_0$ and let $u(\cdot)$ be an optimal control realizing $x(\cdot)$. Let $z = x_0 + \zeta$. We are now going to contradict the dynamic programming principle. Indeed, by convexity of $S(r)$, it is enough to show that there exists $\eta > 0$ such that $x(t) \in B(z, \|\zeta\|)$ for all $t \in (0, \eta)$. In fact this implies that $x(t) \notin S(r)$ for all $t \in (0, \eta)$, i.e., there exists $0 < \bar{t} < T(x_0)$ such that $T(x(\bar{t})) > T(x_0)$, which is against the optimality of $x(\cdot)$. We have

$$\begin{aligned}\frac{d}{dt}\|x(t) - z\|^2 &= \frac{d}{dt}\langle x(t) - x_0 - \zeta, x(t) - x_0 - \zeta\rangle \\ &= 2\langle \dot{x}(t), x(t) - x_0 - \zeta\rangle \\ &= 2\langle \dot{x}(t), x(t) - x_0\rangle - 2\langle Ax(t) + u(t), \zeta\rangle \\ &\leq 2\left(K^2 t - h(x(t), \zeta)\right),\end{aligned}$$

where K is a bound on $\|\dot{x}\|$. According to our hypothesis $h(x(0), \zeta) > 0$, so for small t we have by continuity $\frac{d}{dt}\|x(t) - z\|^2 < 0$, which implies $x(t) \in B(z, \|\zeta\|)$. ∎

LEMMA 3.5. *Let (H1) hold. Let $r > 0$, and let $x_0 \in \mathbb{R}^n$ be such that $T(x_0) = r$. If $(\zeta, 0) \in N^P_{\mathrm{epi}(T)}(x_0, T(x_0))$, then $\zeta \in N_{S(r)}(x_0)$ and $h(x_0, \zeta) \leq 0$.*

Proof. In view of Lemma 3.4, it is enough to show that $\zeta \in N_{S(r)}(x_0)$. To this aim, observe that there exists $\sigma > 0$ such that

$$\langle(\zeta, 0), (y, \xi) - (x_0, T(x_0))\rangle \leq \sigma(\|x_0 - y\|^2 + |T(x_0) - \xi|^2) \tag{3.8}$$

for all $(y, \xi) \in \mathrm{epi}(T)$. In particular, for $y \in S(r)$ and $\xi = r$ the inequality (3.8) yields

$$\langle \zeta, y - x_0\rangle \leq \sigma\|x_0 - y\|^2,$$

and this says that $\zeta \in N^P_{S(r)}(x_0)$. Since $S(r)$ is convex, this fact is equivalent to $\zeta \in N_{S(r)}(x_0)$. The proof is concluded. ∎

LEMMA 3.6. *Let $r > 0$, $x_0 \in S(r)$, $T(x_0) = r$. If $(\zeta, -1) \in N^P_{\mathrm{epi}(T)}(x_0, T(x_0))$, then $h(x_0, \zeta) = -1$.*

Proof. By hypothesis, $\zeta \in \partial_P T(x_0)$; then apply Lemma 3.3. ∎

The following is the main result of the paper.

THEOREM 3.7. *Consider the system* (3.1) *with the assumptions* (H1), (H2). *Then there exists a continuous function φ such that the epigraph of $T_{|\mathcal{R}_\delta}$ is φ-convex.*

Proof. The proof consists of two steps. In the first step we establish an inequality of the type (2.5) for a particular choice of points in $\mathrm{epi}(T)$ by assuming that S is compact. In the second, we show that the inequality proved in the first step holds in general.

Step 1. Let S be compact. We claim that there exists $K = K(\delta) > 0$ with the following property: for all $x_1, x_2 \in \mathcal{R}_\delta$, for all $(\zeta, \xi) \in N^P_{\mathrm{epi}(T)}(x_1, T(x_1))$ with $\xi \in \{0, -1\}$ it holds

$$\langle(\zeta, \xi), (x_2, T(x_2)) - (x_1, T(x_1))\rangle \leq K(\|\zeta\| + |\xi|)(\|x_2 - x_1\|^2 + |T(x_2) - T(x_1)|^2). \tag{3.9}$$

Proof of Step 1. Let $r_1 = T(x_1)$, $r_2 = T(x_2)$. Let u_i be an optimal control steering x_i to $\bar{x}_i \in S$ in time r_i for $i = 1, 2$. Take $(\zeta, \xi) \in N^P_{\mathrm{epi}(T)}(x_1, T(x_1))$.

We have the following possibilities:

1. $\xi = -1$: in this case $\zeta \in \partial_P T(x_1)$ and, by Lemma 3.3, we have $h(x_1, \zeta) = -1$ and there exists $p \in N_{R^{r_1}(x_1)}(\bar{x}_1) \cap [-N_S(\bar{x}_1)]$ such that $\zeta = -e^{A^\top r_1} p$.
2. $\xi = 0$: in this case, by Lemma 3.5 we have $\zeta \in N_{S(r_1)}(x_1)$ and $h(x_1, \zeta) \le 0$, and by Lemma 3.2 there exists $p \in N_{R^{r_1}(x_1)}(\bar{x}_1) \cap [-N_S(\bar{x}_1)]$ such that $\zeta = -e^{A^\top r_1} p$.

In both cases, we have the existence of $p \in N_{R^{r_1}(x_1)}(\bar{x}_1) \cap [-N_S(\bar{x}_1)]$ such that $\zeta = -e^{A^\top r_1} p$. By the Pontryagin maximum principle,

$$\langle p(t), u_1(t) \rangle = \max_{u \in \mathcal{U}} \langle p(t), u \rangle$$

for a.e. t, where $p(t) = e^{A^\top (r_1 - t)} p$, and $\zeta = -p(0) (= -p)$.

Now suppose $r_2 \le r_1$ and define

$$\begin{aligned} y &:= e^{A(r_1 - r_2)} x_1 + \int_0^{r_1 - r_2} e^{A(r_1 - r_2 - t)} u_1(t)\, dt \\ &= e^{-Ar_2} \bar{x}_1 - \int_{r_1 - r_2}^{r_1} e^{A(r_1 - r_2 - t)} u_1(t)\, dt. \end{aligned}$$

We have

$$\begin{aligned} \langle \zeta, x_2 - x_1 \rangle &= \langle p(r_1 - r_2) - p(0), x_2 - x_1 \rangle + \langle -p(r_1 - r_2), x_2 - y \rangle \\ &\quad + \langle -p(r_1 - r_2), y - x_1 \rangle \\ &=: \text{(I)} + \text{(II)} + \text{(III)}. \end{aligned}$$

We estimate separately each term of the above sum:

$$\begin{aligned} |\text{(I)}| &= \left| \langle (e^{A^\top r_2} - e^{A^\top r_1}) p, x_2 - x_1 \rangle \right| = \left| \langle e^{A^\top r_2} (Id - e^{A^\top (r_1 - r_2)}) p, x_2 - x_1 \rangle \right| \\ &\le k_2'(r_1 - r_2) \|p\| \, \|x_2 - x_1\| \le k_2'' \|p\| (\|x_2 - x_1\|^2 + |r_2 - r_1|^2), \end{aligned}$$

where $k_2', k_2'' \in \mathbb{R}$ are positive constants, and Id denotes the identity matrix. Furthermore, observe that $\|p\| \le k \|\zeta\|$, with k independent of ζ, r_1, r_2 because δ is finite. So it holds

$$|\text{(I)}| \le k_2 \|\zeta\| (\|x_2 - x_1\|^2 + |r_2 - r_1|^2),$$

where k_2 is a positive constant independent of $x_2, x_1, r_2, r_1, \zeta$.

Let us now consider (II). First observe that

$$\begin{aligned} x_2 - y &= e^{-r_2 A} (\bar{x}_2 - \bar{x}_1) + \int_{r_1 - r_2}^{r_1} e^{A(r_1 - r_2 - t)} u_1(t)\, dt - \int_0^{r_2} e^{-At} u_2(t)\, dt \\ &= e^{-r_2 A} (\bar{x}_2 - \bar{x}_1) + \int_{r_1 - r_2}^{r_1} e^{A(r_1 - r_2 - t)} (u_1(t) - u_2(t - r_1 + r_2))\, dt. \end{aligned}$$

Then

$$\text{(II)} = \langle -e^{A^\top r_2} p, x_2 - y \rangle = \langle -p, \bar{x}_2 - \bar{x}_1 \rangle + \int_{r_1 - r_2}^{r_1} \langle p(t), u_2(t - r_1 + r_2) - u_1(t) \rangle\, dt.$$

By observing that $-p \in N_S(\bar{x}_1)$ and by the maximum principle, we have that (II) ≤ 0. Let us now consider (III). First, observe that

$$y - x_1 = \int_0^{r_1 - r_2} \dot{x}_1(t)\, dt = \int_0^{r_1 - r_2} (Ax_1(t) + u_1(t))\, dt,$$

where

$$x_1(t) := e^{At}x_1 + \int_0^t e^{A(r_1 - t)}u_1(t)\, dt$$

is the optimal trajectory associated with x_1 and $u_1(t)$.

Let us define

$$k_3'' = \max\{\|A\|\|x\| + \|u\| : x \in \mathcal{R}_\delta,\ u \in \mathcal{U}\}.$$

Then we have that

$$\text{(III)} = \int_0^{r_1 - r_2} \langle p(t) - p(r_1 - r_2), Ax_1(t) + u_1(t)\rangle\, dt + \int_0^{r_1 - r_2} \langle -p(t), Ax_1(t) + u_1(t)\rangle\, dt.$$

We have also the following estimate, valid for all $t \in [0, r_1 - r_2]$:

$$\begin{aligned} |\langle p(t) - p(r_1 - r_2), Ax_1(t) + u_1(t)\rangle| &\leq k_3''\|p(t) - p(r_1 - r_2)\| \\ &\leq k_3'\|p\|(r_1 - r_2) \\ &\leq k_3\|\zeta\|(r_1 - r_2). \end{aligned}$$

So the first integral in (III) can be majorized by

$$k_3\|\zeta\||r_1 - r_2|^2,$$

where k_3 is a positive constant independent of $x_1, x_2, r_1, r_2, \zeta$. By the maximum principle, the second integral in (III) is

$$\int_0^{r_1 - r_2} [\langle -p(t), Ax_1(t)\rangle - \max_{u \in \mathcal{U}} \langle p(t), u\rangle]\, dt.$$

The following estimates hold, for a suitable constant k_4, independent of $x_1, x_2, r_1, r_2, \zeta$:

$$\begin{aligned} \int_0^{r_1 - r_2} \langle -p(t), Ax_1(t)\rangle &= \int_0^{r_1 - r_2} \Big[\langle p(0) - p(t), Ax_1(t)\rangle \\ &\qquad + \langle -p(0), A(x_1(t) - x_1)\rangle\Big]\, dt \\ &\quad + \int_0^{r_1 - r_2} \langle -p(0), Ax_1\rangle\, dt \\ &\leq k_4\|\zeta\||r_1 - r_2|^2 + \int_0^{r_1 - r_2} \langle \zeta, Ax_1\rangle\, dt, \\ \int_0^{r_1 - r_2} -\max_{u \in \mathcal{U}} \langle p(t), u\rangle\, dt &= \int_0^{r_1 - r_2} \min_{u \in \mathcal{U}} \langle p(0) - p(t), u\rangle\, dt \\ &\quad + \int_0^{r_1 - r_2} \min_{u \in \mathcal{U}} \langle -p(0), u\rangle\, dt \\ &\leq k_4\|\zeta\||r_1 - r_2|^2 + \int_0^{r_1 - r_2} \min_{u \in \mathcal{U}} \langle \zeta, u\rangle\, dt. \end{aligned}$$

Therefore,

$$(\mathrm{III}) \le k_4' \|\zeta\| \, |r_1 - r_2|^2 + (r_1 - r_2) h(x_1, \zeta).$$

Now we have to distinguish two cases:

1. If $\xi = -1$, then $h(x_1, \zeta) = -1$ and so putting together the estimates on (I), (II), and (III) we obtain that

$$\langle \zeta, x_2 - x_1 \rangle \le r_2 - r_1 + k' \|\zeta\| |r_1 - r_2|^2 + k'' \|\zeta\| (\|x_2 - x_1\|^2 + |r_2 - r_1|^2),$$

which may be written, for a suitable constant k_5 independent of $x_1, x_2, r_1, r_2, \zeta$, as

$$\langle (\zeta, -1), (x_2, T(x_2)) - (x_1, T(x_1)) \rangle \le k_5 (\|\zeta\| + 1)(\|x_2 - x_1\|^2 + |T(x_2) - T(x_1)|^2)$$

for all $x_1, x_2 \in \mathcal{R}_\delta$ and for all $(\zeta, -1) \in N^P_{\mathrm{epi}(T)}(x_1)$.

2. If $\xi = 0$, then $h(x_1, \zeta) \le 0$ and so

$$(\mathrm{III}) \le k_4' \|\zeta\| \, |r_1 - r_2|^2.$$

Putting the estimates together, we obtain

$$\langle \zeta, x_2 - x_1 \rangle \le k' \|\zeta\| \, |r_1 - r_2|^2 + k'' \|\zeta\| (\|x_2 - x_1\|^2 + |r_2 - r_1|^2),$$

which may be written, for a suitable constant k_5 independent of $x_1, x_2, r_1, r_2, \zeta$, as

$$\langle (\zeta, 0), (x_2, T(x_2)) - (x_1, T(x_1)) \rangle \le k_5 \|\zeta\| (\|x_2 - x_1\|^2 + |T(x_2) - T(x_1)|^2)$$

for all $x_1, x_2 \in \mathcal{R}_\delta$ and for all $(\zeta, 0) \in N^P_{\mathrm{epi}(T)}(x_1)$.

In both cases, we obtain (3.9).

The case $r_2 > r_1$ is similar. Let $u_i(\cdot) \in \mathcal{U}^{r_i}_{\mathrm{ad}}$ be controls steering x_i to $\overline{x}_i \in S$ in the optimal times r_i, $i = 1, 2$, together with adjoint arcs $p_i : [0, r_i] \to \mathbb{R}^n$, $p_i(t) = e^{A^\top (r_i - t)} \overline{p}_i$. Now set $\tilde{p}(t) = e^{A^\top (r_2 - t)} \overline{p}_1$ for $t \in [0, r_2]$ and observe that, for $t \in [r_2 - r_1, r_2]$, $u_1(t - (r_2 - r_1)) \in \mathrm{Argmax}_{u \in \mathcal{U}} \langle \tilde{p}(t), u \rangle$. Choose now, for $t \in [0, r_2 - r_1]$, $\bar{u}(t) \in \mathcal{U}$ such that $\bar{u}(t) \in \mathrm{Argmax}_{u \in \mathcal{U}} \langle \tilde{p}(t), u \rangle$, and set

$$\tilde{u}(t) = \begin{cases} \bar{u}(t), & t \in [0, r_2 - r_1], \\ u_1(t - (r_2 - r_1)), & t \in (r_2 - r_1, r_2]. \end{cases}$$

Define

$$y = e^{-A(r_2 - r_1)} x_1 - \int_0^{r_2 - r_1} e^{-At} \bar{u}(t) \, dt = e^{-A r_2} \bar{x}_1 - \int_0^{r_2} e^{-At} \tilde{u}(t) \, dt.$$

Now the estimates proceed analogously to the previous case $r_2 \le r_1$, with $\tilde{p}, \tilde{u}$ in place of p_1, u_1. The proof of Step 1 is concluded.

Step 2. Let $T : \mathbb{R}^n \to \mathbb{R}^+ \cup \{+\infty\}$ be lower semicontinuous and proper, with a φ-convex domain $D = \{x \in \mathbb{R}^n : T(x) < +\infty\}$ and such that

1. T is continuous on D.
2. For all $R > 0$ there exists $\sigma = \sigma(R) > 0$ such that for all $x, y \in D \cap \bar{B}(0, R)$ and for all $(\zeta, \xi) \in N^P_{\mathrm{epi}(T)}(x, T(x))$ with $\xi \in \{0, 1\}$ it holds

$$\langle (\zeta, \xi), (y, T(y)) - (x, T(x)) \rangle \le \sigma(\|\zeta\| + |\xi|)(\|y - x\|^2 + |T(y) - T(x)|^2).$$

Then there exists a continuous $\overline{\varphi}$ such that $\mathrm{epi}(T)$ is $\overline{\varphi}$-convex.

Proof of Step 2. We have to prove that given $(x,\alpha),(y,\beta)\in \operatorname{epi}(T)$ with $\|x\|,\|y\|\le R$ and $(\zeta,\xi)\in N^P_{\operatorname{epi}(T)}(x,\alpha)$ with $\xi\in\{0,-1\}$, there exists $\sigma'=\sigma'(R)>0$ such that

$$\langle(\zeta,\xi),(y,\beta)-(x,\alpha)\rangle\le\sigma'(\|\zeta\|+|\xi|)(\|y-x\|^2+|\alpha-\beta|^2).$$

Let $\alpha>T(x)$. Two cases may occur:

1. If $(x,\alpha)\in\operatorname{int}\operatorname{epi}(T)$, then $N^P_{\operatorname{epi}(T)}(x,\alpha)=\{(0,0)\}$, and there is nothing to prove.
2. Suppose $(x,\alpha)\in\operatorname{bdry}\operatorname{epi}(T)$. Let $(\zeta,\xi)\neq(0,0)$ be such that $(\zeta,\xi)\in N^P_{\operatorname{epi}(T)}(x,\alpha)$. Without loss of generality, suppose that $\|(\zeta,\xi)\|=1$. Assume that (ζ,ξ) is realized by an r-ball, with $2r\sigma\le 1$. We claim that $\xi=0$. In fact, by contradiction, let $\xi\neq 0$; since (ζ,ξ) is normal to an epigraph, we necessarily have that $\xi<0$. Then there exists $0<\varepsilon<\alpha-T(x)$ such that

$$\|(x,\alpha-\varepsilon)-(x+r\zeta,\alpha+r\xi)\|^2<r^2.$$

This means that $(x,\alpha-\varepsilon)\in B((x+r\zeta,\alpha+r\xi),r)$, which is a contradiction since $(x,\alpha-\varepsilon)\in\operatorname{epi}(T)$. So, if $(x,\alpha)\in\operatorname{bdry}\operatorname{epi}(T)$ and $0\neq(\zeta,0)\in N^P_{\operatorname{epi}(T)}(x,\alpha)$, by the continuity of T on D and the same argument of Lemma 3.5 we have that $\zeta\in N_D$. Since D is φ-convex,

$$\langle\zeta,y-x\rangle\le\varphi(x)\|\zeta\|\,\|y-x\|^2\qquad\text{for all } x,y\in D,$$

and so

$$\langle(\zeta,0),(y,\beta)-(x,\alpha)\rangle\le(\sigma\vee\varphi(x))\|\zeta\|(\|y-x\|^2+|\alpha-\beta|^2).$$

It remains to consider the case $\alpha=T(x)$. Define

$$z=x+\frac{1}{2\sigma}\frac{\zeta}{\|(\zeta,\xi)\|},\qquad\qquad \chi=T(x)+\frac{1}{2\sigma}\frac{\xi}{\|(\zeta,\xi)\|}.$$

Let $(y,\beta)\in\operatorname{epi}(T)$ with $\beta>T(y)$ and $y\neq x$. The segment connecting (z,χ) and (y,β) contains a point (y',β') which lies on the boundary of $\operatorname{epi}(T)$, so $\beta'=T(y')$. Thus we have

$$d((x,T(x)),(z,\chi))<d((y',T(y)),(z,\chi))<d((y,\beta),(z,\chi)).$$

By direct computation the desired inequality follows. By the arbitrariness of R, the proof is concluded. □

Remark. (1) The problem in Example 1 satisfies the assumptions of Theorem 3.7, although STC does not hold.

(2) If (H1) and (H2) are valid in the whole of $\mathcal{R}$, then there exists a continuous function φ such that the epigraph of T is φ-convex. Indeed, it is enough to apply Theorem 3.7 in $\mathcal{R}_\delta$ for all $\delta>0$.

In [12], functions with φ-convex epigraph were studied. As a corollary of the above result, we list some regularity properties of the minimal time function, which are direct consequences of Theorem 3.7 and of [12].

COROLLARY 3.8. *Let the assumption of Theorem* 3.7 *hold. Then,*

1. *for a.e.* $x \in \mathcal{R}_\delta$, *there exists* $\varepsilon = \varepsilon(x)$ *such that* T *is semiconvex on* $B(x, \varepsilon(x))$;
2. *in particular,* T *is twice differentiable a.e. on* $\mathcal{R}_\delta$, *in the sense that for a.e.* $x \in \mathcal{R}_\delta$ *there exists a symmetric* $n \times n$ *matrix* X_x *such that*

$$DT(y) = DT(x) + X_x(y - x) + o(\|y - x\|)$$

for $y \to x$, $y \in \operatorname{dom}(DT)$ *and, as* $y \to x$, $y \in \operatorname{dom}(T)$,

(3.10)
$$\left| T(y) - T(x) - \langle DT(x), y - x\rangle - \frac{1}{2}\langle X_x(y - x), y - x\rangle \right| = o(\|y - x\|^2);$$

3. *for a.e.* $x \in \mathcal{R}_\delta$, *there exist* $\epsilon = \epsilon(x) > 0$ *and* $c = c(x) \geq 0$ *such that for all* $\nu \in \mathbb{R}^n$, *with* $\|\nu\| = 1$, *we have* $\frac{\partial^2 T}{\partial \nu^2} \geq -c$ *in the sense of distributions in* $B(x, \varepsilon)$;
4. *set, for* $1 \leq k \leq n$,

$$\Sigma_k = \{x \in \operatorname{int}\operatorname{dom}(T) : \mathcal{H} - dim(\partial_P T(x)) \geq k\};$$

then Σ_k *is countably* $\mathcal{H}^{n-k}$*-rectifiable;*
5. *let* $\operatorname{int}\operatorname{dom}(T)$ *be nonempty; then, for all open set* $\Omega \subseteq \operatorname{int}\operatorname{dom}(T)$, $T \in BV(\Omega)$; *moreover, for a.e.* $x \in \Omega$, *there exists* $\varepsilon = \varepsilon(x)$ *such that* $DT \in BV(B(x, \varepsilon))$.

Proof. Extend T to $\mathbb{R}^n$ by setting $T(x) = +\infty$ if $x \notin \mathcal{R}_\delta$. By standard arguments, T is lower semicontinuous on $\mathbb{R}^n$. By Theorem 3.7, epi(T) is φ-convex. Then the statements (1)–(5) are direct consequences of corresponding properties proved in [12], to which all the following citations refer. Statement 1 follows from Theorem 6.1; 2 and 3 are Corollaries 6.1 and 6.2, respectively; 4 is Proposition 5.1, while 5 is Propositions 7.1 and 7.2. □

Acknowledgments. The authors are indebted to the referees for constructive criticisms and bibliographical remarks. They also thank P. Cardaliaguet for bibliographical remarks.

REFERENCES

[1] L. AMBROSIO, N. FUSCO, AND D. PALLARA, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford Science Publications, Clarendon Press, Oxford, UK, 2000.
[2] A. BACCIOTTI, *Fondamenti Geometrici Della Teoria Della Controllabilità*, Quad. Unione Mat. Italiana 31, Pitagora Editrice, Bologna, 1986.
[3] M. BARDI AND I. CAPUZZO-DOLCETTA, *Optimal Control and Viscosity Solutions of Hamilton–Jacobi–Bellman Equations*, Birkhäuser, Boston, 1997.
[4] U. BOSCAIN AND B. PICCOLI, *Optimal Syntheses for Control Systems on* 2*-D Manifolds*, Springer, Berlin, 2004.
[5] P. BRUNOVSKÝ, *Every normal linear system has a regular time-optimal synthesis*, Math. Slovaca, 28 (1978), pp. 81–100.
[6] P. CANNARSA AND C. SINESTRARI, *Convexity properties of the minimum time function*, Calc. Var. Partial Differential Equations, 3 (1995), pp. 273–298.
[7] P. CANNARSA AND C. SINESTRARI, *Semiconcave functions, Hamilton–Jacobi Equations, and Optimal Control*, Birkhäuser, Boston, 2004.
[8] P. CARDALIAGUET, *On the regularity of semipermeable surfaces in control theory with application to the optimal exit-time problem*, I, II, SIAM J. Control Optim., 35 (1997), pp. 1638–1671.
[9] L. CESARI, *Optimization—Theory and Applications*, Springer, New York, 1983.

[10] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Nonsmooth Analysis and Control Theory*, Springer, New York, 1998.
[11] F. H. Clarke, R. J. Stern, and P. R. Wolenski, *Proximal smoothness and the lower–$\mathcal{C}^2$ property*, J. Convex Anal., 2 (1995), pp. 117–144.
[12] G. Colombo and A. Marigonda, *Differentiability properties for a class of non-convex functions*, Calc. Var. Partial Differential Equations, 25 (2005), pp. 1–31.
[13] G. Colombo and P. R. Wolenski, *The subgradient formula for the minimal time function with linear dynamics and convex target*, in Proceedings of the Sixth Portuguese Conference on Automatic Control, University of Algarve, Faro, Portugal, 2004, session T2A2.
[14] R. Conti, *Processi di controllo lineari in $\mathbb{R}^n$*, Quad. Unione Mat. Italiana 30, Pitagora Editrice, Bologna, 1985.
[15] M. Degiovanni, A. Marino, and M. Tosques, *General properties of (p, q)-convex functions and (p, q)-monotone operators*, Ricerche Mat., 32 (1983), pp. 285–319.
[16] H. Federer, Curvature measures, Trans. Amer. Math. Soc., 93 (1959), pp. 418–491.
[17] H. Hermes and J. P. LaSalle, *Functional Analysis and Time Optimal Control*, Academic Press, New York, London, 1969.
[18] M. Krastanov and M. Quincampoix, *Local small time controllability and attainability of a set for nonlinear control system*, ESAIM Control Optim. Calc. Var., 6 (2001), pp. 499–516.
[19] A. Marigonda, *Second Order Conditions for the Controllability of Nonlinear Systems with Drift*, Comm. Pure Appl. Anal., to appear.
[20] F. Morgan, *Geometric Measure Theory. A Beginner's Guide, 3rd ed.*, Academic Press, San Diego, 2000.
[21] P. Nistri and M. Quincampoix, *On open-loop and feedback attainability of a closed set for nonlinear control systems*, J. Math. Anal. Appl., 270 (2002), pp. 474–487.
[22] R. A. Poliquin, R. T. Rockafellar, and L. Thibault, *Local differentiability of distance functions*, Trans. Amer. Math. Soc., 352 (2000), pp. 5231–5249.
[23] P. Soravia, *Pursuit-evasion problems and viscosity solutions of Isaacs equations*, SIAM J. Control Optim., 34 (1993), pp. 604–623.
[24] V. Veliov, *Lipschitz continuity of the value function in optimal control*, J. Optim. Theory Appl., 94 (1997), pp. 335–363.
[25] P. R. Wolenski and Z. Yu, *Proximal Aanalysis and the Minimal Time Function*, SIAM J. Control and Optim., 36 (1998), pp. 1048–1072.